ACHI 2026

The Nineteenth International Conference on Advances in Computer-Human Interactions

ISBN: 978-1-68558-383-5

May 24 - 28, 2026

Venice, Italy

ACHI 2026 Editors

Sibylle Kunz, IU Internationale Hochschule, Germany
Claudia Heß, IU Internationale Hochschule, Germany

ACHI 2026

Forward

The Nineteenth International Conference on Advances in Computer-Human Interactions (ACHI 2026), held between May 24, 2026, and May 28, 2026, in Venice, Italy, continued a series of events addressing the most recent achievements and future trends in human interactions with increasingly complex systems. Adaptive and knowledge-based user interfaces, universal accessibility, human-robot interaction, agent-driven human computer interaction, and sharable mobile devices are a few of these trends. ACHI 2026 also brought a suite of specific domain applications, such as e-learning, social, medicine, education, and engineering.

We take here the opportunity to warmly thank all the members of the ACHI 2026 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank the authors who dedicated time and effort to contribute to ACHI 2026. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the ACHI 2026 organizing committee for their help in handling the logistics of this event.

We hope that ACHI 2026 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the area of human-computer interactions.

ACHI 2026 Chairs

ACHI 2026 Steering Committee
Sibylle Kunz, IU Internationale Hochschule, Germany
Lasse Berntzen, University of South-Eastern Norway, Norway
Weizhi Meng, Lancaster University, UK
Flaminia Luccio, University of Venice, Italy
Abdul Khalique, Liverpool John Moores University, UK

ACHI 2026 Publicity Chairs
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

ACHI 2026
Committee

ACHI 2026 Steering Committee

Sibylle Kunz, IU Internationale Hochschule, Germany
Lasse Berntzen, University of South-Eastern Norway, Norway
Weizhi Meng, Lancaster University, UK
Flaminia Luccio, University of Venice, Italy
Abdul Khalique, Liverpool John Moores University, UK

ACHI 2026 Publicity Chairs

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

ACHI 2026 Technical Program Committee

Mark Abdollahian, Claremont Graduate University, USA
Md. Sabbir Ahmed, BRAC University, Bangladesh
Mostafa Alani, Tuskegee University, USA
Obead Alhadreti, Umm Al-Qura University, Al-Qunfudah, Saudi Arabia
Asam Almohamed, University of Kerbala, Iraq
Mehdi Ammi, Univ. Paris 8, France
Anmol Anubhai, Amazon, USA
Prima Oky Dicky Ardiansyah, Iwate Prefectural University, Japan
Mohd Ashraf Bin Ahmad, University Malaysia Pahang, Malaysia
Charles Averill, University of Texas at Dallas, USA
Snježana Babić, Juraj Dobrila University, Croatia
Matthias Baldauf, OST - Eastern Switzerland University of Applied Sciences, Switzerland
Giulio Barbero, Leiden University (Leiden Institute of Advanced Computer Science), The Netherlands
Catalin-Mihai Barbu, University of Duisburg-Essen, Germany
Yacine Bellik, IUT d'Orsay | Université Paris-Saclay, France
Lasse Berntzen, University of South-Eastern Norway, Norway
Ganesh D. Bhutkar, Vishwakarma Institute of Technology (VIT), Pune, India
Cezary Biele, National Information Processing Institute, Poland
Christos J. Bouras, University of Patras, Greece
Christian Bourret, UPEM - Université Paris-Est Marne-la-Vallée, France
Sabrina Bouzidi-Hassini, Ecole nationale Supérieure d'Informatique (ESI), Algeria
James Braman, The Community College of Baltimore County, USA
Justin Brooks, University of Maryland Baltimore County / D-Prime LLC, USA
Pradeep Buddharaju, University of Houston - Clear Lake, USA
Paolo Burelli, IT University of Copenhagen, Denmark

Freddy Alberto Paz Espinoza, Pontificia Universidad Católica del Perú, Peru
Gerald Penn, University of Toronto, Canada
Jorge Henrique Piazentin Ono, New York University - Tandon School of Engineering, USA
Ana C. Pires, Universidade de Lisboa, Portugal
Jorge Luis Pérez Medina, Universidad de Las Américas, Ecuador
Brian Pickering, IT Innovation Centre - University of Southampton, UK
Thomas M. Prinz, Friedrich Schiller University Jena, Germany
Annu Sible Prabhakar, University of Cincinnati, USA
Mike Preuss, Leiden University, Netherlands
Marina Puyuelo Cazorla, Universitat Politècnica de València, Spain
Yuanyuan (Heather) Qian, Carleton University in Ottawa, Canada
Claudia Quaresma, Universidade NOVA de Lisboa, Portugal
Aiswarya R., Tata Consultancy Services, India
Neha Rani, University of Florida, USA
Mariusz Rawski, Warsaw University of Technology, Poland
Carsten Röcker, inIT - Institute Industrial IT / TH OWL University of Applied Sciences and Arts, Germany
Suleiman Saka, University of Denver, USA
Joni Salminen, Qatar Computing Research Institute, Qatar
Sandra Sanchez-Gordon, Escuela Politécnica Nacional, Ecuador
Antonio-José Sánchez-Salmerón, Instituto de Automática e Informática Industrial - Universitat Politecnica de Valencia, Spain
Paulus Insap Santosa, Universitas Gadjah Mada - Yogyakarta, Indonesia
Markus Santoso, University of Florida, USA
Diana Saplacan, University of Oslo, Norway
Hélène Sauzéon, Centre Inria Bordeaux, France
Daniel Schneider, Federal University of Rio de Janeiro, Brazil
Kamran Sedig, Western University, Ontario, Canada
Sylvain Senecal, HEC Montreal, Canada
Yuhki Shiraishi, Tsukuba University of Technology, Japan
Zdzisław Sroczyński, Silesian University of Technology, Gliwice, Poland
Ben Steichen, California State Polytechnic University, Pomona, USA
Han Su, RA - MIT, USA
Federico Tajariol, University Bourgogne Franche-Comté, France
Sheng Tan, Trinity University, Texas, USA
Ranjeet Tayi, User Experience - Informatica, San Francisco, USA
Masashi Toda, Kumamoto University, Japan
Hawi Humnessa Tolera, KAIST, Korea
David Unbehaun, University of Siegen, Germany
Simona Vasilache, University of Tsukuba, Japan
Katia Vega, University of California, Davis, USA
Konstantinos Votis, Information Technologies Institute | Centre for Research and Technology Hellas, Greece
Lin Wang, U.S. Census Bureau, USA
Pinhao Wang, Zhejiang University - College of Computer Science and Technology, Hangzhou, China
Gloria Washington, Howard University, USA
Andreas Wendemuth, Otto-von-Guericke University, Germany
Zhanwei Wu, Shanghai Jiao Tong University, China
Shuping Xiong, KAIST, South Korea

Tong Xue, Beijing Film Academy, China
Rui Yang, Xi'an Jiaotong-Liverpool University, China
Ye Zhu, Cleveland State University, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Is Auditive Communication with ChatGPT an Effective Means of Building Trust Between People and Machines: A Quantitative Study

Marvin Tessitore
Technical University of Applied Sciences Wuerzburg-Schweinfurt
Faculty of Computer Science and Business Informatics
Wuerzburg, Germany
email: Marvin.tessitore.doktoranden@thws.de

Hakan Arda
Technical University of Applied Sciences Wuerzburg-Schweinfurt
Faculty of Computer Science and Business Informatics
Wuerzburg, Germany
email: Hakan.Arda.doktoranden@thws.de

Nicholas Müller
Technical University of Applied Sciences Wuerzburg-Schweinfurt
Faculty of Computer Science and Business Informatics
Wuerzburg, Germany
email: Nicholas.mueller@thws.de

Karsten Huffstadt
Technical University of Applied Sciences Wuerzburg-Schweinfurt
Faculty of Computer Science and Business Informatics
Wuerzburg, Germany
email: Karsten.huffstadt@thws.de

Abstract—**Rapid advances in Artificial Intelligence (AI) have accelerated the integration of conversational agents into everyday tasks. While voice-based interaction is becoming increasingly prevalent, its influence on user trust in AI systems remains insufficiently understood. Existing research has largely focused on text-based interfaces, leaving open whether auditory interaction can enhance or even diminish perceived trustworthiness. This study empirically examines whether the communication modality of ChatGPT (text vs. auditory) affects users' trust in the system. In a controlled experiment, participants with diverse backgrounds interacted with ChatGPT to complete story-based tasks requiring nuanced reasoning. Trust was measured through a nine-item quantitative questionnaire grounded in the Technology Acceptance Model (TAM). The results show that speech-based interaction was associated with significantly higher general trust in technological systems (Q1: $p = 0.019$, $d = 0.59$). No significant differences were found for perceived truthfulness, doubts about system accuracy, usefulness, or ease of use. These findings suggest that trust formation depends less on the interaction channel and more on underlying system qualities, such as accuracy, coherence, and conversational competence. The study provides new insights for designers of AI-driven voice systems: resources should be prioritised toward improving response quality and transparent system behaviour rather than assuming inherent trust benefits from auditory communication.**

Keywords - Human–AI Interaction; User Trust; Voice Interface; Technology Acceptance Model; Conversational AI.

I. INTRODUCTION

Over the past decade, AI has evolved into a central component of everyday digital interaction [1]. Large Language Models (LLMs), such as ChatGPT, in particular, have transformed expectations of conversational systems through their ability to generate coherent, context-sensitive, and human-like responses [2]. The rapid adoption of ChatGPT is illustrated in Figure 1, which shows its weekly active user base growing to over 100 million users within months of its release, underscoring the societal relevance of research into human–AI interaction. As these technologies continue to advance, understanding the human factors that influence their acceptance has become increasingly important, especially with regard to trust [3][4].

Trust is widely recognized as a key determinant of successful human–AI interaction [3][4]. Previous research indicates that trust formation is shaped by factors, such as perceived competence, transparency, contextual relevance, and users' prior attitudes toward AI systems [3][5]. Studies in domains, such as healthcare and education further suggest that interaction modality (text, speech, or multimodal) can influence perceptions of credibility and reliability [6][7][8].

Although voice assistants, such as Amazon Alexa, Google Home, and Apple Siri are widely adopted [9], the impact of auditory interaction on trust in advanced language models remains insufficiently explored [10][11]. Voice-based interfaces offer a more natural mode of communication [6][8], yet it is unclear whether they enhance, reduce, or merely replicate trust dynamics observed in text-based interaction [12].

This study addresses this gap by examining whether communication modality (text versus speech) affects users' trust in ChatGPT. In a controlled experiment, participants completed story-based tasks and evaluated the system using a structured quantitative questionnaire measuring perceived reliability and competence [13].

The aims of this research are twofold: first, to determine whether voice-based interaction influences trust differently than text-based interaction; and second, to provide empirical evidence to support the design of transparent, reliable, and user-centred conversational AI systems.

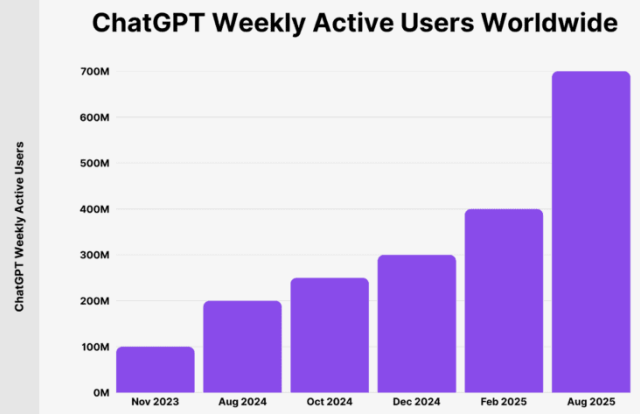


Figure 1. ChatGPT: Weekly Active Users.

The remainder of this paper is organized as follows: Section II reviews related work on trust and human–AI interaction. Section III describes the methodology and theoretical background. Section IV presents the questionnaire design. Section V outlines the experimental setup and participants. Section VI reports the results. Section VII discusses the findings in relation to the research hypotheses. Section VIII presents conclusions and directions for future work.

II. Related Works

A. Trust and Risk in a Digital World

The report "Trust and Risk in a Digital World" [14] examines key mechanisms of trust formation in virtual teams and online environments based on extensive empirical evidence. It identifies several determinants that influence how individuals establish and maintain trust in digitally mediated contexts.

A central finding highlights the restorative role of face-to-face interaction when trust is compromised in electronic communication. This emphasizes that, despite increasing digitalization, direct interpersonal contact remains an important mechanism for rebuilding trust, posing challenges for systems operating exclusively in virtual environments.

Furthermore, the report underscores the influence of interface design on initial trust perceptions. Factors, such as visual aesthetics, usability, and information clarity significantly affect users' evaluations of credibility and reliability, suggesting that trust is shaped not only by social factors but also by design characteristics.

The authors also emphasize the growing importance of the internet in everyday life, arguing that trust considerations must be systematically integrated into digital services that function as primary sources of information and communication.

Although the report does not explicitly address artificial intelligence, its findings are transferable to AI-driven systems. Since conversational agents operate within digital interfaces, principles, such as transparency, clarity, and perceived reliability are likely to influence trust in human–AI interaction.

Overall, this work provides a theoretical foundation for understanding trust in virtual environments and offers relevant implications for the design of trustworthy AI interfaces.

B. Improved Trust in Human–Robot Collaboration with ChatGPT

The study "Improved Trust in Human–Robot Collaboration with ChatGPT" [15] investigates the use of ChatGPT as a conversational control interface in a Human–Robot Collaboration (HRC) setting. The authors analyze how natural language interaction influences operator performance and trust in robotic systems.

In a controlled experiment with 15 participants, users assembled a workpiece using a ChatGPT-based assistant that issued commands to a robotic arm. The study was conducted in a virtual reality environment to enable detailed behavioral observation. Performance was evaluated based on task completion time and self-reported ratings.

Results showed that the ChatGPT-enabled interface significantly improved task efficiency compared to conventional fixed-command systems. Participants reported that the assistant's ability to retain contextual information and adapt to prior interactions contributed to smoother and more intuitive task execution.

Trust and cognitive load were measured using standardized questionnaires. The findings indicate reduced mental effort and increased trust when participants interacted with the conversational interface. These effects were attributed to the system's natural language capabilities and its capacity for contextual continuity.

Overall, the study demonstrates that naturalistic communication with ChatGPT can enhance both performance and trust in collaborative robotic systems. While the focus lies on physical human–robot interaction, the results suggest that conversational AI can positively influence trust formation in technologically mediated environments, which is relevant to the present study.

Prior research has established that conversational interfaces can enhance user trust compared to traditional text-based interaction. Gupta et al. [22] demonstrated that dialogue-based AI agents increase perceived trust in information systems, while Rheu et al. [23] showed that conversational framing positively influences user attitudes toward automated systems. The present study extends this line of research in two important ways: first, by focusing specifically on LLM-based conversational AI (i.e., ChatGPT) rather than general-purpose chatbots or rule-based dialogue systems; and second, by comparing voice-based and text-based modalities within the same AI system using the TAM as a validated measurement framework. This allows for a direct, controlled assessment of modality effects in the context of current-generation generative AI systems, which remains an underexplored area in the existing literature.

III. METHODOLOGY AND IMPLEMENTATION

A. Research Concept

The present study employs a quantitative research design to examine how trust is formed in interactions with ChatGPT. This methodological choice is grounded in the objective of generating an initial empirical understanding of trust-related factors in the early adoption phase of large language models [16]. Quantitative methods are particularly suited for this purpose, as they enable systematic data collection, statistical comparison between groups, and the extraction of generalizable patterns, which are important requirements for exploratory work on emerging technologies [17].

Given that ChatGPT represents a relatively new class of AI systems, a descriptive, data-driven approach is essential to map the trust landscape before more complex theoretical models or causal mechanisms can be meaningfully investigated. Early-stage research benefits from quantification, as it allows researchers to identify trends, user perceptions, and recurring evaluation patterns without relying on preconceived assumptions [18][19].

To address these aims, the study was structured around two experimental conditions representing different communication modalities. Participants were assigned to one of these groups and completed standardized tasks designed to elicit interaction with ChatGPT. Following the experiment, participants' attitudes, perceptions, and trust assessments were captured through a structured questionnaire.

The conceptual basis of this questionnaire draws on the TAM [20], a well-established framework for examining user acceptance of novel technologies. TAM focuses on perceived usefulness and perceived ease of use as central determinants of user attitudes and behavioral intentions. These constructs provide a theoretically grounded lens for assessing trust in AI systems, as trust is closely tied to perceptions of system competence, reliability, and effortlessness of interaction. By integrating TAM into the measurement approach, the study ensures that user trust is evaluated systematically and in alignment with established technology acceptance theory [20].

B. Theoretical Background

The TAM, originally introduced by Davis [20], is one of the most widely applied frameworks for understanding how individuals evaluate, accept, and use technological systems. Over several decades, TAM has demonstrated strong predictive validity across diverse technological contexts and remains a foundational model in information systems research.

At the center of TAM are two core constructs: Perceived Usefulness (PU) and Perceived Ease of Use (PEU) [20].

- Perceived Usefulness refers to the degree to which an individual believes that using a particular system enhances their performance or supports task accomplishment. In the context of ChatGPT, this includes how effectively users feel the system helps them generate information, solve tasks, or achieve specific goals.
- Perceived Ease of Use describes the extent to which an individual believes that interacting with a system requires minimal effort. Applied to ChatGPT, this relates to how intuitively users can formulate prompts, understand responses, and operate the system without cognitive strain.

TAM proposes that both constructs shape users' attitudes toward a technology, which subsequently influence their intention to use and, ultimately, their actual usage behavior [20]. As trust is closely tied to perceptions of system capability, reliability, and predictability, TAM provides a meaningful theoretical basis for examining how perceived usefulness and ease of use contribute to trust in AI-driven conversational systems, such as ChatGPT.

For the purpose of the present study, the original version of TAM was deliberately selected as the theoretical foundation. More recent extensions, such as TAM2, TAM3, and the Unified Theory of Acceptance and Use of Technology (UTAUT), introduce additional constructs including social influence, facilitating conditions, or perceived enjoyment. While these models offer broader explanatory power, their complexity may obscure the specific focus of this study: understanding how basic perceptions of usefulness and ease of use relate to user trust. Future research could build upon the current work by integrating extended TAM versions to capture more nuanced determinants of trust in AI technologies.

As shown in Figure 2, the original TAM structure proposed by Davis [20] is illustrated.

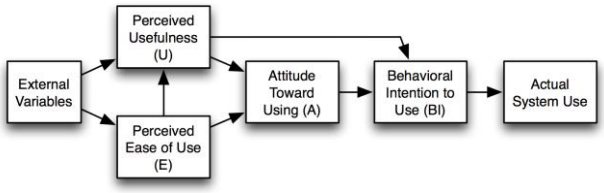


Figure 2. Technology Acceptance Model.

IV. QUESTIONNAIRE

Based on the assumption that trust in technological systems is shaped by perceived usefulness and perceived ease of use [20], the questionnaire was designed to reflect the core constructs of the TAM. The items were formulated to capture participants' evaluations of ChatGPT along these dimensions while including additional questions explicitly targeting trust as an independent psychological factor.

All items were presented as questions and answered on a five-point Likert scale ranging from 1 (strongly disagree) to

5 (strongly agree). This format was chosen to enable standardized quantitative comparison between participants and experimental conditions. Prior to data collection, the questionnaire was reviewed for clarity and consistency to ensure that all items were comprehensible and unambiguous.

The questionnaire consisted of the following items:

1. Do you generally trust machines or systems in your environment?
2. Do you think you have acquired new knowledge?
3. Would you pass on what you have learned to friends or acquaintances?
4. Would you use the information you received for academic work at university or at the TH?
5. Did you feel that your questions were answered competently?
6. Did you feel sufficiently supported by the system?
7. Did you doubt the information provided by the system at any point?
8. How truthful do you consider the generated text to be?
9. Would you use the system again to acquire knowledge?

A substantial proportion of the items was oriented toward perceived usefulness (e.g., Items 2, 3, 5, and 9), as this construct reflects participants' evaluation of the system's relevance and practical value. Trust was measured through targeted items addressing both general attitudes and situational evaluations (Items 1, 7, and 8), enabling respondents to express their confidence in the system's reliability and truthfulness.

Perceived ease of use was captured primarily through Item 6, with partial relevance in Item 9. Since voice-based interaction is generally expected to reduce operational barriers, this construct was intentionally represented in a concise manner. This approach ensured that the questionnaire remained focused while still covering the central TAM dimensions relevant to the present study.

The allocation of questionnaire items to the respective TAM constructs and trust category is summarized in Table I.

TABLE I. ASSIGNMENT OF QUESTIONNAIRE ITEMS TO TAM DIMENSIONS

Q-Number	*TAM*
Q1	Trust
Q2	Usefulness
Q3	Usefulness (private)
Q4	Usefulness (research)
Q5	Usefulness
Q6	Ease of use
Q7	Trust
Q8	Trust
Q9	Ease of use and Usefulness

V. EXPERIMENTAL SETUP

A. Test A: Auditory Interaction with ChatGPT

For the voice-based condition, a custom prototype was developed using a laptop, the ChatGPT model, Microsoft Azure speech services, and C++ integration. Spoken input was converted to text via Azure, processed by ChatGPT, and returned to participants through text-to-speech output.

The system operated in a continuous listening mode without manual activation. An initial feedback loop, caused by the system recognizing its own speech output as input, was resolved through targeted adjustments in the C++ implementation.

All sessions were conducted in a quiet office environment to minimize background noise. Participants entered the room individually and received a printed task description identical to that of the text-based condition. Tasks required them to retrieve information on historical topics using spoken interaction only, ensuring a focused voice-based experience.

Following the interaction, participants completed a structured nine-item TAM-based questionnaire assessing perceived usefulness, ease of use, and trust.

B. Test B: Text-Based Interaction with ChatGPT

For the text-based condition, participants interacted with ChatGPT via the standard interface using keyboard input and received written responses, enabling a continuous text-based dialogue. All sessions were conducted in a controlled, distraction-free environment to ensure consistency and reliable data collection.

Participants entered the room individually and received a printed task description identical to that of the auditory condition, ensuring comparable task requirements and cognitive demands across modalities. The standardized task sheet minimized interface-related variability and supported a consistent experimental procedure.

Following the interaction, participants completed the same nine-item TAM-based questionnaire to assess perceived usefulness, ease of use, and trust, enabling direct comparison with the auditory condition.

C. Participants and Procedure

The final dataset comprised N = 66 participants, evenly distributed across the two conditions (33 per group), with a mean age of 23.51 years. Most participants were students or employees of the Technical University of Applied Sciences Würzburg-Schweinfurt (THWS).

Each session lasted approximately 5–7 minutes. All participants were instructed to ask the same number of questions and were given identical opportunities to interact with the system, ensuring comparability between the speech-based and text-based conditions.

D. Hypothesis

The central aim of this study is to examine whether the modality of interaction with ChatGPT, comparing speech-

based versus text-based interaction, affects the level of trust users place in the system. To investigate this question empirically, the following hypotheses were formulated:

- H0 (Null Hypothesis):
  Speech-based input and output have no effect on the degree of trust users place in ChatGPT.
- H1 (Alternative Hypothesis):
  Speech-based input and output have an effect on the degree of trust users place in ChatGPT.

These hypotheses provide the basis for the comparative analysis between the auditory and text-based interaction conditions and guide the statistical evaluation performed in this study.

VI. RESULTS

To examine whether interaction modality (speech-based vs. text-based) influenced user perceptions, independent two-sample t-tests were conducted for each questionnaire item. The analysis was structured according to three categories derived from the TAM: usefulness, ease of use, and trust [20].

Usefulness reflects participants' evaluations of the system's practical value for information retrieval and learning, ease of use captures perceived interaction effort, and trust represents assessments of credibility and reliability. These categories provide a structured framework for comparing the two experimental conditions.

A. Response Distribution

Figure 3 presents box-and-whisker plots comparing the response distributions for the speech-based and text-based conditions across all nine questionnaire items. The speech-based condition (SP) is shown in blue and the text-based condition (TX) in red.

Each box represents the interquartile range (IQR), encompassing the middle 50% of responses. The horizontal line inside each box indicates the median response. Whiskers extend to the most extreme values within 1.5 times the IQR, and individual data points beyond this range are plotted as outliers.

Across most items, the two conditions show comparable medians and IQRs, indicating broadly similar response patterns. A notable exception is Q1 (general trust in the system), where the speech-based condition yields a visibly higher median and a more compact distribution, consistent with the statistically significant result ($p = 0.019$). Q7 (doubt about information) shows the most divergent distributions: the text-based group is clustered at low doubt scores, while the speech-based group displays greater spread.

Overall, the boxplots confirm that response patterns were largely equivalent across modalities, with the exception of Q1. The directional trends for Q3 and Q4 (marked †) are visible as a slight upward shift in the text-based boxes, though the overlap in IQRs reflects the non-significant p-values.

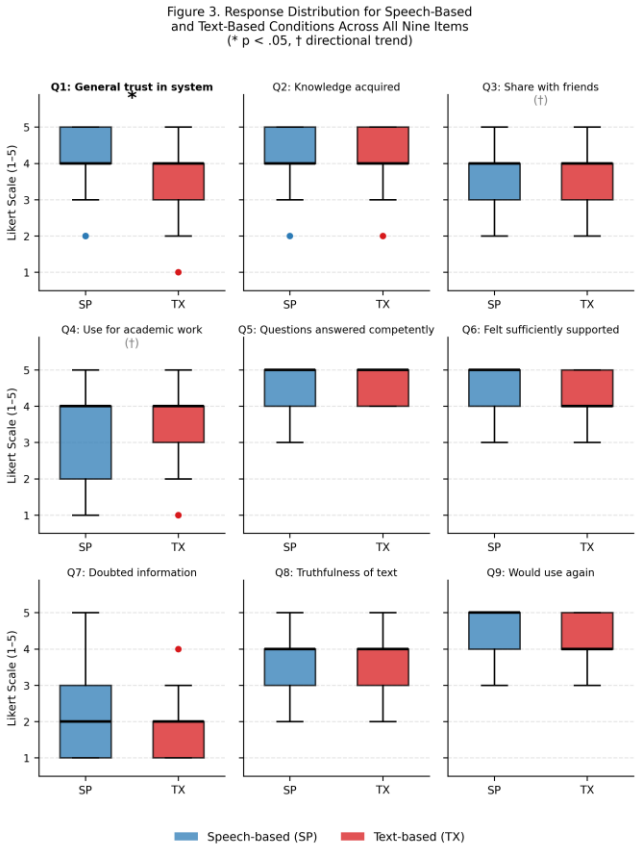


Figure 3. Response Distribution for Speech-Based and Text-Based Conditions Across All Nine Questionnaire Items.

B. Independent Two-Sample t-Test

To examine whether interaction modality (speech-based vs. text-based) influenced participant responses, independent two-sample t-tests were conducted for each questionnaire item using IBM SPSS Statistics.

Statistical analysis revealed a significant difference between the two conditions for Item Q1 only ($t(64) = 2.41$, $p = 0.019$, $d = 0.59$, 95 % CI [0.09, 0.94]), leading to the rejection of the null hypothesis for this item. Items Q3 ($p = 0.131$) and Q4 ($p = 0.066$) showed directional trends that did not reach statistical significance. All remaining items (Q2, Q5–Q9) showed no meaningful differences between conditions (all $p > 0.10$). Table II presents the group means, standard deviations, t-values, p-values, and effect sizes (Cohen's d) for all nine items.

Table II presents descriptive and inferential statistics (means, standard deviations, t-values, p-values, Cohen's d, and 95 % confidence intervals) for all nine items.

TABLE II. DESCRIPTIVE AND INFERENTIAL STATISTICS (N = 66, N = 33 PER CONDITION). SP = SPEECH, TX = TEXT; D = COHEN'S D; CI = 95 % CONFIDENCE INTERVAL FOR MEAN DIFFERENCE. * P < 0.05.

Item	Speech M (SD)	Text M (SD)	t(64)	p	d	95 % CI
Q1	4.00 (0.83)	3.49 (0.91)	2.41	0.019	0.59	[0.09, 0.94]
Q2	4.03 (0.95)	4.03 (0.98)	0.00	1.000	0.00	[−0.48, 0.48]
Q3	3.73 (0.84)	4.03 (0.77)	−1.53	0.131	−0.38	[−0.70, 0.09]
Q4	3.27 (1.31)	3.82 (1.04)	−1.87	0.066	−0.46	[−1.13, 0.04]
Q5	4.61 (0.50)	4.67 (0.48)	−0.51	0.615	−0.12	[−0.30, 0.18]
Q6	4.52 (0.57)	4.36 (0.60)	1.05	0.297	0.26	[−0.14, 0.44]
Q7	2.18 (0.98)	1.79 (0.82)	1.77	0.082	0.44	[−0.05, 0.84]
Q8	4.12 (0.55)	4.12 (0.65)	0.00	1.000	0.00	[−0.30, 0.30]
Q9	4.45 (0.62)	4.55 (0.51)	−0.66	0.515	−0.16	[−0.37, 0.19]

C. *Reliability Analysis: Cronbach's Alpha Test*

To assess the internal consistency of the questionnaire, a Cronbach's Alpha reliability analysis was conducted. The resulting coefficient of $\alpha = 0.547$ indicates a moderate level of internal consistency. Although this value falls below the commonly recommended threshold of 0.70, its interpretation must be considered in light of the conceptual structure of the instrument.

The questionnaire integrates multiple theoretically distinct constructs, including perceived usefulness, perceived ease of use, and trust. As these dimensions capture different aspects of user perception, heterogeneous response patterns are expected. Such multidimensionality typically leads to lower overall alpha values when reliability is calculated across all items, without necessarily indicating insufficient measurement quality [21].

Given the exploratory nature of the present study and its focus on capturing a broad range of user perceptions, a unified reliability analysis was considered appropriate. The moderate alpha value therefore reflects the instrument's multi-construct design rather than methodological inadequacy.

No items were removed, as each question contributes relevant information to the investigation of trust formation in human–AI interaction. Retaining all items preserves the conceptual breadth of the measurement approach and supports the study's aim of providing an initial empirical assessment of modality-dependent trust dynamics.

Reporting sub-scale reliability separately for the Trust (Q1, Q7, Q8), Usefulness (Q2, Q3, Q4, Q5, Q9), and Ease of Use (Q6, Q9) dimensions would provide a more granular assessment of measurement consistency. However, as data were collected using printed questionnaires and only aggregated frequency distributions were recorded, individual response vectors required for sub-scale Cronbach's alpha computation are not available. Future studies should ensure digital data capture to enable sub-scale reliability analysis.

VII. DISCUSSION

The present study employed an A/B testing design and an independent two-sample t-test to investigate whether the modality of interaction with ChatGPT, comparing speech-based versus text-based modalities, influences user perceptions of trust, usefulness, and ease of use. Statistical analysis revealed a significant between-group difference for one item only: Q1 (general trust in machines; $p = 0.019$). All other items did not yield statistically significant differences (all $p > 0.05$). Directional but non-significant trends were observed for Q4 ($p = 0.066$) and Q3 ($p = 0.131$), both favouring the text condition.

To interpret these findings more precisely, the questionnaire was structured into three theoretical categories derived from the TAM: Usefulness, Ease of Use, and Trust. In the following subsections, each category is discussed individually with respect to the corresponding questionnaire items. This structure allows for a more differentiated understanding of how the interaction modality affects distinct dimensions of user perception.

Finally, the discussion concludes with a synthesis across all categories. This integrative perspective highlights how usefulness, ease of use, and trust interact with one another and collectively shape user judgments in human–AI interaction. By connecting these findings, the study aims to provide a holistic interpretation of how communication modality may, or may not, influence the broader acceptance of AI systems, such as ChatGPT.

A. *Category: Trust*

The Trust category comprised the following questionnaire items:

- Q1: Do you generally trust machines or systems in your environment?
- Q7: Did you doubt the information provided by the system?

- Q8: How truthful do you consider the generated text to be?

Analysis using the independent two-sample t-test (see Table II) showed that only Q1 yielded a statistically significant difference between the two groups. Participants in the speech-based condition reported a higher general trust in machines and systems than those in the text-based condition. This suggests that the auditory modality may enhance users' baseline trust in technology during interaction.

In contrast, Q7 (doubt about information: $p = 0.082$, $d = 0.44$) and Q8 (perceived truthfulness: $p = 1.000$, $d = 0.00$) showed no significant differences between the groups. Notably, Q7 approached but did not cross the conventional significance threshold, suggesting a weak trend toward greater expressed doubt in the text condition ($M = 1.79$) compared to the speech condition ($M = 2.18$) that warrants investigation in larger samples.

Taken together, the findings for this category suggest that while specific trust-related evaluations (credibility or doubts) are not affected by modality, generalized trust toward the system appears to be higher when interaction occurs via speech. This indicates that the communicative channel may influence broader perceptions of trust, even if it does not alter assessments of specific system outputs.

B. Category: Usefulness

The Usefulness category included Items Q2, Q3, Q4, Q5, and Q9, addressing perceived knowledge acquisition, information dissemination, academic applicability, response competence, and continued system use.

Statistical analysis did not reveal significant differences between the two conditions for any usefulness item. Item Q2 (knowledge acquisition) showed no difference ($p = 1.000$, $d = 0.00$), confirming that both modalities were equally effective in supporting learning. Item Q9 (willingness to reuse the system) was also comparable across conditions ($p = 0.515$, $d = 0.16$).

Items Q3 and Q4 showed directional trends that did not reach statistical significance. Participants in the text-based condition tended to report higher intentions to share learned content with others (Q3: $M_{text} = 4.03$ vs. $M_{speech} = 3.73$; $p = 0.131$, $d = 0.38$) and to use information for academic purposes (Q4: $M_{text} = 3.82$ vs. $M_{speech} = 3.27$; $p = 0.066$, $d = 0.46$), though these differences remain inconclusive at the conventional $\alpha = 0.05$ level.

These trends are consistent with the intuitive advantage of written text for information retention and reuse: text can be scrolled back, copied, and cited, whereas spoken output is transient. Future research with larger samples should examine whether these tendencies reach statistical significance.

Item Q5 (perceived competence of responses) showed no significant difference between conditions ($p = 0.615$, $d = 0.12$), with both groups rating competence highly ($M_{speech} = 4.61$, $M_{text} = 4.67$). This suggests that the perceived quality of ChatGPT's responses was independent of the interaction modality.

Overall, the data indicate that usefulness perceptions were broadly equivalent across modalities. Neither condition produced systematically higher ratings across the usefulness dimension.

In summary, the hypothesis that modality significantly differentiates usefulness perceptions is not supported by the present data. Both interaction modes were equally effective for knowledge acquisition and willingness to reuse the system. The directional trends for Q3 and Q4 suggest a potential advantage of text for information sharing and academic use, but these require replication with larger, more diverse samples before conclusions can be drawn.

C. Category: Ease of Use

The Ease of Use category comprised Items Q6 and Q9, addressing perceived system support and willingness to use the system again for knowledge acquisition.

Statistical analysis showed highly similar response patterns across the speech-based and text-based conditions for both items, indicating that interaction modality had no significant effect on perceived ease of use.

This finding is noteworthy given that voice-based interaction is generally less familiar to many users than text-based input. Despite this, participants reported that the speech-based system was easy to use and did not introduce additional cognitive or technical barriers.

Both groups expressed comparable willingness to use the system again (Q9), suggesting that neither modality hindered continued engagement. While familiarity may account for positive evaluations of the text-based interface, the similarly favourable assessment of the speech-based variant indicates that participants were able to interact with the voice system without difficulty.

Overall, the results demonstrate that no meaningful differences emerged between the two conditions with respect to ease of use. Speech-based interaction was perceived as equally accessible and user-friendly as text-based communication.

D. Integrated Summary of Findings

The comparative analysis of the three dimensions (Trust, Usefulness, and Ease of Use) provides insights into how interaction modality shapes users' perceptions of ChatGPT.

In the Trust category, speech-based interaction was associated with significantly higher general trust in technological systems (Q1: $M_{speech} = 4.00$ vs. $M_{text} = 3.49$; $t(64) = 2.41$, $p = 0.019$, $d = 0.59$). This constitutes a medium-sized effect and represents the only statistically significant finding in the study. Perceived truthfulness (Q8) and expressed doubt (Q7) did not differ significantly, indicating that the modality effect is limited to

a baseline or dispositional trust in the technology rather than situation-specific credibility judgements.

With respect to Usefulness, no significant differences were found across any of the five items. Both modalities equally supported knowledge acquisition (Q2) and willingness to reuse the system (Q9). Directional trends suggested that text-based output may be associated with greater intentions to share information (Q3) and apply it academically (Q4), but these differences did not reach statistical significance and should be interpreted with caution.

For Ease of Use, no meaningful differences were identified between the two conditions. Both interaction modes were perceived as equally accessible and user-friendly, and participants reported comparable levels of support and willingness to use the system again. This suggests that speech-based interaction does not introduce additional usability barriers.

Overall, the results demonstrate that interaction modality influences specific dimensions of user perception rather than overall evaluations of the system. Speech-based interaction enhances generalized trust in technological systems, whereas text-based interaction shows a non-significant directional trend toward supporting information sharing and academic applicability. Both modalities are regarded as equally easy to use.

Accordingly, the null hypothesis is rejected for Q1 only. For all other items, the data are consistent with the null hypothesis of no modality effect. Overall, the results suggest that interaction modality has a limited and specific influence on user perceptions of ChatGPT: it elevates general trust in the technology without altering assessments of content quality, usefulness, or ease of use.

VIII. Conclusion and Future Work

This study provides empirical insights into interaction modality (speech versus text), which influences user perceptions of ChatGPT with respect to trust, perceived usefulness, and ease of use. Statistical analysis of a nine-item TAM-based questionnaire administered to $N = 66$ participants revealed that only one item yielded a significant between-group difference.

With regard to trust, speech-based interaction was associated with significantly higher general trust in technological systems (Q1: $p = 0.019$, $d = 0.59$). This suggests that auditory interaction may activate a broader sense of confidence in the technology, possibly due to the naturalness and immediacy of spoken dialogue. However, no significant differences were found for perceived truthfulness (Q8: $p = 1.000$) or expressed doubt about system accuracy (Q7: $p = 0.082$), indicating that modality does not influence situation-specific credibility assessments. Developers should therefore not assume that adding a voice interface will make users perceive AI-generated content as more truthful or accurate.

In terms of usefulness, no significant differences were observed across any item. Both modalities equally supported knowledge acquisition (Q2: $p = 1.000$) and willingness to reuse the system (Q9: $p = 0.515$). Directional trends, though not statistically significant, suggested that text-based output may be associated with higher intentions to share information (Q3: $p = 0.131$) and apply it in academic contexts (Q4: $p = 0.066$). These tendencies are consistent with the practical advantage of written text for retention and reuse, but they require replication in larger, more diverse samples before practical recommendations can be derived.

For ease of use, no significant differences emerged between the two conditions. Participants evaluated both interaction modes as equally accessible, demonstrating that speech-based interfaces do not introduce additional usability barriers despite their lower familiarity for many users.

Several limitations should be acknowledged when interpreting these results. The sample comprised $N = 66$ participants, predominantly students and employees of a German university of applied sciences (mean age 23.51 years). This sample is relatively homogeneous with respect to age, educational background, and cultural context. Age and prior experience with AI systems are known to moderate technology acceptance and trust formation [4], and cultural dimensions, such as individualism versus collectivism or uncertainty avoidance, may substantially affect how users respond to voice-based AI interfaces. Future research should replicate this study with larger, more diverse, and cross-cultural samples to improve the generalizability of the findings.

A further limitation concerns the comparability of the two experimental conditions. The voice-based prototype was implemented using a custom integration of Microsoft Azure Text-to-Speech (TTS) and Speech-to-Text (STT) services with the ChatGPT API, operated via a C++ application. This differs substantially from the standard ChatGPT web interface used in the text-based condition. Consequently, observed differences between conditions may partially reflect factors beyond interaction modality alone, such as differences in response latency, voice naturalness, or interface familiarity. Future research should employ more equivalent technical implementations, for instance by leveraging ChatGPT's native voice mode, to isolate modality effects more cleanly. Nonetheless, this limitation is acknowledged transparently here, as it motivates important directions for methodological refinement in subsequent studies.

References

[1] OECD, Ed., Artificial Intelligence in Society. Paris: OECD Publishing, 2019. doi: 10.1787/eedfee77-en.

[2] T. B. Brown et al., "Language Models are Few-Shot Learners," 2020, arXiv. doi: 10.48550/ARXIV.2005.14165.

[3] K. A. Hoff and M. Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," Hum.

Factors, vol. 57, no. 3, pp. 407–434, May 2015, doi: 10.1177/0018720814547570.

[4] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward A Unified View," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, Sep. 2003, doi: 10.2307/30036540.

[5] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," The Academy of Management Review, vol. 20, no. 3, p. 709, Jul. 1995, doi: 10.2307/258792.

[6] C. I. Nass and S. Brave, Wired for speech: how voice activates and advances the human-computer relationship. in Computer-human interaction. Cambridge, Mass.: MIT Press, 2005.

[7] T. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," Journal of Biomedical Informatics, vol. 39, no. 5, pp. 556–571, Oct. 2006, doi: 10.1016/j.jbi.2005.12.004.

[8] S. Oviatt, "Ten myths of multimodal interaction," Commun. ACM, vol. 42, no. 11, pp. 74–81, Nov. 1999, doi: 10.1145/319382.319398.

[9] A. Purington, J. G. Taft, S. Sannon, N. N. Bazarova, and S. H. Taylor, "'Alexa is my new BFF': Social Roles, User Satisfaction, and Personification of the Amazon Echo," in Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver Colorado USA: ACM, May 2017, pp. 2853–2859. doi: 10.1145/3027063.3053246.

[10] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in Artificial Intelligence Applications and Innovations, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., in IFIP Advances in Information and Communication Technology, vol. 584, Cham: Springer International Publishing, 2020, pp. 373–383. doi: 10.1007/978-3-030-49186-4_31.

[11] T. Ueno, Y. Sawa, Y. Kim, J. Urakami, H. Oura, and K. Seaborn, "Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods," in CHI Conference on Human Factors in Computing Systems Extended Abstracts, New Orleans LA USA: ACM, Apr. 2022, pp. 1–7. doi: 10.1145/3491101.3519772.

[12] L. Marconi, L. Longo, and F. Cabitza, "Assessing Interaction Quality in Human–AI Dialogue: An Integrative Review and Multi-Layer Framework for Conversational Agents," MAKE, vol. 8, no. 2, p. 28, Jan. 2026, doi: 10.3390/make8020028.

[13] S. W. T. Ng and R. Zhang, "Trust in AI chatbots: A systematic review," Telematics and Informatics, vol. 97, p. 102240, Feb. 2025, doi: 10.1016/j.tele.2025.102240.

[14] C. Heckersbruch, A. Öksüz, N. Walter, J. Becker, and G. Hertel, "Vertrauen und Risiko in einer digitalen Welt," 2013.

[15] Y. Ye, H. You, and J. Du, "Improved Trust in Human–Robot Collaboration with ChatGPT," 2023, arXiv. doi: 10.48550/ARXIV.2304.12529.

[16] E. M. Rogers, Diffusion of innovations, Fifth edition. New York London Toronto Sydney: Free Press, 2003.

[17] D. Huyler and C. M. McGill, "Book Review: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, by CreswellJohn and CreswellJ. David. Thousand Oaks, CA: Sage Publication, Inc.275 pages, $67.00 (Paperback).," New Horizons in Adult Education and Human Resource Development, vol. 31, no. 3, pp. 75–77, Jun. 2019, doi: 10.1002/nha3.20258.

[18] A. Onwuegbuzie, N. Leech, and K. Collins, "Qualitative Analysis Techniques for the Review of the Literature," TQR, Jan. 2015, doi: 10.46743/2160-3715/2012.1754.

[19] E. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," *Science*, vol. 358, no. 6370, pp. 1530–1534, Dec. 2017, doi: 10.1126/science.aap8062.

[20] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

[21] L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, vol. 16, no. 3, pp. 297–334, Sep. 1951, doi: 10.1007/BF02310555.

[22] A. Gupta, D. Basu, R. Ghantasala, S. Qiu, and U. Gadiraju, "To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System," in Proc. ACM Web Conference 2022 (WWW '22), Lyon, France: ACM, Apr. 2022, pp. 1–10. doi: 10.1145/3485447.3512248.

[23] M. Rheu, J. Y. Shin, W. Peng, and J. Huh-Yoo, "Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design," Int. J. Hum.–Comput. Interact., vol. 37, no. 1, pp. 81–96, 2021. doi: 10.1080/10447318.2020.1807710.

Mapping Vibrotactile Patterns to Emotions:
An Experimental Study on Intuitive Tactile Communication

Lisa Frohwieser*, Marie Herz, Susanna Götz, Nicholas H. Müller, Karsten Huffstadt
Institute of Design and Information Systems
Technical University of Applied Sciences Würzburg-Schweinfurt
Sanderheinrichsleitenweg 20, 97074 Würzburg, Germany
*Corresponding author
E-mail: lisa.frohwieser@thws.de (L. Frohwieser), marie.herz@thws.de (M. Herz),
susanna.goetz@thws.de (S. Götz), nicholas.mueller@thws.de (N. H. Müller), karsten.huffstadt@thws.de (K. Huffstadt)

***Abstract*—Vibrotactile wearables have the potential to discreetly convey nonverbal emotional cues to people with visual impairments. However, many haptic approaches rely on arbitrary codes that require training. This evaluation study examines whether vibrotactile patterns can be intuitively associated with emotion labels without prior training, and assesses the emergence of stable dominant ("winner") emotions for individual patterns, to support accessible affective communication in human-computer interaction. In a controlled within-subjects experiment, 33 participants evaluated 14 patterns presented via a wrist-worn device. After each stimulus, participants selected one of seven emotion labels in a seven-alternative forced-choice task based on Paul Ekman's basic-emotions framework, including contempt. We tested recognition rates (target hits) and winner rates (most frequent label) against chance performance using one-sided exact binomial tests with Holm correction. Dominance was defined as the difference between the most frequent and the second-most frequent label. Three patterns showed a significant target mapping, while six patterns exhibited a significant winner emotion with high dominance. Overall, several vibrotactile patterns showed stable emotion associations without prior training and serve as candidates for further refinement and validation with users with visual impairments.**

Keywords—tactile communication; assistive technology; emotional communication; nonverbal communication; visual impairment.

I. INTRODUCTION

Nonverbal cues, such as facial expressions, gestures, and body language play a central role in the emotional classification of social situations [1]. However, people with visual impairments often have limited or no access to these cues, which can result in a loss of affective information in everyday life. Assistive sensory substitution approaches therefore aim to make visual emotional cues accessible via alternative sensory channels, ideally in a discreet manner and with as little cognitive load as possible [2][3].

The sense of touch is well suited in this context. Vibrotactile output systems can be discreetly integrated into wearables and do not require visual attention [2]. Research on affective haptics and mediated social touch suggests that the perception of haptic stimuli is not solely determined by the vibration itself, but significantly by the parameterisation of these vibrations, such as intensity, temporal structure, and dynamics. Thus, these parameters can systematically modulate perceived affective qualities [4][5].

A notable challenge is that many haptic communication approaches rely on conventionalised or arbitrary codes that require users to learn and stabilise their intended meanings [6][7]. This is especially problematic in situations where no visual reference or feedback channel is available, as it can significantly impair intuitive interpretability and usability [3].

This study examines whether vibrotactile patterns can be consistently assigned to specific emotion categories in a baseline sample without prior training. The conceptual framework of this study is based on the valence–arousal space (circumplex model), which describes affective states in terms of pleasantness (valence) and activation (arousal) [8]. Based on this framework, 14 vibrotactile patterns (two variants per emotion) were constructed to target seven emotion labels in line with Ekman's basic-emotions approach [1].

We conducted a controlled pre-study with sighted participants to establish a baseline mapping of 14 vibrotactile patterns to seven Ekman labels. We report recognition rate (target hits), winner emotion and winner rate (emergent mapping), and dominance Δ as indicators of mapping clarity, and we explore associations with empathy and vibration experience.

This exploratory baseline evaluation was conducted with a sighted sample, a low-fidelity prototype, and a no-training setup under controlled conditions. The findings provide an initial basis for refinement, but they do not yet allow direct conclusions about real-world assistive use.

The remainder of this paper is structured as follows. Section II reviews related work on affective haptics, the valence–arousal model, and Ekman's basic emotions as the conceptual foundation of the study. Section III describes the methodology, including the prototype, vibrotactile patterns, study design, measures, participants, and procedure. Section IV presents the results of the experimental evaluation. Section V discusses the findings, their implications, and limitations. Finally, Section VI concludes the paper and outlines directions for future work.

II. Related Work

A. Affective Haptics and Vibrotactile Communication

Affective haptics refers to haptic interaction design that aims to evoke affective experiences or convey emotional information [4]. In wearables, this is often implemented using vibrotactile signals, since they can be delivered close to the body, are suitable for everyday use, and work independently of visual attention [2][9]. The quality of a pattern is largely shaped by parameters, such as intensity, rhythm/pulse density, pauses, and ramp profiles [10][11].

A recurring challenge concerns the semantics of haptic signals. Many systems rely on tactile "vocabularies" or haptic icons whose meaning is not inherent, but becomes established through convention, training, or repeated use [10][11]. In assistive applications without a visual reference channel, this learning requirement is particularly challenging, as it can limit the immediate, intuitive interpretability of haptic patterns [2][3].

B. The Valence–Arousal Model

Affective perception is often described in a two-dimensional space defined by valence (pleasant–unpleasant) and arousal (activated–calm/sleepy). Russell's circumplex model describes affective states as combinations of these two basic dimensions and offers a continuous framework for positioning and varying stimuli [8].

This framework is well suited to vibrotactile stimuli because parameters such as intensity, pulse density/rhythm, pauses, and dynamics can be operationalised as carriers of arousal and, indirectly, valence. Empirical studies on the emotional effects of haptic parameters suggest that changes in intensity and temporal structure can lead to systematic shifts in ratings within the valence-arousal space [4][5].

C. Ekman's Basic Emotions

In addition to dimensional models, emotions are frequently characterised as discrete categories. Ekman's basic-emotions framework proposes a core set of emotions that are relatively stable and functionally meaningful, with distinct patterns of response and expression that can be modelled as qualitatively different states. The extant literature on the subject commonly refers to six basic emotions: happiness, sadness, anger, fear, surprise, and disgust [1]. Following prior empirical work that treats contempt as a distinct category, we included contempt as an additional label in the target set [12].

In this study, seven emotion labels (including contempt) are used as the target set for assignment in order to test whether vibrotactile patterns can be categorised intuitively at the label level. Valence and arousal are additionally used to locate the perceived effect of each stimulus within a dimensional space and to describe pattern profiles in a comparable way.

Specifically, this study makes two contributions. First, it introduces a structured evaluation framework for training-free vibrotactile emotion mapping based on recognition, emergent winner emotions, and dominance (Δ). Second, it provides empirical baseline evidence on parameter configurations that yield robust above-chance emotion associations.

Against this background, three Research Questions (RQs) are formulated, focusing on target assignment, dominance, and person-specific influencing factors:

RQ 1: Do individual vibrotactile patterns, without prior training, achieve assignment to the intended Ekman target emotion above chance level?

RQ 2: Which emotion category is selected as the dominant choice for each pattern, and how robust is this dominance?

RQ 3: Are trait empathy and experience with vibrotactile feedback associated with individual recognition rates? (exploratory)

III. Methodology

A. Prototype and Vibrotactile Patterns

The study used a low-fidelity prototype as the vibrotactile output system. The device uses two vibration motors integrated into an elastic wristband worn on the non-dominant wrist. One motor is placed on the upper side of the wrist, while the second is positioned on the underside. Control is handled by an ESP32-based microcontroller connected via Universal Serial Bus (USB) to a control unit. An administration unit is used to trigger the predefined patterns during the study. A motor driver ensures that the motors are supplied with sufficient current. The firmware was implemented in the Arduino development environment. Figure 1 shows the wristband prototype.

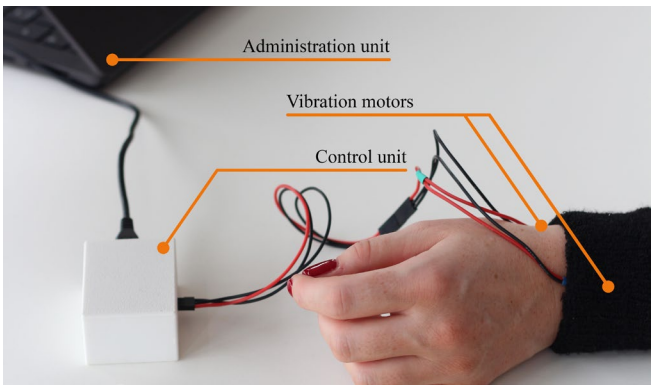


Figure 1. Wristband prototype with control unit, administration unit, and two vibration motors.

The vibrotactile patterns were specified as combinations of intensity, pulse duration, pauses, and the time course of intensity (ramp profile) [6][10]. Intensity was implemented using 8-bit Pulse-Width Modulation (PWM) in a range from 0 to 255. In the context of pattern design, three intensity levels were employed (low: 125 PWM, medium: 190 PWM, high: 255 PWM). Furthermore, separate time intervals were

defined for pulse duration and pauses (short: 300 milliseconds, medium: 600 milliseconds, long: 900 milliseconds). In order to model different dynamics, both linear and exponential ramp profiles were applied.

The stimulus set comprised 14 vibrotactile patterns, with two variants created for each of the target emotions (six basic emotions, plus contempt) [1][12]. The development of the patterns was guided by the dimensions of valence and arousal, which are widely considered core factors of affective states [8]. This broad parameterisation was considered appropriate for an exploratory study, as prior work suggests that the affective interpretation of haptic stimuli is shaped by both stimulus parameters and context [4][5].

Patterns intended to convey higher activation were typically operationalised using higher intensities, shorter intervals, and more pronounced pulse sequences. In contrast, low-activation patterns were characterised by longer, calmer temporal profiles and lower intensities [4][5]. The specifications of all 14 patterns are summarised in Table I.

To illustrate theory-driven parameterisation along the valence–arousal space, three patterns are briefly described. We assume arousal is primarily encoded via intensity, pulse density, and temporal dynamics, whereas valence is influenced more indirectly by qualitative dynamics (e.g., smooth/fading vs. abrupt/jerky) [4][5].

Pattern 4 (Sadness B) was defined as a linear ramp-down from medium intensity (190 PWM) to 0 over 900 ms (R↓lin(M→0,900)). The continuous fade without repeated pulses reflects low arousal and a dampened signal character [4][5][8].

Pattern 9 (Surprise A) consisted of three short pulses with very short pauses (P(S,300)+p150+P(S,300)+p150+P(H,300)), culminating in a high-intensity pulse. The burst-like structure and abrupt intensity increase operationalise high arousal (startle/orienting), while valence remains context-dependent, consistent with surprise being potentially positive or negative [4][5][8].

Pattern 11 (Disgust A) consisted of three ramp-and-pulse segments with a medium – high – medium intensity structure ([R↑lin(0→M,600)+P(M,300)] + p150 + [R↑lin(0→H,600)+P(H,300)] + p150 + [R↑lin(0→M,600)+P(M,300)]). The repeated build-ups and the dominant high-intensity middle segment reflect medium-to-high arousal and a more irregular signal character. The flanking medium-intensity segments before and after the peak were intended to avoid a single alarm-like burst and instead create a structured rise–peak–decline profile, which was considered more suitable for approximating disgust within the valence–arousal framework [4][5][8].

Before the study, the programmed patterns were repeatedly checked on the prototype through direct tactile inspection to confirm perceptible differences and reliable motor function. No external calibration was undertaken to verify a linear relationship between PWM input and physical actuator output.

TABLE I
OVERVIEW OF VIBROTACTILE PATTERNS

ID	*Target emotion*	*Parameters*	*Duration (ms)*
1	Happiness A	2×[2×P(M,300)+p150]+p300	1800
2	Happiness B	R↑lin(0→M,600)+p300+2×P(M,300)+p150	1650
3	Sadness A	P(S,900)+p450+P(S,900)	2250
4	Sadness B	R↓lin(M→0,900)	900
5	Anger A	P(M,300)+p150+P(H,300)+p150+P(S,300)+p150+P(H,300)	1650
6	Anger B	R↑exp(0→H,600)+D(H,900)	1500
7	Fear A	4×[P(H,300)+p150]	1650
8	Fear B	(P(M,300)+p150+P(H,300)) + p150 + (…) + p300 + (…)	2700
9	Surprise A	P(S,300)+p150+P(S,300)+p150+P(H,300)	1200
10	Surprise B	(P(H,300)+p150+P(M,300)) + p300 + (…)	1800
11	Disgust A	[R↑lin(0→M,600)+P(M,300)] + p150 + [R↑lin(0→H,600)+P(H,300)] + p150 + [R↑lin(0→M,600)+P(M,300)]	3000
12	Disgust B	P(H,600)+p300+R↓lin(H→M,600)+p150+R↓exp(M→S,600)	2250
13	Contempt A	R↑lin(0→S,600)+D(S,300)+R↑lin(S→M,600)+D(M,600)	2100
14	Contempt B	R↑exp(0→M,900)+D(M,600)	1500

PWM levels: S = 125 (low), M = 190 (medium), H = 255 (high); P(x,t) = Pulse with intensity x and duration t ms; D(x,t) = Continuous vibration (constant) with intensity x and duration t ms; p = Pause (ms); R↑/R↓ = Ramp up/down; lin/exp = Linear/exponential.

Variations in total duration reflect the underlying parameter combinations (pulse lengths, pauses, and ramp segments) and were not controlled as a separate variable in this exploratory design.

B. Study Design and Measures

1) Study Design

The study used a within-subjects design. Each participant evaluated all 14 vibrotactile patterns, which helped reduce potential sources of variance caused by individual differences, such as general response tendencies or differences in tactile sensitivity [13]. Each trial presented one pattern followed by a standardised questionnaire. The aim was to capture an immediate, experience-based assignment for each pattern. Patterns were presented in a randomised order for each participant to minimise order effects [13]. Participants were allowed one optional repeat per pattern if needed.

2) Measures and Operationalisation

a) Measurements per trial

The primary dependent variable was the emotion assignment in a forced-choice format (ekman_choice). Participants were instructed to select one of seven emotion labels (happiness, sadness, anger, fear, surprise, disgust, and contempt) [1][12]. This response format was chosen to capture the assignment as a categorical decision and to test pattern recogni-

tion statistically against a defined chance level [14]. The order of response options was randomised to minimise position and order effects [15].

To encourage intuitive responses, the emotion category was collected first in each trial. Participants then rated their confidence in the assignment using a 5-point Likert-type scale (choice_confidence; 1 = very unsure to 5 = very sure) [16]. This measure adds a metacognitive judgement to the forced-choice assignment and indicates whether patterns are selected with low or high conviction [17].

Next, the affective dimensions valence and arousal were collected using two 5-point scales adapted from the Self-Assessment Manikin (SAM), but without the dominance dimension [18]. Valence was assessed with the question "How pleasant or unpleasant was the vibration pattern?" (valence; 1 = very unpleasant to 5 = very pleasant). Arousal was assessed with "How calm or arousing/activating was the vibration pattern?" (arousal; 1 = very calm/sleepy to 5 = very arousing/activating). Omitting the dominance dimension reduced the number of items per trial and lowered participant burden in a repeated-measures design [14]. In addition, many models treat valence and arousal as the central dimensions for describing affective states [8].

Final trial order: (1) ekman_choice → (2) choice_confidence → (3) valence → (4) arousal

b) Measurements per test subject

Two additional criteria were collected to describe the sample and to capture potential moderating or explanatory variables: (i) experience with vibrotactile feedback in everyday life and (ii) empathy. Experience with vibrotactile feedback was measured using five Likert-type items (including one reverse-coded) to capture everyday familiarity as a potential covariate. Empathy was assessed using the Single Item Trait Empathy Scale (SITES) to provide a compact measure of a dispositional trait that may relate to processing affective information [19]. In addition, demographic variables (age, gender) were recorded to describe the sample and to identify potential confounding factors (e.g., age-related differences in tactile perception) descriptively.

Pre-trial measures: (1) gender → (2) age

Post-trial measures: (1) vibration_experience → (2) empathy

3) Control of Potential Confounding Factors

To minimise order effects (learning, fatigue, or contrast effects), the sequence of the 14 patterns was presented in a randomised order for each participant [13].

Before the trials, the experimenter checked fit and motor placement and provided standardised instructions to ensure comparable stimulation conditions. Furthermore, participants were permitted only one repeat per pattern to avoid distorting the original intuitive assignment through frequent replays. Repeats were recorded and used as a quality metric (e.g., potential indications of low perceptibility for particular patterns).

C. Participants and Session Procedure

a) Participants

The sample consisted of N = 33 sighted participants. A baseline sample of sighted participants was employed in order to establish statistically robust stimulus-emotion associations under controlled conditions, prior to validation with visually impaired users. Participants were recruited as a convenience sample of sighted adults for this baseline study. No further selection protocol or stratification based on specific target characteristics was applied. The study was conducted as a supervised experiment, with an experimenter present throughout. Participants were aged 21 to 62 years; 45.45% identified as female and 54.55% as male. Each session lasted approximately 20–30 minutes per participant.

b) Study Setting and Setup

The data were collected in individual sessions. The prototype was fitted on the non-dominant wrist to ensure the dominant hand remained available for the completion of the questionnaire. The two vibration motors were positioned above and below the wrist, and stabilised using an elastic fixation similar to that of a sweatband.

The stimuli were triggered via a laptop using a Wizard-of-Oz setup, allowing the experimenter to play the predefined patterns in a controlled manner [20].

c) Session Procedure

At the beginning of the experiment, the subjects were given an introduction to the aim of the study and the procedures they would be expected to follow. They were also given information on the voluntary nature of their participation, the right to withdraw at any time, and the confidentiality of their data. The subjects then provided their signature on an informed consent form. In the subsequent phase of the experiment, the researcher affixed the prototype to the non-dominant wrist and conducted a meticulous evaluation of its fit.

The participants initially completed a brief questionnaire that solicited demographic information. Two neutral practice trials were conducted to familiarise participants with the sensation and response format and excluded from the analyses. Subsequent to this, the 14 vibrotactile patterns were presented in a randomised order. In each trial, participants initially made a forced-choice assignment to an Ekman emotion, followed by a confidence rating, and subsequently valence and arousal ratings. If needed, each pattern could be repeated once. Repeats were logged to capture potential perceptual issues. After completing all trials, participants answered questions about their experience with vibration signals and an empathy item. Finally, the prototype was removed and the session ended.

IV. RESULTS

A. Dataset and Analysis Strategy

The analysis is based on N = 33 participants who each evaluated 14 vibrotactile patterns (462 test trials). Two neutral practice trials were excluded from all analyses. After data collection, the dataset was screened for completeness

and plausibility, then cleaned and coded. Per trial, the emotion label was treated as a categorical variable, while confidence, valence, and arousal ratings were coded as Likert-type values (1–5) [16]. For analysis, we derived pattern ID, target emotion, and a correctness indicator (target hit). In addition, we summarised valence and arousal at the pattern level using means and standard deviations and reported repeat frequency as an indicator of perceptibility. For the binomial tests, significance was evaluated one-sided because effects were hypothesised as above-chance mappings. Effect sizes are reflected by the observed proportions (recognition/winner rates) and the dominance gap Δ. Data preparation and metric calculation were conducted in Microsoft Excel; statistical analyses were performed in Jamovi.

To address RQ1 and RQ2, we computed two pattern-level metrics: recognition rate as the proportion of target hits ($k_{target}/33$) and winner rate as the proportion of the most frequently selected label ($k_{winner}/33$; emergent mapping). Both were tested against chance level $p0 = 1/7$ using one-sided exact binomial tests [14]. Holm correction was applied across the 14 patterns separately for recognition tests (p_{rec}) and winner tests (p_{win}) to control the family-wise error rate at $\alpha = .05$ [21]. Mapping clarity is reported via the runner-up label and dominance $\Delta = n_{winner} - n_{runner\text{-}up}$. Repeats were analysed descriptively. Exploratory Spearman correlations tested mean confidence vs. recognition rate (pattern level) and participant recognition vs. empathy/vibration experience (participant level) [14].

B. Recognition Rate: Assignment to the Target Emotion

After Holm correction, three patterns exceeded chance-level target assignment: Sadness A (ID 3: 51.52%, 17/33), Sadness B (ID 4: 57.58%, 19/33), and Surprise A (ID 9: 36.36%, 12/33) (Table II). Aggregated across variants, sadness showed the highest target assignment (54.5%, 36/66), whereas disgust (10.6%, 7/66) and fear (12.1%, 8/66) were lowest.

C. Winner Emotion and Mapping Robustness

To capture emergent mappings, we report the most frequently selected label per pattern (winner emotion) and its winner rate (Table II). Winner emotions can diverge from the intended target. After Holm correction, six patterns showed winner rates above chance (IDs 2, 3, 4, 5, 8, 9). Mapping clarity was quantified using dominance Δ. Sadness patterns were strongly dominant (ID 3: Δ = 11; ID 4: Δ = 14), whereas ID 8 (anger; Δ = 6) and ID 9 (surprise; Δ = 5) showed moderate dominance. ID 13 showed no clear winner (Δ = 0).

D. Repeats and Choice Confidence

Repeats were rare (7/462 trials; 1.5%) and occurred once each for seven different patterns (IDs 1, 2, 4, 5, 11, 12, 14), indicating no systematic perceptibility issues. Pattern-level confidence was not associated with recognition rate (Spearman $\rho = -0.306$, $p = .288$; $n = 14$ patterns).

E. Affective Dimensions: Valence and Arousal

In addition to the emotion label, participants rated each pattern on 5-point scales for valence and arousal. Table III reports pattern-level means and standard deviations. Overall, arousal ratings varied more strongly across patterns than valence.

F. Participant Characteristics and Recognition Rate

At the participant level (N = 33), recognition rate (across all 14 patterns) was not correlated with empathy (SITES; $\rho = -0.006$, $p = .975$) or vibrotactile experience ($\rho = 0.013$, $p = .943$; Spearman).

TABLE II
RESULTS FOR VIBROTACTILE PATTERNS

ID	*Target emotion*	*Recognition Rate*	p_{rec}	p_{rec}*(Holm)*	*Winner emotion*	*Winner Rate*	p_{win}	p_{win}*(Holm)*	*Δ*	*Runner-up emotion*
1	Happiness	21.21% (7/33)	0.183	1.000	Surprise	27.27% (9/33)	0.038	0.264	2	Anger
2	Happiness	24.24% (8/33)	0.089	0.977	Surprise	36.36% (12/33)	0.001	0.015*	4	Happiness
3	Sadness	51.52% (17/33)	< .001	< .001***	Sadness	51.52% (17/33)	<.001	< .001***	11◆	Happiness
4	Sadness	57.58% (19/33)	< .001	< .001***	Sadness	57.58% (19/33)	<.001	< .001***	14◆	Surprise
5	Anger	18.18% (6/33)	0.330	1.000	Happiness	33.33% (11/33)	0.005	0.042*	2	Surprise
6	Anger	21.21% (7/33)	0.183	1.000	Surprise	24.24% (8/33)	0.089	0.444	1	Anger
7	Fear	15.15% (5/33)	0.518	1.000	Anger	30.30% (10/33)	0.014	0.113	2	Happiness
8	Fear	9.09% (3/33)	0.869	1.000	Anger	42.42% (14/33)	<.001	< .001***	6◇	Happiness
9	Surprise	36.36% (12/33)	0.001	0.016*	Surprise	36.36% (12/33)	0.001	0.015*	5◇	Happiness
10	Surprise	15.15% (5/33)	0.518	1.000	Happiness	24.24% (8/33)	0.089	0.444	1	Anger
11	Disgust	12.12% (4/33)	0.713	1.000	Anger	21.21% (7/33)	0.183	0.549	1	Contempt
12	Disgust	9.09% (3/33)	0.869	1.000	Surprise	21.21% (7/33)	0.183	0.549	1	Fear
13	Contempt	21.21% (7/33)	0.183	1.000	Happiness	21.21% (7/33)	0.183	0.549	0	Contempt
14	Contempt	12.12% (4/33)	0.713	1.000	Surprise	27.27% (9/33)	0.038	0.264	4	Fear

Recognition Rate = Proportion of correct assignments to the target emotion ($k_{target}/33$); Winner emotion/Winner Rate = Most frequently selected emotion/($k_{winner}r/33$); Runner-up Emotion = Second most frequently selected emotion; p-values: One-sided exact binomial tests against chance level $p_0=1/7$; Holm correction within each respective test family (m=14 Patterns) applied separately for p_{rec} and p_{win}.
* p(Holm) < .05, ** p(Holm) < .01, *** p(Holm) < .001.; $\Delta = n_{winner} - n_{runner\text{-}up}$; ◇ $\Delta \geq 5$ (moderately dominant), ◆ $\Delta \geq 8$ (strongly dominant), in case of a tie $\Delta = 0$ (e.g., ID 13).

TABLE III
RESULTS FOR VALENCE AND AROUSAL

ID	*Target emotion*	*Valence, M (SD)*	*Arousal, M (SD)*
1	Happiness	3.36 (1.08)	3.85 (0.91)
2	Happiness	3.24 (1.00)	3.30 (0.95)
3	Sadness	3.42 (1.09)	2.36 (1.25)
4	Sadness	3.82 (1.07)	1.79 (1.08)
5	Anger	3.42 (1.06)	4.03 (0.88)
6	Anger	3.24 (1.12)	3.55 (1.03)
7	Fear	2.78 (1.24)	4.61 (0.56)
8	Fear	2.85 (1.37)	4.45 (0.62)
9	Surprise	3.33 (1.27)	3.91 (0.98)
10	Surprise	3.09 (1.10)	3.97 (0.77)
11	Disgust	3.12 (1.02)	3.73 (1.04)
12	Disgust	2.82 (1.21)	3.91 (1.07)
13	Contempt	3.15 (1.23)	3.06 (1.30)
14	Contempt	3.42 (1.09)	2.88 (1.17)

M = Mean; SD = Standard deviation.

V. DISCUSSION

This pilot study assessed the association of vibrotactile patterns with discrete emotion labels without prior training and examined the emergence of stable "winner" emotions at the pattern level. Overall, only a subset of patterns supported target-consistent mappings, yet several stimuli converged reliably on a dominant label. Given the current parameter space, a purely intuitive one-to-one mapping to seven emotion categories appears difficult, whereas data-driven selection of robust signals remains feasible for building an initial haptic vocabulary [6][7].

The present study should be interpreted as an exploratory screening of a relatively broad pattern space. This made it possible to compare multiple candidate patterns across seven target emotions, but it does not yet allow a precise assessment of which individual parameters were responsible for the observed effects.

Target-consistent assignments occurred primarily for patterns whose temporal-energy profile was distinctive. Overall, sadness appeared to be the most robust category in the present dataset, as both sadness patterns performed above chance and showed strong dominance in the winner analysis. In particular, the two sadness patterns were consistently recognized, aligning with their low-activation design (longer, calmer envelopes and reduced abrupt peaks) [5]. At the same time, dimensional ratings indicate that arousal is captured more directly by intensity, pulse density, and dynamics, whereas valence appears harder to encode [4]. The valence and arousal ratings further suggest that arousal was conveyed more clearly than valence in the present dataset. This is consistent with the circumplex perspective, in which categories in the same arousal region can be difficult to separate when a stimulus primarily communicates activation rather than pleasantness [8]. The recurring emergence of surprise as a winner label is compatible with this mechanism: salient, abrupt patterns may be interpreted as an orienting/startle-like signal when valence cues are weak [5][8]. From a design perspective, the winner-based view is practically relevant. Even when a pattern misses its intended target, a stable emergent mapping can still be leveraged as a reliable carrier of meaning in a tactile vocabulary [6][7]. A useful selection heuristic is the combination of winner rate and dominance Δ, because it captures both preference strength and separation from the runner-up. Patterns with high winner rate and moderate-to-high Δ are plausible candidates for a core set, while low-Δ patterns should be reparameterised to increase distinctiveness (e.g., stronger contrasts in timing, clearer dynamic "signatures," or additional cues beyond global intensity/rhythm) [10][11]. If the goal remains a seven-label emotion set, future iterations likely need more explicit valence coding or an interaction concept that provides contextual framing or brief familiarisation to stabilise meanings [6][7].

Nevertheless, several limitations should be considered. Results were obtained in a controlled pre-study with sighted participants using a low-fidelity wrist-worn prototype and a forced-choice task without training, which may limit transfer to real-world assistive use and to people with visual impairments [3]. In addition, overlapping categories (especially among negative, high-arousal labels) may be amplified by the seven-alternative forced-choice format when stimuli mainly convey arousal [8].

VI. CONCLUSION AND FUTURE WORK

This study provides a baseline mapping of 14 vibrotactile patterns to discrete emotion labels under no-training conditions and introduces a pragmatic screening approach based on winner rate and dominance Δ. The results indicate that some patterns can support stable, training-free associations, while many mappings are dominated by arousal-related interpretations, suggesting that valence is comparatively harder to convey via the current parameterisation. Further work should address four directions. First, the stimulus set should be iteratively optimised using the observed winner/Δ profiles: retain robust patterns, redesign ambiguous ones, and explicitly test parameter changes intended to improve separability (e.g., timing contrasts, dynamic ramps, or additional distinguishing cues). This next step should reduce the pattern space and test simplified stimulus families before moving to more complex combinations. Second, validation with people with visual impairments is required, including calibration (fit and intensity thresholds) and evaluation in context-relevant scenarios where meaning is used rather than only judged. Third, the role of minimal familiarisation should be tested systematically (e.g., short onboarding vs. none), measuring learning curves, retention, and potential cognitive load trade-offs in repeated sessions. Fourth, future work may revisit the methodology used to select vibrotactile patterns and examine whether emotion interpretations are shaped not only by the stimulus itself, but also by the situation, social context, or recent stimulus history.

REFERENCES

[1] P. Ekman and E. L. Rosenberg, Eds., What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), 2nd ed. New York, NY, USA: Oxford Univ. Press, 2005.

[2] D. T. V. Pawluk, R. J. Adams, and R. Kitada, "Designing Haptic Assistive Technology for Individuals Who Are Blind or Visually Impaired," *IEEE Transactions on Haptics*, vol. 8, no. 3, pp. 258–278, 2015, doi: 10.1109/TOH.2015.2471300.

[3] P. Bach-y-Rita and S. W. Kercel, "Sensory substitution and the human–machine interface," *Trends in Cognitive Sciences*, vol. 7, no. 12, pp. 541–546, 2003, doi: 10.1016/j.tics.2003.10.013.

[4] M. A. Eid and H. Al Osman, "Affective Haptics: Current Research and Future Directions," *IEEE Access*, vol. 4, pp. 26–40, 2016, doi: 10.1109/ACCESS.2015.2497316.

[5] C. Rognon, B. Stephens-Fripp, J. Hartcher-O'Brien, B. Rost, and A. Israr, "Linking Haptic Parameters to the Emotional Space for Mediated Social Touch," *Frontiers in Computer Science*, vol. 4, Art. no. 826545, pp. 1–14, 2022, doi: 10.3389/fcomp.2022.826545.

[6] L. M. Brown, S. A. Brewster, and H. C. Purchase, "Multidimensional tactons for non-visual information presentation in mobile devices," in *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '06)*, Helsinki, Finland, 2006, pp. 231–238, doi: 10.1145/1152215.1152265.

[7] S. Zhao, A. Israr, F. W. Lau, and F. Abnousi, "Coding Tactile Symbols for Phonemic Communication," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, Montréal, QC, Canada, 2018, Paper 392, pp. 1–13, doi: 10.1145/3173574.3173966.

[8] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[9] V. Hayward, O. R. Astley, M. Cruz-Hernandez, D. Grant, and G. Robles-De-La-Torre, "Haptic interfaces and devices," *Sensor Review*, vol. 24, no. 1, pp. 16–29, 2004, doi: 10.1108/02602280410515770.

[10] M. Azadi and L. A. Jones, "Evaluating Vibrotactile Dimensions for the Design of Tactons," *IEEE Transactions on Haptics*, vol. 7, no. 1, pp. 14–23, 2014, doi: 10.1109/TOH.2013.2296051.

[11] T. D. Nyasulu, S. Du, N. Steyn, and E. Dong, "A Study of Cutaneous Perception Parameters for Designing Haptic Symbols towards Information Transfer," *Electronics*, vol. 10, no. 17, Art. no. 2147, pp. 1–17, 2021, doi: 10.3390/electronics10172147.

[12] P. Ekman and W. V. Friesen, "A New Pan-Cultural Facial Expression of Emotion," *Motivation and Emotion*, vol. 10, no. 2, pp. 159–168, 1986, doi: 10.1007/BF00992253.

[13] D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Boston, MA, USA: Houghton Mifflin, 1963.

[14] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. London, U.K.: SAGE Publications Ltd., 2017.

[15] J. A. Krosnick and D. F. Alwin, "An evaluation of a cognitive theory of response-order effects in survey measurement," *Public Opinion Quarterly*, vol. 51, no. 2, pp. 201–219, 1987, doi: 10.1086/269029.

[16] G. M. Sullivan and A. R. Artino, Jr., "Analyzing and interpreting data from Likert-type scales," *Journal of Graduate Medical Education*, vol. 5, no. 4, pp. 541–542, 2013, doi: 10.4300/JGME-5-4-18.

[17] S. M. Fleming and H. C. Lau, "How to measure metacognition," *Frontiers in Human Neuroscience*, vol. 8, Art. no. 443, pp. 1–9, 2014, doi: 10.3389/fnhum.2014.00443.

[18] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994, doi: 10.1016/0005-7916(94)90063-9.

[19] S. Konrath, B. P. Meier, and B. J. Bushman, "Development and validation of the single item trait empathy scale (SITES)," *Journal of Research in Personality*, vol. 73, pp. 111–122, 2018, doi: 10.1016/j.jrp.2017.11.009.

[20] M. A. Chavarria *et al*., "Challenges and Opportunities of the Human-Centered Design Approach: Case Study Development of an Assistive Device for the Navigation of Persons With Visual Impairment," *JMIR Rehabilitation and Assistive Technologies*, vol. 12, Art. no. e70694, pp. 1–20, 2025, doi: 10.2196/70694.

[21] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

Semantic Segmentation of Extremely Small Defects in Sliced Apples

Yueying Shi and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, *Iwate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
E-mail: s236w002@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—Semantic segmentation of extremely small defect regions in food inspection remains a challenging task due to severe foreground–background class imbalance and the high cost of missed detections. In the inspection of sliced apples, remaining skin and core fragments often occupy only a few pixels, making them prone to being overlooked despite high overall segmentation accuracy. This study systematically investigates semantic segmentation strategies for detecting extremely small defects in ultraviolet (UV) images of sliced apples. A stepwise experimental framework is employed to isolate and evaluate the effects of loss functions, encoder architectures, and decoder designs under identical training conditions. Quantitative results demonstrate that region-based loss functions, particularly Tversky and Focal Tversky losses, provide superior spatial consistency and recall compared to pixel-wise reweighting approaches. Furthermore, lightweight encoders, such as MobileNetV2, combined with UNet-based decoders, achieve more stable and robust performance in preserving fine-grained defect regions across different random seeds. These findings provide practical design guidelines for recall-oriented semantic segmentation in food inspection tasks, where reliable detection of extremely small defects is critical for quality assurance and safety.**

Keywords*—*Machine vision; semantic segmentation; small defect regions; extreme class imbalance.

I. INTRODUCTION

Automated visual inspection is a key component of modern food processing industries, where high-throughput production requires consistent quality control and strict safety assurance. Apples are widely processed in sliced form in industrial food production [1][2]. During these mechanical operations, residual skin or core fragments may remain on sliced apples due to natural variations in fruit size, shape, and internal structure. Although such defects are typically small, their presence can negatively affect product appearance and may raise safety concerns, making reliable inspection an important requirement in apple processing lines.

In current industrial practice, inspection of sliced apples is still largely dependent on manual visual checking by human operators. Machine vision–based systems [3][4] and deep learning enable pixel-level defect localization through semantic segmentation. However, detecting defects in sliced apples presents a particularly challenging scenario. Remaining skin and core fragments often occupy less than one percent of the image area and appear as extremely small and sparse regions against a dominant background. Under such severe foreground–background class imbalance, segmentation models may achieve high overall accuracy while still failing to detect critical defect regions, where missed detections are especially problematic in food inspection.

UV imaging has emerged as an effective modality for enhancing the visibility of subtle surface features in food inspection. Under UV illumination, residual skin and core fragments on sliced apples exhibit higher contrast than surrounding flesh, enabling more reliable visual discrimination compared to conventional RGB imaging [5]. When combined with semantic segmentation, UV-based imaging provides a promising approach for detecting extremely small defects at the pixel level. Nevertheless, segmentation performance under such extreme class imbalance remains highly sensitive to both network architecture and optimization strategy, and inappropriate design choices can easily suppress rare defect regions during training.

Although previous studies have investigated apple defect detection using various imaging modalities and deep learning models, most have focused on object-level classification or segmentation of relatively large defects. Systematic evaluation of semantic segmentation strategies for extremely small defect regions remains limited, particularly in terms of isolating the effects of loss functions, encoder architectures, and decoder designs under identical experimental conditions. As a result, practical design guidelines for recall-oriented segmentation in sliced apple inspection are still insufficient.

To address this gap, this study systematically investigates semantic segmentation of extremely small defects in sliced apples using UV images. A stepwise experimental framework is adopted to independently evaluate loss functions, encoder architectures, and decoder designs under severe class imbalance. By emphasizing recall-oriented and overlap-based evaluation metrics, this work aims to identify practical design guidelines that minimize missed detections of small defect regions. The findings of this study provide useful insights for designing robust semantic segmentation models for food inspection tasks, where reliable detection of extremely small defects is critical for quality assurance and safety.

The remainder of this paper is organized as follows. Section 2 reviews related work and outlines key challenges. Section 3 presents the proposed framework. Section 4 describes the experimental setup, followed by results in Section 5. Section 6 concludes the paper and discusses future work.

II. Related Work

Early studies on apple inspection relied on handcrafted features extracted from RGB images to identify surface defects, while recent deep learning–based approaches have significantly improved detection accuracy through object detection and semantic segmentation models [5].

To enhance defect visibility beyond RGB imaging, various modalities, such as near-infrared, short-wave infrared, and X-ray imaging have been explored [6]. These approaches are effective for detecting internal or early-stage defects but often focus on object-level detection or relatively large defect regions and require specialized hardware. In contrast, inspection of sliced apples introduces additional challenges, as defects, such as remaining skin and core fragments, are extremely small, irregularly shaped, and sparsely distributed, making pixel-level semantic segmentation more suitable than object detection.

UV imaging has recently attracted attention for sliced apple inspection because it enhances the contrast of residual skin and core fragments relative to surrounding flesh. Previous studies have shown that UV illumination, when combined with deep learning–based segmentation, enables more reliable detection of such subtle defects. However, semantic segmentation of extremely small defect regions remains difficult due to severe foreground–background class imbalance, where standard pixel-wise loss functions tend to bias optimization toward the dominant background class [5].

To address this issue, imbalance-aware loss functions, such as Dice [7], Tversky [8], and focal variants [9] have been proposed to improve sensitivity to minority regions by optimizing spatial overlap. In addition, network architecture plays a significant role in preserving fine-grained spatial information, as excessive down-sampling can easily suppress small target regions. While these techniques have been extensively studied in medical image segmentation, systematic evaluation of their combined effects in industrial food inspection, particularly for UV-based inspection of sliced apples, remains limited.

Overall, existing studies demonstrate the potential of deep learning and advanced imaging modalities for apple defect detection, yet practical design guidelines for recall-oriented semantic segmentation of extremely small defects in sliced apples are still insufficient. This study addresses this gap by systematically evaluating loss functions and network architectures under identical experimental conditions.

III. Materials and Method

A. Overview of the Framework

This study addresses semantic segmentation of extremely small defect regions in sliced apples using UV images, where remaining skin and core fragments appear as small dark regions occupying only a few pixels.

To systematically analyze this problem, a stepwise experimental framework was adopted (Figure 1). Loss functions, which are computed by comparing the decoder output with the corresponding ground truth and used to optimize the network via backpropagation, encoder architectures, which extract hierarchical features from the input image and compress them into latent representations, and decoder designs, which progressively restore spatial resolution and generate pixel-wise predictions, were evaluated independently under identical training conditions. This framework isolates the impact of each component on segmentation performance.

B. Image Acquisition

UV image acquisition was conducted in a controlled environment to ensure stable and reproducible illumination conditions. Sliced apple samples were placed on a flat platform inside a black enclosure to suppress ambient light and external reflections. Multiple UV light sources were arranged around the enclosure to uniformly illuminate the apple slices and enhance the contrast between defect regions and surrounding flesh.

Images were captured from multiple views to account for variations in defect appearance. For samples containing remaining skin, images were acquired from four directions, while samples containing remaining core were captured from three directions due to limited contrast under bottom-view illumination (Figure 2). This multi-view acquisition strategy increases dataset diversity and improves robustness against viewpoint-dependent variations.

C. Dataset and Processing

Ground-truth annotation was performed manually at the pixel level using a dedicated annotation tool. Three semantic classes were defined: background, remaining core, and remaining skin. The background class includes apple flesh and non-defect regions, while the remaining core and skin classes correspond to defect regions observed under UV illumination.

The annotated dataset was divided into training, validation, and test sets following a fixed ratio to ensure fair evaluation. To mitigate overfitting caused by limited data, data augmentation techniques, such as random rotation and horizontal or vertical flipping were applied to the training set. These augmentations preserve defect characteristics while increasing the effective size of the dataset.

The dataset consists of 339 images collected under a multi-view acquisition setup, where each apple is captured from multiple viewpoints. Images of the same apple are grouped and assigned to a single split (train/validation/test) to prevent data leakage. Each of the remaining core and skin classes

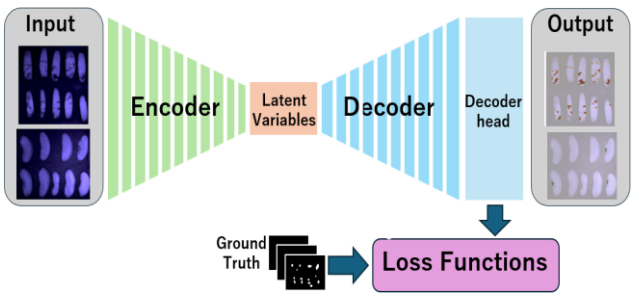


Figure 1. Overall architecture of the encoder-decoder segmentation framework.

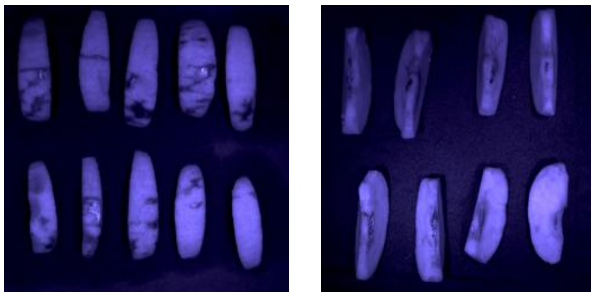


Figure 2. UV images of apple samples containing remaining skin (left) and remaining core (right).

occupies less than 1% of the total image area, confirming the extreme class imbalance in this task.

D. Data Processing

All images were resized to a fixed spatial resolution to standardize input dimensions across experiments. Input images were normalized using the ImageNet mean and standard deviation to ensure compatibility with ImageNet-pretrained encoders. This normalization stabilizes optimization and accelerates convergence during training.

Segmentation masks were resized using nearest-neighbor interpolation to avoid label distortion at class boundaries. This extreme imbalance highlights the necessity of carefully designed loss functions and architectures capable of preserving fine-grained spatial information.

E. Model Architecture

A multi-class semantic segmentation model based on an encoder–decoder architecture was employed. The encoder is responsible for extracting hierarchical feature representations, while the decoder reconstructs spatial resolution and generates dense pixel-wise predictions. ImageNet-pretrained encoders were used in all experiments to improve feature generalization under limited training data.

Multiple encoders (ResNet-34, ResNet-50, MobileNetV2, EfficientNet-B0) [10]-[12] with different depths and computational complexities were evaluated to analyze their ability to preserve small defect features. Lightweight encoders are expected to retain fine-grained spatial details, whereas deeper encoders provide richer contextual information. Several decoder designs were also compared, focusing on their ability to fuse multi-scale features and accurately reconstruct extremely small and fragmented defect regions.

F. Decoder Design

The decoder plays a crucial role in semantic segmentation of extremely small defect regions, where minor spatial reconstruction errors can easily result in missed detections. Its primary function is to restore spatial resolution from encoder feature maps and accurately reconstruct fine-grained defect structures under severe foreground–background class imbalance.

In this study, several representative decoder architectures were evaluated, including UNet, UNet++, DeepLabV3+, and Pyramid Attention Network (PAN)–based decoders [13]-[16]. These architectures differ in their approaches to multi-scale feature fusion and spatial detail recovery. Among them, a UNet-based decoder was adopted as the primary design in the proposed model.

The UNet decoder employs skip connections that directly transfer high-resolution feature maps from the encoder to corresponding decoder stages, effectively mitigating information loss caused by down-sampling. This property is particularly important for preserving boundary and location information of extremely small and sparse defect regions. All decoder variants were integrated with the same encoder backbone and trained under identical conditions to ensure fair comparison.

Pre-experimental results showed that the UNet-based decoder provided more stable and accurate reconstruction of fine-grained defect regions than other decoder designs. Therefore, it was selected as the final decoder configuration used in this study.

G. Loss Function and Training Strategy

Semantic segmentation of extremely small defect regions is strongly affected by severe foreground–background class imbalance, where defect pixels account for only a small fraction of the image. To address this issue, both pixel-wise and region-based loss functions were evaluated in this study, with particular emphasis on reducing false negatives.

Specifically, the following loss functions were examined: Weighted Cross-Entropy loss, Dice loss, Tversky loss, and Focal Tversky loss. Weighted Cross-Entropy assigns larger weights to minority classes based on class frequency and serves as a representative pixel-wise reweighting approach. Dice loss and Tversky loss are region-based overlap losses that directly optimize spatial agreement between predictions and ground-truth masks, making them more suitable for highly imbalanced segmentation tasks. In Tversky loss, the weighting parameters were set to emphasize false negatives, reflecting the importance of recall in defect inspection. Focal Tversky loss further introduces a focusing parameter to emphasize hard-to-segment regions and improve sensitivity to extremely small defects.

All models were trained under identical conditions to ensure fair comparison across loss functions and architectures. The same dataset split, optimizer, learning rate schedule, batch size, and number of training epochs were used throughout the experiments. Training samples were shuffled at each epoch to stabilize optimization, and all networks were trained end-to-end using backpropagation.

This strategy enables systematic analysis of loss functions under controlled conditions. By combining recall-oriented loss functions with controlled training conditions, the proposed framework aims to reliably detect extremely small defect regions in sliced apple inspection.

H. Evaluation Metrics

Segmentation performance was evaluated using recall-oriented and overlap-based metrics that are suitable for extreme class imbalance. Recall measures the ability to detect defect pixels, while Intersection-over-Union (IoU) evaluates spatial consistency between predictions and ground truth. The F2-score was adopted to place greater emphasis on recall than

precision, reflecting the importance of minimizing missed detections in food inspection tasks [17].

These metrics provide a complementary evaluation of segmentation performance for extremely small defects.

IV. Experimental Setup

All experiments were conducted to evaluate semantic segmentation performance for extremely small defect regions in sliced apples under severe foreground–background class imbalance. UV images and corresponding pixel-level annotations were divided into training, validation, and test sets using a fixed split. Data augmentation, including rotation and flipping, was applied only to the training set. All input images were resized to a fixed resolution of 224 × 224 pixels and normalized using ImageNet mean and standard deviation. Segmentation masks were resized using nearest-neighbor interpolation.

A stepwise experimental design was adopted to ensure fair comparison. Loss functions, encoder architectures, and decoder designs were evaluated independently, with only one component varied at a time while all other settings were fixed. ImageNet-pretrained weights were used for all encoders to ensure consistent initialization.

All models were trained for 200 epochs with a batch size of 8 using the same optimizer and learning rate schedule across all experiments. Training samples were shuffled at each epoch, and model parameters were optimized end-to-end using backpropagation. Experiments were implemented using PyTorch and executed on a system equipped with an NVIDIA GeForce RTX 4080 Laptop GPU.

V. Results and Discussion

Following the experimental protocol described above, quantitative and qualitative results are presented in a stepwise manner to clarify the effects of loss functions, encoder architectures, and decoder designs on segmentation performance. Unless otherwise specified, results reported in Tables 1–4 correspond to single training runs, while multi-seed evaluation is conducted for representative configurations to assess robustness.

A. Loss Function Comparison

The first set of experiments examined the influence of loss function design under a fixed encoder–decoder configuration (Table 1). Among the evaluated loss functions, Tversky loss achieved the highest F2-score for remaining core defects (0.82), while Focal Tversky loss yielded the highest F2-score for remaining skin defects (0.578). Dice loss demonstrated relatively balanced performance across both defect types, whereas Weighted Cross-Entropy achieved high recall but substantially lower IoU and F2-scores, indicating degraded spatial consistency.

Pixel-wise reweighting approaches show a fundamental limitation under extreme class imbalance. Although Weighted Cross-Entropy increases the contribution of minority-class pixels, it does not explicitly enforce spatial coherence, leading to fragmented predictions and reduced overlap with ground-truth regions. In contrast, region-based overlap losses directly optimize spatial agreement, making them inherently more suitable for segmenting extremely small and sparse defect regions.

The difference between Tversky and Focal Tversky losses further suggests that defect morphology influences optimal loss design. Remaining core defects tend to form compact regions, benefiting from the false-negative suppression emphasized by Tversky loss, whereas remaining skin defects are often irregular and fragmented, where the focusing mechanism of Focal Tversky loss improves sensitivity to hard-to-segment pixels. This observation underscores the importance of aligning loss function design with defect characteristics rather than relying on a single generic objective.

B. Encoder Comparison under Fixed Loss Functions

1) Core Defect Segmentation under Fixed Tversky Loss

Table 2 compares encoder performance for core defect segmentation under a fixed Tversky loss. MobileNetV2 achieved the highest core F2-score (0.868) and IoU (0.811), outperforming deeper encoders, such as ResNet-34 and ResNet-50. EfficientNet-B0 exhibited the lowest performance among the evaluated encoders.

Lightweight encoders are more effective than deeper architectures when target regions are extremely small. Excessive network depth and repeated downsampling may suppress fine-grained spatial cues associated with small core defects, even if high-level contextual features are well captured. In contrast, MobileNetV2 preserves spatial detail through its compact architecture and reduced parameterization, which appears advantageous under severe class imbalance.

2) Skin Defect Segmentation under Fixed Focal Tversky Loss

Table 3 presents encoder comparison results for skin defect segmentation under a fixed Focal Tversky loss. Again, MobileNetV2 achieved the highest skin F2-score (0.707) and IoU (0.628). Although its core F2-score (0.829) was slightly lower than that obtained under fixed Tversky loss, overall performance remained competitive. ResNet-34 showed moderate performance, while EfficientNet-B0 consistently yielded the lowest scores for both defect types.

Encoder effectiveness depends on preserving spatial resolution rather than representational depth. Lightweight encoders, particularly MobileNetV2, demonstrate strong robustness across defect types and loss configurations, making them well suited for recall-oriented segmentation of extremely small defects [11].

C. Decoder Comparison with Fixed Encoder and Loss Function

Decoder architectures were compared under a fixed MobileNetV2 encoder and Tversky loss, as shown in Table 4. UNet and UNet++ achieved the highest overall segmentation performance among the evaluated models, with UNet achieving a core F2-score of 0.879 and a skin F2-score of 0.678. UNet++ slightly improved skin defect segmentation (F2-score of 0.711) at the cost of a marginal reduction in core performance.

TABLE 1. LOSS FUNCTION COMPARISON UNDER A FIXED ENCODER AND DECODER

Loss Function	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
Tversky	0.895	0.75	0.811	**0.828**	0.772	0.403	0.492	0.469
Focal Tversky	0.868	0.65	0.707	0.708	0.773	0.516	0.605	**0.578**
Dice	0.851	0.752	0.806	0.801	0.724	0.488	0.577	0.535
Weighted CE	0.902	0.546	0.609	0.651	0.817	0.396	0.489	0.523

TABLE 2. CORE FOCUSING ENCODER COMPARISON UNDER FIXED TVERSKY LOSS AND DECODER

Encoder	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
MobileNetV2	0.866	**0.811**	0.872	**0.868**	0.758	**0.636**	0.724	**0.711**
ResNet34	0.895	0.75	0.811	0.828	0.772	0.403	0.492	0.469
ResNet50	0.905	0.735	0.788	0.795	0.792	0.53	0.617	0.593
EfficientNet-B0	0.87	0.694	0.753	0.759	0.751	0.363	0.454	0.441

TABLE 3. SKIN FOCUSING ENCODER COMPARISON UNDER FIXED FOCAL TVERSKY AND DECODER

Encoder	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
MobileNetV2	0.89	0.756	0.821	0.829	0.756	**0.628**	0.717	**0.707**
ResNet50	0.895	0.832	0.889	0.892	0.767	0.511	0.601	0.572
ResNet34	0.868	0.65	0.707	0.708	0.773	0.516	0.605	0.578
EfficientNet-B0	0.85	0.58	0.636	0.637	0.722	0.424	0.513	0.477

TABLE 4. DECODER COMPARISON UNDER A FIXED LOSS FUNCTION AND ENCODER

Decoder	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
UNet	0.881	0.815	0.878	**0.879**	0.715	0.614	0.708	0.678
UNet++	0.866	0.811	0.872	0.868	0.758	0.636	0.724	**0.711**
PAN	0.879	0.773	0.841	0.862	0.677	0.597	0.682	0.677
DeepLabV3+	0.87	0.712	0.78	0.8	0.703	0.566	0.651	0.649

PAN demonstrated slightly lower performance across metrics, while DeepLabV3+ yielded the lowest IoU and F2-scores for both defect types. Decoders relying on deep, low-resolution features are less effective for reconstructing extremely small and fragmented regions.

The superior performance of UNet-based architectures highlights the importance of skip connections that directly transfer high-resolution spatial features from the encoder to the decoder. Such connections mitigate information loss caused by down-sampling and are particularly critical when target regions consist of only a few pixels. These findings confirm that decoder design is a decisive factor in small defect segmentation under extreme class imbalance [14].

D. Reproducibility Analysis across Random Seeds

To assess robustness, reproducibility analysis was conducted using three different random seeds for representative model configurations. The MobileNetV2–UNet model with Tversky loss achieved mean core and skin F2-scores of 0.870 ± 0.022 and 0.734 ± 0.035, respectively, indicating both high performance and low variance. In contrast, the UNet++ model with the same configuration showed lower mean performance and higher variance.

Furthermore, the UNet model with a ResNet-50 encoder and Focal Tversky loss exhibited larger fluctuations, particularly for skin defects. Although full multi-seed evaluation was not performed for all configurations due to computational constraints, these results suggest that lightweight encoders combined with skip-connected decoders tend to achieve both higher accuracy and more stable optimization. The observed performance differences are larger than the corresponding variances, indicating that the overall trends are consistent across random initializations.

E. Qualitative Results

Qualitative analysis revealed that false positive predictions primarily appeared as small, isolated regions distributed across background areas. In addition to background noise caused by illumination artifacts or water droplets, this behavior is an expected consequence of optimizing recall-oriented metrics, such as the F2-score, where reducing false negatives is prioritized over suppressing minor false positives under extreme class imbalance [18].

As illustrated in Figure 3, the proposed models successfully localized most remaining core and skin defect regions. Occasional false negatives were observed near ambiguous boundaries or low-contrast regions, reflecting the inherent difficulty of segmenting extremely small defects. From a practical perspective, this trade-off is acceptable, as missing defects are more critical than minor false positives, which can be handled in downstream inspection.

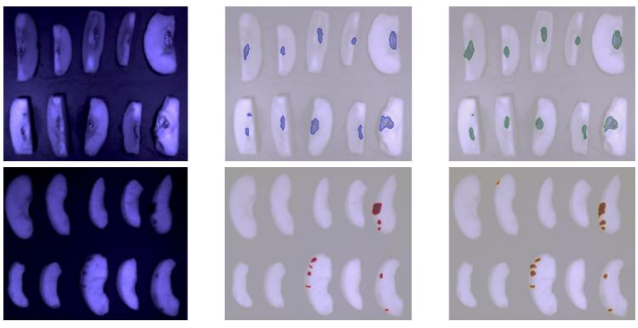


Figure 3. Qualitative segmentation results under UV imaging. From left to right: input images, ground truth masks, and model predictions. The top row corresponds to remaining core, and the bottom row corresponds to remaining skin.

While this study focuses on widely used segmentation architectures and loss functions under controlled conditions, more specialized techniques such as boundary-aware losses, hard-example mining, and high-resolution refinement strategies may further improve performance for extremely small defect regions. In particular, boundary-aware approaches better preserve fine spatial details. These approaches are promising directions for future work and may complement the findings of this study.

VI. Conclusion and Future Work

This study investigated semantic segmentation of extremely small defect regions in sliced apples under severe foreground–background class imbalance. Focusing on remaining skin and core fragments in UV images, a stepwise experimental framework was adopted to systematically evaluate the effects of loss functions, encoder architectures, and decoder designs on recall-oriented segmentation performance.

Experimental results demonstrated that region-based overlap loss functions, particularly Tversky and Focal Tversky losses, consistently outperformed pixel-wise reweighting approaches in terms of spatial consistency and F2-score. These findings indicate that explicitly optimizing spatial overlap is more effective than class-frequency reweighting alone when defect regions occupy only a few pixels.

In terms of network architecture, lightweight encoders, especially MobileNetV2, achieved superior and more stable performance compared to deeper models. This suggests that preserving fine-grained spatial features is more critical than increasing network depth for extremely small defect segmentation. Furthermore, UNet-based decoders with skip connections proved highly effective in reconstructing small and fragmented defect regions, highlighting the importance of direct feature transfer from encoder to decoder under extreme class imbalance.

Reproducibility analysis across multiple random seeds further confirmed that the combination of a lightweight encoder and a skip-connected decoder provides not only high accuracy but also stable optimization, which is essential for practical deployment in industrial food inspection systems. Qualitative results supported these findings by demonstrating reliable localization of most defect regions, with an acceptable trade-off between recall and minor false positives.

Overall, this work provides practical design guidelines for recall-oriented semantic segmentation of extremely small defects in sliced apples. The proposed framework and insights are directly applicable to food inspection tasks where minimizing missed detections is critical for quality assurance and safety. Future work will explore higher-resolution training, advanced post-processing strategies, and integration with real-time inspection pipelines to further improve robustness and deployment readiness.

References

[1] A. B. Oyenihi, Z. A. Belay, A. Mditshwa, and O. J. Caleb, “An apple a day keeps the doctor away: The potentials of apple bioactive constituents for chronic disease prevention,” Journal of Food Science, vol. 87, no. 6, pp. 2291–2309, 2022.

[2] Elsevier, “Apple — an overview,” ScienceDirect Topics, 2025.

[3] K. B. Patel, “A review: Machine vision and its applications,” International Journal of Computer Applications, vol. 70, no. 10, pp. 28–32, 2013.

[4] H. Zhao, “Advances and prospects in machine vision: A critical review based on CiteSpace,” IEEE Access, vol. 8, pp. 12345–12360, 2020.

[5] J. Rahmawan and O. Prima, “Quality inspection of processed apple based on ultraviolet imaging,” Computers and Electronics in Agriculture, vol. 162, pp. 89–97, 2019.

[6] A. Tempelaere et al., “Deep learning for apple fruit quality inspection using X-ray imaging,” in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 4203–4212, 2023.

[7] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, “Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Proc. DLMIA, pp. 240–248, 2017.

[8] S. Lu, F. Gao, C. Piao, and Y. Ma, “Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data,” in Proc. International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), pp. 176–181, 2019.

[9] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3D fully convolutional deep networks,” in Proc. MICCAI Workshops, pp. 379–387, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, pp. 770–778, 2016.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. CVPR, pp. 4510–4520, 2018.

[12] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, pp. 6105–6114, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, “UNet: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, pp. 234–241, 2015.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested UNet architecture for medical image segmentation,” IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856–1867, 2020.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with atrous separable convolution for semantic image segmentation,” in Proc. ECCV, pp. 801–818, 2018.

[16] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” in Proc. British Machine Vision Conference (BMVC), pp. 285–295, 2018.

[17] Z. Wang et al., “Revisiting Evaluation Metrics for Semantic Segmentation: Optimization and Evaluation of Fine-grained Intersection over Union,” in Proc. NeurIPS Datasets and Benchmarks Track, 2023.

[18] J. Tian, N. Mithun, Z. Seymour, H.-P. Chiu, and Z. Kira, “Striking the Right Balance: Recall Loss for Semantic Segmentation,” in Proc. International Conference on Learning Representations (ICLR), 2021.

Comparative Evaluation of Single- and Multi-Marker Pose Estimation for Freehand 3D Ultrasound Reconstruction

Syahid Al Irfan and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, *Iwate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
Email: s236x002@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—This study evaluates marker-based pose estimation methods for freehand three-dimensional (3D) ultrasound reconstruction by comparing single-marker and multi-marker configurations and analyzing their effects on reconstruction accuracy and pose stability. Although vision-based fiducial markers provide a low-cost and compact alternative to optical or electromagnetic tracking systems, their performance can degrade due to occlusion, motion blur, and variations in viewing angle. To improve robustness, multi-marker rigs arranged on polyhedral structures maintain visibility across multiple faces. In this study, both configurations were calibrated to a shared coordinate system and applied to breast-phantom data containing spherical inclusions with known diameters (5 mm and 10 mm). Ultrasound images and marker-tracking videos were recorded simultaneously, and each frame was paired with its corresponding six-Degree-of-Freedom (6-DoF) pose for volumetric reconstruction. The reconstructed inclusions were segmented and compared with ground-truth volumes, and statistical analysis was conducted using two-way Analysis of Variance (ANOVA) with object diameter and tracking methods as factors. The results show that multi-marker tracking consistently reduced volumetric error and variability, achieving the lowest mean error of 12.1 mm³ for 5-mm inclusions, whereas the single-marker configuration produced larger errors, particularly for 10-mm inclusions. ANOVA revealed significant main effects for both diameter and tracking methods. These findings indicate that multi-marker configurations improve pose robustness and enhance reconstruction accuracy in freehand 3D ultrasound systems.**

Keywords-Freehand 3D ultrasound; Pose estimation; Multi-marker tracking; Volumetric reconstruction.

I. INTRODUCTION

3D ultrasound reconstruction generates volumetric datasets by spatially assembling sequential two-dimensional brightness mode (B-mode) images according to the position and orientation of the ultrasound probe [1]. In freehand 3D ultrasound systems, accurate pose estimation is essential to ensure that each acquired frame is correctly positioned within a global coordinate system. The overall reconstruction accuracy, therefore, depends not only on image quality but also critically on the precision and stability of the probe-tracking method.

Various tracking technologies have been employed in freehand 3D ultrasound, including optical infrared systems, electromagnetic sensors, and vision-based tracking approaches [2]. Although optical and electromagnetic systems can provide high accuracy, they often involve high cost, complex installation, and susceptibility to environmental interference. Vision-based fiducial marker tracking using printed markers, such as ArUco offers a compact and low-cost alternative [3]. However, single-marker tracking is vulnerable to occlusion, motion blur, and variations in viewing angle or illumination [4], which may result in unstable corner detection and cumulative pose drift during reconstruction [5]. Such instability can directly degrade volumetric reconstruction accuracy.

To mitigate these limitations, multi-marker fiducial configurations have been proposed. By arranging multiple markers on different faces of a three-dimensional object, multi-marker rigs increase the probability that at least one marker remains visible under challenging viewing conditions [6]. Systems such as DodecaPen have demonstrated that multi-marker configurations can achieve highly accurate six-Degree-of-Freedom (6-DoF) tracking using a monocular camera [7]. Despite these advantages, the effectiveness of multi-marker tracking for improving volumetric reconstruction accuracy in freehand 3D ultrasound has not been systematically evaluated, particularly for small, cancer-like targets.

Mammography remains the standard imaging modality for breast cancer screening. However, its diagnostic sensitivity is significantly reduced in women with dense breast tissue, where overlapping fibroglandular structures can obscure lesions [8]. To address this limitation, Automated Breast Ultrasound (ABUS) has been introduced as a standardized, operator-independent three-dimensional (3D) ultrasound technique [9]. Clinical studies have demonstrated that supplementing mammography with ABUS improves cancer detection in women with dense breasts, reporting an additional detection rate of 2.4 cancers per 1,000 screened women [10] and an increase in diagnostic performance, with the Area Under the Curve (AUC) improving from 0.72 to 0.82, and sensitivity increasing by 29% [11]. These findings highlight the clinical importance of volumetric ultrasound imaging.

Consequently, this study seeks to compare single-marker and multi-marker pose estimation approaches in the context of freehand 3D ultrasound reconstruction. We hypothesize that multi-marker configurations provide significantly improved reconstruction accuracy and stability compared to single-marker tracking. To test this hypothesis, experiments were conducted using breast phantoms containing spherical inclusions of known diameters (5 mm and 10 mm). Reconstruction accuracy was evaluated by volumetric error analysis and statistically examined using two-way Analysis of Variance (ANOVA).

By quantitatively analyzing the influence of marker configuration and object size, this study provides

experimental evidence regarding the effectiveness of multi-marker tracking for low-cost freehand 3D ultrasound systems.

Relevant literature on the topic of implementing marker detection for 3D reconstruction, along with the necessary information to investigate for a deeper understanding of the topic are summarized in Section II. The details on the data collection methods and how the two types of markers for 3D ultrasound reconstruction are evaluated, along with their implementation, are explained in Section III. Finally, the results obtained from the experiment, the conclusions that can be drawn, and the discussions are summarized in Sections IV to VI.

II. Related Works

Several studies on the use of markers as anchors for pose estimation have shown that the use of multi-marker systems provides advantages, especially in terms of reducing detection ambiguity and improving the stability of jittery marker detection [12]. In another study related to the use of multi-marker [13], it is shown that marker configurations arranged in a 3D non-coplanar structure, such as in a dodecahedron-shaped object, provide improved accuracy and stability of pose estimation compared to planar configurations. This is due to the ability of the 3D configuration to provide geometric constraints from multiple orientations, thereby reducing the ambiguity that commonly occurs in coplanar arrangements, even when multiple markers are used with uniform orientation relative to the camera.

Studies that utilize markers as pose-estimation tools and apply them for 3D ultrasound reconstruction have been conducted, demonstrating strong potential. One work directly related to the fundamental development of marker-based pose estimation is the research by Wu et al. [7] in which they developed a system called DodecaPen. The system uses multiple ArUco markers arranged on a dodecahedron structure and projects the global reference points to the stylus-tip position. Their results showed that they achieved high measurement accuracy, with the lowest translation-vector error reaching 0.34 mm. However, the study did not provide a comparative evaluation between multi-marker and single-marker approaches for tip projection.

Another relevant work related to 3D ultrasound reconstruction was conducted by Léger et al. [14]. In their study, they used only a single ArUco marker attached to the ultrasound probe. To enhance marker-detection accuracy, they used a RealSense RGB camera as the marker-capturing sensor. For evaluation, they designed a phantom experiment using LEGO blocks and wires placed between them. The phantom was then submerged in water as the ultrasound medium. The reconstructed wire length was compared to the ground-truth wire length, producing errors of 2.64 mm, 1.50 mm, and 13.83 mm along the x, y, and z axes, respectively. However, the study did not evaluate reconstruction performance on small objects representing cancer-like lesions.

Marker-based pose estimation for 3D ultrasound was also implemented by De Sanctis et al. [3]. They designed a 3D ultrasound system utilizing a dodecahedron-shaped marker as the pose reference. To obtain a reference measurement for comparison, an infrared tracking system was employed. The marker and infrared tracker were positioned on opposite sides of the ultrasound beam to ensure visibility. To assess performance, they conducted reconstruction tests on phantoms constructed using 3D-printed bone structures and tissue-mimicking material. The system achieved average errors of 0.857, 0.453, and 2.689 mm along the x, y, and z axes, respectively. However, like previous works, they did not investigate reconstruction performance on small cancer-representative objects.

III. Method

In the process of 3D reconstruction using marker-based ultrasound, the stability of pose measurement through marker detection becomes crucial in determining the accuracy of the reconstruction results. Although a single marker is easy to implement, it has limitations at extreme angles. Study shows that multi-marker approaches are more robust under extreme angles [13], but they introduce additional complexity in terms of design and implementation. In this study, an experimental workflow will be conducted to evaluate the accuracy of reconstruction results by collecting quantitative data from both single-marker and multi-marker approaches. The overall workflow of the proposed experimental framework is illustrated in Figure 1.

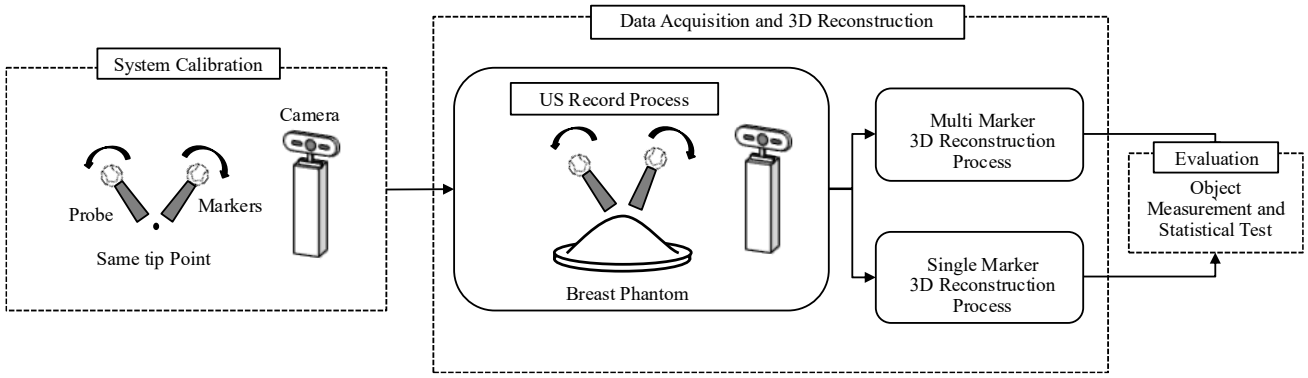
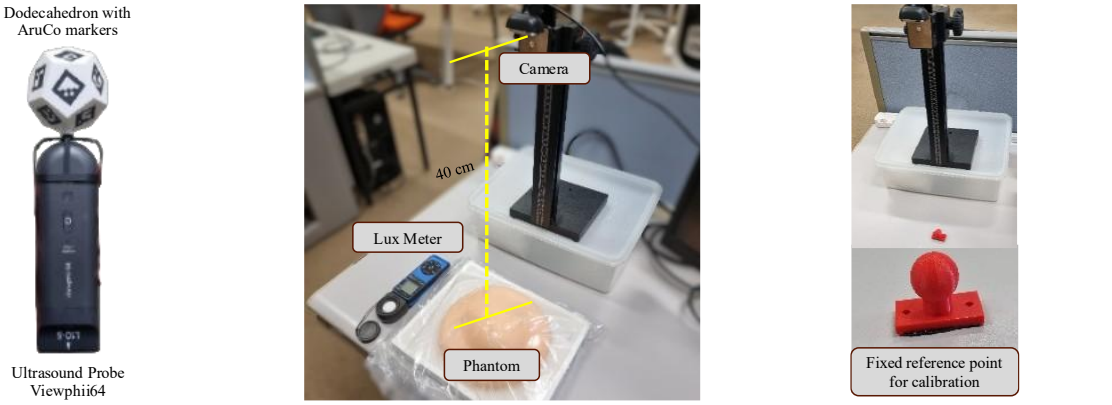


Figure 1. Overview of the proposed experimental workflow.



(a) Viewphii64 + Dodecahedron (b) Camera + Phantom + Lux Meter (c) Reference point for calibration

Figure 2. Experimental setup for freehand 3D ultrasound reconstruction.

A. Experimental Setup

The experimental system consisted of an ultrasound probe equipped with either a single ArUco marker or a dodecahedron-shaped multi-marker structure. In the single-marker configuration, one marker from the same marker set used in the multi-marker structure was selected to maintain geometric consistency between conditions. The ultrasound imaging device used in this study was the Viewphii64 system (Socionext, Japan) [15], which provides real-time B-mode imaging suitable for freehand acquisition. A monocular RGB camera [16] was mounted approximately 40 cm above the acquisition area to capture marker motion during scanning. Illumination was maintained at approximately 500 Lux to minimize detection variability caused by lighting changes.

A commercial breast phantom [17] containing spherical inclusions with known diameters of 5 mm and 10 mm was used as the reconstruction target. Ultrasound B-mode images were acquired while the camera simultaneously recorded marker motion for pose estimation. This synchronized acquisition enabled pairing each ultrasound frame with its corresponding 6-DoF pose.

The detailed hardware configuration is shown in Figure 2. As illustrated in Figure 2(a), the fiducial marker structure was rigidly attached to the probe on the side opposite to the ultrasound emission surface to avoid acoustic interference while ensuring continuous visibility from the overhead camera. Figure 2(b) presents the overall data acquisition environment, where the RGB camera was positioned vertically above the phantom to provide a stable field of view and minimize perspective distortion during tracking.

In addition, Figure 2(c) shows the fixed reference point used during the calibration procedure. This reference point was placed within the acquisition area and served as a spatial anchor for probe-tip projection validation. During calibration, the probe tip was repeatedly aligned with this fixed reference while the marker was moved in various orientations. Consistent projection of the probe-tip position to the same global coordinate confirmed proper alignment between the tracking system and the ultrasound image plane. The inclusion of this fixed reference ensured that calibration accuracy was evaluated under reproducible geometric conditions.

B. Calibration Procedure

To ensure a fair comparison between configurations, each tracking method was calibrated independently while being aligned to a common global coordinate system. The calibration process included camera intrinsic calibration, marker size calibration, and probe-tip offset calibration. The probe-tip position was defined relative to the marker coordinate system to establish a consistent spatial relationship between pose estimation and the ultrasound image plane.

Calibration accuracy was verified by projecting the probe tip to a fixed reference point in the acquisition area. Consistent projection results across repeated probe movements indicated successful calibration. This procedure ensured that any differences observed in reconstruction accuracy were attributable to the tracking configuration rather than misalignment or calibration bias.

C. Data Acquisition

Following calibration, data acquisition was conducted under controlled scanning conditions. For each configuration, ultrasound B-mode images and corresponding marker-tracking video were recorded simultaneously. Each ultrasound frame was associated with its estimated 6-DoF pose derived from marker detection.

To reduce variability in reconstruction results, probe motion was performed slowly and consistently in a single direction. By minimizing the spatial gap between consecutive frames, this approach ensures denser sampling, which is crucial for improving interpolation accuracy and overall 3D reconstruction quality [18].

D. 3D Reconstruction and Volume Measurement

Three-dimensional reconstruction was performed by placing each B-mode image frame in 3D space according to its estimated pose. All processing was conducted using 3D Slicer [19] with Python scripting to ensure reproducibility. The spherical inclusions within the phantom were segmented semi-automatically, and volumetric measurements were computed from the segmented regions.

Segmentation results were visually inspected to identify potential errors. When necessary, manual corrections were

applied to ensure measurement reliability. This hybrid segmentation approach was adopted to balance automation and accuracy.

Reconstruction accuracy was evaluated by comparing the measured volume of each inclusion with its known ground-truth volume. The volumetric error for object *i* was defined as

$$e_i = |V_{measured,i} - V_{actual,i}|$$

where $V_{measured,i}$ denotes the reconstructed volume and $V_{actual,i}$ represents the true volume. Smaller error values indicate higher reconstruction accuracy.

E. Statistical Analysis

To determine whether marker configuration and object size significantly affected reconstruction accuracy, a two-way ANOVA was performed. The independent variables were marker configuration (single-marker versus multi-marker) and object diameter (5 mm versus 10 mm), while the dependent variable was volumetric reconstruction error. Each configuration is tested with 10 repetitions, where smaller error values indicate better performance. The data collection was performed using one type of breast phantom, carried out by a single operator who does not have a medical background.

Statistical significance was defined at a threshold of $p < 0.05$. If significant main effects or interactions were observed, further analysis was conducted to interpret differences between experimental conditions. Through this structured analysis, the study systematically evaluates the influence of tracking configuration and object size on volumetric reconstruction performance.

IV. Results

A. Calibration Accuracy

Prior to reconstruction experiments, calibration accuracy was evaluated to ensure that both tracking configurations were properly aligned within the same global coordinate system. The average probe-tip projection error under approximately 500 Lux illuminations was (0.64, 0.02, 0.63) mm for the multi-marker configuration and (0.22, 0.23, 0.91) mm for the single-marker configuration.

B. Volumetric Reconstruction Accuracy

After calibration, 3D reconstruction experiments were conducted for spherical inclusions with diameters of 5 mm and 10 mm. Volumetric errors were calculated by comparing reconstructed volumes with ground-truth values. Table I shows our volumetric reconstruction errors for single-marker and multi-marker configurations.

For the 5-mm inclusions, the multi-marker configuration achieved a mean volumetric error of 12.1 mm³, whereas the single-marker configuration produced substantially larger errors with greater variability. The standard deviation was notably lower in the multi-marker condition, indicating improved stability.

For the 10-mm inclusions, reconstruction errors increased in both configurations. However, the difference between tracking methods became more pronounced. The single-marker configuration exhibited the largest mean error (132.7 mm³) and higher variability, while the multi-marker configuration maintained comparatively lower error values.

Overall, the results demonstrate that multi-marker tracking consistently reduced volumetric error and improved measurement stability across object sizes.

C. Statistical Analysis

To determine whether the observed differences were statistically significant, a two-way ANOVA was performed with marker configuration and object diameter as independent factors (Table II).

The analysis revealed a highly significant main effect of object diameter (F = 86.02, p < 0.001), indicating that reconstruction accuracy differed substantially between 5-mm and 10-mm inclusions. A significant main effect of marker configuration was also observed (F = 6.34, p = 0.016), demonstrating that tracking method significantly influenced volumetric reconstruction accuracy.

TABLE I. Volumetric Reconstruction Errors for Single-Marker and Multi-Marker Configurations

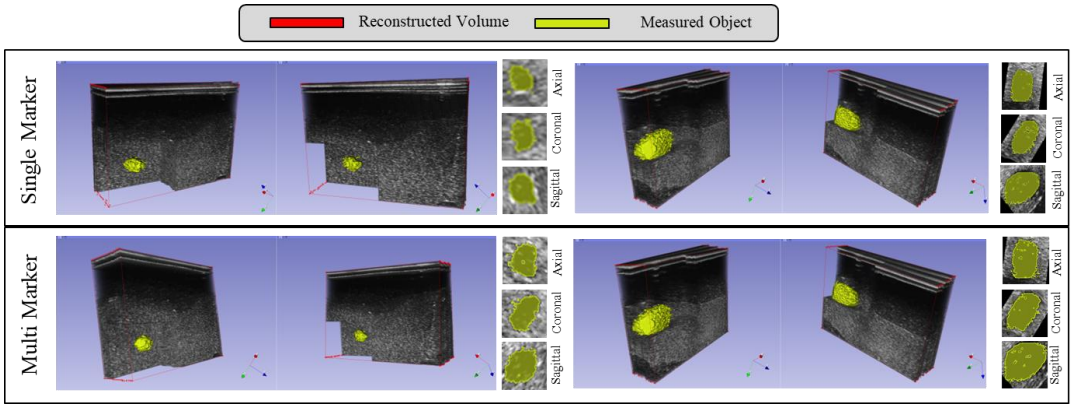
Diameter (mm)	Volume (mm³)	Error (Measured - Actual)					
		Multi Marker (mm³)			*Single Marker (mm³)*		
		e	*Avg*	*StdDev*	*e*	*Avg*	*StdDev*
5	65.5	8.9	12.1	9.7	23	18.5	9.1
		24.9			24		
		15.1			29.4		
		31.3			27.9		
		3.2			20.9		
		7.9			10.8		
		10.1			21.5		
		14.7			19.4		
		0.3			3.7		
		4.7			4.8		
10	523.6	143.2	87.6	29.1	227.6	132.7	56.2
		92.8			223.5		
		59.3			78.9		
		73.3			132.4		
		71.4			85.3		
		69.9			91.6		
		65.2			77.7		
		134.9			148.6		
		92.2			153.5		
		74.2			107.8		

TABLE II. Comparison of Volumetric Reconstruction Errors Between Tracking Configurations

Source	*DF*	*F*	*p*-value
Diameter	1	86.02	<0.001 ***
Method	1	6.34	0.016 *
Diameter × Method	1	3.57	0.067
Error	36	—	—

*: $p < 0.05$, **: $p < 0.01$ (), ***: $p < 0.001$

The interaction between object diameter and marker configuration approached significance (F = 3.57, p = 0.067) but did not reach the 0.05 threshold. This suggests that while reconstruction error increases with object size, the relative advantage of multi-marker tracking remains consistent across diameters.



(a) 5-mm spherical inclusion (b) 10-mm spherical inclusion

Figure 3. Representative 3D reconstruction results.

These statistical findings quantitatively confirm that multi-marker tracking provides measurable improvements in reconstruction accuracy.

D. Qualitative Reconstruction Comparison

Representative reconstruction results are presented in Figure 3 for both 5-mm and 10-mm spherical inclusions. For the 5-mm inclusions, both tracking configurations were able to reconstruct the general spherical structure; however, the multi-marker configuration exhibited smoother surface boundaries and more consistent spatial alignment. In contrast, the single-marker reconstruction showed minor boundary irregularities and slight positional inconsistencies, reflecting small pose fluctuations during scanning.

For the 10-mm inclusions, the qualitative differences between configurations became more apparent. The multi-marker reconstruction maintained a relatively uniform surface shape and coherent volumetric structure. Conversely, the single-marker reconstruction displayed noticeable surface distortions and irregular volume expansion, suggesting the accumulation of pose estimation errors over the scanning trajectory. These distortions are consistent with the larger quantitative errors observed in the volumetric measurements.

V. Discussion

A. Interpretation of Reconstruction Accuracy

The experimental results demonstrate that the multi-marker configuration significantly improved volumetric reconstruction accuracy compared to the single-marker configuration. This improvement can be primarily attributed to enhanced pose stability during freehand scanning. In single-marker tracking, pose estimation relies on the visibility and geometric consistency of a single planar marker. When the marker experiences unfavorable viewing angles, corner detection accuracy degrades, leading to instability in pose estimation and accumulated reconstruction error.

In contrast, the multi-marker configuration provides redundant geometric constraints. Because markers are distributed across multiple faces of a polyhedral structure, at least one marker remains visible under most viewing conditions. This redundancy stabilizes pose estimation and reduces jitters in 6-DoF tracking, an effect that is supported by the observed reduction in variability in volumetric measurements.

B. Influence of Object Size

The results also show that reconstruction error increased significantly for 10-mm inclusions, indicating greater sensitivity of larger objects to cumulative pose drift during scanning. Despite this size dependence, the multi-marker configuration consistently reduced errors, with no significant interaction effect, demonstrating robust performance across object diameters and spatial scales.

C. Engineering Implications

From an engineering perspective, pose stability is critical in freehand 3D ultrasound, as small angular errors can accumulate into significant spatial inaccuracies. Multi-marker configurations provide a low-cost, robust solution to improve tracking stability using simple hardware, making them suitable for portable and resource-limited clinical settings.

D. Clinical Relevance

Although this study was conducted using a breast phantom, the findings have implications for clinical ultrasound imaging. In breast cancer screening, accurate volumetric reconstruction may improve lesion visualization and measurement reproducibility. While automated systems such as ABUS provide standardized volumetric imaging, freehand 3D ultrasound remains attractive due to its flexibility and lower hardware complexity. Improving tracking robustness through multi-marker configurations could help bridge the gap between low-cost freehand systems and more sophisticated automated platforms.

E. Limitations

Several limitations should be considered when interpreting the results. First, the experiments were conducted under controlled lighting conditions with approximately 500 Lux illuminations. Marker detection performance may vary under different clinical lighting environments. Second, only two object sizes were evaluated, limiting the analysis of scale-

dependent reconstruction behavior. Third, segmentation included manual correction, which may introduce minor observer-dependent variability. The fourth limitation lies in the type of phantom used and the limited number of data points per testing type, which is only 10, resulting in low data variability and limiting the precision of the findings across several scenarios.

Another limitation comes from a Human-Computer Interaction (HCI) perspective, where the design places a 50 g 3D-printed model on top of the ultrasound probe shifts the center of gravity of that device. This creates a unique challenge in ensuring that the added weight does not impose extra burden or discomfort on the operator during use. Another HCI-related challenge involves maintaining the marker-equipped dodecahedron within an optimal distance of 10–15 cm from the camera to ensure stable pose estimation.

VI. Conclusion and Future Work

This study quantitatively compares single-marker and multi-marker pose estimation methods for freehand three-dimensional (3D) ultrasound reconstruction using a controlled breast-phantom experiment. The results demonstrated that multi-marker tracking significantly improves volumetric reconstruction accuracy and stability compared to single-marker tracking, with the smallest error observed for the 5-mm inclusions reaching 12.1 mm³. Similar findings related to using multiple sensor configurations can be found in other studies [20], which shows that using multiple sensors of the same type has better accuracy and stability for 3D ultrasound reconstruction.

Two-way ANOVA confirmed significant main effects of both object diameter and tracking configuration on reconstruction error, providing statistical evidence that marker arrangement directly influences volumetric accuracy. Although reconstruction errors increased for larger objects, the relative advantage of multi-marker tracking remained consistent across object sizes.

These findings highlight the importance of pose robustness in freehand 3D ultrasound systems. Even when calibration precision is comparable, dynamic tracking stability plays a critical role in reconstruction accuracy. Multi-marker configurations provide a practical and low-cost solution to enhance pose estimation without requiring expensive tracking hardware.

References

[1] P.-W. Hsu, R. W. Prager, A. H. Gee, and G. M. Treece, "Freehand 3D Ultrasound Calibration: A Review," in *Advanced Imaging in Biology and Medicine*, C. W. Sensen and B. Hallgrímsson, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 47–84. doi: 10.1007/978-3-540-68993-5_3.

[2] W. He *et al.*, "Freehand 3D Ultrasound Imaging Based on Probe-mounted Vision and IMU System," *Ultrasound in Medicine & Biology*, vol. 50, no. 8, pp. 1143–1154, 2024, doi: https://doi.org/10.1016/j.ultrasmedbio.2024.03.021.

[3] L. De Sanctis, A. Carnevale, C. Antonacci, E. Faiella, E. Schena, and U. G. Longo, "Six-Degree-of-Freedom Freehand 3D Ultrasound: A Low-Cost Computer Vision-Based Approach for Orthopedic Applications," *Diagnostics*, vol. 14, no. 14, p. 1501, Jul. 2024, doi: 10.3390/diagnostics14141501.

[4] W. He *et al.*, "Freehand 3D Ultrasound Imaging Based on Probe-mounted Vision and IMU System," *Ultrasound in Medicine and Biology*, vol. 50, no. 8, pp. 1143–1154, Aug. 2024, doi: 10.1016/j.ultrasmedbio.2024.03.021.

[5] C. A. Adriaans, M. Wijkhuizen, L. M. Van Karnenbeek, F. Geldof, and B. Dashtbozorg, "Trackerless 3D Freehand Ultrasound Reconstruction: A Review," *Applied Sciences*, vol. 14, no. 17, p. 7991, Sep. 2024, doi: 10.3390/app14177991.

[6] P. García-Ruiz, F. J. Romero-Ramirez, R. Muñoz-Salinas, M. J. Marín-Jiménez, and R. Medina-Carnicer, "Fiducial Objects: Custom Design and Evaluation," *Sensors*, vol. 23, no. 24, p. 9649, Dec. 2023, doi: 10.3390/s23249649.

[7] P.-C. Wu *et al.*, "DodecaPen: Accurate 6DoF Tracking of a Passive Stylus," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, Québec City QC Canada: ACM, Oct. 2017, pp. 365–374. doi: 10.1145/3126594.3126664.

[8] W. A. Berg, E. A. Rafferty, S. M. Friedewald, C. B. Hruska, and H. Rahbar, "Screening Algorithms in Dense Breasts: *AJR* Expert Panel Narrative Review," *American Journal of Roentgenology*, vol. 216, no. 2, pp. 275–294, Feb. 2021, doi: 10.2214/AJR.20.24436.

[9] I. Boca (Bene), A. I. Ciurea, C. A. Ciortea, and S. M. Dudea, "Pros and Cons for Automated Breast Ultrasound (ABUS): A Narrative Review," *JPM*, vol. 11, no. 8, p. 703, Jul. 2021, doi: 10.3390/jpm11080703.

[10] B. Wilczek, H. E. Wilczek, L. Rasouliyan, and K. Leifland, "Adding 3D automated breast ultrasound to mammography screening in women with heterogeneously and extremely dense breasts: Report from a hospital-based, high-volume, single-center breast cancer screening program," *European Journal of Radiology*, vol. 85, no. 9, pp. 1554–1563, Sep. 2016, doi: 10.1016/j.ejrad.2016.06.004.

[11] M. L. Giger *et al.*, "Automated Breast Ultrasound in Breast Cancer Screening of Women With Dense Breasts: Reader Study of Mammography-Negative and Mammography-Positive Cancers," *American Journal of Roentgenology*, vol. 206, no. 6, pp. 1341–1350, Jun. 2016, doi: 10.2214/AJR.15.15367.

[12] G. Čepon, D. Ocepek, M. Kodrič, M. Demšar, T. Bregar, and M. Boltežar, "Impact-Pose Estimation Using ArUco Markers in Structural Dynamics," *Exp Tech*, vol. 48, no. 2, pp. 369–380, Apr. 2024, doi: 10.1007/s40799-023-00646-0.

[13] P. Oščádal *et al.*, "Improved Pose Estimation of Aruco Tags Using a Novel 3D Placement Strategy," *Sensors*, vol. 20, no. 17, p. 4825, Aug. 2020, doi: 10.3390/s20174825.

[14] É. Léger, H. E. Gueziri, D. L. Collins, T. Popa, and M. Kersten-Oertel, "Evaluation of Low-Cost Hardware Alternatives for 3D Freehand Ultrasound Reconstruction in Image-Guided Neurosurgery," in *Simplifying Medical Ultrasound*, vol. 12967, J. A. Noble, S. Aylward, A. Grimwood, Z. Min, S.-L. Lee, and Y. Hu, Eds., in Lecture Notes in Computer Science, vol. 12967. , Cham: Springer International Publishing, 2021, pp. 106–115. doi: 10.1007/978-3-030-87583-1_11.

[15] "ViewPhii64." [Online]. Available: https://viewphii.com/viewphii64/about/index.html 2026.04.14

[16] "Microsoft LifeCam HD-3000." [Online]. Available: https://www.microsoft.com/en-au/d/lifecam-hd-3000/8Q49LGBW0R58/?activetab=pivot:overviewtab 2026.04.14

[17] "Breast Phantom from OST Coorporation." Accessed: Apr. 14, 2026. [Online]. Available: https://www.ost-jp.com/

[18] Q. Huang, M. Lu, Y. Zheng, and Z. Chi, "Speckle suppression and contrast enhancement in reconstruction of freehand 3D ultrasound images using an adaptive distance-weighted method," *Applied Acoustics*, vol. 70, no. 1, pp. 21–30, Jan. 2009, doi: 10.1016/j.apacoust.2008.02.002.

[19] "3D Slicer." [Online]. Available: https://www.slicer.org 2026.04.14

[20] M. Luo *et al.*, "Multi-IMU with Online Self-consistency for Freehand 3D Ultrasound Reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, vol. 14220, H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, Eds., in Lecture Notes in Computer Science, vol. 14220. , Cham: Springer Nature Switzerland, 2023, pp. 342–351. doi: 10.1007/978-3-031-43907-0_33.

Memory-Driven Person ReID for Identity Consistency in Multi-Object Tracking

Tista Pal, Trinh Quoc Nguyen and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, I*wate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
Email: s231x018@s.iwate-pu.ac.jp, g236v201@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—Many traditional Multi-Object Tracking methods primarily emphasize detection accuracy and short-term trajectory continuity, often overlooking long-term identity consistency, which is crucial for robust person Re-Identification (ReID). This paper presents a ReID focused Multi Object Tracking (MOT) framework designed to improve long-term identity preservation through memory-driven embedding refinement rather than detector-centric enhancements. The proposed framework focuses on a ReID-focused MOT framework aimed at improving identity preservation across extended temporal spans. The framework integrated a You Only Look Once (YOLO)-based detector, DeepSORT for motion-aware association, and a Global Identity Memory module that maintains and refines identity embedding over time through memory driven fusion. In addition, a Filtered IDF1 metric is proposed to evaluate identity consistency by focusing solely on detected instances, providing a fairer assessment of long-term identity retention. To investigate the impact of feature extraction quality, representative backbones, OSNet and ResNet50, are evaluated independently under identical MOT17 benchmark conditions. Experimental results demonstrate that the proposed GlobalID framework consistently improves identity retention across different feature extractors, demonstrating that segmentation based embedding refinement combined with memory driven fusion effectively enhances robust, identity-consistent tracking in surveillance and autonomous systems.**

Keywords-Person ReID; Multi-Object Tracking; Identity Consistency; GlobalID; ResNet50; OSNet; DeepSORT.

I. INTRODUCTION

Person tracking with Re-IDentification (ReID) aims to maintain consistent identity labels for individuals across video frames, and across camera views, forming a crucial component in surveillance systems, public safety systems and intelligent transportation systems [1]. In this work, we focus on single camera MOT, where the goal is to detect individuals and preserve their identities over time within a single continuous video stream.

Maintaining identity consistency remains challenging due to real-world visual variability. Changes in illumination, pose, orientation, occlusion and camera motion can break the appearance continuity of a person, leading to misidentification when they reappear after temporary disappearance [2]. Moreover, individuals who are wearing similar clothes introduce ambiguity in feature associations, often resulting in ID switches or identity merging [3]. These failure cases highlight the need for robust appearance modelling and reliable identity preservation mechanisms in single camera MOT systems.

Many traditional tracking systems, including Kalman filter-based motion models [4] and appearance-augmented approaches, such as SORT [5] and DeepSORT [6] have made progress in reducing fragmented trajectories. However, these methods typically rely on frame-wise feature matching, lacking mechanisms to maintain long-term identity memory or adaptively update identity embeddings. Consequently, they struggle to preserve stable identity assignments under dynamic conditions. Recent research has shown that deep feature embeddings play a crucial role in identity consistency [7]. Networks, such as ResNet50 [8] and Omni-Scale Network (OSNet) [9] have demonstrated strong discriminative power in static ReID benchmarks, but their performance within continuous tracking scenarios remains underexplored.

To address these limitations, this study focuses on improving long-term identity consistency in ReID-enhanced multi-object tracking. We hypothesize that the stability and discriminative capacity of feature embeddings are key factors in maintaining consistent identity assignment over time. Accordingly, our framework introduces three key innovations. First, we employ robust person detection and appearance feature extraction to obtain identity embeddings suitable for temporal refinement. Second, we propose a Global Identity Memory (GlobalID) module that persistently updates identity embeddings using Exponential Moving Averages (EMA), inspired by memory-based representation learning in unsupervised ReID [10], enabling adaptation to gradual appearance changes and reducing identity fragmentation. Third, we integrate cosine similarity, EMA-based embedding fusion, and Intersection over Union (IoU)-based motion cues into a unified association cost to improve robustness under occlusion and motion drift.

Additionally, we investigate segmentation-guided embedding refinement as an auxiliary enhancement to analyze its impact on identity preservation. Experimental results indicate that comparable identity consistency can be achieved without segmentation when memory-driven fusion is applied, highlighting the dominant role of global identity memory in long term tracking.

To more accurately evaluate identity stability in tracking, we introduce a Filtered IDF1 metric, designed to isolate the effect of identity association mechanisms from detection errors. Unlike the standard IDF1 score, which is influenced by missed detections and false positives, the proposed metric evaluates only the successfully detected instances, providing a more interpretable measure of temporal identity consistency. Comprehensive experiments on the MOT17 benchmark, we demonstrate that these enhancements effectively reduce ID switching and surpass the accuracy of conventional DeepSORT-based systems.

 ISBN: 978-1-68558-383-5

The remainder of this paper is organized as follows. Section II reviews related work on multi-object tracking, person re-identification, and memory-based tracking methods. Section III describes the proposed framework, including the object detection, feature extraction, and GlobalID module. Section IV presents experimental results and evaluation on the MOT17-Scale-Dependent Pooling (SDP) benchmark. Section V concludes the paper and outlines directions for future work.

II. RELATED WORK

A. Evolution of Multi-Object Tracking

Early MOT primarily relied on handcrafted features and motion-based prediction models, such as Kalman filters [4] and optical flow [11] to estimate object trajectories across frames. While computationally efficient, these approaches were highly sensitive to occlusion, camera motion, and appearance changes, often producing fragmented trajectories and frequent identity switches which reduced their reliability in crowded or dynamic scenes.

The introduction of Simple Online and Real-time Tracking (SORT) [5] and later DeepSORT [6] marked a significant milestone in the evolution of MOT. SORT employed Kalman filtering with bounding-box IoU association, achieving impressive speed but limited identity preservation. DeepSORT improved upon this by integrating a deep appearance descriptor trained for ReID, enabling the tracker to associate detections using both motion and visual similarity. This enhancement substantially reduced ID switches and improved long-term association stability.

In parallel, object detection frameworks such as YOLO based architecture [12] [13] have gained popularity for their high accuracy and real time performance. This framework achieves high precision and frame-rate efficiency through single-stage detection pipelines. Segmentation variants further improved localization by generating pixel-level masks that effectively suppress background interference and improving feature extraction for ReID integration. However, segmentation incurs additional computational cost and its effectiveness in improving long-term identity preservation within MOT pipelines remains an open research question.

B. Advances in ReID

ReID plays a central role in enhancing tracking reliability by providing appearance-based cues for identity matching. Early Convolutional Neural Network (CNN)-based ReID models, such as ResNet-50 [8], focused on learning global appearance features, offering a robust baseline for visual representation. However, such models often struggled with fine-grained local variations, such as changes in pose or partial occlusion. Recent architecture has introduced multi-scale or part-aware learning to overcome these limitations. OSNet [9] efficiently captures both local fine-grained and global structural information, achieving strong performance with minimal computational overhead and making it suitable for real-time tracking contexts. Other studies, such as PCB (Part-based Convolutional Baseline) [3] and Multi-Granularity Network (MGN) [14] emphasize structured feature decomposition to better handle pose and viewpoint variations. Additionally, Transformer-based ReID models [15] have recently shown promise in modeling long-range dependencies and improving context awareness across scenes.

Despite these advancements, most ReID models are trained and evaluated in static conditions (e.g., Market-1501, DukeMTMC, MSMT17) and are not directly optimized for temporal identity consistency within continuous video sequences. When integrated into tracking pipelines, they still face challenges with dynamic background interference, motion blur, and lighting fluctuations.

C. Embedding-Driven and Memory-Based Tracking

In recent years, research has shifted toward embedding-driven and memory-based MOT frameworks [3], [10]. These approaches accumulate temporal embeddings to maintain consistency across frames, allowing for adaptive feature matching that extends beyond immediate temporal windows. For example, Tracktor++ [16] and FairMOT [17] integrate detection and ReID into joint frameworks, improving both tracking precision and speed. Similarly, Joint Detection and Embedding (JDE) [18] introduced end-to-end training for simultaneous detection and embedding extraction, enabling efficient real-time inference.

However, even with these advances, most existing systems emphasize short-term association and lack explicit mechanisms to ensure long-term identity preservation. Few studies address global identity management, where accumulated embeddings are updated or refined dynamically over time to mitigate ID drift caused by gradual appearance changes or occlusion. Approaches such as self-adaptive galleries [19] or open-world ReID memory systems [20] have made progress toward continuous identity learning but remain limited in maintaining stable embedding representations within unified MOT pipelines.

D. Gap in Literature

While combining object detection and ReID embeddings has improved online MOT performance, existing frameworks still lack mechanisms to preserve identity coherence over longer temporal spans. Identity inconsistency typically arises when individuals undergo pose or orientation changes, partial occlusions, or lighting variations, leading to repeated ID fragmentation. Only a limited number of studies explicitly incorporate a feature embedding-based global identity memory that evolves over time and actively guides association decisions.

To address this gap, we propose an Integrated Global Identity Memory (GlobalID) that dynamically updates identity embeddings through EMA while applying similarity and IoU thresholds to ensure stable associations. Unlike conventional short-term embedding buffers, GlobalID provides a persistent, adaptive memory structure that bridges local frame-level tracking and long-term identity preservation. Implemented within a YOLOv8 + DeepSORT + OSNet/ResNet50 pipeline, the proposed framework effectively reduces ID switching and enhances overall tracking reliability in dynamic visual environments.

III. METHODOLOGY

This section describes the design of the embedding-driven, ReID-enhanced MOT framework, which emphasizes long-term identity consistency through memory-based embedding refinement and global identity management. The framework supports two operating configurations, with and without segmentation-based embedding refinement allowing systematic analysis of the impact of feature isolation versus computational efficiency.

A. Framework Overview

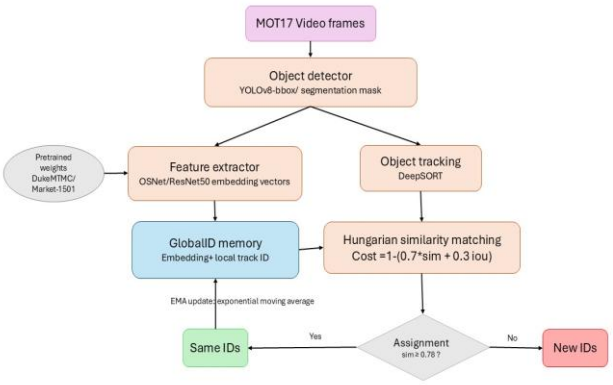


Figure 1. Proposed Framework.

The proposed framework unifies object detection, feature extraction, MOT, and ReID to study how feature quality and identity-embedding management influence long-term ID preservation. Its goal is to assess how embedding refinement and global ID strategies affect MOT performance under challenging conditions such as occlusion, illumination changes, and viewpoint variation [1].

As shown in Figure 1, the pipeline includes a YOLOv8 detector (two configurations are compared: segmentation-based setup that masks background regions before feature extraction, and a detection-based setup that uses raw bounding boxes), DeepSORT for short-term association, ReID backbones (OSNet and ResNet50) for embedding extraction, and a Global Identity Memory (GlobalID) module that maintains long-term consistency using exponential moving averages and adaptive cosine matching.

Overall, the framework provides a controlled way to analyze how embedding refinement and global identity memory contributes to stable ID assignments in realistic MOT scenarios.

B. Dataset

Experiments were conducted on the MOT17 dataset [21]-[23], which consists of multiple pedestrian video sequences captured under varying illumination, crowd density, and camera motion conditions. Each sequence provides ground-truth bounding boxes and identity annotations in the MOTChallenge format. The selected subset - MOT17-02-SDP, MOT17-04-SDP, MOT17-09-SDP, MOT17-10-SDP, and MOT17-11-SDP - these five sequences were selected to represent a diverse range of environmental and motion conditions. MOT17-02 and MOT17-04 feature static cameras with high crowd density, MOT17-09 and MOT17-11 involve low-to-medium density scenes with moderate occlusion, and MOT17-10 includes a moving camera with dynamic viewpoint shifts. Together, they provide a balanced evaluation of identity preservation under varied real-world challenges without introducing redundant or overlapping conditions.

This benchmark is widely used for MOT evaluation due to its complexity and standardization, allowing fair comparison with prior works such as DeepSORT [6] and FairMOT [17].

C. Object Detection and Segmentation

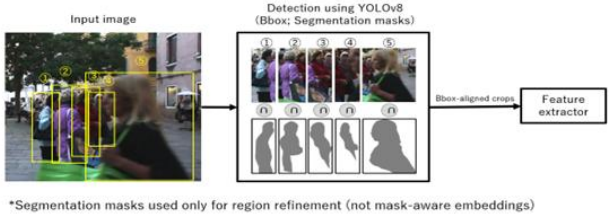


Figure 2. Person detection and segmentation using YOLOv8.

Person detection was performed using YOLOv8 architecture (Figure 2), with two configurations evaluated to study the impact of region refinement on embedding quality. For the segmentation-based configuration, YOLOv8-seg was employed to generate instance-level masks alongside bounding boxes. The segmentation masks were used to isolate person regions by suppressing background pixels before feature extraction. In contrast, the detection only configuration uses standard YOLOv8 bounding boxes directly for feature extraction, offering significantly faster inference. For both the configurations, only class 0 (person) detections were retained, using a confidence threshold of 0.3 and an IoU threshold of 0.6.

D. Feature extraction and ReID

Two representative ReID backbones, ResNet50 [8] and OSNet [9], were employed to extract feature embeddings. Each was initialized with pretrained weights from large-scale ReID datasets (Market-1501 and DukeMTMC), ensuring strong feature generalization. Person crops were resized (256×128 for OSNet, 224×224 for ResNet50) and normalized using ImageNet mean-std statistics and encoded into L2-normalized feature vectors (512-D and 2048-D, respectively)

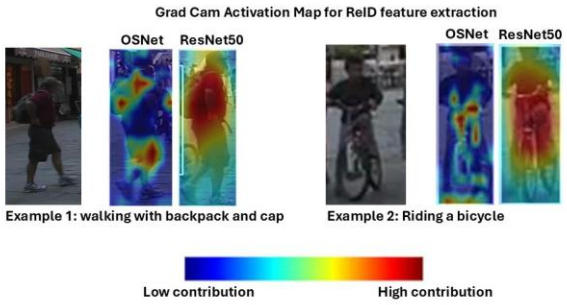


Figure 3. Activation map of features extracted by OSNet and ResNet50.

for cosine-similarity matching. These embeddings are stored in memory and on disk.

Using both backbones allows a controlled comparison between lightweight omni scale features and deeper global representations. To further examine their behavior, Grad-CAM visualization highlights the spatial regions contributing to each model's embeddings. As shown in Figure 3, OSNet attends to multiple fine-grained body regions, whereas ResNet50 focuses more on global cues such as the silhouette and torso.

E. MOT with DeepSORT

DeepSORT [6] was employed as the short-term tracking component, combining Kalman-filter-based motion prediction with appearance-based matching. Although DeepSORT effectively reduces short-term identity switches, it relies on limited temporal embedding buffers and is prone to identity drift under prolonged occlusion. In proposed framework, DeepSORT generates local track identities, which are subsequently refined and stabilized by the GlobalID module to achieve sequence-wide identity consistency. Tracking parameters used in this study are summarized in Table I.

TABLE I. DEEPSORT TRACKING PARAMETERS USED IN THIS STUDY

Parameter	Value	Description
Max_age	30	Max. No. of frames to keep a lost track alive.
nn_budget	200	Max. size of the appearance descriptor gallery.
Max_cosine_distance	0.2	Threshold for matching appearance embeddings.
Max_iou_distance	0.7	IoU threshold for matching.
n_init	3	No. of consecutive detections before confirming a new track.

F. GlobalID

To maintain identity consistency across sequences and handle occlusions or re-entries, Global Identity Memory (GlobalID) was developed. Unlike DeepSORT's limited local gallery, GlobalID functions as a persistent global memory that stores and updates identity embeddings throughout the sequence. Each new detection is compared against stored identity embeddings using cosine similarity, defined as:

$$cos(\theta) = (a \cdot b)/(\|a\|\|b\|) \quad (1)$$

Where, a and b are two feature vectors representing embeddings of the current and stored detections, respectively.

To adapt to gradual appearance changes, embeddings are updated using EMA:

$$f_t = (1-\alpha)f_{\{t-1\}} + \alpha f_{new}, \ \alpha = 0.95 \quad (2)$$

where the most recent embedding is represented by f_{new} and f_t is the updated smoothed feature vector. The chosen smoothing factor is α.

Additionally, for each identity, a maximum of 20 feature vectors were retained to prevent memory overflow and reduce noise accumulation. New identities were only confirmed after appearing consistently for 5 consecutive frames, introducing a hysteresis effect that suppresses fake ID creation.

This memory driven embedding enables the system to maintain coherent identities across long-term sequences, significantly reducing ID fragmentation and false associations compared with conventional trackers [18]-[20].

Figure 4 demonstrates the working of the GlobalID module.

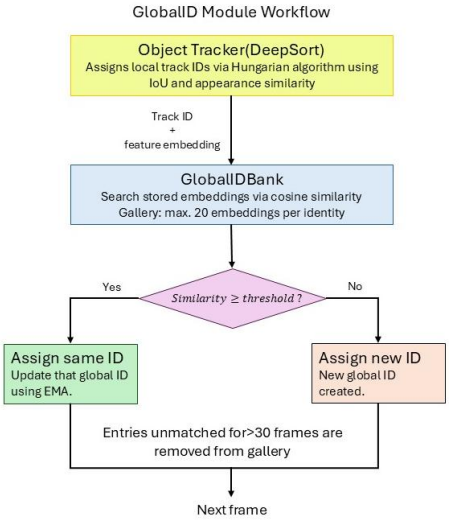


Figure 4. GlobalID Module.

IV. RESULTS

All experiments were conducted on sequences from the MOT17 dataset, which provides diverse real-world tracking scenarios involving frequent occlusions, appearance changes, and dynamic camera motion. We report both the standard MOTChallenge metric (IDF1) [22] and the proposed Filtered IDF1(F-IDF1), which evaluates identity consistency exclusively among successfully detected instances. This dual evaluation enables a clearer distinction between detection accuracy and identity-preservation capabilities.

A. Filtered IDF1 metric

Identity association performance in multi-object tracking is commonly evaluated on Association Accuracy (AssA), which measures how effectively a tracker preserves object identities across frames [22], [23]. This metric is represented by the IDF1 score, the harmonic mean of identity precision and recall, defined as:

$$F-IDF1 = (2 \times IDTP\_det) / (2 \times IDTP\_det + IDFP\_det + IDSW\_det) \quad (3)$$

where the subscript det indicates frames with valid detections, and $IDSW$ quantifies identity changes among detected objects.

This refinement provides a more focused view of ReID and memory integration effects, particularly in frameworks where detection quality is already saturated by high-performance detectors such as YOLOv8 [24].

(a) Baseline (YOLOv8 + DeepSORT) (b) The proposed method

Figure 5. The proposed method should have better consistency, not more switches.

B. Results on MOT17-SDP

To assess the contribution of appearance embeddings and the GlobalID module, pretrained ReID backbones (OSNet_x1_0 and ResNet50) were integrated into the YOLOv8 + DeepSORT pipeline. Each backbone was evaluated using Market-1501 [25] and DukeMTMC [22] pretrained weights, under two configurations: segmentation based embedding refinement and detection only embedding refinements. As a baseline, the standard YOLOv8 + DeepSORT configuration was evaluated using DeepSORT's built-in CNN appearance descriptor, without any external ReID backbone or GlobalID module. This represents the conventional tracking approach that our framework aims to improve upon.

Across both configurations, integrating ReID backbones with GlobalID consistently improved identity preservation on MOT17-SDP (Table II). OSNet-DukeMTMC achieved the strongest results, reaching IDF1 $\approx$ 0.39 and F-IDF1 $\approx$ 0.66 - roughly a $25-35\%$ improvement over the baseline - demonstrating that refined embeddings and global memory updates reduce ID fragmentation. F-IDF1 further shows that identity association becomes more reliable even when detection quality remains unchanged, as illustrated in Figure 5.

The segmentation assisted configurations introduced a substantial computational overhead, reducing inference speed to $1.38-2.46$ FPS. This bottleneck arises primarily from two sources, the additional forward pass required by YOLOv8-seg to generate instance masks, and the per-frame background suppression applied before feature extraction. Importantly, the ReID backbones themselves were not designed for mask-aware inputs, meaning the segmentation step adds cost without being fully exploited by the downstream feature extractors. Without segmentation, both backbones achieve significantly faster speeds ($5.73-7.13$ FPS), representing a more practical operating point for real-time applications.

TABLE II. PERFORMANCE COMPARISON ON THE MOT17-SDP BENCHMARK

Backbone	Pretrained weights	Detection Mode	Performance comparison		
			IDF1	*F-IDF1*	*FPS*
DeepSORT CNN	-	YOLOv8	0.2916	0.6412	8.51
OSNet	DukeMTMC	YOLOv8-seg	0.3918	0.6571	1.38
OSNet	Market-1501	YOLOv8-seg	0.3871	0.6583	1.44
ResNet50	DukeMTMC	YOLOv8-seg	0.3479	0.6354	2.23
ResNet50	Market-1501	YOLOv8-seg	0.3479	0.6354	2.46
OSNet	DukeMTMC	YOLOv8	0.3783	0.6514	7.13
OSNet	Market-1501	YOLOv8	0.3864	0.6531	6.55
ResNet50	DukeMTMC	YOLOv8	0.3084	0.6102	6.03
ResNet50	Market-1501	YOLOv8	0.3140	0.6203	5.73

V. CONCLUSIONS AND FUTURE WORKS

This study introduced a Re-ID-focused multi-object tracking framework that integrates YOLOv8 detection (with optional segmentation), DeepSORT association, and pretrained ReID backbones (OSNet and ResNet50) for appearance embedding. The key contribution, the GlobalID Memory module, provides a persistent, memory-driven identity refinement mechanism that maintains consistent identities across frames through EMA fusion and cosine similarity. Experiments on the MOT17-SDP benchmark demonstrate consistent improvements over the YOLOv8 + DeepSORT baseline in both IDF1 and the proposed Filtered IDF1 metric, validating the effectiveness of ReID-driven association independent of detection quality. OSNet-DukeMTMC achieved the best identity consistency (IDF1 = 0.3918, F-IDF1 = 0.6571), representing a $25-35\%$ improvement over the baseline. A few conclusions can be drawn First, segmentation-based embedding refinement improves identity consistency by suppressing background noise, but its benefit is constrained when using backbones pretrained on bounding box crops rather than mask-aware inputs, the feature extractors were not trained to exploit the cleaner segmented regions. Second, the EMA smoothing factor ($\alpha = 0.95$) proved effective for gradual appearance adaptation. Third, the gallery confirmation threshold of five consecutive frames successfully suppressed false identity

creation in crowded scenes but introduced a slight delay in registering fast moving individuals who briefly exit and re-enter the frame.

The primary limitation of the current framework is its computational cost. Segmentation assisted configurations operate at only $1.38 - 2.46$ FPS, making real-time deployment impractical. Additionally, the framework was evaluated on a subset of five MOT17-SDP sequences, and generalization to other benchmarks or camera setup remains to be verified.

Future work will focus on three directions. First, integrating mask aware ReID architectures to better exploit segmentation cues. Second, exploring lightweight architectures and embedding alternatives to reduce inference overhead to achieve real-time performance. Third, extending the GlobalID module to multi-camera settings, which represents a natural and challenging extension of the current framework.

REFERENCES

[1] W. Luo, J. Xing and X. Zhang, "Multiple object tracking: A review," arXiv preprint arXiv:1409.7618, 2014.

[2] H. Wang, S. Ullah, D. Li and Y. Liu, "Recent advances in deep learning-based person re-identification," Applied Sciences, vol. 9, no. 8, p. 1535, 2019.

[3] Y. Sun, L. Zheng, Y. Yang, Q. Tian and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in Proc. European Conf. Computer Vision (ECCV), 2018, pp. 269-286.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering, vol. 82, no. 1, pp. 35-45, 1960.

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2016, pp. 2956-2960.

[6] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, 2017, pp.3645-3649.

[7] Z. Zhang, L. Sun, Q. Leng and S. Liao, "Towards real-time multi-object tracking with adaptive appearance models," Pattern Recognition Letters, vol. 136, pp. 213-220, 2020.

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp.770-778.

[9] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-scale feature learning for person re-identification," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019, pp.3701-3711.

[10] Y. Li, X. Zhu and S. Gong, "Unsupervised person re-identification with stochastic training strategy," IEEE Trans. Image Process., vol. 31, pp. 4240–4250, 2022.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 1981, pp.674-679.

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp.779-788.

[13] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[14] G. Wang, Y. Yuan, X. Chen, J. Li and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2018, pp.274-282.

[15] S. He, H. Luo, P. Wang, F. Wang, H. Li and W. Jiang, "TransReID: Transformer-based object re-identification," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2021, pp.15013-15022.

[16] P. Bergmann, T. Meinhardt and L. Leal-Taixé, "Tracking without bells and whistles," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019, pp.941-951.

[17] Y. Zhang, C. Wang, X. Wang, W. Zeng and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," Int. J. Comput. Vis., vol. 129, no. 11, p. 3069–3087, 2021.

[18] Z. Wang, L. Zheng, Y. Liu, Y. Li and S. Wang, "Towards real-time multi-object tracking," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 107–122.

[19] L. Jin, Z. Zheng and Y. Sun, "Learning a self-adaptive gallery for unsupervised person re-identification," IEEE Trans. Image Process., vol. 31, p. 5282–5294, 2022.

[20] Z. Zheng, L. Zheng and Y. Yang, "Open-world person re-identification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, p. 2630–2647, 2021.

[21] A. Milan, L. L. Taixé, I. Reid, S. Roth and K. Schindler, "MOT16: A benchmark for multi-object tracking," arXiv spreprint arXiv:1603.00831, 2016.

[22] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. European Conf. Comput. Vis. (ECCV), 2016, pp. 17–35.

[23] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. L. Taixé and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," Int. J. Comput. Vis., vol. 129, p. 548–578, 2021.

[24] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116-1124.

Vision-Based Estimation of PM2.5 from Surveillance Images

Dipti Mitra and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, I*wate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
email: s231x025@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—The paper proposes a vision-based approach for measuring fine Particulate Matter (PM2.5) concentrations by utilizing environmental images as input. A dataset was created by acquiring surveillance camera-captured images from multiple locations in Japan, forming a pair of collected outdoor images and Ground Truth (GT) PM2.5 observation data obtained from monitoring stations. Two preprocessing steps (image dehazing and semantic segmentation) were used to enhance prediction accuracy under varying atmospheric and meteorological conditions. The dehazing method mitigates visual degradation caused by haze, while semantic segmentation extracts and determines object-level information and the coverage amount of extracted objects relevant to PM2.5 estimation. The proposed image-based system combines the dehazed images and segmentation masks, which are then input into a deep learning-based regression model to predict PM2.5 concentrations. The experimental results demonstrate that integrating dehazed images and segmentation masks reduces prediction errors and produces more consistent estimates compared to using original images and other input configurations alone or in combination. The findings indicate that combining enhanced visual representations and segmented objects with deep learning models can serve as an effective and scalable complement to traditional air quality monitoring systems.**

Keywords-computer vision; dehazing; semantic segmentation; PM2.5 forecasting; regression.

I. INTRODUCTION

Despite substantial improvements in air quality over recent decades, air pollution remains a crucial environmental threat and public health issue in Japan. Rapid industrialization and urban development in the postwar period led to severe air pollution problems. However, strict environmental rules and continuous monitoring have significantly mitigated many conventional pollutants. Nevertheless, an air pollutant, fine Particulate Matter (PM2.5), defined as particles with an aerodynamic diameter of less than 2.5 μm, continues to pose a concern, particularly in urban and industrial areas [1]. Particle components are generally classified as solid and liquid substances released into the air from domestic sources, such as indoor activities, industrial emissions, volcanic eruptions, and others. These particles are transmitted through the air and eventually return to the ground [2]. PM2.5 is small enough to enter the human red blood cell, which is 7.8 μm in diameter, and as a result, it can penetrate deeply into the respiratory system. It is linked to adverse health effects, including cancer, cardiovascular diseases, and asthma. According to the World Health Organization (WHO), seven million people die as o result of PM2.5 particles each year globally [2]. Therefore, it is essential to explore various methods to accurately measure PM2.5 particles. Currently, computer vision technologies are being applied due to their high spatial coverage, scalability, sensitivity to changes in visibility, real-time capabilities, easy deployment, and low cost.

Japan is an island country, where ambient PM2.5 concentrations are routinely monitored using ground-based air quality monitoring stations across its forty-seven prefectures. These stations rely on PM2.5 sensors that measure concentrations of particles. Laser scattering and Infrared Rays (IR) are used to detect particles, which provide accurate point-based observations. Nevertheless, the monitoring stations within each prefecture limit their spatial representativeness as a single or a small number of sensors cannot capture the PM2.5 spatial variability over enormous and diverse geographic locations for detection. Although deploying additional sensors could improve coverage, such an approach is costly and logistically demanding. The sensors cannot provide accurate reading due to different factors and small sensitivity. Therefore, they need to be estimated precisely in different domains of each prefecture as they are scattered around the air.

In this paper, vision-based PM2.5 forecasting is proposed. To identify particulates in the air, we have utilized geographical images of Japan's various prefectures as input. We have performed image dehazing and semantic segmentation, and an AI technique called the regression method to accurately estimate PM2.5 concentration, especially in the presence of haze and objects in the scene. Regression approaches have been applied to visual input to model the relationship between environmental image features and their associated PM2.5 values. When images are used as inputs, the model extracts visual information, reflecting haze density, changes of visibility, contrast degradation, and color attenuation, and regresses these features to predict numerical PM2.5 concentrations. Moreover, it enables quantitative estimation of particulate matter levels, making it suitable for continuous monitoring and forecasting tasks rather than categorical classification. Experiments have been conducted using original images, dehazed images, and segmented objects. A regression model is used to examine the actual relationship between images/masks and their corresponding PM2.5 values. The testing results demonstrate lower correlation in terms of original images, dehazed images, semantic segmentation, and combined input, and an improved correlation when applying dehazed and object segmentation.

The paper is divided into five sections. In Section I, an overview of PM2.5, along with the research problems and objectives, is presented. In Section II, the existing literature on vision-based PM2.5 estimation and prediction approaches is reviewed. In Section III, the proposed methodology, including data collection, dataset construction, pre-processing steps, and the architecture of the regression model, is described. In Section IV, the experimental results are presented, and their implications on PM2.5 prediction are discussed. Lastly, in Section V, the paper is concluded and outlines the directions for future work.

II. LITERATURE REVIEW

Numerous research papers have applied vision-based approaches for forecasting PM2.5 air pollutants as a complementary alternative to traditional sensor-based monitoring systems. Previous studies demonstrated the viability of retrieving air quality information directly from the visual cues present in environmental scenes. For example, the vision-based techniques [3] and [4] rely on extracting haze-related features by quantifying saturation, contrast degradation, color attenuation, and information losses from manually designed statistical images to represent the haze-related visual degradation. These methods provide interpretability and relatively low computational cost. However, their dependence on handcrafted image features limits their robustness in complex outdoor environments under varying illumination, lighting, meteorological conditions, and camera conditions.

With the rapid development of deep learning technologies, researchers started adopting Convolutional Neural Network-based architectures to automatically learn discriminative visual characteristics from the images. Studies, such as a hybrid architecture Deep Neural Network model [5], presented a base model named Convolutional Neural Network (CNN), while an output layer named Long Short-Term Memory (LSTM) network was employed. They demonstrated that the two integrated models can automatically capture spatial features and temporal dependencies within the extracted feature sequences, such as pollution-related patterns, by utilizing the hourly images of the sky and surrounding environment in Bangkok, Thailand. Similarly, another approach, deep learning-based image analysis to predict PM2.5 concentrations [6], leverages existing surveillance infrastructure to achieve wide-area monitoring. In this study, a ResNet-based image analysis method was utilized, turning an existing traffic camera into a PM2.5 sensor. To create a dataset, hourly traffic images and their PM2.5 values were attained over a six-month period from a traffic camera and the nearest monitoring station. In the first phase, the neural network model ResNet50 was used to train the acquired dataset. Moreover, a second phase model, Random Forest, was used, where the outputs of the neural network are utilized as input to predict overall hourly PM2.5 values. While these CNN-based approaches enhance prediction accuracy, they remain sensitive to scene-specific factors, such as camera viewpoints, background complexity, and atmospheric visibility, which can degrade model performance across locations.

Several studies proposed time-series modeling methods to address temporal dependencies in particulate matter dynamics. The methods [7] and [8] combined visual information with temporal learning frameworks, such as Long Short-Term Memory and encoder-decoder architectures, to predict PM2.5. The integrated dual-channel model [7] learned intuitive spatiotemporal features from a series of surveillance images and temporal information from atmospheric conditions, meteorological conditions, and temporal data for precise time-series forecasting of PM2.5 and PM10 concentrations. Likewise, the spatio-temporal model [8] captured and measured feature correlation and loss in particular locations by using an image-like technique at a country-wide level for PM2.5 prediction. These approaches improved the accuracy of prediction by modeling both spatial and temporal correlations. However, these methods often require large, continuous datasets and assume stable visual quality, making them sensitive to haze, clouds, and adverse weather.

More recent studies proposed advanced architecture and multimodal learning to improve the robustness of prediction. A Vision Transformer-based model presented in [9] can effectively process and learn complex data with spatiotemporal dependencies and deep features from image data, even in the absence of extensive labeled data, when leveraging the model. In parallel, the multimodal approach Contrastive Learning–Image Pre-training (CLIP) [10] employs transformers as backbones to learn the combined visual features and contextual information by leveraging 2D image data acquired from mobile devices. The model was trained on a Graphics Processing Unit (GPU) and Single-Board Computer (SBC) to enhance the accuracy and scalability of air quality monitoring. Although the methods demonstrate improved accuracy, they often require greater computational resources and depend on access to various data modalities.

On the contrary, in domain-specific applications [11], computer vision technology and a regression model are used to extract the real-time traffic volume and street-view information from the traffic images and to predict the road concentration of PM2.5, which is trained on meteorological conditions, traffic volume, and building variables. This approach demonstrates the effectiveness of vision-based models in complex urban microenvironments. Although these methods achieve promising results in localized settings, their generalization to diverse geographical areas and broader atmospheric conditions remains a challenge.

The existing vision-based PM2.5 forecast approaches demonstrated the efficacy of deep learning and temporal modeling in extracting particulate matter-related visual patterns. However, most prior work relies on raw image data to capture visual information related to atmospheric conditions, such as haze and visibility degradation, meteorological conditions, particulate matter-related predictors, and the development of deep learning technologies to map these visual features to PM2.5 concentrations. These

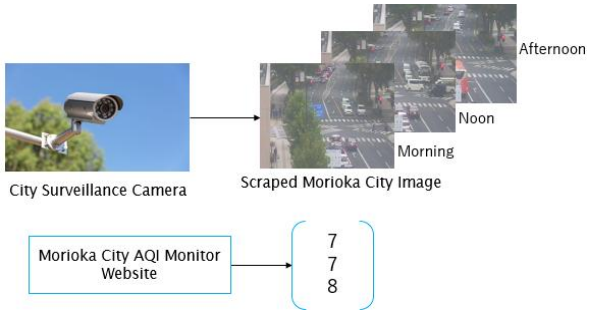


Figure 1. Collecting data using Web-scraping method.

limitations motivate the need for using dehazed images integrated with object segmentation images, which can provide clearer visual cues to atmospheric conditions, while object segmentation enables the model to focus on semantically meaningful regions in the scene for consistent PM2.5 prediction. This proposed approach will be discussed in the following section.

III. VISION-BASED PM2.5 FORECASTING

A. Data Collection Approach

In this study, Web-scraping is employed as a data acquisition method, as shown in Figure 1, to attain environmental scene images along with their associated PM2.5 concentration values. The images are obtained from publicly available city surveillance camera footage from multiple locations, whereas the corresponding PM2.5 observation data are retrieved from official air quality index monitoring websites (www.aqi.in). Data collection is

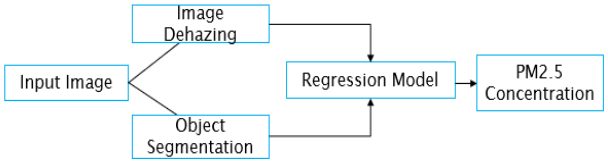


Figure 2. Block diagram of proposed system.

conducted at regular intervals during three daily time periods: morning (9:00 ~ 12:00), noon (12:00 ~ 13:00), and afternoon (14:00 ~ 17:00) from October 2024 to March 2025. The visual data and particulate matter measurements are temporally aligned to ensure consistency between image observations and Ground Truth PM2.5 concentrations.

B. Data Description

The dataset utilized in this work consists of paired environmental images and corresponding PM2.5 values obtained from various urban locations in Japan. Each data sample comprises an outdoor image and its associated PM2.5 concentration, creating a supervised dataset for regression-based particulate matter prediction. In addition to the original images, the dataset is further extended with derived representations, namely dehazed images and their associated segmentation masks, which are used to improve visual quality and extract object-level cues.

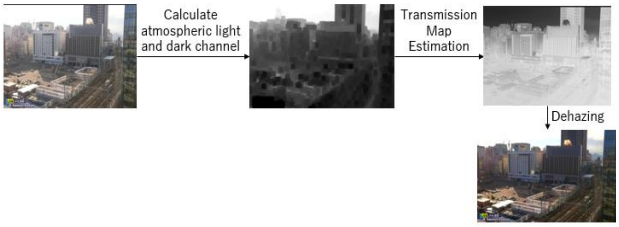


Figure 3. Process of dehazing the Sapporo city image.

The images capture a wide range of environmental conditions, including variations in weather, such as sunny, cloudy, and snowy scenarios. These conditions introduce differences in visibility, illumination, and atmospheric appearance, and these are significant factors for vision-based PM2.5 estimation. The images are collected at fixed resolutions, relying on the camera sources, ensuring that within each location, while preserving sufficient visual details for feature extraction.

In the dataset, each sample is represented by multiple inputs, including the original image, the dehazed version of the image, and the corresponding segmentation mask, along with the Ground Truth PM2.5 concentration. This multi-representation structure empowers the model to manipulate both enhanced visual information and semantic scene cues. Moreover, the dataset has temporal variation by

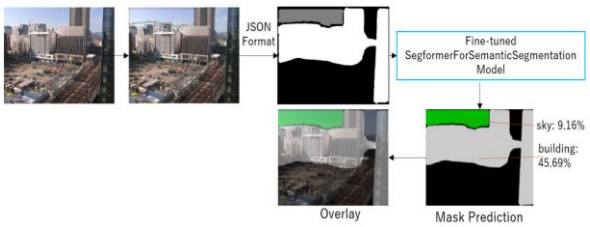


Figure 4. Model fine-tuning and object coverage percentage estimation.

incorporating samples obtained at different times of the day, as well as spatial variability across different cities with distinct urban and atmospheric conditions.

C. Data Pairing

The data pairing is performed to correspond to each environmental image with a Ground Truth PM2.5 value. The corresponding PM2.5 concentration is assigned based on both spatial and temporal alignment for each attained image. Firstly, a camera-to-station correspondence is considered by identifying the nearest air quality monitoring station to each camera location, utilizing Google Maps. The approximate distance between the camera and station is obtained from the map, and only stations located within a threshold distance of 10 kilometers have been considered. This ensures that the selected particulate matter concentrations reasonably reflect the atmospheric conditions seen in the images. Furthermore, temporal alignment is performed by matching each image with the closest available PM2.5 measurement based on its timestamp. This stage verifies consistency between the captured visual scene and the recorded PM2.5 measurement. Finally, each valid image is paired with its associated particulate matter value to create a labeled data sample.

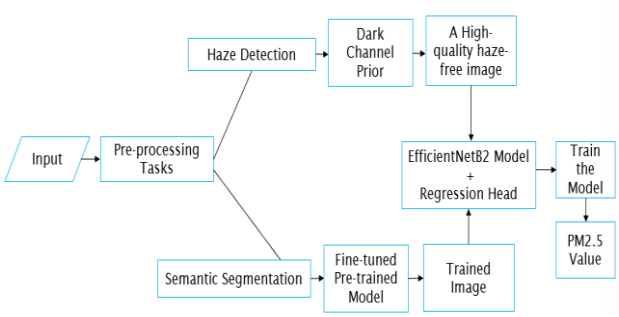


Figure 5. Flowchart of proposed work.

For data cleaning, repeated images and images captured from altered camera positions are excluded to establish consistency in scene representation. Initially, around 200–300 images are obtained per city; after the filtering process, the remaining dataset is reduced to approximately 70–100 images per city. Eventually, the filtered images, along with their corresponding PM2.5 values, are exploited to construct the final dataset.

D. System Model

The proposed image-based system's block diagram is illustrated in Figure 2, where environmental images are used as input to the proposed framework. A dehazing process is applied to enhance image quality by reducing the effects of atmospheric scattering. In addition, object segmentation is subsequently performed to identify the different key scene components and then compute the object coverage amount. These refined visual representations and the extracted object features are fed into the regression model. The model learns the pattern between these two inputs, along with the Ground Truth PM2.5 numerical data, to predict PM2.5 concentrations precisely.

E. Image Dehazing using Dark Channel Prior Algorithm

The first pre-processing step, the dark channel prior approach [12], is employed, as depicted in Figure 3, to remove the influence of atmospheric haze from the retrieved environmental images. This method is based on the observation that at least one color channel exhibits very low intensity values in most non-sky regions of haze-free outdoor images. Based on this assumption, the dark channel for each input image is determined first to estimate the spatial distribution of haze. Afterwards, the atmospheric light is calculated by identifying the brightest pixels in the dark channel, which represent areas with the highest haze concentration. Moreover, a transmission map is estimated by using the determined atmospheric light and dark channel information to illustrate the portions of the scene radiance that reach the camera sensor. Lastly, the dehazed image is recovered by restoring the scene radiance using the estimated parameters, including atmospheric light and transmission map, resulting in a dehazed image that exhibits enhanced visibility and contrast, providing much clearer visual features for further analysis.

F. Semantic Segmentation

Semantic segmentation constitutes the second pre-processing phase of the proposed system architecture, as described in Figure 4, to obtain fine-grained, object-level cues from the environmental scenes. The aim of this stage is to identify and distinguish semantically meaningful areas or classes, such as sky, buildings, roads, vegetation, water, and other structural components within the different urban scenes that are potentially correlated with PM2.5 concentration. The segmentation process provides further contextual information beyond global image appearance by explicitly modeling the spatial distribution of those objects.

The Ground Truth segmentation masks are constructed from the original images by manually annotating object classes, including sky, building, water, road, etc., exploiting the LabelMe annotation tool. The annotations are converted into JSON format to generate the corresponding segmentation masks.

These masks are utilized to fine-tune a deep learning model based on the SegFormer architecture on the acquired dataset. The model is adopted due to its robust performance in capturing both local and global contextual cues, allowing it to adapt to the visual characteristics of the scenes. After fine-tuning the model, the trained model generates pixel-wise segmentation masks for each input image. Eventually, the predicted masks are subsequently utilized to determine the coverage amount of each segmented object class in the scene, providing quantitative object-level features to accurately estimate PM2.5.

The performance of the segmentation model is evaluated using standard metrics, such as pixel accuracy, mean Intersection over Union (mIoU), F1-score, and recall. The results specify that the model obtains high segmentation performance across all cities. Specifically, Sapporo and Chofugaoka exhibit the highest performance with pixel accuracy and F1-score of 0.99, along with mIoU values of 0.97 and 0.99, respectively. Aomori and Kagoshima also demonstrate strong performance, with pixel accuracy values of 0.98 and mIoU values of 0.93. These results verify that the SegFormer model effectively captures both global and local contextual features of environmental scenes.

G. PM2.5 Prediction

We have employed a backbone named EfficientNet-B2, a Convolutional Neural Network-based regression model, shown in Figure 5, for extracting visual features from the input data. The proposed approach utilizes dehazed images combined with segmentation masks as inputs, where a feature-level fusion strategy is applied. Specifically, the masks are integrated with the dehazed images through channel-wise concatenation to create a unified multi-channel input, enabling the model to simultaneously learn high-quality visual information and learn trained object-level cues derived from semantic segmentation.

These fused inputs are fed into the EfficientNet-B2 backbone to learn patterns and hierarchical visual features, ranging from low-level texture and color information to high-level semantic representations related to atmospheric conditions. The output features are generated by the

TABLE 1. EXPERIMENTAL RESULTS USING FIVE INPUT CONFIGURATIONS

Dataset	Model A			Model B			Model C			Model D			Model E		
	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*	R^2	*MSE*	*MAE*
Sapporo	0.16	1.95	1.72	0.29	6.21	1.92	0.01	6.70	1.81	0.22	2.37	1.81	0.18	4.93	1.70
Kagoshima	0.14	5.28	1.99	0.20	6.73	2.09	0.08	7.23	2.20	0.27	3.82	1.58	0.24	5.29	1.79
Aomori	0.19	8.69	2.49	0.26	7.82	2.32	0.04	7.47	2.16	0.43	7.39	2.40	0.31	8.34	2.32
Chofugaoka	-0.21	13.06	3.02	0.18	14.91	3.00	0.04	5.33	1.94	0.15	13.35	2.95	0.11	8.56	2.03

backbone and then are applied to a regression head designed to predict continuous PM2.5 values. The regression head comprised fully connected layers that map the extracted visual cues to a numerical output, enabling end-to-end learning of the relationship between the images and PM2.5 numerical data. For training, the regression head and the backbone network are optimized in an end-to-end manner using a Graphics Processing Unit (GPU). The dataset is split into training and testing sets with an 8:2 ratio. The model is trained utilizing a regression loss function, including Mean Squared Error (MSE), to minimize the difference between predicted and Ground Truth PM2.5 concentrations. An adaptive optimization algorithm named Adam optimizer is used to update the network parameters, with an appropriate learning rate of 0.0003, batch size of 32, and epochs of 30 to ensure convergence and improved prediction stability and generalization performance.

IV. RESULTS AND DISCUSSION

In this section, a comprehensive assessment of the proposed vision-based particulate matter prediction approach has been demonstrated, utilizing datasets from multiple cities in Japan. The experimental results are reported using five phases. Standard regression metrics, namely the coefficient of determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE), are used to evaluate the model's performance. Additionally, we have conducted comparative analyses to examine the effect of the five stages on prediction accuracy. A detailed discussion of the quantitative results and their implications for PM2.5 prediction is provided in the following subsections.

A. Experimental results

In this study, we train a computer vision model, EfficientNet-B2, as a feature extractor along with the regression head on Japan's different cities datasets. The experiments are conducted using environmental images of various cities, including Sapporo, Kagoshima, Aomori, and Chofugaoka. The prediction results for each city are presented in Table 1 using five input configurations. Model A utilizes original environmental images as input to the regression model. Model B utilizes dehazed images to assess the effect of visual enhancement on prediction performance. Model C uses semantic segmentation masks derived from the original images to verify the contribution of object-level scene cues. Model D integrates dehazed images and their associated masks, indicating the proposed method that combines both improved visual features and semantic information. Finally, Model E uses the combination of all inputs.

As shown in Table 1, all alternative configurations have demonstrated varying changes in prediction accuracy across cities, compared to Model A. The dehazed images enhance R^2 values compared to the original images across all cities, demonstrating the effectiveness of visibility improvement in capturing pollution-related features. On the other hand, segmentation masks exhibit inconsistent performance, with generally lower correlation values, indicating that segmentation is insufficient solely for reliable prediction.

Model D, which integrates dehazed images and segmentation masks without original images, exhibits more consistent and competitive performance across multiple datasets. It achieves the strongest R^2 values, which are 0.27 and 0.43, respectively, in Kagoshima and Aomori, while also maintaining relatively lower MAE values of 1.58, particularly for Kagoshima. These results suggest that combining dehazed visual information with semantic features can significantly capture pollution-related patterns even in the absence of raw images.

In comparison, Model E, which incorporates original images along with dehazed images and segmentation masks, does not consistently outperform Model D. Although Model E achieves lower MAE in some cases, including 1.70 and 2.03 for Sapporo and Chofugaoka, respectively. Its R^2 values are lower than those of Model D in key datasets, including Kagoshima and Aomori. This suggests that the inclusion of original images does not necessarily lead to enhanced prediction.

Overall, the comparative analysis demonstrates that Model D provides a balanced and robust performance across cities.

B. Discussion

The experimental results show that employing dehazed images and semantic segmentation masks improves PM2.5 prediction performance across most cities, yielding a higher R^2 value and lower error metrics compared to the original image results. In addition, the performance of the regression model is influenced by dataset characteristics, including sample size, segmented predictors, weather conditions, camera resolution, and distance between the camera and station.

The performance of the Sapporo dataset achieves relatively stable performance across configurations. This can be attributed to the moderate dataset size of 72 samples, the extraction of semantic features, such as sky and building, consistent sunny weather conditions, and a comparatively short camera-station distance of 3.4 kilometers, which ensures reliable spatial alignment. The uniform camera resolution of

960 x 540 provides consistent feature extraction, further enabling enhanced prediction accuracy.

In Kagoshima, although the dataset contains 79 samples, the semantic information sky and mountain, appearance of sunny and cloudy weather, and a longer camera-station distance of 4.5 kilometers add additional variability. Despite this, Model D achieves the best performance, indicating that the integration of dehazing and semantic segmentation helps reduce the effects of environmental variation.

For Aomori, Model D obtains the highest R^2 value of 0.43, indicating a strong correlation between predicted and observed PM2.5 values. This indicates that the integration of dehazed images and segmentation masks effectively captures the overall pollution trends in this dataset. However, the error metrics MSE and MAE remain relatively higher compared to other cities, suggesting that although the model captures the general pattern well, prediction deviations still exist. This behavior may be attributed to snowy conditions, which introduce visual complexity and weak semantic feature cues, thereby limiting precise prediction. While the camera-to-station distance is 3.5 kilometers, which supports reasonable spatial alignment, the challenging environmental conditions limit accurate estimation.

For Chofugaoka, despite having a relatively larger dataset, which contains 93 samples and the shortest distance between the camera and station is 2 kilometers, the performance is comparatively lower. It indicates that factors, including scene complexity and variability in sunny and cloudy weather, and the presence of sky and water, have a stronger impact than spatial proximity alone. The higher error values specify that visual obscurity and weak feature correlations limit the model's effectiveness.

Model D demonstrates more consistent performance across cities, specifically under varying environmental conditions. Notably, it reaches competitive or superior results without utilizing original images. In contrast, Model E does not consistently enhance performance, suggesting that raw images may introduce redundancy or noise rather than useful information.

V. CONCLUSION AND FUTURE WORK

The aim of this study is to develop an image-based system framework for predicting particulate matter concentrations using geographical images collected from various geographical locations of Japan. The urban-level datasets are constructed by pairing outdoor city images with the associated Ground Truth PM2.5 observations obtained from the cities' monitoring stations. We have employed image dehazing techniques to enhance the reliability of visual information under varying atmospheric conditions for reducing haze effects. Semantic segmentation is proposed to extract meaningful object-level characteristics from the urban scenes. These proposed pre-processing stages aim to emphasize visual features, such as sky conditions, buildings, and water, that are closely related to air pollution, thus improving the input data quality for prediction.

The proposed system integrates two complementary inputs, such as dehazed images and object segmentation masks, along with PM2.5 measurements, which are fed into a regression method and trained to predict PM2.5 concentrations. Experimental results demonstrate satisfactory performance in terms of using dehazed images and segmentation masks in vision-based air quality estimation compared to the other input configurations.

In future work, we would like to apply our proposed method to other existing PM2.5 vision datasets to evaluate its performance under diverse environmental conditions.

ACKNOWLEDGMENT

We would like to convey our special thanks to Swannack Raymond Amaki for providing continuous support in the data collection process.

REFERENCES

[1] T. Ohara and M. Ono, "An Overview of PM2.5 Pollution Research Conducted in ERTDF Projects since 2011," Global Environmental Research, vol. 22, pp. 3–12, Dec. 2018.

[2] D. Mitra and A. Saha, "IoT-Based Air Pollution Detection, Monitoring and Controlling System," Journal of Discrete Mathematical Sciences and Cryptography, vol. 25, pp. 2173–2182, Dec. 2022, doi: 10.1080/09720529.2022.2133254.

[3] K. Zhang, Z. Chen, and Y. Xiang, "Vision-Based Particulate Matter Estimation," Deep Learning Applications, pp. 3–17, 2023, doi: 10.1142/9789811266911_0001.

[4] G. Wang et al., "Vision-Based PM2.5 Concentration Estimation with Natural Scene Statistical Analysis," IEEE Transactions on Artificial Intelligence, vol. 5, pp. 2805–2815, Oct. 2023, doi: 10.1109/TAI.2023.3324892.

[5] S. Laohakiat, S. Klerkkidakan, and N. Wiwatwattana, "Visually Estimating and Forecasting PM2.5 Levels Using Hybrid Architecture Deep Neural Network," Current Applied Science and Technology, vol. 24, p. e0258074, Dec. 2023, doi: 1055003/cast.2023.258074.

[6] Y. Liu et al., "Applying Traffic Camera and Deep Learning-Based Image Analysis to Predict PM2.5 Concentrations," Science of The Total Environment, vol. 912, p. 169233, Dec. 2024, doi: 10.1016/j.scitotenv.2023.169233.

[7] Y. Wu, X. Wang, M. Wang, X. Liu, and S. Zhu, "Time-Series Forecasting of PM2.5 and PM10 Concentrations Based on the Integration of Surveillance Images," Sensors, vol. 25, pp. 95–113, Dec. 2024, doi: 10.3390/s25010095.

[8] N. Sirisumpun, K. Wongwailikhit, P. Painmanakul, and P. Vateekul, "Spatio-Temporal PM2.5 Forecasting in Thailand Using Encoder-Decoder Networks," IEEE Access, vol. 11, pp. 69601–69613, Jul. 2023, doi: 10.1109/ACCESS.2023.3293398.

[9] T. Zhao and M. Qu, "VDMS: An Improved Vision Transformer-based Model for PM2.5 Concentration Prediction," Applied Sciences, vol. 15, pp. 7346–7362, Jun. 2025, doi: 10.3390/app15137346.

[10] H. Madokoro and S. Nix, "Multimodal Particulate Matter Prediction: Enabling Scalable and High-Precision Air Quality Monitoring Using Mobile Devices and Deep Learning Models," Sensors, vol. 25, pp. 4053–4076, Jun. 2025, doi: 10.3390/s25134053.

[11] Z. Fan et al., "Enhancing Urban Real-Time PM2.5 Monitoring in Street Canyons by Machine Learning and Computer Vision Technology," Sustainable Cities and Society, vol. 100, p. 105009, Jan. 2024, doi: 10.1016/j.scs.2023.105009.

[12] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, pp. 2341–2353, Dec. 2011, doi: 10.1109/TPAMI.2010.168.

User Attention in the Interface: Comparative Eye-Tracking Analysis of Website Buttons

Piotr Izydor Tokarski
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: p.tokarski@pollub.pl

Małgorzata Plechawska-Wójcik
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: m.plechawska@pollub.pl

Karol Łazaruk
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: k.lazaruk@pollub.pl

Mariusz Dzieńkowski
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: m.dzienkowski@pollub.pl

***Abstract*—This study presents a comprehensive comparative analysis of two web application interface variants that differ in button design, specifically graphical buttons and text-based buttons. The research was conducted in the context of universal design, using eye-tracking technology, which enables objective assessment of users' visual behaviour during task performance. A range of measures was analysed, including the number and duration of fixations, the distribution of attention points, and heat maps. Furthermore, the temporal parameters of task completion and the quality of execution were considered. The findings reveal substantial disparities in the manner in which users navigate the interface, thereby substantiating the efficacy of integrating quantitative and visual analyses in the development of accessible web applications.**

Keywords-eye tracking; user interface; universal design; accessibility; gaze fixations.

I. INTRODUCTION

The progressive evolution of web applications has led to an escalating emphasis on the calibre of user interfaces and their accessibility to a broad audience [1]. Universal design is predicated on the creation of solutions that are useful and comprehensible for users irrespective of age, experience, or perceptual limitations [2][3]. Buttons represent a fundamental component of the interface, serving as the primary mechanism through which users interact with the system [4]. However, designing button interfaces that are both intuitive and universally accessible remains a non-trivial challenge.

In the domain of design practice, the utilisation of both text and graphic buttons is a prevalent phenomenon. Despite the potential of icons to expedite the identification of functions, their interpretation can be ambiguous. A key challenge in interface design is balancing visual efficiency with clarity of interpretation, particularly for users with diverse levels of experience and perceptual abilities. Graphic buttons may accelerate recognition for familiar users, but can introduce ambiguity, while text buttons improve clarity at the cost of increased processing time. Conversely, text buttons necessitate a higher level of cognitive engagement, yet frequently yield enhanced precision in interaction [5][6]. The objective of this article is to provide a comparative analysis of both solutions using eye-tracking research, which allows for the assessment of actual user behaviour, rather than merely their declarations [7]-[9]. Although prior research has investigated eye tracking in the context of interface evaluation, relatively little attention has been given to the direct comparison of text and graphic buttons using both objective eye-tracking data and subjective usability measures. This study aims to fill this gap by integrating these two perspectives in a unified analysis. This study focuses on three research questions. First, it investigates whether the type of button (text versus graphic) influences visual attention patterns. Second, it examines the effect of button type on task completion efficiency. Third, it analyses how button type impacts perceived usability.

The remainder of the paper is structured as follows. Section II presents the literature review. Section III describes the methodology and experimental design. Section IV outlines the research plan. Section V presents the results of the study. Section VI discusses the findings. Finally, Section VII concludes the paper and outlines directions for future work.

II. LITERATURE REVIEW

Issues of usability and accessibility of user interfaces represent a significant research domain within the broader field of human-computer interaction [10][11]. In the context of the proliferation of web applications, there is an increasing emphasis on the design of interfaces that are not only aesthetically pleasing, but also intuitive and accessible to users with varying perceptual abilities. In this context, interactive elements, such as buttons, which play a pivotal

role in the communication process between the user and the system, are of particular importance [12][13].

Conventional usability evaluation methodologies, encompassing task tests, surveys and heuristic analyses, furnish valuable insights concerning the quality of an interface. However, these approaches are predominantly reliant on user declarations. Consequently, there is an increasing call to supplement these methods with approaches that facilitate the recording of actual, often unconscious, user behaviour. One of the most frequently employed techniques of this nature is eye tracking, which facilitates the analysis of eye movements during interaction with the interface [14]-[16].

Eye-tracking studies are utilised for a variety of purposes, including the assessment of the visual hierarchy of interfaces, the identification of key elements for task completion, and the detection of problem areas. The analysis of the number and duration of fixations allows for the drawing of conclusions regarding the cognitive load experienced by users, while the utilisation of heat maps facilitates a visual assessment of attention distribution. In the context of universal design, these techniques are of particular importance because they allow the identification of perceptual barriers that may not be revealed in studies based solely on subjective assessments [17]-[19].

A significant area of research pertains to the comparison of diverse forms of interface element presentation, encompassing graphic and text buttons. It has been posited by certain authors that the employment of graphic icons can facilitate more expeditious recognition of functions and enhance the visual appeal of the interface [20]. It is emphasised that the effectiveness of these systems is contingent upon the clarity of the symbols employed and the experience of the users. In some cases, this may result in interpretative uncertainty. Conversely, text buttons are frequently regarded as offering enhanced clarity and predictability, although they may necessitate a more protracted information processing duration [21].

Research on interface usability increasingly emphasises the importance of combining eye-tracking analyses with usability survey results. This methodological approach facilitates the comparison of objective measures of visual behaviour with subjective assessments of comfort, readability and ergonomics of the interface [22]. The findings of preceding studies suggest that a thorough evaluation of interfaces must encompass both considerations to enhance the efficacy of the design decision-making process [23]. However, existing studies often focus on either subjective usability or isolated eye-tracking metrics, without integrating both perspectives, which limits the comprehensiveness of their conclusions.

III. METHODOLOGY

Issues of accessibility and usability of user interfaces are widely discussed in the relevant literature. The authors of numerous works have indicated that classic evaluation methods, such as surveys or heuristic tests, should be supplemented with techniques that enable the recording of unconscious user reactions. One such methodology is eye tracking, which facilitates the analysis of eye movements.

Eye-tracking studies are utilised for a variety of purposes, including the assessment of the visual hierarchy of interfaces, the identification of key elements, and the identification of problem areas. The analysis of fixations and saccades enables conclusions to be drawn about the cognitive load and intuitiveness of design solutions. In the context of universal design, it is emphasised that interfaces should minimise the need to interpret symbols and reduce the risk of incorrect user decisions.

A corpus of previous studies comparing text and graphic elements indicates that the superiority of one solution over the other depends on the context of the task and the experience of users. This article provides a detailed expansion on the aforementioned research, with a focus on a comprehensive analysis of visual perception patterns.

A. Participants

The study involved 10 participants (4 females, 6 males), aged between 23 and 25 years. All participants were students of Computer Science at the Lublin University of Technology who were recruited on a voluntary basis. They reported normal or corrected to normal vision. The respondents had extensive experience with web layouts (web design and development), although they had no prior experience with the tested interfaces. The participation in the study was voluntary. Informed consent was obtained, and the participants were informed about the purpose of the study.

B. Apparatus

The Gazepoint GP3 HD eye tracker characterised by a sampling rate of 150 Hz with an accuracy of 0.5-1° was used to conduct the study. The device was connected to the Acer Nitro 5 AN517-41-R48Y laptop with a 17.3-inch screen with a resolution of 1920 × 1080 pixels. The experiment was performed under identical conditions in a laboratory with stable light conditions to ensure the accuracy of the measurements.

C. Experimental Design

The research plan was developed in such a way as to enable a reliable comparison of the two interface variants while maintaining repeatable experimental conditions. This can be seen in Figure 1.

The study was divided into two primary sections, corresponding to two methods of interface presentation: method A, which employed graphic buttons, and method B, which utilised text buttons. Participants were randomly assigned to one of the two interface variants, thus allowing the experimenters to avoid both the learning effect and the transfer of experience between application versions. The applied design ensured independent observations and a more reliable comparison of the two interface variants.

A comprehensive dataset was meticulously collected for each task performed during the study. It encompassed two main categories of information: eye-tracking metrics, which captured detailed aspects of participants' visual behaviour, and temporal information, providing insight into how much

time was devoted to each activity. Every participant completed a total of 10 tasks. With ten participants involved in the study, the experiment yielded 100 task-level observations (10 participants multiplied by ten tasks). This enabled a direct comparative analysis between the two interface versions under investigation. Once the experimental phase had concluded, the collected data was carefully organised and prepared for further analysis. This analysis consisted of two complementary components: a comparison of average metric values obtained for each interface version, and an assessment of visual representations of users' visual behaviour. Together, these approaches provided both a quantitative and a qualitative perspective on the results.

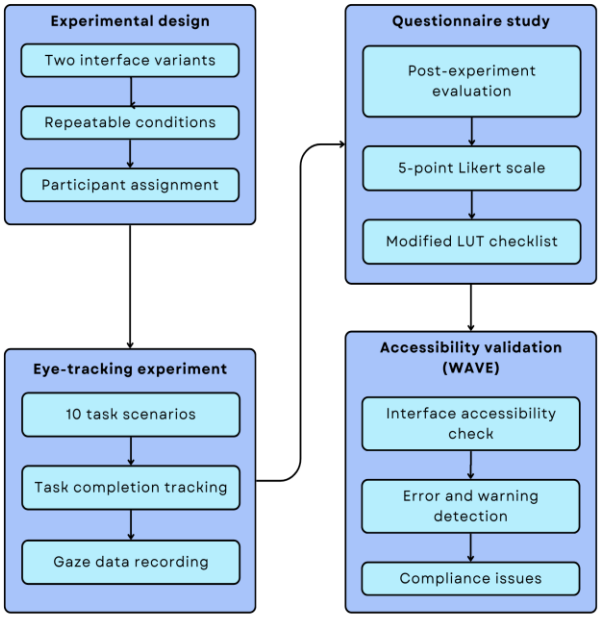


Figure 1. Research plan of the study.

D. Procedure

The participants in the study performed a series of tasks involving the identification and utilisation of particular application functions. The tasks included: (1) locating a specific function within the interface, (2) navigating to a selected section of the application, and (3) completing predefined actions requiring interaction with the available buttons. These tasks were designed to reflect typical user scenarios and to ensure comparability between the two interface variants. The tasks were meticulously designed to necessitate active searching for interface elements and the formulation of interactive decisions, thereby facilitating the observation of natural visual exploration strategies. The imposition of a time limit during the execution of the tasks was deemed unnecessary, as this could have resulted in the introduction of undue pressure, which might have influenced the user behaviour in an unnatural manner.

E. Data Collection and Measures

The experiment involved the collection of both quantitative and qualitative data. Quantitative analyses encompassed task completion time, the number and average duration of fixations, and the number of errors made. The qualitative data comprised heat maps and visualisations of fixation distribution, which facilitated the assessment of the areas of the interface that attracted the most attention from users. The analysis of fixations was given particular emphasis, since their number and duration are widely recognised as indicators of cognitive load and the degree of clarity of the information presented.

F. Data Analysis

The collected data was then aggregated and subjected to a comparative analysis between the two versions of the interface. This methodology enabled a multidimensional assessment of the impact of button design on interface perception, task performance efficiency, and interaction accuracy, while maintaining the principles of universal design. For each participant, mean values were calculated for all tasks across the individual metrics analysed. Due to the small sample size, a non-parametric Mann-Whitney U test was applied instead of parametric tests to compare the two interface variants. Effect sizes (r) were also calculated.

IV. RESULTS

The research results include both quantitative and qualitative analysis of data obtained during the eye-tracking experiment. The integration of these two perspectives yielded a more comprehensive understanding of the variations in the perception of interfaces with graphic and text buttons.

The initial aspect analysed was the number and duration of fixations. The data presented in Table 1 illustrate the mean fixation values for both methods. In the case of the interface with graphic buttons, a greater number of fixations was observed, with a shorter duration. This configuration is indicative of a thorough scanning of the interface, necessitating the interpretation of the significance of the icons. In the case of the text interface, the number of fixations was lower, while their average duration was longer. This suggests that the subjects were focusing more intently on the information presented, and that they found it clearer. These results may suggest that the text-based interface imposed a lower cognitive load, allowing for more efficient information processing. However, this interpretation should be treated with caution, as cognitive load was not directly measured using dedicated physiological or validated workload assessment methods. The interface with graphic buttons required greater interpretative effort from users, which was reflected in the higher number of shorter fixations.

The Mann-Whitney U test did not reveal statistically significant differences between the two interface variants for each analysed metric ($p > 0.05$). However, an analysis of effect sizes revealed significant differences in user behaviour. It should be noted that a large effect was observed for the fixation count ($r = 0.56$) and a moderate effect for the completion time ($r = 0.50$). A moderate effect was also found for fixation duration ($r = 0.36$), while in the case of saccade duration and saccade count, only small effects ($r \approx 0.10$) can be observed.

TABLE I. MEAN NUMBER OF FIXATIONS AND MEAN NUMBER OF SACCADES ACROSS INDIVIDUAL TASKS FOR BOTH METHODS

Task	*Method*	*Fixation count*	*Saccade count*
1	Method A	28.20	108.00
	Method B	18.80	55.75
2	Method A	11.20	38.20
	Method B	10.60	43.80
3	Method A	55.20	184.40
	Method B	27.20	106.60
4	Method A	24.80	113.80
	Method B	15.20	59.00
5	Method A	35.20	111.40
	Method B	12.20	56.60
6	Method A	7.60	32.40
	Method B	6.80	34.60
7	Method A	21.20	87.80
	Method B	13.80	80.40
8	Method A	25.80	98.00
	Method B	11.60	78.60
9	Method A	12.75	35.25
	Method B	7.80	35.60
10	Method A	16.75	55.75
	Method B	14.20	76.60

A key component of the analysis involved the heat maps generated for Task 2, as shown in Figures 2 and 3. In the case of method A, the heat maps reveal the dispersion of users' attention between several areas of the interface, including elements that are irrelevant to the task at hand. In contrast, method B has been shown to result in clear clusters of fixation within the text buttons, which are directly related to the task. This finding indicates that the correct element is recognised more quickly.

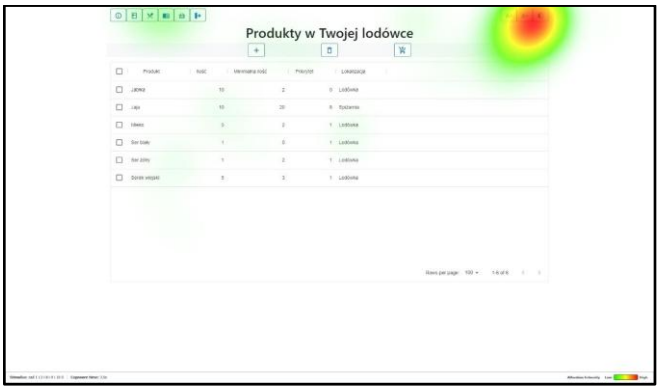


Figure 2. Heat map for task 2 for method A.

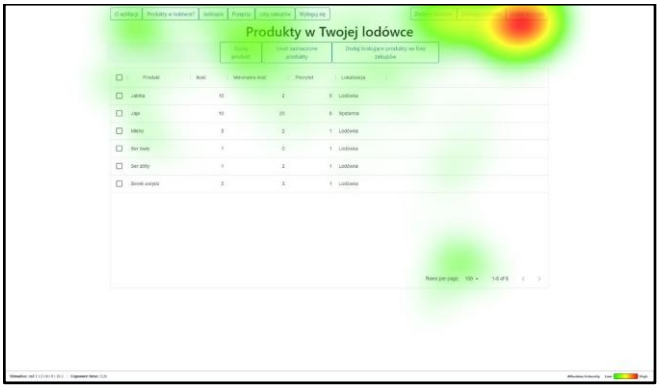


Figure 3. Heat map for task 2 for method B.

Another parameter analysed was the task completion time. The data is presented in Figure 4.

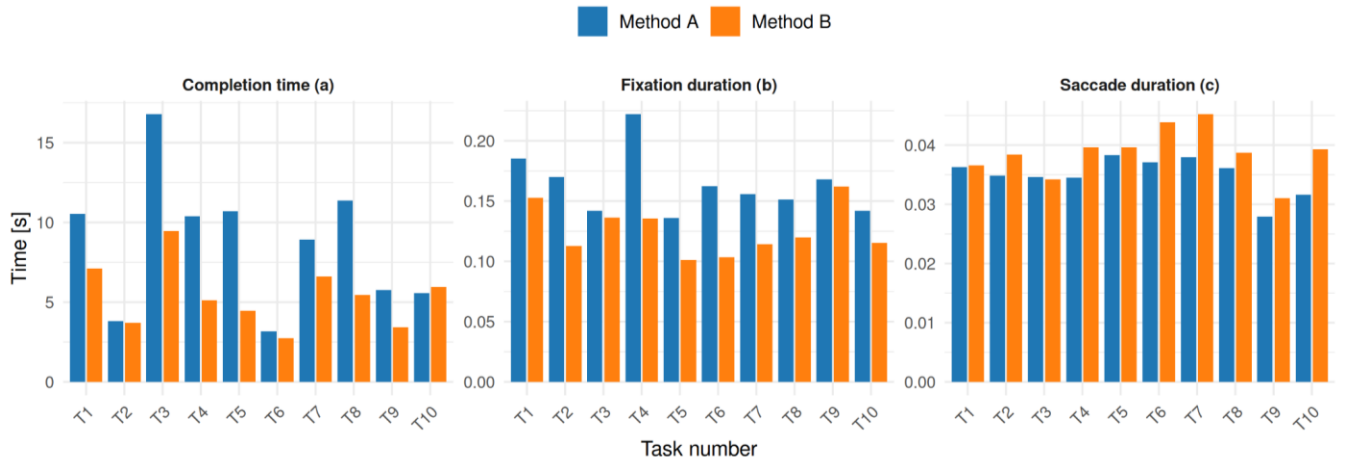


Figure 4. Mean values of eye-tracking metrics for AOI (Area of Interest) for each task: mean task completion time (a), mean fixation duration (b), mean saccade duration (c).

A thorough examination of the graph reveals that the utilisation of graphic buttons resulted in a reduction of task completion times. However, a discernible correlation between this reduction and the accuracy of completion was not observed. The text variant was characterised by a marginally extended completion time, yet concurrently exhibited enhanced stability in the results obtained by participants.

An additional element of the analysis was a chart based on the results of the LUT (Lublin University of Technology) survey, which can be seen in Figure 5. This was used to subjectively assess the usability of the interface by participants after completing the tasks. The survey incorporated a series of statements evaluated on an ordinal scale, encompassing, inter alia, aspects, such as the readability of the interface, the ease of locating functions, and the overall ergonomics of the application. The presentation of the responses in the form of a chart facilitated a direct comparison of the perception of both interface variants.

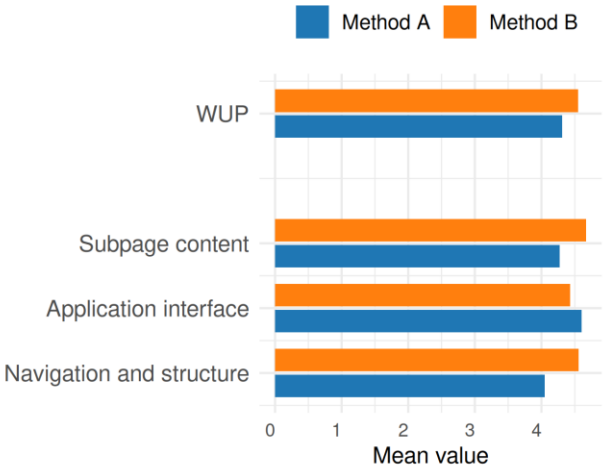


Figure 5. Mean WUP (Weighted Usability Points) scores by evaluation area and final score for Methods A and B (LUT survey).

An analysis of the LUT survey chart reveals that the interface with text buttons received higher ratings in areas related to the clarity and comprehensibility of functions. The overall WUP rating for method B was significantly higher, at around 4.6, whilst for the graphical version it remained just above 4.3. It was evident that respondents frequently asserted a high degree of certainty with regard to the function of interface elements. In contrast, there was a notable decrease in the number of instances requiring guesswork regarding the purpose of buttons. The advantages of the text-based interface were particularly evident in terms of navigation and content (in both cases, the scores ranged from 4.6 to 4.7, whilst the graphical version scored around 4.0-4.2 points). Conversely, the graphic variant was regarded as more visually appealing and dynamic with slightly higher ratings in the application interface category, approximately 4.6 points for method A vs. 4.4 for method B. However, respondents frequently highlighted the necessity for additional interpretation of the icons. The results of the LUT survey are consistent with the observations from the eye-tracking data and complement them significantly, combining an objective perspective with the subjective assessment of users.

The WAVE tool (Web Accessibility Evaluation Tool) [24] was used to assess compliance with the Web Content Accessibility Guidelines (WCAG) 2.0 guidelines [25]. It is an automated tool that detects accessibility issues, such as errors, contrast problems, and structural elements, enabling a systematic evaluation of the interface. No significant problems or warning messages were identified during testing. A summary of the results obtained is presented in Table 2.

TABLE II. RESULTS OF THE ANALYSIS OF TWO VERSIONS OF THE INTERFACE USING THE WAVE TOOL

WAVE Category	*Method*	*Frequency*
Errors	Method A	0
	Method B	0
Contrast errors	Method A	0
	Method B	0
Alerts	Method A	0
	Method B	0
Features	Method A	1
	Method B	1
Structural elements	Method A	8
	Method B	17
Accessible Rich Internet Applications (ARIA)	Method A	8
	Method B	20

V. DISCUSSION

The analysis of the results confirms that the form of presentation of buttons in the web application interface has a significant impact on both visual perception and the effectiveness of user interaction. The amalgamation of eye-tracking data with the results of the LUT survey facilitated a multidimensional evaluation of the solutions that were tested, thereby complementing both objective and subjective perspectives.

The interface incorporating graphic buttons encouraged faster action, which was reflected in shorter task completion times, as it was suggested in study [20]. Concurrently, the examination of fixations and heat maps suggests that users predominantly engaged their visual attention in interpreting the significance of icons. The dispersion of fixation points and the increased frequency of brief glances suggest an elevated cognitive load, particularly in circumstances necessitating unambiguous identification of the interface's functions. Moreover, the analysis of effect sizes is consistent with the visual patterns observed in the heat maps and confirms the interpretation that using an interface designed with graphic buttons demands greater cognitive effort.

The variant with text buttons demonstrated a distinct behavioural pattern. It is evident that longer and more concentrated fixations, in addition to a focus on key areas of the interface, serve to indicate greater clarity and predictability of the system's operation, which is consistent with the findings of the study [21]. Despite the slightly protracted task completion time in this instance, users demonstrated a reduced incidence of errors and exhibited a heightened sense of confidence in their decisions, a finding that is corroborated by the LUT survey results.

Analysis of data from the WAVE tool indicates that there are no errors, contrast issues, or alerts in either interface variant, which confirms their fundamental technical correctness. However, the noticeable difference in the number of structural elements and ARIA attributes (Method B has a higher number) are due to the different ways of implementing semantics and accessibility, despite compliance with WCAG 2.0.

These results are significant in the context of universal design, which aims to minimise cognitive barriers and ensure that interfaces are accessible to as wide a range of users as possible. From this standpoint, the lucidity of communications and the predictability of interactions may be of greater significance than the maximum velocity of operations. The analysis suggests that text buttons are better suited to meeting these requirements, especially in the context of utility applications.

Concomitantly, it is imperative to acknowledge that a compromise solution may be to utilise hybrid interfaces that combine graphic icons with short text labels. This observation is consistent with previous findings suggesting that graphical elements support rapid recognition, while text enhances interpretability [20][21]. This approach has the potential to address the interpretation challenges observed while maintaining the visual benefits of graphical interfaces. The findings suggest that eye tracking is an effective tool for supporting design decisions and identifying subtle usability issues that are not always revealed by traditional evaluation methods, as demonstrated in [17].

VI. Conclusion And Future Work

The article presents the findings of comparative tests conducted on two variants of a web application interface, distinguished by the form of the buttons utilised – graphic and text. The employment of eye-tracking technology, in conjunction with a LUT survey, facilitated a comprehensive evaluation of both the objective visual behaviour exhibited by users and their subjective sentiments regarding the usability of the interface. The findings suggest that there are substantial variations in visual exploration strategies, task completion time and user confidence levels contingent on the design solution adopted.

The findings of the present study demonstrate that interfaces based on text buttons promote greater clarity and accuracy of interaction, a factor that is of particular importance from the point of view of universal design. Conversely, graphic buttons can facilitate expeditious orientation within the interface; however, their judicious selection is contingent upon the selection of appropriate symbols and the context of utilisation. The conclusions presented herein have the potential to provide practical support for user interface designers. Further research could focus on the analysis of hybrid solutions and the expansion of the research group to include users with diverse perceptual needs. Increasing the sample size would allow for a more reliable statistical verification of the observed differences. Although statistical tests were conducted in this study, no statistically significant differences were observed ($p > 0.05$); however, effect size analysis revealed significant differences between the compared interface variants.

References

[1] V. Panwar, “Web evolution to revolution: navigating the future of web application development,” International Journal of Computer Trends and Technology, vol. 72, no. 2, pp. 34–40, 2024.

[2] S. Goldsmith, Universal design. London: Routledge, 2007.

[3] R. Mace, “What is universal design,” The Center for Universal Design at North Carolina State University, vol. 19, 1997.

[4] K. Rao, M. W. Ok, and B. R. Bryant, “A review of research on universal design educational models,” Remedial and Special Education, vol. 35, no. 3, pp. 153–166, 2014.

[5] M. F. Story, “Principles of universal design,” in Universal Design Handbook, 2nd ed. New York: McGraw-Hill, 2001.

[6] M. F. Story, “Maximizing usability: the principles of universal design,” Assistive Technology, vol. 10, no. 1, pp. 4–12, 1998.

[7] J. P. P. Hansen, P. Bækgaard, D. Valgeirsdottir, and S. Beier, “Universal design of gaze interactive applications for people with special needs,” in Proc. 2023 Symp. Eye Tracking Research and Applications, May 2023, pp. 1–7.

[8] K. Conklin, A. Pellicer-Sánchez, and G. Carrol, Eye-tracking. Cambridge: Cambridge University Press, 2018.

[9] B. T. Carter and S. G. Luke, “Best practices in eye tracking research,” International Journal of Psychophysiology, vol. 155, pp. 49–62, 2020.

[10] J. Preece et al., Human-computer interaction. Reading, MA: Addison-Wesley Longman, 1994.

[11] H. Bansal and R. Khan, “A review paper on human computer interaction,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 8, no. 4, p. 53, 2018.

[12] H. Sørum, K. N. Andersen, and R. Vatrapu, “Public websites and human–computer interaction: an empirical study of measurement of website quality and user satisfaction,” Behaviour & Information Technology, vol. 31, no. 7, pp. 697–706, 2012.

[13] Y. Yang and W. Bao, “Application of human-computer interaction technology in remote language learning platform,” International Journal of Human–Computer Interaction, vol. 41, no. 3, pp. 1751–1761, 2025.

[14] D. Li et al., “An eye-tracking-based approach to evaluate the usability of government portal websites in pilot smart cities,” Engineering, Construction and Architectural Management, vol. 32, no. 4, pp. 2369–2396, 2025.

[15] A. I. Molina et al., “Eye tracking-based evaluation of accessible and usable interactive systems: tool set of guidelines

and methodological issues," Universal Access in the Information Society, vol. 24, no. 4, pp. 3085–3108, 2025.

[16] S. Röhrl, F. Hauser, T. Ezer, L. Grabinger, and P. D. J. Mottok, "WebGazeTrack: a web-based eye tracking tool," in Proc. 6th European Conf. Software Engineering Education, June 2025, pp. 125–134.

[17] T. Strandvall, "Eye tracking in human-computer interaction and usability research," in Proc. IFIP Conf. Human-Computer Interaction. Berlin, Germany: Springer, Aug. 2009, pp. 936–937.

[18] M. A. Just and P. A. Carpenter, "A theory of reading: from eye fixations to comprehension," Psychological Review, vol. 87, no. 4, p. 329, 1980.

[19] K. Rayner, "Eye movements in reading and information processing: 20 years of research," Psychological Bulletin, vol. 124, no. 3, p. 372, 1998.

[20] J. J. Foster, "Graphical symbols: test methods for judged comprehensibility and for comprehension," ISO Bulletin, pp. 11–13, 2001.

[21] E. Tenner, "The design of everyday things by Donald Norman," Technology and Culture, vol. 56, no. 3, pp. 785–787, 2015.

[22] J. Nielsen, Usability engineering. San Francisco, CA: Morgan Kaufmann, 1994.

[23] W. Albert and T. Tullis, Measuring the user experience. San Francisco, CA: Morgan Kaufmann, 2010, pp. 195–216.

[24] WebAIM (Web Accessibility In Mind). *WAVE Web Accessibility Evaluation Tool*. [Online]. Available from: https://wave.webaim.org/ 2026.05.18

[25] B. Caldwell, M. Cooper, L. G. Reid, and G. Vanderheiden, Eds. *Web Content Accessibility Guidelines (WCAG) 2.0, W3C Recommendation*. [Online]. Available from: https://www.w3.org/TR/WCAG20/ 2026.05.18

Visual Accessibility and Readability in User Interfaces: An Eye-Tracking Study

Karol Łazaruk
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: k.lazaruk@pollub.pl

Piotr Izydor Tokarski
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: p.tokarski@pollub.pl

Małgorzata Plechawska-Wójcik
Department of Computer Science
Lublin University of Technology
Lublin, Poland
email: m.plechawska@pollub.pl

***Abstract*—The aim of this work is to investigate how specific visual features of graphical interfaces influence the efficiency of information localization by users. This study focuses on examining the role of readability in visual characteristics, such as contrast, text formatting, and element highlighting, in shaping the speed and accuracy with which users can identify relevant information. The experiment involved comparing two versions of a museum interface: one designed in accordance with the principles of universal design and the second one with reduced readability. Eight stimuli were used in the study, which was conducted with 15 participants using eye-tracking technology. The qualitative and quantitative analyses were performed. Among the analysed features, task completion time, number of fixations, time to first fixation, and fixation dwell time were considered. Moreover, heat maps were also included in the study. The results indicate that high contrast, larger font size, and a well-structured content layout significantly reduce the number of fixations and shorten the time needed to locate information. Additionally, larger graphic elements with intense colours attract users' attention more effectively, and their highlighting facilitates localization. The findings are expected to contribute to a deeper understanding of user interaction with visual interfaces and to provide guidelines for designing more effective and user-friendly digital environments.**

Keywords-eye tracking; readability; visual accessibility; user interfaces.

I. INTRODUCTION

The User Interface (UI) is the main point of human-computer interaction [1]. It represents the perceptual layer, which often directly influences visual attention patterns even before cognitive processing occurs. Thus, there is a particular need to investigate how digital accessibility elements, such as visual accessibility and readability, affect user behaviour. There is also a lack of analyses that take into account how users truly perceive interfaces, rather than assessing this aspect only in terms of compliance with guidelines.

In the context of digital systems, UX refers to the user experience, which includes aspects such as usability, intuitiveness, efficiency, satisfaction, and accessibility [2][3]. The UI directly impacts the quality of the UX and, more importantly, serves as its visual and interactive element. That is why it is so important for designed interfaces to be accessible, i.e., usable by the largest possible group of users regardless of their age, background, skill level or other limitations [4], and usable, i.e., the use of the interface should be effective and easy [5].

Both usability and accessibility are the subject of numerous studies that explicitly emphasize the great importance of inclusive interface design [6][7]. Research shows that failure to comply with basic principles in this area can lead to a decline in the quality of the user experience, which also translates into, among other things, a decline in satisfaction, efficiency, and engagement [8][9]. The solution is to apply universal design principles, which define a set of guidelines that should be followed during the design process, allowing for the developed solution to be used, to the greatest extent possible, by all people, regardless of their abilities or disabilities [10].

One element of accessibility is visual accessibility, which covers the aspect of visual perception. It determines the extent to which content is presented in a readable and understandable way for the widest group of users [11][12]. Aspects of visual presentation increase visual accessibility and apply to all users [13], not just those with disabilities.

An element of usability is readability, which refers to the ease of reading content [14]. Key aspects affecting readability are contrast, as well as appropriate font type and size [15][16]. Furthermore, the appropriate structure and layout of elements increase user engagement and task completion efficiency [17][18].

Ready-made solutions in the form of heuristics [19] and accessibility standards not only facilitate interface design, but can also be used to evaluate existing solutions. Furthermore, visual accessibility and readability can be tested using automated tools, such as WAVE [20]. A tool that allows for an objective assessment of both perceptual and cognitive aspects is eye tracking [21].

The aim of this paper is to examine how different aspects of visual accessibility and readability affect the effectiveness in terms of users' ability to locate information in two versions of a museum application interface. To achieve this goal, an eye-tracking study was performed.

Unlike previous studies that primarily formulate general usability and accessibility guidelines, this study provides a quantitative validation of selected visual accessibility principles within a specific application domain, namely museum interfaces. Moreover, the research contributes by linking eye-tracking metrics (such as fixation count, dwell time, and time to first fixation) with concrete interface design variables, including contrast, typography, and layout structure. This allows not only confirmation of known principles, but also their measurable impact on user attention and search efficiency in realistic interaction scenarios.

The remainder of this paper is structured as follows. In Section II, the materials and methods used in the study are described. Section III presents the results of the eye-tracking analysis. Section IV discusses the findings and their implications. Finally, Section V concludes the paper and outlines directions for future research.

II. MATERIALS AND METHODS

Fifteen participants (14 males and 1 female) took part in the study, with an average age of 23.2 years (SD = 0.54). Prior to the study, informed consent was obtained from all participants (the opinion from the Scientific Research Ethics Committee of the Lublin University of Technology No. 1/2024 of 19 February 2024). All participants had experience in using websites.

The research object was a proprietary online museum website created using the Angular library. Two alternative versions of the interface were implemented: the first, designed based on universal design principles (Interface 1), and the second, in which no good practices were applied and readability was intentionally limited (Interface 2).

Eye movements were recorded using a Gazepoint GP3 HD eye tracker (Gazepoint, Vancouver, Canada; 150 Hz, accuracy 0.5-1°) on a 17.3" screen (1920 × 1080 pixels).

The experimental procedure involved a short introduction to the experiment, in which the participants were informed about the aim and procedures. The actual eye tracking study was conducted on a group of participants who were asked to complete a total of 13 tasks (commands): five tasks (1-5) in two versions of the application interface, and three additional commands unrelated to any version of the interface (6-8), which were utilised for additional verification of the readability aspect. Table 1 presents an overview of the tasks performed by the respondents, along with the visual accessibility and readability issues that were the subject of their assessment. To minimize potential learning effects associated with the repeated exposure to similar interface structures, the order of interface presentation was counterbalanced across participants. Half of the participants started with Interface 1, while the other half began with Interface 2. Additionally, the order of tasks was randomized within each interface version. This approach was applied to reduce bias related to familiarity and learning effects in the paired experimental design.

TABLE I. RESEARCH TASKS WITH CORRESPONDING ISSUES OF VISUAL ACCESSIBILITY AND READABILITY

Task number	Task scope	Area of accessibility/readability
1	Locating the link to the museum's social media	Visual contrast of the interface
2	Locating the question in the FAQ section	Aesthetics, formatting, and grouping of text content
3	Locating the link to the latest exhibition	Organisation and order of elements
4	Locating the button/link to the museum's rules and regulations	Distinguishability of buttons and links
5	Locating the link to the FAQ section	Visibility and location
6-8	Locating the image	Readability of typography (font type, colour, size), visibility of visual elements against the background, relationship between captions and graphic elements, and visual hierarchy of elements

Each task was associated with a particular visual stimulus in the form of a screenshot of the application and required participants to localise specific interface elements that were formulated as commands (e.g., 'Locate the museum's social media button'). The participants could move on to the next tasks by pressing the space bar. The tasks were displayed one at a time.

Heat maps were exported for the purpose of qualitative analysis. The heat maps allow for a visual representation of the aggregated level of visual attention, which is represented by colours applied to the visual stimulus. This makes it possible to locate areas that attract attention as well as those that are ignored by users [22].

An Area of Interest (AOI) was defined for the collected eye tracking data. The AOI data was used in qualitative analysis, which was conducted on the basis of proprietary scripts in the R language. The following eye tracking metrics were selected and considered in the context of the defined AOIs:

- Number of fixations - the total number of fixations in the AOI, which is negatively correlated with the performance [23].
- Time To First Fixation (TTFF) - the time elapsed from the appearance of the stimulus to the first fixation in the AOI. This measure indicates how quickly a particular element attracts the user's attention - the shorter the time, the faster the observer's attention is attracted [24].
- Fixation dwell time - the total time that the user spent in the AOI. It indicates the level of interest - the longer the time, the greater the interest [25].

An additional metric was the task completion time, which refers to the time taken to complete a particular task. It measures the performance of completed tasks - shorter times indicate that specific tasks are easier to complete and that the interface design is better [26].

Normality and homogeneity of variance tests were performed within the quantitative eye-tracking data to verify the applicability of Student's t-test. As the assumptions were not met, a non-parametric equivalent of Student's t-test, i.e., Wilcoxon signed-rank (paired) test, was performed.

III. RESULTS

Based on eye-tracking data, the heatmaps were generated. Their content made it possible to verify that, for tasks 1-5 in Interface 1, areas receiving a high level of attention covered a smaller region and were significantly fewer in number compared to the corresponding heat maps for Interface 2.

Figures 1 and 2 illustrate sample heat maps associated with task 2, which involved finding a specific question in the FAQ section. In Figure 1, it can be observed that in Interface 1, the participants' gaze was concentrated most intensely on the element representing the target solution of the task, but it is also worth noting the increased intensity of attention within another section of the FAQ, where the participants focused their gaze when reading the content of the preceding section. Several areas of low attention intensity within the navigation bar and other questions can be seen, but their intensity is significantly lower. In the case of Figure 2, it can be observed that the area of increased visual attention has a much larger surface area than in Interface 1, where attention is concentrated on only two relatively small, dense points. Furthermore, this area has a generally higher level of attention intensity, indicating that in Interface 2, the subjects had to put more effort into finding the target element.

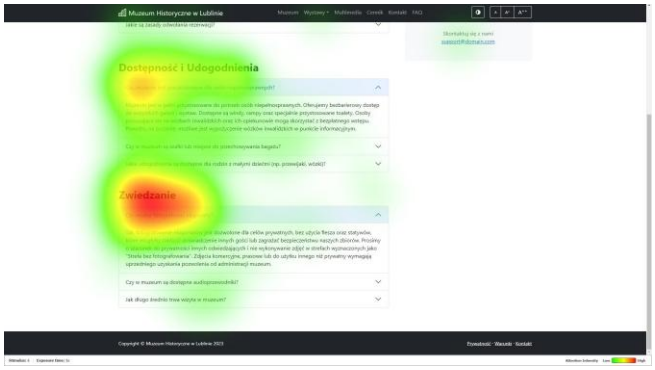


Figure 1. Example of a heatmap of task 2 in the case of Interface 1.

Figure 2. Example of a heatmap of task 2 in the case of Interface 2.

In the case of three additional commands (6-8), heat maps clearly show that the visual attention of the participants was focused primarily on the highlighted elements. In the case of graphics, these were large objects in bright colours, while the text was visually appealing due to its clear, sans-serif font in bright but well-contrasting colours. Similarly, smaller elements were more difficult to locate, resulting in a greater number of areas of moderate or increased attention intensity.

Figure 3 shows a heat map associated with task 8, which shows that the large highlighted element with a bright contrasting font attracts the most attention compared to other less prominent elements.

Figure 3. Example of a heatmap of task 8.

For the purposes of qualitative analysis, mean values, medians, and confidence intervals were calculated for the individual tasks in both interfaces: task completion time and (in the context of AOI) fixation dwell time, TTFF, and number of fixations.

Figure 4 presents a comparison of mean task completion times for individual tasks for both interfaces. As can be easily seen, the mean times for Interface 1 were shorter than those for Interface 2 in each command. The difference between the times ranged from 0.6 s to approximately 7.0 s on average.

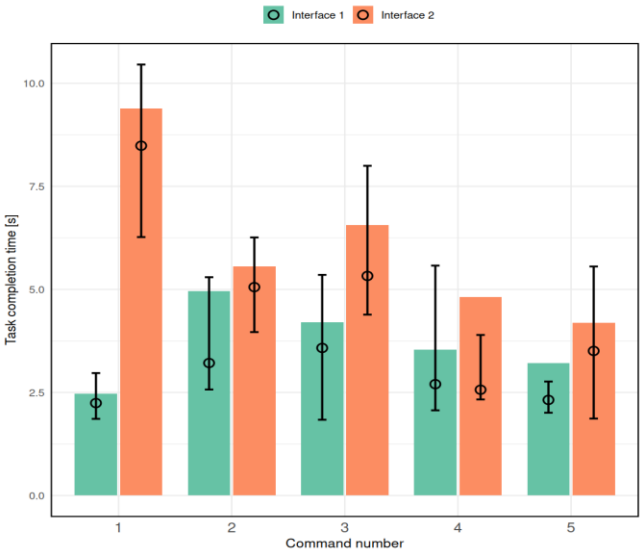


Figure 4. Mean task completion time per task according to the interface obtained for all participants.

Figure 5 presents a comparison of mean fixation dwell times (AOI) in individual tasks for both interfaces. Similar to the mean task completion time, the mean dwell times for Interface 1 were also shorter than those for Interface 2 for each command. Only for task 2, the mean results are almost identical, while in the other commands, the differences between the times ranged up to approximately 1.6 s on average.

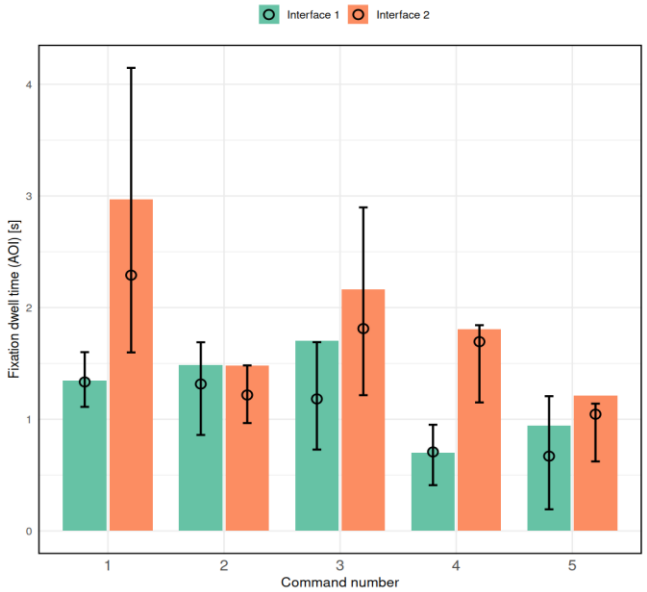


Figure 5. Mean fixation dwell time (AOI) per task according to the interface obtained for all participants.

In the case of the TTFF (AOI) measure, for which the mean values for individual commands are presented in Figure 6, it can be seen that the TTFF time for tasks 2, 3, and 5 was greater in the case of Interface 2, averaging approximately 0.5 s to 1.5 s. For task 1, the TTFF times are similar, while for task 4, the discrepancy is clearly noticeable and is approximately 0.9 s less for Interface 2.

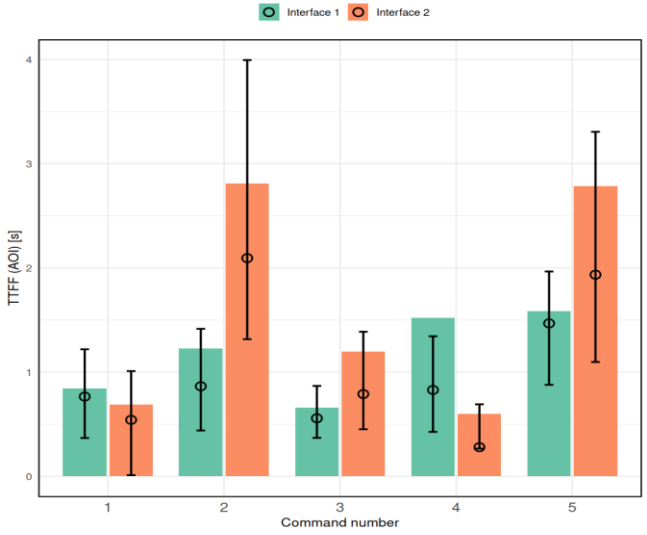


Figure 6. Mean TTFF (AOI) per task according to the interface obtained for all participants.

A comparison between the mean fixation counts for both interfaces across all commands presented in Figure 7 demonstrates that, except for task 5, the mean fixation count was higher, ranging from approximately 4.4 to 11 fixations for Interface 2. In the case of command 5, this number was very close for both interfaces.

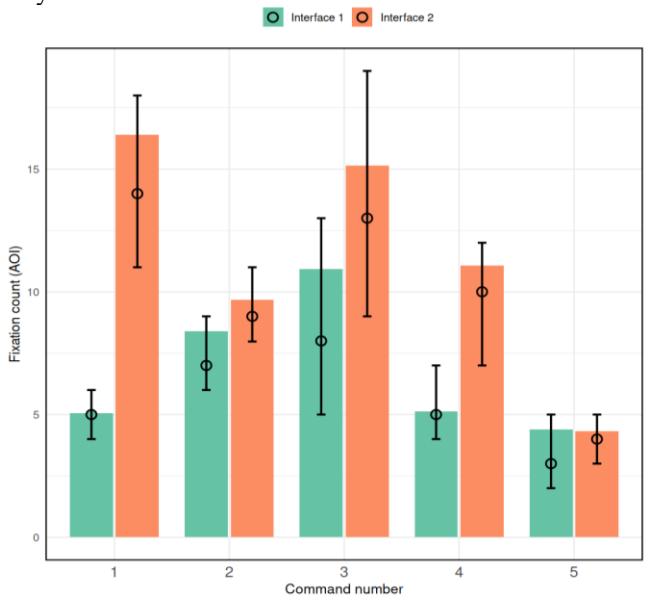


Figure 7. Mean number of fixations (AOI) per task according to the interface obtained for all participants.

Normality and variance heterogeneity tests were performed for every command for each measure for both interfaces. The test results indicated that for some data, the assumptions of the Student's t-test were not met, so non-parametric Wilcoxon signed-rank (paired) tests were performed. The tests showed that there are statistically significant differences between interfaces that are not applicable to some commands. In the case of the task completion time, the statistically significant differences were observed for command 1, and similarly for dwell time for commands 1 and 4. Regarding TTFF, a significant difference between interfaces was observed for command 2. In the case of fixation count, the differences were observed for commands 1, 3 and 4. All values met the criterion of $p < 0.05$.

IV. DISCUSSION

As can be observed from the heat maps, participants' attention is more clearly focused on key areas in Interface 1 than in Interface 2. These areas are smaller, yet they attract more visual attention. In Interface 2, however, participants' attention is less concentrated, which is reflected by a more even distribution of attention and the presence of areas with lower fixation intensity. This suggests that, even without prior experience with the application, users can intuitively locate the most important functions in Interface 1. The results also confirm the conclusions presented in [17], which state that elements located in the top, middle, and right sections of the interface attract the most attention.

Similar observations can be made regarding heat maps relating to contrast, font size, the order of elements, and the readability of navigational elements. Participants were able to use the website more easily in high contrast mode, and elements with larger font sizes were more visible. Furthermore, Interface 1, with clearly organised and logically ordered elements, was associated with higher levels of visual attention. Key navigational components, such as clearly visible FAQ buttons in the navigation bar, attracted more attention than less prominent links in the footer. These observations are consistent with the results presented in article [16], which found that participants preferred text written in larger font sizes, as well as high contrast between the font and background colours.

In the case of heat maps related to visual aspects and interface readability, such as layout, element highlighting, and typography, it has been proven that larger, emphasized or intensely coloured images attract the user's attention much more than smaller, less prominent, or less noticeable elements. Heat maps suggest that the participants looked at smaller areas, indicating greater intuitiveness and simplicity of Interface 1. These results are consistent with the conclusions of the study [15], in which the authors demonstrated that bright, saturated colours effectively attract users' attention. Furthermore, a study conducted in [18] shows that large images, photos of celebrities, minimal text, and search functions attract attention, and this study proves the same with regard to large images and minimal text.

In the context of quantitative analysis, for both task completion time and fixation dwell time, it can be observed that Interface 1 was characterised by shorter values compared to Interface 2. This leads to the conclusion that increasing visual accessibility and readability significantly improves the efficiency with which the participants are able to find specific elements within the interface. However, it is worth noting that the differences in fixation dwell time between the interfaces were smaller than the differences in task completion time. This is because fixation dwell time is only part of the total task completion time within the AOI and does not include, for example, the time needed for navigation, information processing, and decision-making by the participants.

It should also be noted that in task 2, the fixation dwell times were very similar, which may be due to the fact that in both interfaces, the subjects had to analyse the same text fragments, and the only difference was in the aesthetics, formatting, and grouping of the text content. Therefore, these visual changes did not significantly affect fixation time, although they still facilitated orientation in the content and influenced the overall efficiency of task performance, which is clearly noticeable in the other measures (task completion time, TTFF, and number of fixations).

The TTFF measure revealed that the times associated with tasks 2, 3, and 5 were notably faster for Interface 1 compared to Interface 2. This suggests that Interface 1 has superior visibility, aesthetics, and content organisation, enhancing the perceptibility of its key areas. Furthermore, the shorter TTFF time suggests that Interface 1 is more effective at managing user attention, is more readable, and is visually clear.

By contrast, it is worth noting the opposite trend for tasks 1 and 4, for which TTFF times were shorter for Interface 2. This is because, in both cases, the page content included a map that took up most of the screen and attracted participants' attention. In case of Interface 1, the map was placed above the target text in task 1, making it slower to locate the necessary information than in Interface 2, where the map was located below the target text. In case of Interface 1 in task 4, only the text content was displayed for the participants to analyse. These results suggest that compliance with universal design principles does not always guarantee shorter TTFF, and that the layout and nature of page content also significantly impact the effectiveness of information retrieval.

The mean number of fixations confirmed that in almost all cases, the number of fixations for Interface 1 was lower than for Interface 2. This suggests that better organisation and readability of interface elements, as well as contrast, content aesthetics, and the use of buttons instead of links, contribute to more effective content searching.

However, it is worth noting that, in task 5, locating the FAQ link in Interface 1 took less time than in Interface 2 but required almost the same number of fixations. This suggests that the greater readability and better placement of the element (the link in the navbar with the larger font) enabled it to be identified more quickly, despite the similar number of fixations. In Interface 2, however, the link in the footer with the smaller font required a similar number of fixations but took more time to be located.

V. CONCLUSION AND FUTURE WORKS

The aim of the study was to assess the impact of visual accessibility and interface readability on the effectiveness of information location by users, based on a comparison of two versions of a museum web application. Eye-tracking and task-related metrics were utilised to perform the analysis. The study contributes to the field by providing quantitative evidence of how specific visual design features influence user attention and performance within a domain-specific interface.

The study indicates that users are more likely to notice graphic elements that are larger, have intense colours, or are highlighted or emphasised. Additionally, visual interface features, such as high contrast, larger font size, aesthetic formatting, and logical content organisation significantly improve visual accessibility and reduce the time needed to find information. These results demonstrate that an effective layout and visual distinction of elements can enhance the efficiency of navigating the interface and provide valuable insights for creating more user-friendly digital environments.

While the study provides relevant findings and clear conclusions, it should be noted that it has certain limitations. First, the relatively small sample size may limit the generalizability of the results. Non-parametric Wilcoxon signed-rank (paired) tests revealed no statistically significant differences between the two interfaces in relation to certain assigned tasks and measures. Future research plans to examine a significantly larger group of respondents and expand the number of visual stimuli to investigate more

aspects that could affect visual accessibility and readability. Second, the within-subject (paired) experimental design introduces a potential risk of learning effects, although counterbalancing and task randomization were applied to mitigate this issue. Third, the study relies solely on eye-tracking data, without incorporating predictive models, such as visual saliency analysis. An important extension of the presented study could involve the use of computational visual saliency models. Such models allow prediction of areas that are likely to attract user attention prior to actual interaction. Comparing predicted saliency maps with empirical eye-tracking data could provide deeper insight into discrepancies between expected and observed gaze behavior. This would further strengthen the evaluation of interface effectiveness and support more predictive design approaches. Future work should address these limitations by increasing the sample size, integrating computational attention models, and exploring additional interface types and domains.

REFERENCES

[1] A. Pitale and A. Bhumgara, "Human computer interaction strategies—designing the user interface," Proc. 2019 Int. Conf. Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 752–758, doi:10.1109/ICSSIT46314.2019.8987819.

[2] O. Badran and S. Al-Haddad, "The impact of software user experience on customer satisfaction," J. Management Information and Decision Science, vol. 21, pp. 1–20, 2018.

[3] M. Hassenzahl and N. Tractinsky, "User experience—A research agenda," Behaviour & Information Technology, vol. 25, pp. 91–97, 2006, doi:10.1080/01449290500330331.

[4] O. Sohaib, W. Hussain, and M. Khalid, "User experience (UX) and the web accessibility standards," Int. J. Computer Science Issues, vol. 8, pp. 584–609, 2011.

[5] M. Matera, F. Rizzo, and G. Carughi, "Web usability: principles and evaluation methods," in Web Engineering, 2006, pp. 143–180, doi:10.1007/3-540-28218-1_5.

[6] J. Gómez-Delgado, C. Marín-Palacios, and L. Moreno, "Integrating accessibility, usability and UX in web project management: a systematic review," Univ. Access Inf. Soc., vol. 25, p. 28, in press, doi:10.1007/s10209-025-01298-0.

[7] V. Nasr, M. Benden, and M. Zahabi, "An approach for assessing usability and accessibility of assistive technology for persons with disabilities," Applied Ergonomics, vol. 129, art. no. 104581, 2025, doi:10.1016/j.apergo.2025.104581.

[8] T. Tullis and W. Albert, Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, 2nd ed. Burlington, MA: Morgan Kaufmann, 2008.

[9] M. Ekin et al., "Impact of web accessibility on cognitive engagement in individuals without disabilities: evidence from a psychophysiological study," PLoS One, vol. 20, no. 7, art. e0328552, July 2025, doi:10.1371/journal.pone.0328552.

[10] J. Lazar, D. Goldstein, and A. Taylor, "Evaluation methods and measurement," in Ensuring Digital Accessibility Through Process and Policy, San Francisco, CA: Morgan Kaufmann, 2015, pp. 139–159.

[11] G. E. Legge, "Visual accessibility: a challenge for low-vision research," Optometry and Vision Science, vol. 91, no. 7, pp. 696–706, July 2014, doi:10.1097/OPX.0000000000000310.

[12] K. Cornish, J. Goodman-Deane, K. Ruggeri, and P. J. Clarkson, "Visual accessibility in graphic design: a client–designer communication failure," Design Studies, vol. 40, pp. 176–195, 2015, doi:10.1016/j.destud.2015.07.003.

[13] H. Petrie, F. Hamilton, and N. King, "Tension, what tension? website accessibility and visual design," Proc. 2004 Int. Cross-Disciplinary Workshop on Web Accessibility (W4A '04), New York, NY, USA: ACM, 2004, pp. 13–18, doi:10.1145/990657.990660.

[14] A. Miniukovich, A. De Angeli, S. Sulpizio, and P. Venuti, "Design guidelines for web readability," Proc. 2017 Conf. Human Factors in Computing Systems, 2017, pp. 285–296, doi:10.1145/3064663.3064711.

[15] A. Lewandowska, A. Olejnik-Krugly, J. Jankowski, and M. Dziśko, "Subjective and objective user behavior disparity: towards balanced visual design and color adjustment," Sensors, vol. 21, pp. 1–18, 2021, doi:10.3390/s21248502.

[16] L. Rello and M.-C. Marcos, "An eye tracking study on text customization for user performance and preference," Proc. 8th Latin American Web Congress, 2012, pp. 64–70, doi:10.1109/LA-WEB.2012.13.

[17] M. Țichindelean, M. T. Țichindelean, I. Cetină, and G. Orzan, "A comparative eye tracking study of usability—towards sustainable web design," Sustainability, vol. 13, pp. 1–31, 2021, doi:10.3390/su131810415.

[18] S. Djamasbi, M. Siegel, and T. Tullis, "Generation Y, web design, and eye tracking," Int. J. Human-Computer Studies, vol. 68, pp. 307–323, 2010, doi:10.1016/j.ijhcs.2009.12.006.

[19] R. da Silva de Queiroz Pierre, "Heuristics in Design: A Literature Review," Procedia Manufacturing, vol. 3, pp. 6571–6578, 2015, doi: 10.1016/j.promfg.2015.07.961.

[20] D. Mothy, A. P. Reddy, C. W. Cai, H. S. Choudhry, and M. H. Dastjerdi, "Assessing the readability, quality, and visual accessibility of patient education websites for laser refractive surgery," Ophthalmic Epidemiology, vol. 32, no. 6, pp. 704–710, 2025, doi:10.1080/09286586.2025.2500014.

[21] S. Eraslan, Y. Yesilada, and S. Harper, "Eye tracking scanpath analysis techniques on web pages: a survey, evaluation and comparison," J. Eye Movement Research, vol. 9, pp. 1–19, 2016, doi:10.16910/jemr.9.1.2.

[22] O. Špakov and D. Miniotas, "Visualization of eye gaze data using heat maps," Elektronika ir Elektrotechnika, vol. 115, pp. 55–58, 2007.

[23] M. M. Porras, C. A. N. K. Campen, J. J. González-Rosa, F. L. Sánchez-Fernández, and J. I. N. Guzmán, "Eye tracking study in children to assess mental calculation and eye movements," Scientific Reports, vol. 14, art. no. 18901, Aug. 2024, doi:10.1038/s41598-024-69800-x.

[24] P. Wlekły, M. Nermend, M. Gryczka, M. Borawski, and H. Shabani, "Use of eye-tracking in the design and optimisation of training materials for future drivers," in Proc. 28th European Conference on Artificial Intelligence (ECAI 2025), 2025, pp. 83–96, doi:10.18276/978-83-8419-053-1-6.

[25] A. Andrychowicz-Trojanowska, "Basic terminology of eye-tracking research," Applied Linguistics Papers, vol. 2, no. 2, pp. 123–132, 2018, doi: 10.32612/uw.25449354.2018.2.pp.123-132.

[26] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," Decision Support Systems, vol. 62, pp. 1–10, 2014, doi:10.1016/j.dss.2014.02.007.

Mobile Apps for Students: Usability Without Barriers?

An Analysis of the Usability and Accessibility of Selected Apps for Students

Piotr Izydor Tokarski
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: p.tokarski@pollub.pl

Jakub Podgórski
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: jakub.podgorski@pollub.edu.pl

Karol Łazaruk
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: k.lazaruk@pollub.pl

Jakub Posikata
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: s95533@pollub.edu.pl

Małgorzata Plechawska-Wójcik
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: m.plechawska@pollub.pl

Mariusz Dzieńkowski
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: m.dzienkowski@pollub.pl

Abstract— **The aim of the paper is to evaluate the usability and accessibility of three mobile apps designed for university students, including those with various types of disabilities. The following apps from Polish universities were selected for analysis: Kampus Pollub, PW Navi, and SmartUMED. The study employed multiple methods to assess usability, including eye-tracking technology, the System Usability Scale (SUS), the User Version of the Mobile App Rating Scale (uMARS), and an automated accessibility evaluation tool. The study was conducted with a group of 30 participants. The results indicate that contemporary mobile apps intended for students generally achieve a satisfactory level of usability, although variability in user evaluations was observed. Significant differences were found between objective and subjective quality assessments within the uMARS questionnaire, with users rating objective aspects more favorably than their overall subjective experience. Furthermore, no statistically significant relationship was identified between SUS scores and uMARS ratings, suggesting that usability and perceived app quality represent distinct dimensions of user experience. Eye-tracking analysis revealed trends indicating higher visual effort for less usable interfaces; however, these differences were not statistically significant.**

Keywords-mobile apps for students; usability; accessibility; eye tracking; SUS; uMARS; Accessibility Scanner.

I. INTRODUCTION

Nowadays, mobile apps constitute an integral part of the everyday functioning of most users. Their large number and diversity contribute to improving the performance of daily activities in almost every area of life. The widespread availability of mobile devices makes it possible to use these apps anytime and anywhere. A particularly important group of mobile app users are young people, including students, for whom such apps serve as a significant tool supporting the organization of the educational process. They enable, among other things, the optimization of study schedules, quick access to essential information, assistance in planning travel to the university or locating teaching facilities, and above all, they allow for time savings and the reduction of potential difficulties resulting from the lack of specific information.

Regardless of functionality and the scope of provided content, an important aspect of mobile apps quality is the intuitiveness and simplicity of the user interface, which enables effective use. Equally important is adapting apps to the needs of people with various types of disabilities, which still does not constitute a universally applied standard.

Despite the wide range of mobile apps dedicated to students, many of them are not fully adapted to the needs of diverse user groups due to the design solutions applied. The lack of essential features, such as options for configuring and personalizing the user interface, as well as insufficient accessibility support for users with mobility impairments, means that some apps do not meet the requirements of full functionality.

Usability evaluation includes the verification of ease of use, intuitiveness, and efficiency of app usage in the context of achieving user goals. Accessibility evaluation, in turn, focuses on determining whether the app is adapted to the needs of people with various types of disabilities, including visual, auditory, motor, and cognitive impairments. Such studies are conducted based on the Web Content Accessibility Guidelines (WCAG), which constitute an international standard developed by the World Wide Web Consortium (W3C) and define principles for designing and creating websites and mobile apps [1].

The evaluation of a mobile app is conducted to improve its quality, reduce errors, and enhance user satisfaction. One common approach involves user-based testing in the form of experiments, in which participants perform predefined task scenarios. During task execution, user interactions with the app are analyzed using selected research methods and tools.

The remainder of the paper is organized as follows. Section II reviews selected studies presenting various approaches to evaluating mobile apps for diverse user groups. Section III defines the research aim and formulates the hypotheses. Section IV describes the methodology, including the evaluated apps, research procedure, participant characteristics, research environment and instruments, and study scenarios. Section V presents the results, including SUS and uMARS questionnaire outcomes, eye-tracking measures, scenario performance metrics, and qualitative accessibility assessment. Finally, Section VI discusses the findings, verifies the hypotheses, outlines limitations, and suggests directions for future work.

II. RELATED WORK

The literature presents a variety of approaches to evaluating mobile apps designed for diverse user groups. Usability studies often adopt mixed-method strategies that combine subjective techniques, such as questionnaires and interviews, with objective methods, including eye tracking. Eye-tracking technology, in particular, is widely recognized as an effective tool for analyzing user–interface interactions and identifying potential usability issues related to interface design [2].

Examples of such approaches can be found in recent studies. A usability evaluation of an educational mobile app incorporating gamification elements, intended for children with type 1 diabetes, their caregivers, and diabetes educators, applied a combination of eye tracking, the thinking-aloud technique, the System Usability Scale (SUS) questionnaire, and focus group interviews [3]. A similar methodological framework was used to assess the usability and interface quality of the "DiagNurse" app, designed for practicing nurses and nursing students [4].

In another study, eye-tracking data were combined with the results of the User Experience Questionnaire (UEQ), allowing for a comprehensive evaluation of mobile app usability through the integration of objective and subjective data [5]. In turn, studies focusing on the quality and functionality of nutrition-related mobile apps employed the standardized Mobile App Rating Scale (MARS) as an evaluation tool [6]. In the usability assessment of the "DiaCare" app, which supports self-management of diabetes, the User Version of the Mobile App Rating Scale (uMARS) was used. This instrument is designed for end users and enables the evaluation of aspects such as engagement, functionality, aesthetics, and information quality [7].

III. AIM AND HYPOTHESES

The aim of the article is to evaluate the usability and accessibility of three selected mobile apps intended for students, while also providing broader insights with implications for the scientific community. Accordingly, the following research hypotheses were formulated:

H1: Mobile apps designed for students demonstrate a measurable level of usability, which varies depending on interface design characteristics.

H2: Statistically significant differences exist between the mean scores of the objective quality dimensions (sections A–D) and the subjective quality dimension (section E) of the uMARS questionnaire in mobile app evaluation.

H3: There is a significant positive correlation between usability scores obtained using the SUS and interface quality scores obtained using the uMARS.

H4: Differences in usability of mobile app interfaces are associated with corresponding trends in eye-tracking metrics, such as fixation count, fixation duration, and saccade count.

IV. METHODOLOGY

A. Selection of Apps

Before selecting the apps, criteria were defined, including availability on Android and iOS, free access, key functionalities (e.g., campus maps and navigation), and suitability for users with various disabilities, allowing adaptation to individual needs.

Three mobile apps for Polish university students were selected:

- Kampus Pollub (Lublin University of Technology)
- PW Navi (Warsaw University of Technology)
- SmartUMED (Medical University of Łódź)

Their main interfaces are shown in Figure 1.

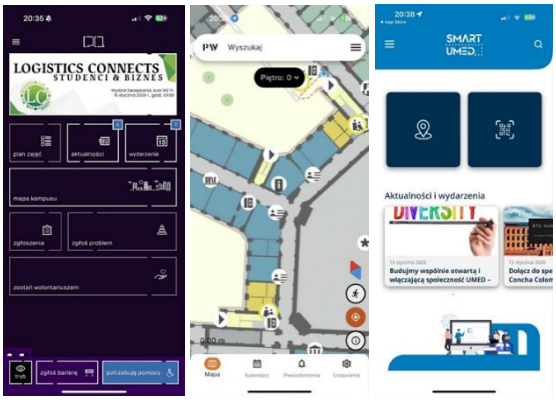


Figure 1. Main panels of the selected apps analyzed in the study.

B. Procedure

To conduct the study, the eye-tracking technique was employed alongside the SUS and uMARS questionnaires, consisting of 10 and 23 items, respectively. The research procedure comprised the following stages:

- **Study preparation**: selection of apps for analysis, development of research scenarios, recruitment of the study group, preparation of the testing environment, and conduct of a preliminary study.
- **Eye-tracking study**: participants completed four predefined test scenarios for a given app.
- **Questionnaires**: participants completed two questionnaires: SUS and uMARS.

- **Data processing**: analysis of eye-tracking data and calculation of SUS and uMARS scores.
- **Results analysis**: interpretation of the obtained results and formulation of conclusions.

C. Participants

The study involved a group of 30 participants aged between 20 and 25 years and included both first-year students and students at the final stage of their academic education. Their experience and proficiency in using mobile apps varied, ranging from highly experienced users to individuals with relatively limited familiarity with such solutions.

A. Research Environment and Instruments

The study was conducted in a laboratory at the Department of Computer Science, Lublin University of Technology, under controlled conditions to ensure participant comfort and concentration. A Motorola Moto G73 5G smartphone with the evaluated apps installed was mounted in a fixed desk holder, providing stable positioning at an appropriate angle and distance from the participant.

Participants' actions and eye movements were recorded using Pupil Invisible eye-tracking glasses connected to a OnePlus 8 smartphone via the Pupil Companion app. The system enables binocular gaze tracking (up to 200 Hz for eye cameras and ~30 Hz for the scene camera) and supports natural movement. Data were transmitted to Pupil Cloud and analyzed using iMotions 11 on an Acer Nitro 5 laptop.

The eye-tracking study began with informing each participant about the purpose and course of the study, followed by obtaining their informed consent to take part in the experiment. Next, the participant's position was adjusted to ensure proper posture and optimal conditions for eye-tracking data collection, and a one-point calibration was performed to ensure measurement accuracy. During the main phase, participants completed tasks defined in the research scenario while wearing the eye-tracking glasses, which recorded both the visual scene and eye-movement activity. The experimental workstation and example views of the evaluated mobile apps with recorded fixation points and saccades during task execution are illustrated in Figure 2.

The study was conducted under the supervision of an experienced moderator, whose role was to oversee the proper course of the experiment and ensure that all procedures were carried out consistently across participants. Task scenarios were provided on paper at the workstation, allowing participants to refer to them freely whenever needed and minimising additional eye movements that would otherwise occur when switching between the task description and the evaluated application on the same screen. After completing the tasks, participants were asked to fill in two standardised questionnaires: the System Usability Scale (SUS) and the user version of the Mobile Application Rating Scale (uMARS), which together provided subjective feedback complementing the objective eye-tracking measurements.

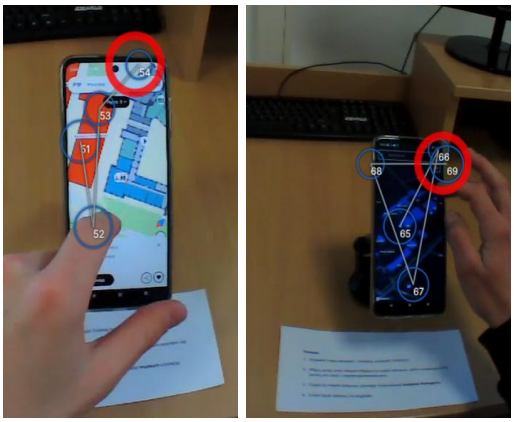


Figure 2. Example views from the eye-tracking study.

The SUS questionnaire consists of 10 items rated on a five-point Likert scale, producing a score from 0 to 100, where values above 68 indicate above-average usability [8]. The uMARS questionnaire, designed for non-expert users, evaluates app quality across 20 items in five sections: engagement, functionality, aesthetics, information quality, and subjective quality, also using a five-point Likert scale [9].

D. Research Scenarios

Tables I–III present the research scenarios, including tasks designed specifically for the SmartUMED, Kampus Pollub, and PW Navi apps, respectively.

TABLE I. SCENARIO FOR SMARTUMED APP

No.	Tasks
1	Display the campus map and locate the Sędziowska Campus in the Bałuty district.
2	Enable the route option adapted for users with disabilities.
3	Using the search function, find the Rectorate building on the UMED campus map.
4	Change the app language to English.

TABLE II. SCENARIO FOR KAMPUS POLLUB APP

No.	Tasks
1	Display the campus map and locate the CENTECH building.
2	Enable the option that indicates on the campus map locations with various barriers for people with disabilities.
3	Using the search function, find the Pentagon building on the campus map.
4	Change the app language to English.

TABLE III. SCENARIO FOR PW NAVI APP

No.	Tasks
1	Display the campus map and locate a men's restroom adapted for people with disabilities.
2	Set the navigation option to bypass stairs for users with mobility difficulties.
3	Using the search function, find the museum on the Warsaw University of Technology campus map.
4	Change the app language to English.

V. RESULTS

To compare the apps, the study utilized the following metrics:

- SUS score – for a reliable assessment of the app's usability, including user satisfaction and memorability.
- uMARS score – which allows evaluation of the overall quality of mobile apps.
- Scenario performance metrics: scenario completion time, scenario execution correctness.
- Eye-tracking measures: fixation count, fixation duration, saccade count.
- Qualitative accessibility assessment – evaluates how well a mobile app meets the needs of users with disabilities using accessibility features and the automated Accessibility Scanner tool.

A. SUS Evaluation

Based on the data collected using the SUS questionnaire from 30 participants, a SUS score was calculated for each participant and each app. The results were then averaged across the three tested apps and presented in Figure 3, together with standard deviations. All apps achieved scores above 68 points, which represents the commonly accepted threshold for satisfactory usability [10].

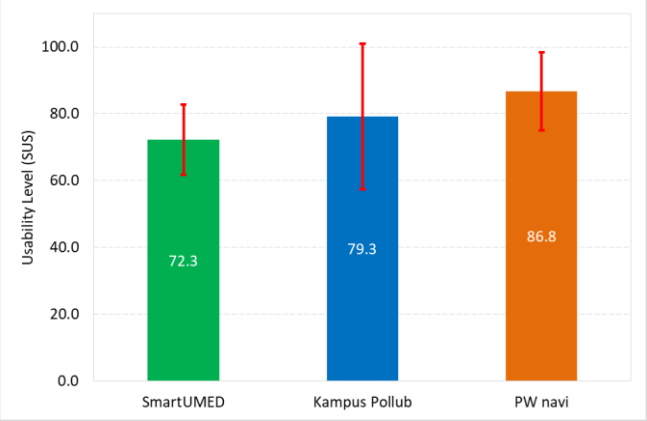


Figure 3. Results of SUS evaluation.

A one-sample t-test was conducted to compare the mean SUS scores of each app against the benchmark value of 68 in order to assess H1. The results showed that the Kampus Pollub ($M = 79.25$, $SD = 11.73$, $p = 0.014$) and PW Navi ($M = 86.75$, $SD = 10.54$, $p < 0.001$) apps achieved usability scores significantly above the benchmark, indicating high usability. In contrast, although the SmartUMED obtained a mean score above 68 ($M = 72.25$, $SD = 21.81$), this difference was not statistically significant ($p = 0.553$), likely due to greater variability in user evaluations.

B. uMARS Evaluation

The uMARS results were used to compare the three apps. Figure 4 presents mean scores and standard deviations for each section: four objective dimensions: engagement (A), functionality (B), aesthetics (C), and information quality (D), and one subjective dimension assessing overall app quality (E).

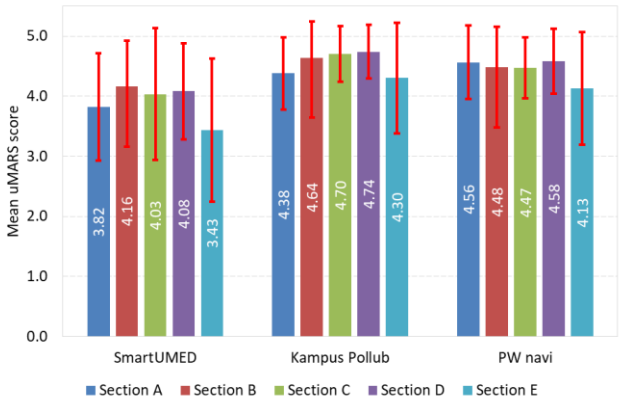


Figure 4. uMARS assessment by sections A–E.

The results show that Kampus Pollub achieved the highest scores in functionality, aesthetics, information quality, and subjective evaluation, while PW Navi scored highest in engagement. SmartUMED received the lowest ratings across the evaluated apps.

In addition, Figure 5 presents a chart showing the average scores calculated for sections A–D (left bars) and section E (right bar).

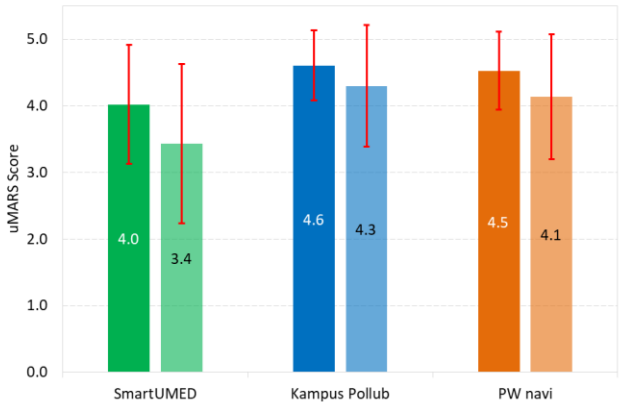


Figure 5. Average uMARS scores for sections A–D and section E.

A Wilcoxon signed-rank test was conducted to compare the objective quality scores (sections A–D) and subjective quality scores (section E) obtained from the uMARS in order to assess H2. The results revealed a statistically significant difference between the two measures ($W = 26.0$, $p < 0.001$), indicating that users evaluated the apps differently in terms of objective quality and subjective experience.

The relationship between SUS usability scores and uMARS interface quality ratings was examined to assess H3. Due to the non-normal distribution of the data, Spearman rank correlation was applied. The analysis revealed no statistically significant relationship between SUS and either the objective quality dimensions (A–D) ($\rho = -0.23$, $p = 0.23$) or subjective quality (section E) ($\rho = -0.28$, $p = 0.13$), indicating that usability scores were not associated with perceived app quality.

C. Scenario Performance Evaluation

Figure 6 shows results of evaluation scenario performance using mean completion time and execution correctness. Correctness exceeded 80% for all apps. Kampus Pollub achieved the shortest completion time (17.6 s) and the highest execution correctness (100%), while SmartUMED required the longest time (21.0 s) and exhibited lower correctness (80.9%). These differences suggest variations in interaction efficiency between the apps.

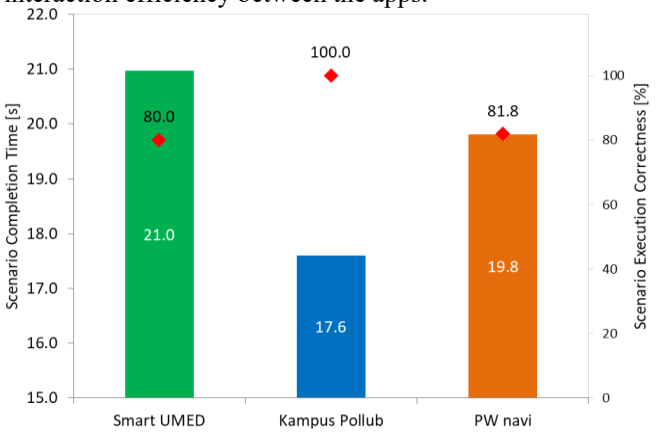


Figure 6. Average results of scenario performance evaluation.

D. Eye-tracking Measures Evaluation

Figures 7–9 present the results for three eye-tracking metrics: fixations count (Figure 7), saccade count (Figure 8), and fixation duration (Figure 9). These metrics enable the identification of usability issues and differences in interaction efficiency between interfaces [2].

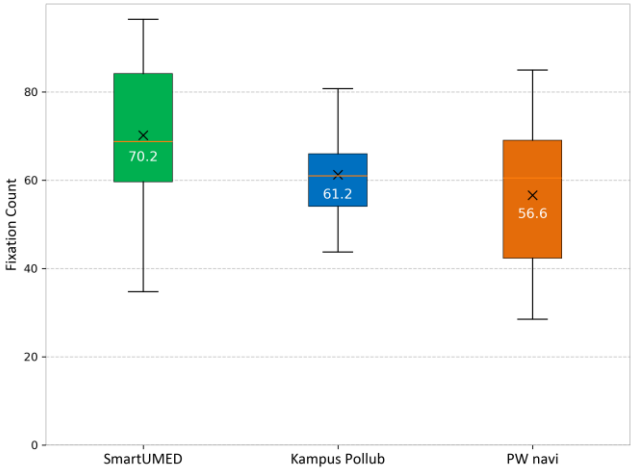


Figure 7. Average fixation count.

Across all three metrics, the highest values were observed for the SmartUMED app. In contrast, PW Navi and Kampus Pollub achieved lower and relatively comparable values. The elevated metrics for Smart UMED suggest that users devoted more visual attention to its interface, which may indicate increased cognitive load or lower interface intuitiveness [11]. Conversely, the lower values recorded for Kampus Pollub and PW Navi point to more efficient visual processing and easier task navigation [12]. However, Kruskal–Wallis tests revealed no statistically significant differences between the apps for fixation count ($H = 1.95$, $p = 0.378$), saccade count ($H = 0.40$, $p = 0.817$), or fixation duration ($H = 2.55$, $p = 0.280$). These observations relate to H4, indicating that while trends in eye-tracking metrics correspond to differences in usability, they do not reach statistical significance.

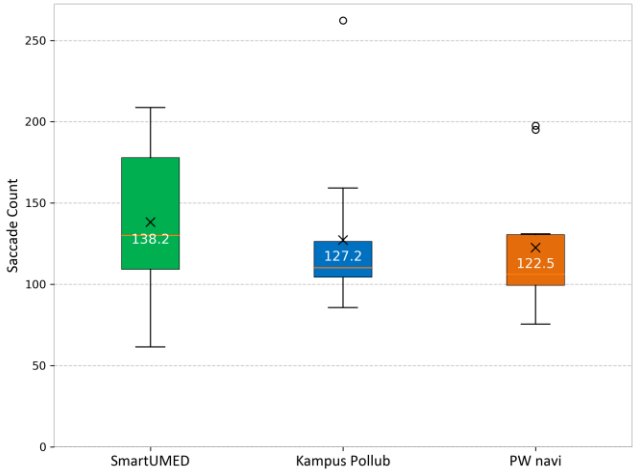


Figure 8. Average sacade count.

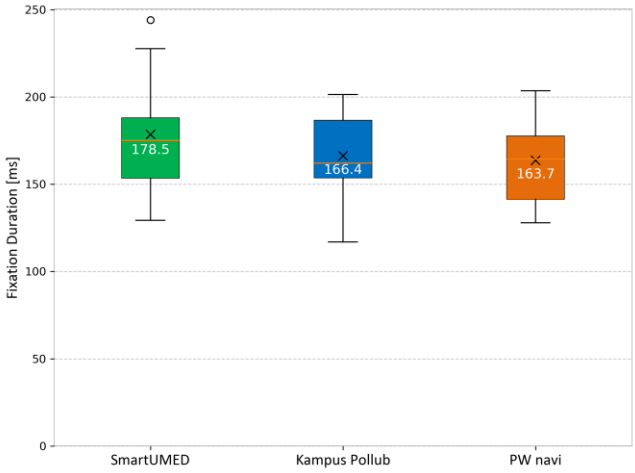


Figure 9. Average fixation duration.

E. Accessibility Evaluation of Mobile Apps

To assess the accessibility of the evaluated apps, both built-in accessibility features and automated analysis results were examined. Table IV presents the additional accessibility settings of each app. These settings enable users with limited abilities to use the apps more easily. The PW Navi offers a significant number of accessibility options compared to the others, which have only a few settings. Accessibility analysis using the Accessibility Scanner revealed that most issues occurred with maps (Table V), where the contrast between the background and interactive elements was often insufficient. Additionally, some interactive elements were too small, potentially hindering use for users with visual or motor impairments.

TABLE IV. ACCESSIBILITY FEATURES PRESENT IN THE APPS

App	Accessibility Features
SmartUMED	**Adaptive route option** – automatically selects a route that avoids obstacles hindering movement for users with mobility impairments. **SOS function** – enables quick contact with campus services supporting users with disabilities.
Kampus Pollub	**Theme change** (light, dark, high contrast) **Ability to report and locate barriers** for wheelchair users
PW Navi	**Visual support options** ● Voice-guided navigation points ● Alternative views adapted for visually impaired users ● Always display location information ● Show warnings when approaching obstacles ● Use vibration notifications ● Path planning based on segments for visually impaired users **Hearing support options** ● Present content using sign language ● Disable voice prompts ● Use vibrations instead of sounds ● Use vibration notifications **Mobility support options** ● Avoid stairs in navigation ● Use wheelchair-accessible navigation

TABLE V. ACCESSIBILITY ANALYSIS RESULTS WITH THE MAP VIEW

App	Panel	Potential issues
SmartUMED	Home	Image contrast
	Map	Low contrast Buttons too small
	Settings	-
Kampus Pollub	Home	-
	Map	Repeated element descriptions Low contrast Problematic element for screen readers
	Settings	-
PW Navi	Home	-
	Map	Low contrast
	Settings	Buttons too small

VI. CONCLUSION AND FUTURE WORK

The aim of this analysis was to evaluate the usability and accessibility of three mobile apps designed for students, using eye-tracking metrics, the SUS and uMARS questionnaires, and an automated accessibility assessment tool.

The usability assessment based on the SUS questionnaire indicated that all evaluated apps achieved good or higher ratings, with scores ranging from 72.3 to 86.8 – well above the standard 68-point threshold. The uMARS evaluation showed that Kampus Pollub achieved the highest scores in most sections, except for engagement (Section A), where PW Navi scored highest. In contrast, SmartUMED received the lowest scores across all categories and exhibited the largest standard deviations, indicating a wider variability in user experiences. This suggests that PW Navi and Kampus Pollub provided more consistent user experiences, whereas SmartUMED elicited diverse responses, possibly reflecting a less refined interface.

Performance metrics, which are among the most commonly used usability measures [2], also supported these findings. Kampus Pollub demonstrated the shortest mean scenario completion times and the highest scenario execution correctness, indicating efficient and accurate task performance [13]. Although SmartUMED users required longer task completion times, scenario correctness remained comparable to PW Navi. Eye-tracking data revealed that SmartUMED elicited a higher number of fixations and saccades, as well as longer mean fixation durations, likely reflecting increased cognitive load or less intuitive interface design [11], [14].

All tested apps exhibited a small number of accessibility-related issues. Based on the results, it is recommended to increase button sizes, improve their visibility, and enhance the contrast of certain interface elements to ensure better legibility for users. It is also worth noting that the apps differ in the range of accessibility support features. PW Navi stands out, offering the most comprehensive set of accessibility functions for users with various types of disabilities.

The analysis of the collected data enabled the verification of the formulated research hypotheses and provided insights into the relationships between usability, perceived quality, and user interaction patterns. The results of this analysis led to the following conclusions:

- H1 was partially supported: The analysis of SUS scores indicates that mobile apps designed for students generally achieve usability levels above the accepted benchmark. However, not all apps reached statistical significance, likely due to variability in user evaluations. Overall, these findings suggest that contemporary student-oriented mobile apps tend to provide a satisfactory level of usability.
- H2 was supported: The analysis revealed statistically significant differences between the mean scores of the objective quality dimensions (sections A–D) and the subjective quality dimension (section E) of the uMARS questionnaire. These findings indicate that users tend to evaluate mobile apps more favorably in terms of objective quality aspects than in terms of their overall subjective experience. This suggests that functional and structural qualities of apps do not necessarily translate into equally positive user perceptions, highlighting the importance of considering both objective and subjective measures in mobile app evaluation.
- H3 was not supported: The analysis did not reveal a statistically significant relationship between usability scores obtained using the SUS and interface quality assessments measured with the uMARS questionnaire. These findings indicate that subjective usability and perceived app quality represent distinct dimensions of user experience and should therefore be considered complementary in evaluation.
- H4 was only partially supported: The analysis indicated observable differences in eye-tracking metrics between the evaluated apps, with lower usability interfaces generally associated with higher fixation counts, longer

fixation durations, and increased saccade activity. However, these differences were not statistically significant.

The study demonstrates that PW Navi and Kampus Pollub provide high levels of usability and accessibility, with PW Navi particularly excelling in accessibility support, whereas SmartUMED requires further improvements to achieve comparable user satisfaction and interface intuitiveness. These findings highlight the value of combining eye-tracking techniques, standardized usability questionnaires, and automated accessibility assessments to obtain a comprehensive evaluation of mobile apps [15].

Although all analyzed apps meet the essential usability and accessibility criteria, further interface optimization is recommended. Among them, PW Navi stands out as a strong example of accessibility-oriented design.

Finally, the limitations of the present study should be acknowledged, including the homogeneity of the study group, which consisted of 30 participants aged 20–25 years. Future research should consider including a more diverse sample of participants, including older individuals and people with varying degrees of disability, in order to increase the generalizability of the findings.

REFERENCES

[1] Web Content Accessibility Guidelines (WCAG) 2.2. [Online]. Available from: https://www.w3.org/TR/WCAG22/ 2026.02.06

[2] J. Š. Novák, J. Masner, P. Benda, P. Šimek, and V. Merunka, "Eye tracking, usability, and user experience: A systematic review," Int. J. Hum.-Comput. Interact., vol. 40, no. 17, pp. 4484–4500, 2024, doi: 10.1080/10447318.2023.2221600.

[3] M. Miłosz, M. Plechawska-Wójcik, and M. Dzieńkowski, "Testing the Quality of the Mobile Application Interface Using Various Methods—A Case Study of the T1DCoach Application," Appl. Sci., vol. 14, no. 15, p. 6583, 2024, doi: 10.3390/app14156583.

[4] G. J. Nowicki et al., "Development and pre-evaluation of a 'DiagNurse' mobile app to support nurses in clinical diagnosis using the ADDIE model," Sci. Rep., vol. 14, no. 1, p. 29765, 2024, doi: 10.1038/s41598-024-81813-0.

[5] I. Julian, D. F. Murad, and R. Y. Riva'i, "Combining UEQ and Eye-Tracking Method as Usability Evaluation for Mobile Apps," in Proc. 2021 3rd Int. Conf. Cybern. Intell. Syst. (ICORIS), Makassar, Indonesia, 25–26 Oct. 2021, pp. 1–6.

[6] B. Taszarek, E. Książek, and K. Konikowska, "Functional and Quality Analysis of Nutrition-Related Mobile Apps Available in Poland Using the MARS Scale," J. Food Eng. Sci., vol. 41, pp. 42–54, 2025, doi: 10.15611/nit.2025.41.05.

[7] A. Kurnia, F. M. Said, and S. L. Paduragan, "User-Centered Usability Evaluation of the DiaCare App for Diabetes Self-Management: A uMARS Analysis," Al-Rafidain J. Med. Sci., vol. 7, no. 2, pp. 171–176, 2024, doi: 10.54133/ajms.v7i2.1499.

[8] J. Sauro and J. R. Lewis, Quantifying the User Experience: Practical Statistics for User Research. San Francisco, CA: Morgan Kaufmann, 2016.

[9] S. R. Stoyanov, L. Hides, D. J. Kavanagh, and H. Wilson, "Development and validation of the user version of the Mobile Application Rating Scale (uMARS)," JMIR mHealth uHealth, vol. 4, no. 2, p. e72, 2016, doi: 10.2196/mhealth.5849.

[10] J. Brooke, "SUS: A Retrospective," J. Usability Stud., vol. 8, pp. 29–40, 2013.

[11] A. R. Idrees et al., "Exploring the usability of an internet-based intervention and its providing eHealth platform in an eye-tracking study," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 7, pp. 9621–9636, 2023, doi: 10.1007/s12652-023-04635-4.

[12] K. Holmqvist et al., Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford, U.K.: Oxford University Press, 2011.

[13] A. N. Tuch, J. A. Bargas-Avila, K. Opwis, and F. H. Wilhelm, "Visual complexity of websites: Effects on users' experience, physiology, performance, and memory," Int. J. Hum.-Comput. Stud., vol. 67, no. 9, pp. 703–715, 2009, doi: 10.1016/j.ijhcs.2009.04.002.

[14] J. Falkowska, J. Sobecki, and M. Falkowski, "Utilization of Eye-Tracking Metrics to Evaluate User Experiences—Technology Description and Preliminary Study," Sensors, vol. 25, p. 6101, 2025, doi: 10.3390/s25196101.

[15] K. Łazaruk, P. Tokarski, K. Rybak, M. Plechawska-Wojcik, and M. Dzieńkowski, "The Force of Habit: Comparing Graphical User Interfaces of Popular Operating Systems Using Eye-Tracking Analysis," IEEE Access, pp. 1–28, 2025, doi: 10.1109/ACCESS.2025.3614018.

Impact Analysis of Microinteractions on User Experience in User Interfaces

Karol Łazaruk, Natalia Prażmo, Karolina Rybak, Mariusz Dzieńkowski, Piotr Tokarski, and Małgorzata Plechawska-Wójcik
Department of Computer Science
Lublin University of Technology
Lublin, Poland
e-mail: k.lazaruk@pollub.pl, s95537@pollub.edu.pl, k.rybak@pollub.pl, m.dzienkowski@pollub.pl, p.tokarski@pollub.pl, m.plechawska@pollub.pl

***Abstract*—Microinteractions constitute a crucial element in enhancing the usability and overall user experience of digital interfaces. The objective of this study is to empirically investigate the impact of incorporating microinteractions into an e-commerce interface by comparing two versions: one with microinteractions and one without them. The experimental procedure employed eye-tracking technology in conjunction with the Short version of the User Experience Questionnaire (UEQ-S), which was extended with additional author-developed items. The results revealed a substantial improvement in usability, a higher intention to reuse the system, and approximately double the levels of excitement and enjoyment when interacting with the interface enriched with microinteractions. Notably, participants using the version without microinteractions reported a perceived need for these features more frequently, potentially indicating heightened frustration due to their absence. Although not all observed differences reached statistical significance, the overall findings support the conclusion that microinteractions have a positive influence on the quality of the user experience.**

Keywords-microinteractions; user interface; user experience; usability; eye tracking; UEQ-S questionnaire.

I. Introduction

User eXperience (UX) and User Interface (UI) design are central to modern digital systems, with microinteractions, such as animations and feedback, playing an increasingly important role in shaping user perception. As defined by Saffer, microinteractions consist of triggers, rules, feedback, and loops [1], which together enhance usability and engagement.

Research on microinteractions focuses primarily on their impact on UX and design approaches within user interfaces. McDaniel [2] defines microinteractions as structured units composed of triggers, rules, feedback, and loops that enhance usability and system efficiency. Gonzales et al. [3] emphasize the role of observing user behavior in interface design, while Boyd and Bond [4] demonstrate that microinteractions positively affect perceived usability in studies using System Usability Scale (SUS) and UEQ metrics. Falkowska et al. [5] show that rapid feedback improves form completion efficiency and user satisfaction. In mobile health applications, microinteractions increase accessibility and satisfaction [6], and similar benefits have been observed on academic platforms [7]. Ahn et al. [8] link personalization and visual interactivity to user agency and intention to recommend, while Reyneke [9] demonstrates their role in emotional attachment and brand loyalty. Positive emotional responses have also been associated with long-term engagement [10].

In the context of UX animation, Burge [11] reports that animated microinteractions enhance credibility, particularly among older users, while Lomakina [12] applies Disney's animation principles to support intuitive interface design. On platforms such as TikTok, effective microinteraction design is associated with increased retention and loyalty [13], while positive effects on satisfaction and usability have also been reported in e-commerce contexts [14]. Sosa-Tzec and Stolterman [15] emphasize the semiotic role of animations in clarifying functionality, and Jergović et al. [16] underline the importance of feedback-rich microinteractions for engagement. Smooth transitions between interface states have been shown to reduce disorientation [17], while appropriate animation timing supports UX fluency [18][19].

Research on wearable devices indicates that microinteraction timing and complexity influence cognitive load and user comfort [20]. Betz and Hall [21] show that optimized microinteractions increase satisfaction and adoption in institutional repositories, while Avila-Munoz et al. [22] and Antal [23] highlight the importance of balancing aesthetic and functional aspects to improve feedback, accessibility, and interface coherence.

A/B testing is commonly used to evaluate usability through controlled comparisons of interface variants [24]. Eye tracking complements this approach by providing objective measures, such as fixation duration and gaze trajectories [25], with machine learning increasingly used to support adaptive interface design [26]. Among subjective methods, questionnaires remain essential, with the UEQ-S validated as a reliable tool for assessing pragmatic and hedonic UX dimensions [27].

In summary, microinteractions significantly influence usability, emotional engagement, and user satisfaction. Their effectiveness depends on intentional design supported by rigorous evaluation using complementary research methods.

The rest of the paper is structured as follows: Section II describes the research methods, Section III presents the research results, Section IV discusses the results, and Section

 ISBN: 978-1-68558-383-5

V concludes the paper.

II. Materials and Methods

The study analyzed the impact of microinteractions on the interface and UX, using eye-tracking technology and an evaluation questionnaire. Data were collected using an eye-tracking device, which enabled the recording and analysis of eye movements to evaluate how users interacted with the interface. Key eye-tracking metrics included Fixation Count, Fixation Duration (Dwell Time), Peak Saccade Velocity, Time to First Fixation (TTFF), as well as visualizations, such as heatmaps and scanpaths.

Complementarily, the UEQ-S was used to evaluate the usability and attractiveness of the interface [27], extended with additional items developed by the authors to allow for a more nuanced assessment.

A. Research plan

The research design comprised several stages (Figure 1) and was preceded by the development of two interface versions: one with microinteractions and one without. Participants were randomly assigned to one of two groups, each interacting with a different version of the system. Additional questionnaire items were included to enable a more detailed assessment of usability and user perception.

The experimental procedure consisted of two identical stages for both groups: eye-tracking analysis of responses to visual stimuli and an evaluation of the overall interface experience. After data validation, quantitative and qualitative analyses were performed.

B. Research object

For the purposes of the experiment, a simple website in the form of a store for technology items, such as electronics and accessories was implemented. The site was designed as a browser-based application in two versions: one included microinteractions, the other did not. The store consisted of five main views: home page, product list, product details, shopping cart, and registration page. Users could perform basic actions, such as registering, browsing, searching for products, and adding and removing them from the shopping cart.

The main difference between the interface versions was the use of microinteractions, such as button animations, user guidance, and feedback on actions performed. The application was designed for intuitive use and its interface provided

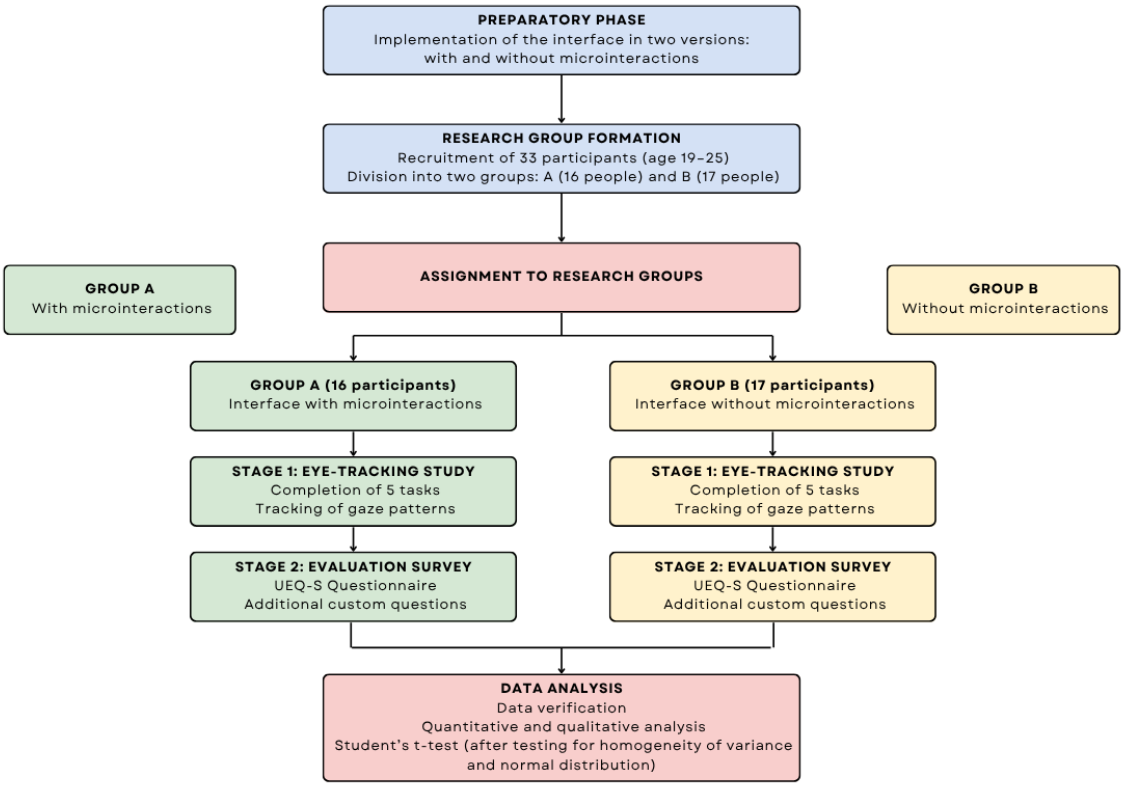


Figure 1. Study design.

readability and user-friendliness, allowing participants to focus on the tasks at hand.

C. Research group

The study included 33 participants aged 19 to 25. Their mean age was 21.9 with a standard deviation of 1.06. The study group mainly consisted of men (30 males, 3 females). Most participants had normal or corrected vision. Minor uncorrected visual impairments were present in three cases but did not affect the results due to the task nature and short viewing distance. All the respondents were computer science students with advanced digital literacy and prior experience using online stores. In accordance with the A/B testing [24], participants were randomly assigned to two equivalent experimental groups, which tested two alternative versions of the interface that differed in the presence (A, $n = 16$) or absence (B, $n = 17$) of microinteractions, allowing for a reliable assessment of their impact on the UX under experimental conditions.

D. Research stand

The experiment was conducted under controlled laboratory conditions at the Lublin University of Technology, with appropriate lighting and ergonomic workstations. Eye movements were recorded using a Gazepoint GP3 HD remote eye tracker (150 Hz, 0.5–1° accuracy). Data were recorded and analyzed using iMotions software (v. 9.1).

E. Experiment description

During the eye-tracking phase, participants located and interpreted information across key website views, including promotional content on the main page (Task 1), product pricesand specifications on product list (Task 2) and detail pages (Task 3), interface controls in the shopping cart (Task 4), and form errors during registration (Task 5).

After completing the tasks, participants filled in an evaluation survey comprising the UEQ-S questionnaire [27] and 14 additional statements grouped into three dimensions. The usability dimension (S1–S7) assessed ease of information retrieval (S1), visual clarity (S2), task orientation (S3), action awareness (S4), feedback quality (S5), sense of control (S6), and perceived functional value of microinteractions (S7). The acceptability dimension (S8–S11) measured willingness to reuse the website (S8), preference for microinteractions (S9), perceived negative impact (S10), and influence on visual appeal (S11). The user experience dimension (S12–S14) evaluated enjoyment (S12), perceived system responsiveness (S13), and navigation smoothness (S14).

Data were collected individually and compared between groups to assess the impact of microinteractions.

F. Research metrics

To assess visual attention and cognitive load during UI interaction, standard eye-tracking metrics were applied, providing quantitative insight into visual processing and navigation. The analyzed measures included Time to First Fixation, Fixation Count, Dwell Time, Peak Saccade Velocity, and gaze visualizations in the form of heatmaps and scanpaths. These metrics were compared across interface variants to evaluate the impact of microinteractions on attention, search efficiency, and cognitive workload.

III. Research Results

This section presents the study's results, including the most used eye-tracking metrics for evaluating usability, as well as users' responses to the questionnaire.

A. Eye-tracking results

Table I presents average values for TTFF, Dwell Time, Fixation Count, and Peak Saccade Velocity for interfaces with and without microinteractions. The version with microinteractions showed shorter TTFF, higher Dwell Time, increased Fixation Count, and higher Peak Saccade Velocity.

TABLE I. Average values for chosen metrics with the standard deviation

Metric	Micro.	No Micro.
TTFF [s]	1.88 ± 1.06	2.34 ± 0.79
Dwell Time [%]	25.48 ± 8.00	18.28 ± 5.38
Fixation Count	5.36 ± 2.47	4.06 ± 1.82
Peak Sac. Vel. [deg/s]	80.68 ± 6.25	72.46 ± 4.34

Average TTFF values (Figure 2) were generally lower for the interface with microinteractions, except in Task 3. The largest differences occurred in Tasks 2 and 4.

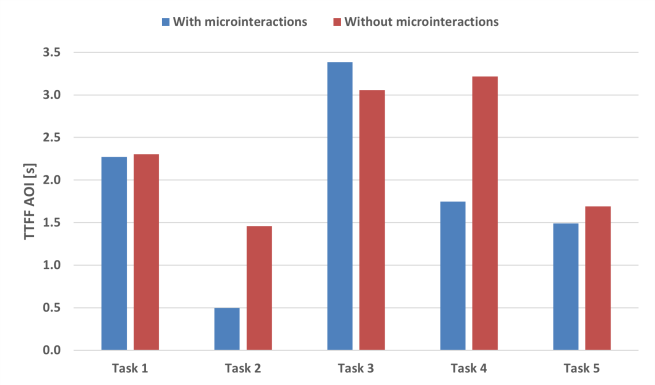


Figure 2. Average TTFF values for each task.

Average Dwell Time in the Area Of Interest (AOI) (Figure 3) was higher for the interface with microinteractions across all tasks except Task 4, where values were comparable.

The interface with microinteractions showed higher Fixation Count values and greater variance before entering the AOI compared to the non-interactive interface. Higher Peak Saccade Velocity values were also observed for the interactive version, with the distribution shifted toward higher velocities. In contrast, the non-interactive interface exhibited a more uniform distribution of viewing behavior and fewer high-velocity outliers.

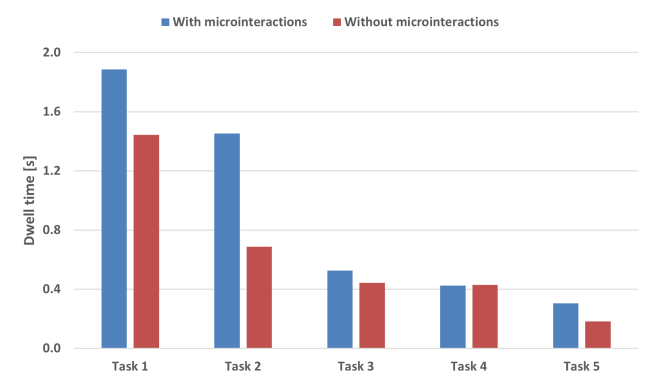


Figure 3. Average Dwell Time on AOI for each task.

Heatmaps for Task 3 are presented in Figure 4. For the interface with microinteractions (accordion – collapsible content), gaze concentrations were more localized around the expanded content section containing the searched information. In contrast, the non-interactive interface showed a more dispersed gaze distribution across the interface.

Similar behaviour can be noted for Task 5 (Figure 5). In the case of the interface with microinteractions (highlighting, error icon, and clear feedback), the strongest user attention is focused on the form field containing the error and the accompanying message. As for the version without microinteractions, users' gaze is primarily focused on the field above the actual error, which may indicate difficulty in locating it or a possible shift caused by measurement limitations. There is also a clear dispersion of attention, represented by green-yellow cold areas.

The scanpaths for Task 3 are shown in Figure 6. For the interface with microinteractions, scanpaths consisted of fewer fixations and shorter saccades. The non-interactive interface exhibited scanpaths with a higher number of fixations and longer saccades.

Similar observations can be made in the case of Task 5. Selected scanpaths, which are presented in Figure 7, show that in the interface with micro-interactions, the scanning pattern is significantly less chaotic. The number of fixations has decreased by more than half, and the number of long saccades has also been clearly reduced.

B. Questionnaire results

The UEQ-S questionnaire comprises eight bipolar items rated on a 7-point scale, assessing pragmatic quality (usability and functionality) and hedonic quality (aesthetics, stimulation, and attractiveness).

UEQ-S results (–3 to 3; Figure 8) show higher ratings for the interface with microinteractions across most dimensions, particularly usability, efficiency, perspicuity, supportiveness, and stimulation. The largest differences were observed in usability-related dimensions, indicating improved navigation and information processing. Higher stimulation and attractiveness scores suggest increased user engagement, while novelty showed minimal differences, implying a limited effect of microinteractions on perceived innovation.

Average ratings (1–7; Figure 9) show consistently higher scores for the interface with microinteractions across overall, pragmatic, and hedonic scales. The largest difference was observed for pragmatic quality, indicating a positive effect on perceived functionality, while the smallest difference occurred for hedonic quality, which nonetheless favored the microinteractions version.

Statistical significance between the two interface versions was assessed using UEQ-S results from an A/B test with two independent groups (n = 16 and n = 17).

As shown in Table II, independent samples t-tests revealed a statistically significant difference in pragmatic quality between the two interface versions ($p = 0.004$), whereas no significant difference was observed for hedonic quality ($p = 0.328$). The interface with microinteractions received higher ratings for supportiveness, efficiency, and perspicuity. The difference in the overall UEQ-S score did not reach statistical significance ($p = 0.063$).

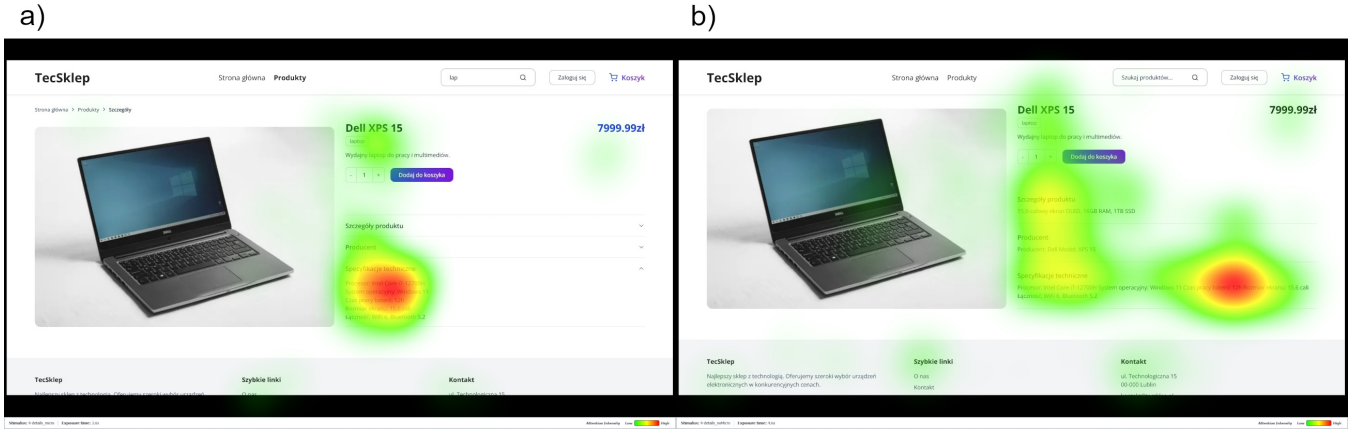


Figure 4. Heat map for Task 3: a) with microinteractions, b) without microinteractions.

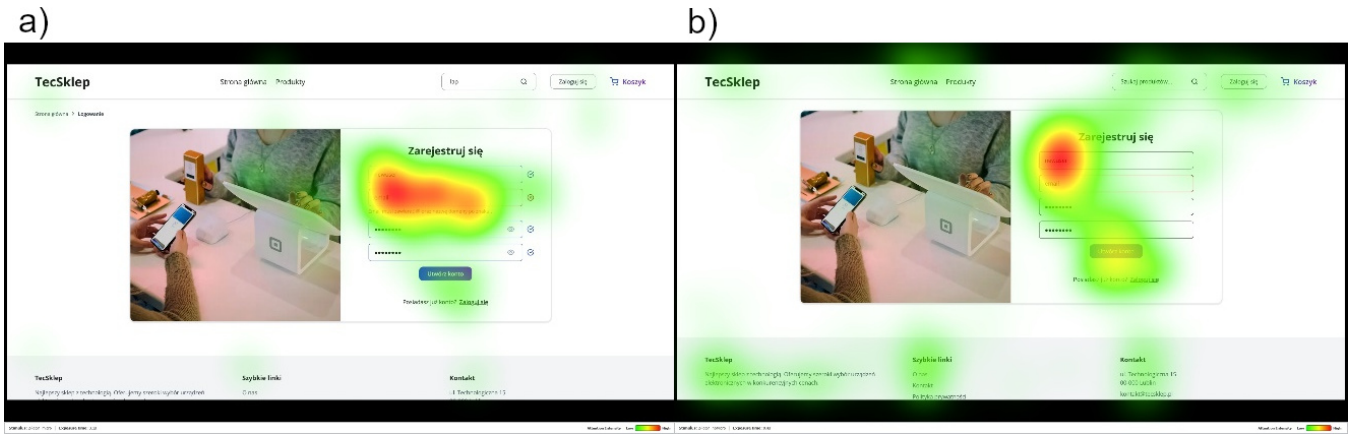


Figure 5. Heat map for Task 5: a) with microinteractions, b) without microinteractions.

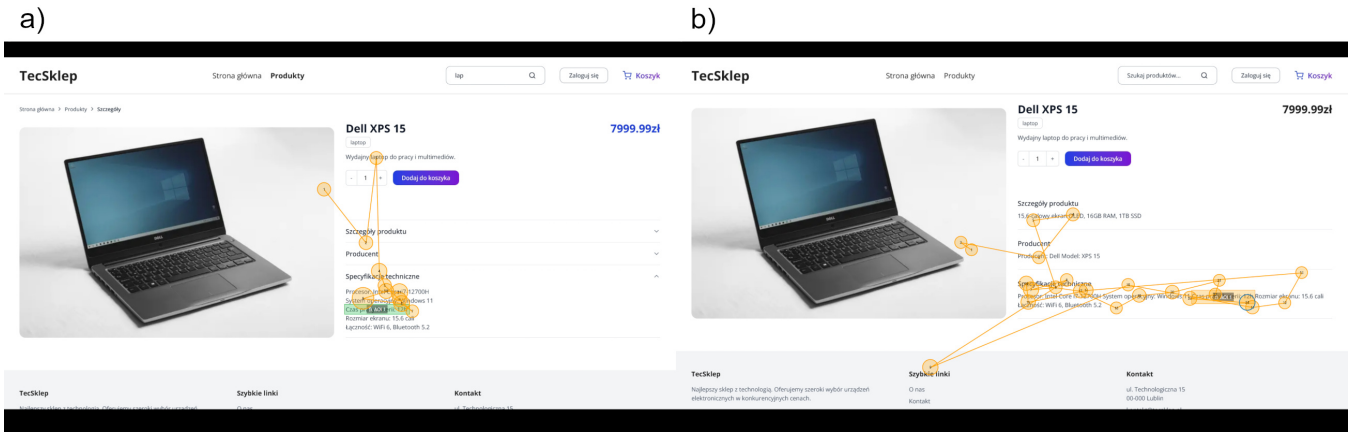


Figure 6. Scanning paths for Task 3: a) with microinteractions, b) without microinteractions.

The second part of the questionnaire comprised 14 statements evaluated on a 7-point Likert scale. Twelve statements were positively worded, while two were negatively worded. For the purpose of aggregated analyses, responses to the negatively worded statements were reverse-coded so that higher values consistently reflected more positive evaluations across all items.

Figure 10 presents the mean scores for individual statements for both interface versions. In the majority of cases, higher scores were observed for the version with microint-

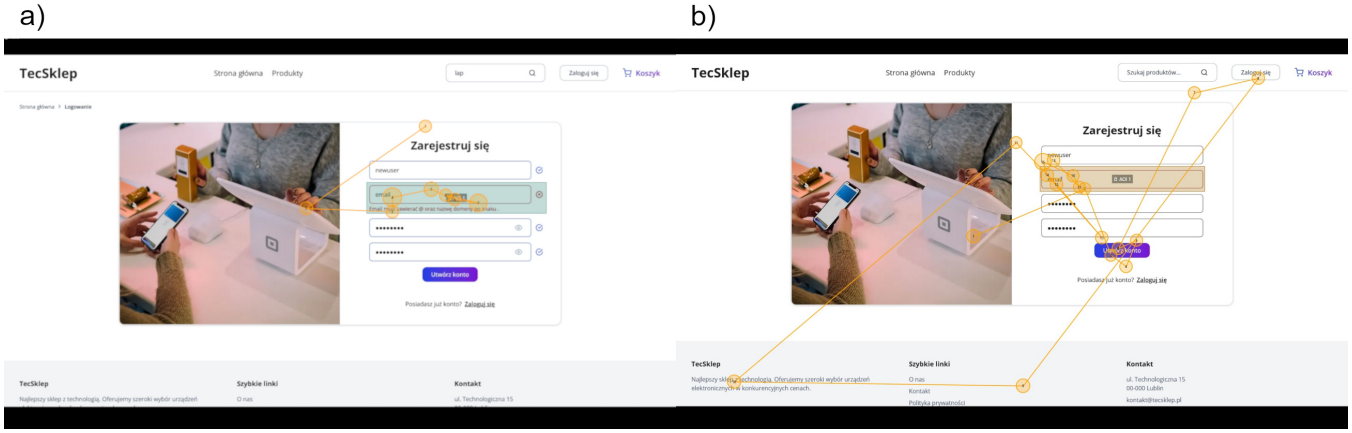


Figure 7. Scanning paths for Task 5: a) with microinteractions, b) without microinteractions.

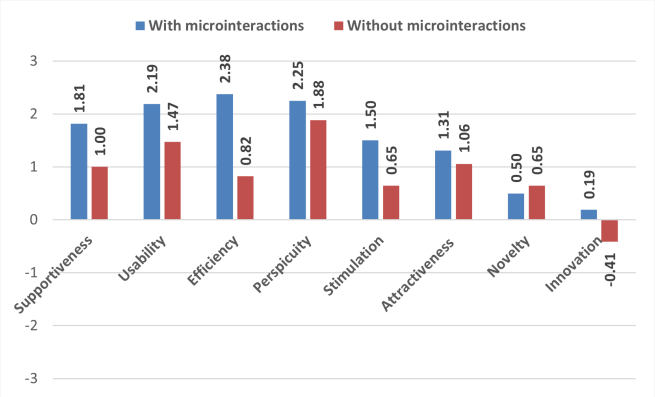


Figure 8. Average user UEQ-S ratings for both interface versions.

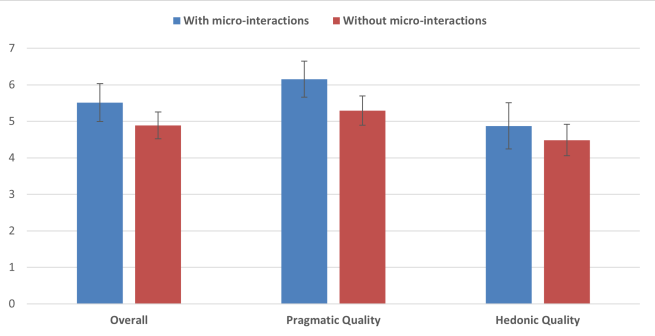


Figure 9. Average UEQ-S ratings by dimension.

TABLE II. STATISTICAL TEST RESULTS FOR UEQ-S FOR EACH DIMENSION

Scale	Test result	Stat. sig. diff.
UEQ-S overall	0.063	None
Pragmatic quality	0.004	Present
Hedonic quality	0.328	None

eractions. The largest differences between the two versions were recorded for statements S4–S6 (usability dimension) and S12–S13 (user experience dimension).

For statements S2, S7, and S9–S11, higher mean ratings were observed for the interface without microinteractions. Overall, the independent samples t-test results (Table III) revealed statistically significant differences between the two interface versions for usability, acceptability, and overall user experience.

Within the usability dimension (S1–S7), statistically significant differences in favor of the interface with microinteractions were found for action awareness (S4: $p = 0.007$), feedback clarity (S5: $p < 0.001$), and sense of control (S6: $p < 0.001$). The difference for information findability did not reach statistical significance (S1: $p = 0.059$). For statement S7, higher ratings were recorded for the interface without microinteractions ($p = 0.027$).

In the acceptability dimension (S8–S11), the interface

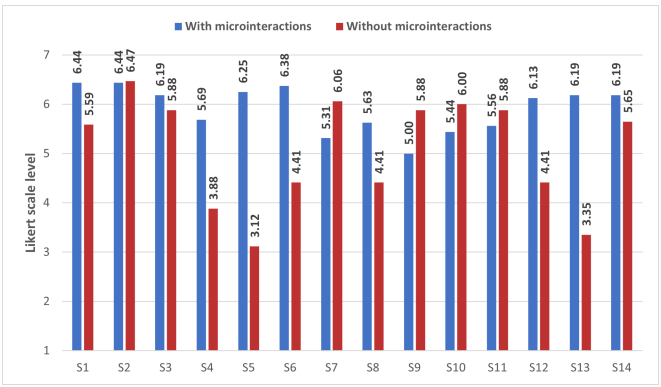


Figure 10. Average user ratings for additional statements.

TABLE III. STATISTICAL TEST RESULTS FOR ADDITIONAL QUESTIONS

Statement	S1	S2	S3	S4	S5
p-value	0.059	0.915	0.527	0.007	0.000
Statement	**S6**	**S7**	**S8**	**S9**	**S10**
p-value	0.001	0.027	0.021	0.033	0.154
Statement	**S11**	**S12**	**S13**	**S14**	**Total**
p-value	0.423	0.004	0.000	0.139	0.017

with microinteractions received higher ratings for willingness to reuse the interface (S8: $p = 0.021$). No statistically significant difference was observed for perceived distraction (S10: $p = 0.154$). Higher ratings for the interface without microinteractions were found for statement S9 ($p = 0.033$).

For the user experience dimension (S12–S14), statistically significant differences in favor of the interface with microinteractions were observed for enjoyment (S12: $p = 0.004$) and perceived responsiveness (S13: $p < 0.001$). The difference for navigational fluidity was not statistically significant (S14: $p = 0.139$).

Aggregated analysis across all statements showed a statistically significant difference between the two interface versions ($p = 0.017$).

IV. DISCUSSION

The study demonstrates that microinteractions positively influence interface usability, User eXperience (UX), and overall acceptability. Objective eye-tracking measures and subjective assessments (UEQ-S and additional Likert items) provide converging evidence of their impact.

Eye-tracking results indicate that interfaces with microinteractions achieved shorter average TTFF, higher Fixation Count, longer Dwell Time in the AOI, and higher Peak Saccade Velocity, suggesting increased engagement and more dynamic visual exploration. Analyses of heatmaps and scanpaths further confirm that microinteractions improved visual focus on key elements, reduced attention dispersion, and supported faster localization of relevant information. In more

complex interface structures, they led to fewer fixations and shorter scanpaths, indicating more efficient visual processing.

UEQ-S results confirmed the positive effect of microinteractions on perceived supportiveness, usability, and efficiency. Additional questionnaire items supported these findings, with users rating the interactive version as more responsive and more effective in providing feedback. The statistically significant difference in pragmatic quality between interfaces further reinforces the positive influence of microinteractions on usability. These results are consistent with prior studies showing that microinteractions improve engagement, feedback clarity, and interaction efficiency [2][5][7]. Research by [3][6][22] similarly emphasizes their role in enhancing interface intuitiveness, reflected in the improved perspicuity scores observed in this study. An exception was novelty, which was rated higher for the non-interactive version, likely due to the widespread presence of microinteractions in contemporary interfaces, such as TikTok [13].

The results also demonstrate a positive impact of microinteractions on UX. Questionnaire data showed significant improvements in stimulation, enjoyment, perceived responsiveness, and sense of control. These findings align with studies indicating that animation-based microinteractions enhance engagement and intuitiveness [12][16]. Microinteractions also increased perceived interface attractiveness and users' willingness to reuse the system. Eye-tracking data corroborated these outcomes by revealing more focused and less chaotic navigation patterns. Importantly, no negative impact on user comfort was observed, consistent with findings that microinteractions enhance UX by capturing attention and stimulating interest [4][10][15]. Users of the non-interactive version more often indicated that adding microinteractions could improve functionality and expressed a stronger preference for their inclusion in future interfaces, suggesting that their absence was noticeable [3][16][21].

Microinteractions did not significantly affect navigation fluidity, potentially due to the need for more refined animation timing, as noted in prior work [18][19]. Nevertheless, consistent with earlier research [8]–[10][15][17], microinteractions elicited positive emotional responses, leading to higher hedonic ratings.

In conclusion, the findings indicate that microinteractions significantly enhance usability, UX, and interface acceptability, confirming their value as an effective design tool in user interface development.

A. Limitations of the study

Certain subjective evaluations, including preferences for the version without microinteractions and assessments of interface originality, revealed discrepancies between participants. This suggests that individual expectations and prior experience may significantly influence the perception of microinteractions, thereby limiting the conclusiveness of findings regarding their perceived attractiveness and innovativeness.

Another limitation may be the laboratory-based experimental setting, which does not fully reflect naturalistic user behavior in real-world environments. The tasks were short and strictly defined (e.g., information search, shopping cart interaction, registration form completion), which may reduce the validity of the findings.

Additionally, the sample size was relatively small (n = 33) and homogeneous, consisting exclusively of computer science students with high digital literacy and prior experience with e-commerce systems. This limits the generalizability of the results to broader and more diverse user populations.

Finally, the analysis was restricted to two interface variants (with and without microinteractions), without considering different types or combinations of microinteractions, which may exert varying effects on attention allocation and cognitive load.

V. Conclusion and Future Work

The study confirmed that microinteractions significantly enhance interface usability and UX, as evidenced by improved eye-tracking metrics and questionnaire results indicating greater engagement, clarity, supportiveness, enjoyment, sense of control, and willingness to reuse the interface.

However, some results, such as mixed ratings of originality and occasional preference for the non-interactive version, suggest that user expectations and prior experience influence the perception of microinteractions, particularly in terms of innovativeness.

Future research should examine interactive website implementations, broader usage scenarios, and diverse types of microinteractions, while including participants with varying levels of technological proficiency to better understand their overall impact.

References

[1] D. Saffer, Microinteractions. O'Reilly, 2013, p. 151, ISBN: 9781449342807.

[2] R. McDaniel, "Understanding microinteractions as applied research opportunities for information designers," Communication Design Quarterly, vol. 3, pp. 55–62, 2 Mar. 2015, ISSN: 2166-1642. DOI: 10.1145/2752853.2752860

[3] S. Gonzales et al., "User testing with microinteractions," Information Technology and Libraries, vol. 40, 1 Mar. 2021, ISSN: 2163-5226. DOI: 10.6017/ital.v40i1.12341

[4] K. Boyd and R. Bond, "Can micro interactions in user interfaces affect their perceived usability?" In European Conference on Cognitive Ergonomics 2021, ACM, Apr. 2021, pp. 1–5, ISBN: 9781450387576. DOI: 10.1145/3452853.3452865

[5] J. Falkowska, B. Kilijańska, J. Sobecki, and K. Zerka, "Microinteractions of forms in web based systems usability and eye tracking metrics analysis," in 2019, pp. 164–174. DOI: 10.1007/978-3-319-94334-3_18

[6] P. Bajda, R. Baliński, and M. Dzieńkowski, "Research on user experience during interactions with mobile applications for diabetics," Journal of Computer Sciences Institute, vol. 29, pp. 333–340, Dec. 2023, ISSN: 2544-0764. DOI: 10.35784/jcsi.3779

[7] Z. Shwany, C. Salh, H. Abdulrahman, and K. Khoshnaw, "Evaluating the impact of micro-interactions on user engagement," Qalaai Zanist Scientific Journal, vol. 9, pp. 1468–1485, 2 Jul. 2024, ISSN: 25186558. DOI: 10.25212/lfu.qzj.9.2.54

[8] J. Ahn, J.-M. Park, W.-H. Lee, and G.-Y. Noh, "Website interactivity and processing: Menu customization and sense of agency are keys to better interaction design," International Journal of Human-Computer Studies, vol. 147, p. 102 581, Mar. 2021, ISSN: 10715819. DOI: 10.1016/j.ijhcs.2020.102581

[9] R. Reyneke, "Improving interactive user experience with microinteractions: An application of biometric and affect detection systems on landing pages [Master's thesis]," Brigham Young University, 2019.

[10] J. Y. Ma and C.-C. Chen, "Evaluating user perception and emotion of microinteractions using a contradictory semantic scale," Journal of the Society for Information Display, vol. 30, pp. 103–114, 2 Feb. 2022, ISSN: 1071-0922. DOI: 10.1002/jsid.1075

[11] L. F. Burge, "The impact of website interface micro animations on user perceptions of trust, credibility, and design quality [Undergraduate honors project]," University of Texas at Arlington, 2024.

[12] M. Lomakina, "Defining microinteractions: Animation in UX [Bachelor's thesis]," University of Applied Sciences, 2017.

[13] J. Cai and Y. Luo, "Micro-interaction in SNS interface: Research on the design of micro-interaction in SNS interface [Master's thesis]," Jönköping University, 2022.

[14] E. S. Ari, "Cognitive effect of product-based animation on interface aesthetics of web-based stores," Entertainment Computing, vol. 52, p. 100 895, Jan. 2025, ISSN: 18759521. DOI: 10.1016/j.entcom.2024.100895

[15] O. Sosa-Tzec and E. S. Bergqvis, "Delight by motion: Investigating the role of animation in microinteractions," in the Motion Design Educators Summit Conf Proc, 2021, pp. 10–13.

[16] M. Jergović, N. S. Loknar, T. K. Ivančević, and A. A. Cmrk, "Micro-interactions within user interfaces," in Proceedings - The Twelfth International Symposium GRID 2024, University of Novi Sad Faculty of Technical Sciences Department of Graphic Engineering and Design 21000 Novi Sad, Trg Dositeja Obradovića 6, Nov. 2024. DOI: 10.24867/GRID-2024-p23

[17] M. Donati, G. Mori, and F. Paternò, "Understanding the transitions between web interfaces designed to stimulate specific emotions," Universal Access in the Information Society, vol. 19, pp. 391–407, 2 Jun. 2020, ISSN: 1615-5289. DOI: 10.1007/s10209-019-00649-y

[18] Y. Ge et al., "User perception of animation fluency: The effect of time duration in different phases of animated transitions during application usage," International Journal of Human-Computer Studies, vol. 186, p. 103 257, Jun. 2024, ISSN: 10715819. DOI: 10.1016/j.ijhcs.2024.103257

[19] F. Huth, T. Blascheck, S. Koch, and T. Ertl, "Studies and design considerations for animated transitions between small-scale visualizations," Journal of Visualization, vol. 26, pp. 1421–1443, 6 Dec. 2023, ISSN: 1343-8875. DOI: 10.1007/s12650-023-00937-z

[20] X. Yan, Y. Li, B. Huang, S. Y. Park, and M. W. Newman, "User burden of microinteractions: An in-lab experiment examining user performance and perceived burden related to in-situ self-reporting," in Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction, ACM, Sep. 2021, pp. 1–14, ISBN: 9781450383288. DOI: 10.1145/3447526.3472046

[21] S. Betz and R. Hall, "Self-archiving with ease in an institutional repository: Microinteractions and the user experience," Information Technology and Libraries, vol. 34, pp. 43–58, 3 Sep. 2015, ISSN: 2163-5226. DOI: 10.6017/ital.v34i3.5900

[22] R. Avila-Munoz, J. Clemente-Mediavilla, and P.-L. Perez-Luque, "Communicative functions in human-computer interface design: A taxonomy of functional animation," Review of Communication Research, vol. 9, pp. 119–146, 2021, ISSN: 22554165. DOI: 10.12840/ISSN.2255-4165.030

[23] A. Antal, "Micro-interactions and animations in UX design for mobile applications," MASTERCOM – Politehnica Graduate Student Journal of Communication, vol. 7, pp. 32–44, 2 2022.

[24] F. Quin, D. Weyns, M. Galster, and C. C. Silva, "A/B testing: A systematic literature review," Journal of Systems and Software, vol. 211, p. 112 011, May 2024, ISSN: 01641212. DOI: 10.1016/j.jss.2024.112011

[25] M. García and S. Cano, "Eye tracking to evaluate the user eXperience (UX): Literature review," in 2022, pp. 134–145. DOI: 10.1007/978-3-031-05061-9_10

[26] J. Š. Novák, J. Masner, P. Benda, P. Šimek, and V. Merunka, "Eye tracking, usability, and user experience: A systematic review," International Journal of Human–Computer Interaction, vol. 40, pp. 4484–4500, 17 Sep. 2024, ISSN: 1044-7318. DOI: 10.1080/10447318.2023.2221600

[27] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Design and evaluation of a short version of the user experience questionnaire (UEQ-S)," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, p. 103, 6 2017, ISSN: 1989-1660. DOI: 10.9781/ijimai.2017.09.001

Co-Designing A Low-Barrier Digital Platform for Culturally Diverse Communities

Lauren W. Forbes, Sai Meenakshi Hariharan,
Venkat Sai Subash Panchakarla, Venkata Guna Sundhar Grandhe, Siddharth Urankar, Annu Prabhakar
University of Cincinnati
Cincinnati, Ohio, USA
Emails: forbesln@ucmail.uc.edu,harihasi@mail.uc.edu, panchavh@mail.uc.edu, grandhvh@mail.uc.edu,
urankasj@mail.uc.edu, annu.prabhakar@uc.edu,

***Abstract*—This paper discusses the design, development, and deployment of an iteratively designed information management and visualization platform known as the People's Market Dashboard System (PMDS) created for a subsidy-based farmers market. The platform seeks to enhance the administrative capacity, service delivery, and data-driven decision-making of participating public service organizations; improve market management quality and sustainability; and advance broader health and social equity objectives of market sponsors and managers. Exploratory interviews, participatory observations, iterative prototyping, research memos, and in-situ testing were used to design a low-barrier technology for use within predominantly low-resource, racial minority, and immigrant and refugee populations. Disrupting traditional power dynamics between institutions and community partners, the platform reflects participatory and adaptive co-design practices that put community priorities at the center.**

Keywords-participatory design; co-design; community informatics; digital innovation; public service system.

I. Introduction

Digital innovations are typically developed for mainstream, middle-class users by for-profit start-ups and established technology businesses. These innovations emerge within well-resourced contexts of relative predictability with efficiency centered performance objectives and formalized organizational policies that structure the context of development and implementation. However, the development and deployment of digital innovations within the public service sector—which consists of contexts often characterized by administrative uncertainty, organizational and cultural diversity, and resource scarcity—may be particularly important for Human-Computer Interaction (HCI) researchers to consider.

In this study, we examine the co-design, development, and implementation of a novel, low-barrier information management and point-of-sale system developed for use within a grassroots, subsidy-based community farmers' market. The market is supported through a community-institution partnership between a county government entity and a grassroots social enterprise that works with marginalized youth. The scientific goal of this applied study is to identify and examine key consideration factors that shape the process of digital technology co-design and development within low-resource communities.

The platform, referred to as the People's Market Dashboard System (PMDS), aggregates data on various market processes and interactions and displays these data as interactive visuals within a set of integrated dashboards, facilitating standardized data collection and data validity processes while democratizing the value of those data for decision-making among both high- and low-power system users. The rest of the paper is structured as follows. Section II presents the previous research that contributed to the methodology and design of the PMDS. Section III describes the context of the project and an overview of the platform. Section IV details the different phases in the co-design and development of the PMDS. Section V discusses system deployment. Section VI presents the key findings. Finally, section VII concludes the paper and outlines directions for future work.

II. Related work

A. Co-Design

Co-design is closely intertwined with Participatory Design (PD), a design practice that originated in Scandinavia in the 1970s-1980s where workers, unions, and researchers collaborated to create workplace technologies. Participatory design aims to ensure that people affected by technology have a voice in its development, through the lens of participation as empowerment [1][2]. Co-design utilizes a variety of toolkits and techniques that facilitate stakeholder collaboration. They help translate their lived experiences, knowledge, hopes, and priorities into actionable design insights. These methods are designed to lower barriers to participation. Toolkits such as sorting, collage making, design games allow participants express their creativity and needs in technology [3]. Through storyboards, sketches or cutout paper interfaces, low-fidelity prototyping enables iterative exploration, as non-technical environment makes participants more comfortable in critiquing and modifying ideas [4][5]. Buxton emphasizes the importance of sketching as a design practice, highlighting its role in generating ideas rather than validating them [4].

Conversations are themselves a powerful co-design techniques, enabling stakeholders to articulate experiences, negotiate meanings and surface tacit knowledge that may not be captured through formal methods. Conversations serve a dual role; they elicit input and cultivate a sense of ownership, making stakeholders active contributors rather than passive informants. Paired with low-fidelity artifacts (such as sketches or mock-ups), conversations can become framework for co-design, grounding abstract ideas in tangible form, sparking discussions and iteration [4]. In our work, we use sketching as a starting point for design conversation with key system users to bring to life ideas that emerge in the co-design process (Figure 1).

 ISBN: 978-1-68558-383-5

B. Asset-aware Design Approach

Recent work in HCI has advocated for moving beyond deficit-oriented framing of resource-constrained communities toward asset-aware design approaches. This approach situates the design on local strengths rather than deficit. Wyche et al. introduced asset-based design by emphasizing how community resilience, social network, and informal innovation serve as foundations for sustainable technology in resource-limited setting [6].

C. Adaptive, In-situ Testing

When designing technologies for resource-constrained populations, it is often necessary to use methods that extend beyond traditional laboratory-based evaluations, since community members may face limitations related to time, transportation, or financial resources, as well as challenges like limited or unstable connectivity, shared devices, or low literacy. By refining features in context, technologies become more usable, resilient, and aligned with the lived realities of marginalized communities [7]. Previous HCI research shows that in situ methods surface contexual factors that remain visible in controlled settings, such as workarounds that users develop to cope with resource limitations [6][8].

D. Low-barrier Technology Design

The development of low-barrier technology has emerged in HCI as a critical approach to designing systems that are accessible to communities facing socioeconomic, linguistic, or infrastructural constraints [9]–[12]. This approach prioritizes simplicity, affordability, low entry threshold for participation in digital technology. The research of Talhouk et al. with refugees underscores the need for minimal hardware and literacy demands. They designed health technologies in Lebanon to operate on low-cost mobile platforms and to enable multilingual interaction [7].

III. Context and Platform Overview

In this research project, we co-designed and implemented PMDS as a novel, low-barrier information management and point-of-sale system developed for use within a grassroots, subsidy-based community farmers' market. The market is supported through a community-institution partnership between a government entity (primary funder) and a grassroots, social enterprise (hereafter referred to by their role as "market manager") that works with marginalized youth for agricultural and entrepreneurial skill-building. The purpose of the farmers' market is to expand access to fresh, affordable and culturally-specific foods for residents living in a limited food access, high poverty region of the city. Corollary purposes of the market include creating economic opportunities for small businesses of all kinds with limited market access and reducing nutrition-related health inequities in the region. The platform has several system user groups, including the market manager staff and volunteers who manage all check-in and registration processes for vendors and attendees as well as subsidy reimbursement processes for vendors (see Figure 2). Vendor types are subdivided based on what they sell at the market and, concomitantly, the types of market subsidies they can receive in payment from attendees.

The platform aggregates data about various market processes (e.g., registration, transaction), inputs them into a database through a series of interconnected HTML forms completed by various market actors such as the market manager staff and vendors, and displays this data as interactive visuals within a set of Tableau dashboards. In this way, the designed platform facilitates standardized data collection and data validity processes which concomitantly supports market compliance and sustainability. It also democratizes the value of that data for decision-making by enabling all system users to access these dashboards.

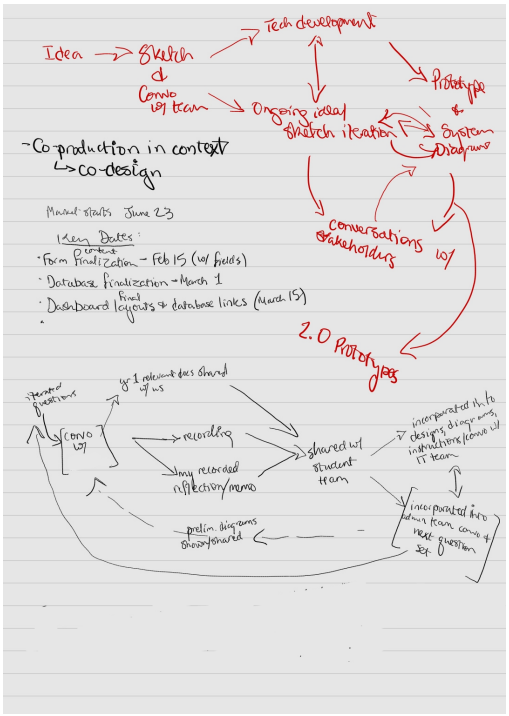


Figure 1. Initial sketch of PMDS co-design process illustrating early system ideation and stakeholder collaboration.

IV. Co-design and Development Process

A. Needs Assessment

Co-design of the platform began through a series of exploratory, unstructured interviews with the market manager. The initial needs assessment revealed many pain points in the current market processes that could be aided through a digital platform. We will discuss two of those challenges here. First, there were serious communication challenges across the many community-institution partners involved in market implementation. Language barriers were a concern among market vendors with limited English proficiency; however, it was difficult to determine the extent to which language barriers would continue to be a problem as related to system use given that the market manager reassured our team that language

barriers were not a major concern. Another important challenge was the lack of clear communication among stakeholders and the inconsistent implementation of market policies. We began to envision opportunities for the platform to support greater policy transparency and enforcement between managers and vendors.

The second identified issue was that vendors had previously not always received timely payment for items purchased by attendees through subsidy-based benefits. Vendors who did not receive payment at the end of the market day or shortly thereafter were left in very difficult situations affecting their personal finances, as most operated on tight margins. This reimbursement challenge was due to several contingent reasons; however, we believed that our system could support more timely repayment and greater accuracy in tracking those balances owed to vendors.

These challenges were initially revealed through a virtual conversation involving the project sponsor, market manager, and market vendors. However, this conversation revealed that ongoing conversations with market vendors would not be possible due to both their time constraints, the optional nature of their participation in dialogue with our team, and the significant language barriers of the vendors with the highest likelihood of participation. Their input was later incorporated into the iterative design of the high fidelity prototype during the training and early implementation processes.

Several other operational challenges were identified through a series of follow-up conversations with the market manager and review of both observational and secondary data from the previous year's market. The notes and memos generated from these conversations were collectively analyzed and the policy design considerations were concurrently translated into prospective technical functionalities.

This preliminary information gathering process was extraordinarily important because it allowed the development team to understand the system pain points and to begin to determine what types of system designs would be needed. Working within many constraints including a relatively short design-to-deployment timeline, our team leveraged its technical, administrative, and food systems expertise to conceptualize and create a mobile wallet based system which would enable point-of-sale tracking and standardization of the market benefits allocation processes while accounting for important differences in the allocation policies. This mechanism became the basis on which PMDS was designed and developed, ultimately revealing the high level of interdependence between market policy design and the technical design of the platform. The timeline for this project was approximately 8-10 months with the first stage (needs assessment) lasting about 5 months and occurring concurrently with the low-to-moderate fidelity prototyping (approximately 2 months).

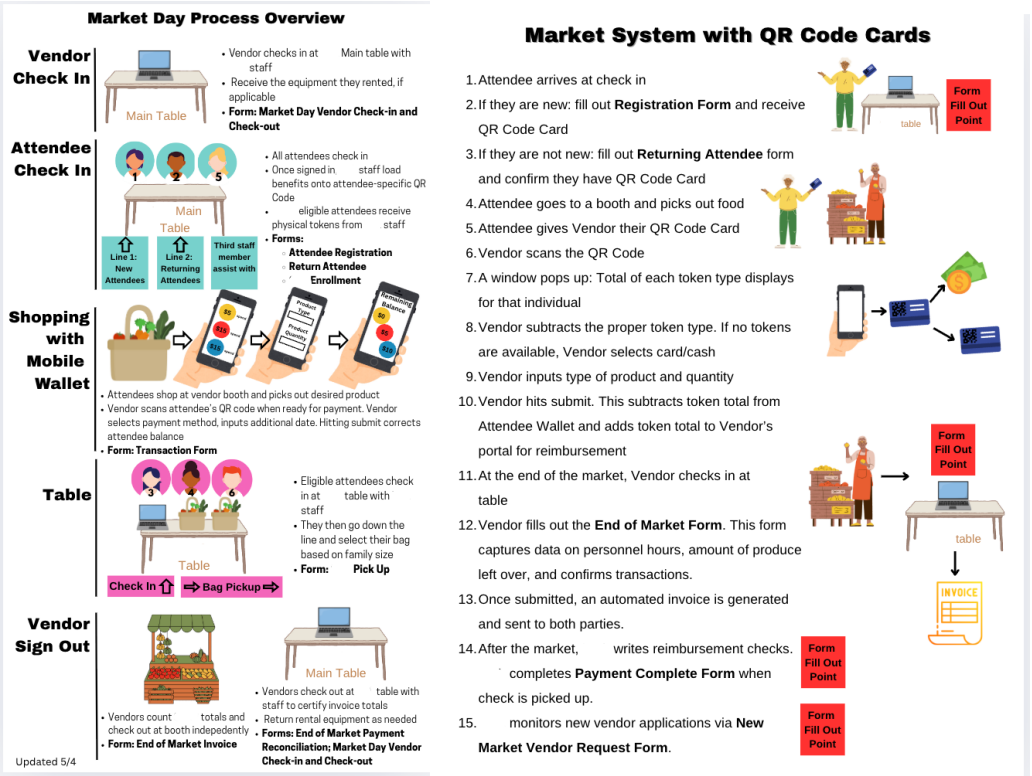


Figure 2. Overview of key market processes and system interactions.

B. Low-Fidelity Prototype

Informed by the insights from the initial needs assessment, the team created early system diagrams to model information flows, user interactions, and user access levels. These diagrams served both as internal tools for development clarity and as communication artifacts for aligning with non-technical system users including the primary funder and market manger. These initial diagrams were primarily based on the expected data needs and interests of the key system users, which enabled a backwards mapping from data visuals to needed fields and forms. Apart from the technical expertise of our team, we have both theoretical and practical food systems expertise within our team that informed these preliminary sketches, which were iteratively modified through conversation with the system users. The ongoing communication of the development team with the market manager about their needs and the evolving decisions and contingencies of market management enabled iterative development of both the preliminary sketches and the platform through Spring 2025.

Simultaneously, the database schema was iteratively constructed using Smartsheet, with particular focus on relational logic between vendors, market attendees, and transaction records (see Figure 4). This database was selected because the primary funder already used this software, and the original plan was to develop and deploy the system for use within the funder's IT environment. The process emphasized modularity to accommodate evolving market workflow and market management policies. Tableau was selected as the data visualization software because it was familiar to the project director, highly rated for its user-friendly interface, and its practicality as our team has institutional access to a secure Tableau server which enabled cost containment for our end users.

Next, the development process moved into early-stage prototyping and database schema design. This phase was characterized by the development of low- to mid -fidelity prototypes created to visualize and refine potential system functionality, which was guided by the preliminary system architecture and design requirements (see Figure 3). Components of these sketches reflected market manager objectives and plans for the upcoming market expressed in the unstructured interviews. For example, graphics selected for the dashboards addressed expressed data needs of the market manager required for funder compliance and future grant writing purposes. The types of graphics used were based on data visualization best practices for low-moderate data literacy communities, and the data fields and forms used were and mapped onto market workflow processes.

During this phase, the technical team also began sketching interface components as reflected in the ongoing conversations with the market manager, including the intake form fields and dashboard structures, based on the following:

- Potential data input fields expressed or implied by market manager as data needs and/or interests
- Anticipated data categories and field values
- Administrative workflows

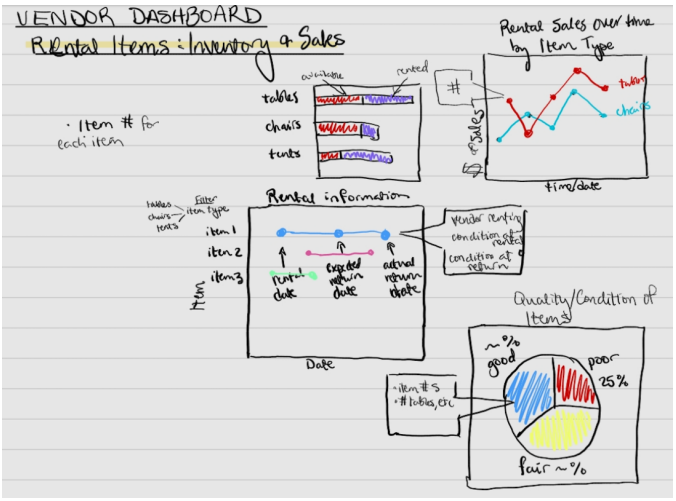


Figure 3. Low-fidelity prototype sketch of a vendor dashboard.

Continuing the co-design process, the team organized a design feedback session with stakeholders (manager and primary sponsor), aimed at introducing stakeholders to the emerging features and functionalities of the platform under design. Vendors were not present at this feedback session due largely to scheduling and timing difficulties. This session provided a crucial sounding board for exploring the system's scope and usability. Although not a technical walk-through, the session helped align expectations and generated feedback on high-level goals and constraints. It also improved system users' ability to understand what the platform was that the team was developing. Up until this point, system users with less direct involvement in the development process had a very limited understanding of what this prospective digital platform was. The feedback session provided system users with greater clarity about the system structure and its purposes.

C. High-Fidelity Prototype

This phase marked the evolution from conceptual sketches and diagrams to functional system modules (see Figure 5). The technical team built interactive prototypes that incorporated the following:

- Dynamically validated forms that ensured data accuracy at point-of-entry
- Dynamically updated visualization dashboards to convey market trends
- Unique identifiers to track vendor and attendee records between system components
- Role-based access flows

1) The Dashboard: In order to facilitate the ongoing dialogue with system users and co-design activities, our team designed image-based dashboard mockups derived from earlier sketches. The dashboard design process involved navigating the tension between accurate *data representation* (e.g., specifying exact charts, metrics, and data sources) and *data comprehensibility* (including data visualization best practices regarding visual layouts, color composition, data literacy and clear communication standards). These decisions were made

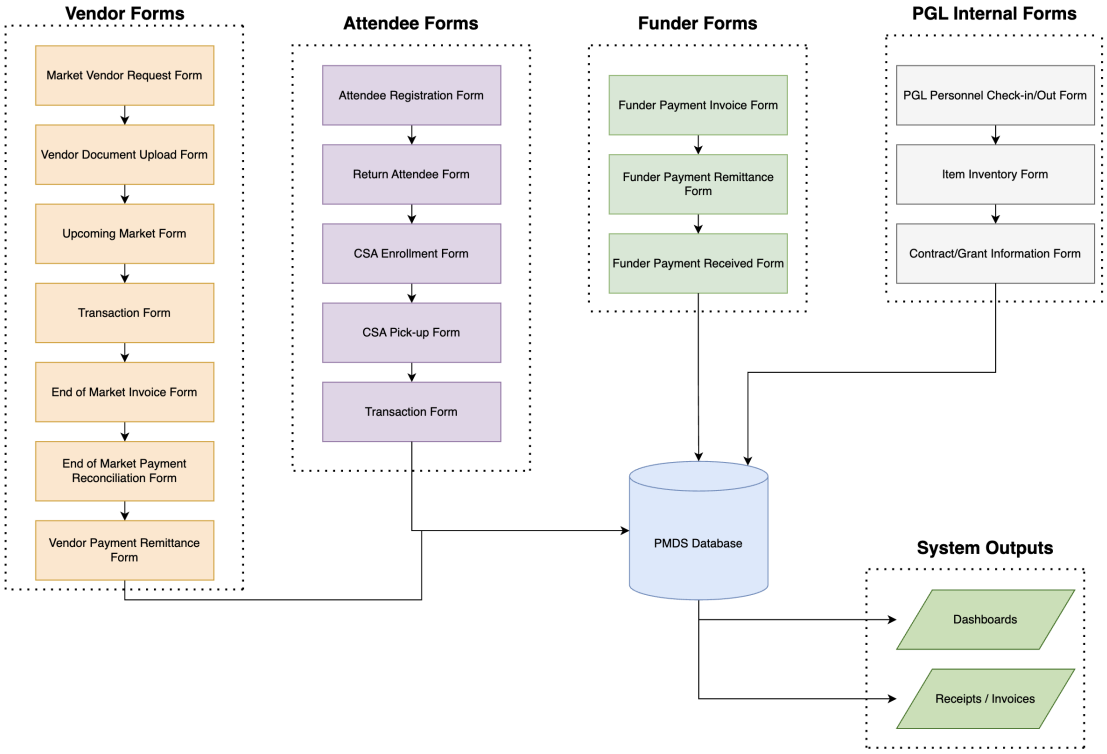


Figure 4. System diagram illustrating the interaction between input forms, database schema, dashboards, and receipt generation modules.

collaboratively with the expertise of both the technical and practical (e.g. public administration, food systems) members of our team.

After refining the dashboard layout and features, the team developed Tableau visualizations for process-related outcomes such as market performance, payment remittance, and benefits utilization. The backend development included establishing connection between the HTML forms and Smartsheet through Google Apps Scripts. Tableau was ultimately selected as the data visualization software because it was a practical solution that produces high quality, dynamic data visuals with a wide range of functionalities and graphic types available. System users would be accessing all dashboards through the website, so there was no concern about inaccessibility or functional use challenges when viewing dashboards through a mobile app. The website was determined to be the most appropriate modality for housing the PMDS given our study timeline and budget. High fidelity prototyping and iterative development lasted approximately 4 months, while initial user training and implementation (beta testing) occurred over a 3-6 month period. These two study phases where overlapping and included copious amounts of informally gathered user feedback and observational data (back-end observation of database entry and "front-end" observation onsite at the market) gathered by our team.

We also developed and deployed a web interface for a unified appearance and a single access point of the platform that supported convenient access for system users (see Figure 6). This, along with the logo that we developed, helped to create an polished identity for the platform that system users could learn to recognize and see themselves reflected in.

There were many cycles of internal testing and ongoing iterative feedback solicitation from system users involving the following areas:

- Form interfaces (e.g., auto-fill logic, field hiding based on conditions)
- Dashboard hierarchy (e.g., utility of filters and information visibility)
- System documentation (e.g., tooltips, form labels)
- "Virtual ID card" (QR code generated to access mobile wallet)

D. Beta-Testing

Once the working platform was almost complete, the focus shifted to beta testing and capacity building. In the weeks leading up to the market opening, the team tested the system with a limited group of stakeholders. The development team members and stakeholders participated in structured walkthroughs to assess usability and data integrity across different system entry points.

Initial testing focused on form functionality and the mobile wallet mechanism, with the goal of identifying usability issues and testing edge cases. As the testing progressed, the emphasis shifted toward robustness testing — purposely attempting to "break" the system through unanticipated inputs, duplicate entries, or timing-based conflicts. Our team continuously looked

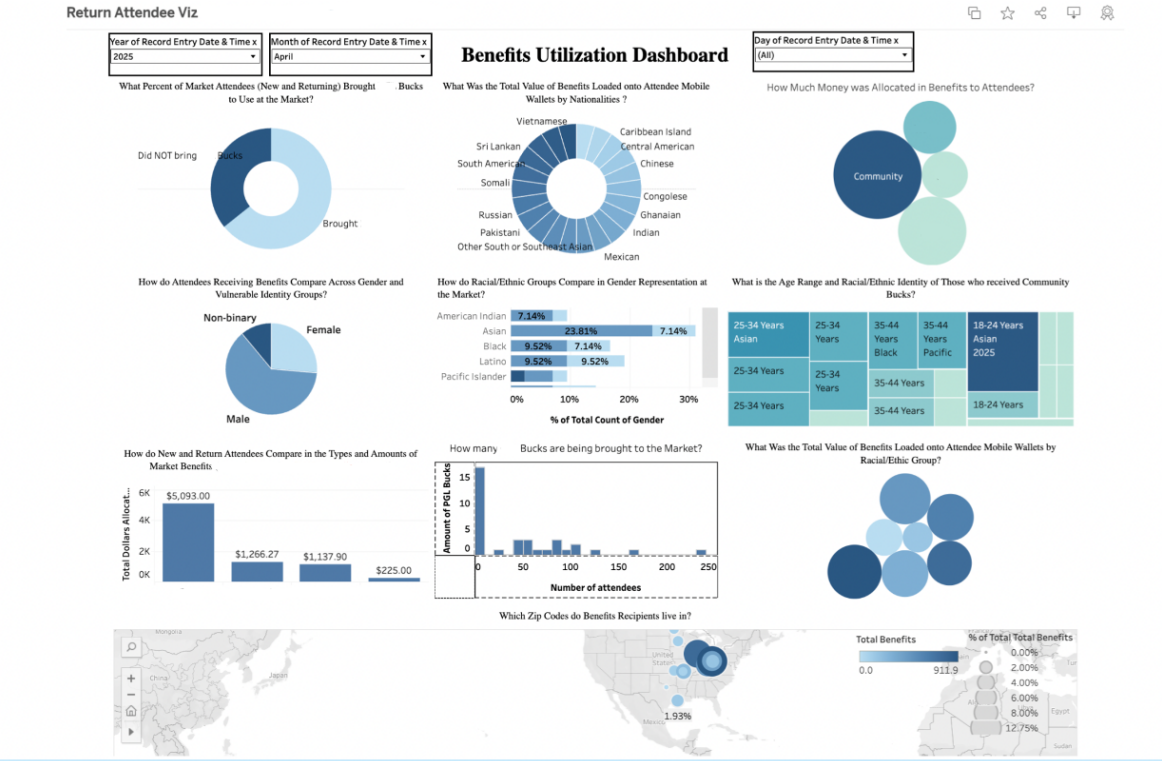


Figure 5. High-fidelity dashboard prototype illustrating refined visualization modules and reduced cognitive load (redacted for confidentiality).

for ways to enhance the platform for user needs and the User Experience (UX). For example, we created a live mobile wallet view of current balance of benefits for each customer, and later a similar vendor wallet providing the same so vendors could track their expected reimbursement income during the market.

Concurrently, training materials and hands-on walkthroughs were provided for stakeholders, focusing on operational workflows such as:

- Registering new attendees and vendors
- Processing market transactions and reimbursing vendors for subsidy-based transactions
- Monitoring benefit distributions in real-time

The final rounds of beta-testing were constrained by real-world operational timelines. As the market season approached, safeguards were implemented to prevent test data from mixing with actual attendee and vendor information, which limited certain test scenarios. In these cases, validation relied on simulated runs using dummy datasets. System performance, feature gaps, and final usability concerns were logged for post-season reflections and revisions.

V. Deployment and Adaptations

The platform was deployed approximately one week before the first day of the market. Onsite participatory observation occurred during the first market days, which were characterized by a hectic market environment as system users adjusted to using a new platform and adapted market policies and procedures as technical challenges arose.

One key issue identified was the lack of a strong, central Wi-Fi connection, which hindered registration and transaction processing and resulted in initial data loss. Adaptive modifications were made to the form submission process, including disabling submit buttons, and adding success messages. However, the negative impact of the initial data loss on system users' confidence in the platform was significant. Gradually, their confidence in the platform and their ability to use it successfully was rebuilt through continuous practice, ongoing onsite technical assistance, and iterative system modifications.

VI. Key Findings

Several important findings about the opportunities and limitations of co-designing digital innovations within limited-resource community contexts were identified in this study.

A. User Engagement in Co-Design and Implementation

First, close attention to the needs, capacities, and concerns of all types of system users during the co-design and testing phases can greatly support implementation success. Due to logistical and linguistic constraints, we were not able to engage with vendors during the co-design process; we primarily engaged them during training and in situ testing during implementation. Earlier and ongoing exposure of the vendors to the platform would have helped to avoid many of the technical challenges experienced and would likely have fostered early confidence in the system among end users. In turn, these factors would likely have reduced the heavy reliance of system users on technical assistance provided by our team. These considerations are

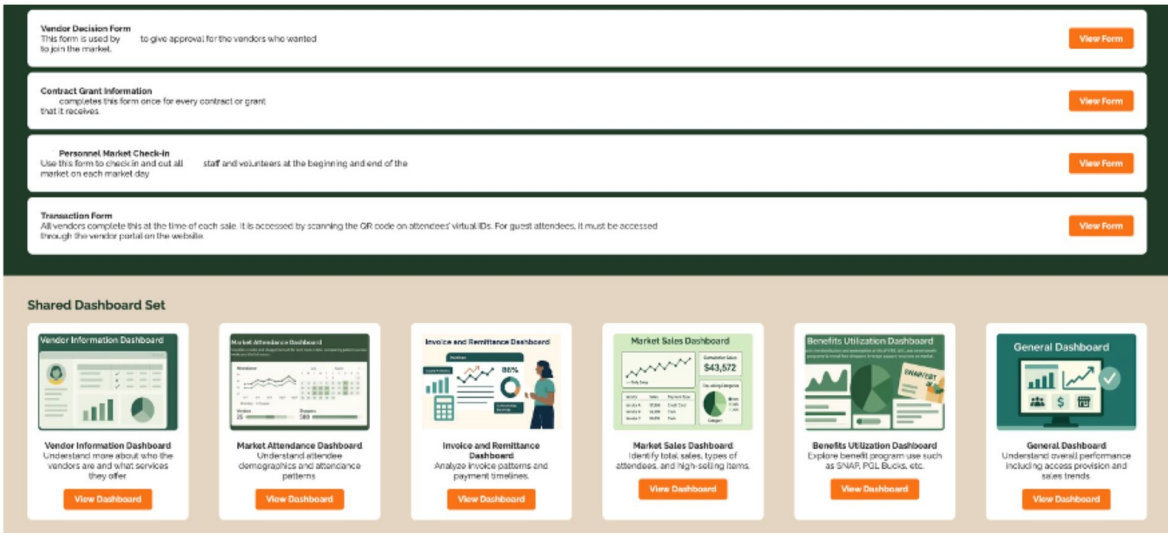


Figure 6. Web interface portal view providing a unified access point for system users.

being thoroughly incorporated within the platform design and implementation timeline of the next phase of this study.

B. User Trust and Proximate Support During Implementation

Additionally, trust and proximity were essential elements of the success of this implementation, despite the myriad challenges experienced. When implementation challenges arose in the first weeks of the market, it was invaluable that the lead author had an established relationship with the market manager and was present onsite helping system users navigate through those challenges in the first weeks of the market. She was able to help distinguish challenges that were caused by technical glitches from those that were merely system use errors. Together, she and the development team were able to determine which of the latter warranted technical modifications and adjustments to meet system users' needs and thus distinguish those needs from supplemental user training on correct system use. Intensive technical assistance was provided through phone calls, text messages, and later through a Discord-based "help desk" chat between the market manager and our student-based developer team. These regular, direct interactions between our team and the market manager helped to minimize the physical, temporal (time zone), cultural, and expertise differences between their team and ours. These interactions also provided invaluable learning opportunities for both groups, which collectively enhanced the use and improvement of the platform.

C. Prioritizing User Privacy and Security by Design

Ensuring privacy and security, even when not explicitly requested by system users, was central to maintaining trust and platform integrity. Our team prioritized encryption and other data security processes to ensure the security and privacy of data at all times. This included awareness of the socio-political context in which both the platform and the market were being implemented, which directly impacts system design and project related decision-making. Consideration of these contextual factors and their implications on system users sense of safety and security continue to be of the utmost importance to our team.

VII. CONCLUSION AND FUTURE WORK

This project illustrates how understanding dynamic public service contexts as common situational parameters can elicit effective co-design and adaptive development practices. Digital innovations designed for resource-limited and culturally diverse communities should always be rooted in proximity, alignment of priorities, and trusting relationships with system users. These socio-relational factors enable digital tools to be effectively designed to support end users' goal achievement and sustainable use. Co-designing digital innovations with system users requires direct engagement of development teams within social and administrative contexts that are often characterized by frequent changes, contingent decision-making, and multi-layered politics. Future work on this platform includes adapting PMDS into a multi-user system to support other types of related programs for similar networks of diverse local food systems actors. This includes expanding the platform into a mobile application, building a technical assistance and training infrastructure with dynamic, gamified learning elements, and redesigning the data architecture—along with input forms—to enable AI-driven insights and seamless system integrations. Engaging interdisciplinary development teams with diverse technical and domain expertise enables the proactive development of solutions to address these and other implementation challenges. Without these relational elements, digital innovation risks exacerbating mistrust and reinforcing existing social inequities. This work offers practical insights for researchers and practitioners designing human-centered digital platforms for public service delivery.

REFERENCES

[1] E. Björgvinsson, P. Ehn, and P.-A. Hillgren, "Agonistic participatory design: Working with marginalised social movements", *CoDesign*, vol. 8, no. 2-3, pp. 127–144, 2012.

[2] S. Bødker, C. Dindler, O. Iversen, and R. Smith, *Participatory Design* (Synthesis Lectures on Human-Centered Informatics). Morgan & Claypool Publishers, 2021, vol. 14. DOI: 10.2200/S01115ED1V01Y202107HCI051.

[3] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design", *CoDesign*, vol. 4, no. 1, pp. 5–18, 2008. DOI: 10.1080/15710880701875068.

[4] B. Buxton, *Sketching User Experiences: Getting the Design Right and the Right Design*. San Francisco, CA: Morgan Kaufmann, 2007, ISBN: 978-0-12-374037-3.

[5] S. Bødker and O. S. Iversen, "Staging a professional participatory design practice", in *Proceedings of the Second Nordic Conference on Human-Computer Interaction (NordiCHI)*, ACM, 2002, pp. 11–18. DOI: 10.1145/572020.572023.

[6] S. Wyche, T. T. Dillahunt, P. Ferreira, and R. Grinter, "Asset-based design: Towards a new hci research agenda for designing with resource-constrained communities", *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–20, 2019. DOI: 10.1145/3359315.

[7] R. Talhouk et al., "Syrian refugees and digital health in lebanon: Opportunities for participatory design", in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 331–342. DOI: 10.1145/2858036.2858330.

[8] J. Vines, R. Clarke, P. Wright, J. McCarthy, and P. Olivier, "Configuring participation: On how we involve people in design", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, ACM, 2013, pp. 429–438. DOI: 10.1145/2470654.2470716.

[9] N. Sambasivan, E. Cutrell, K. Toyama, and B. Nardi, "Intermediated technology use in developing communities", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 2583–2592. DOI: 10.1145/1753326.1753718.

[10] I. Medhi, A. Sagar, and K. Toyama, "Exploring the potential of voice interfaces for low-literate users", in *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, ACM, 2015, pp. 1035–1044. DOI: 10.1145/2702123.2702169.

[11] W. Thies et al., "99dots: A low-cost approach to monitoring and improving medication adherence", in *Proceedings of the 2019 International Conference on Information and Communication Technologies and Development (ICTD)*, ACM, 2019, pp. 1–12. DOI: 10.1145/3287098.3287110.

[12] A. Vashistha, N. Kumar, and R. Anderson, "Ekichabi: Informal digital payments for community-based savings groups", in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, pp. 1–11. DOI: 10.1145/3173574.3174030.

Cortical Activation Patterns During Visual and Vibrotactile Emotion Stimulation: A Comparative fNIRS Study

Lena Schubart, Marie Herz*, Susanna Götz, Karsten Huffstadt, Nicholas H. Müller
Institute of Design and Information Systems
Technical University of Applied Sciences Würzburg-Schweinfurt Sanderheinrichsleitenweg 20, 97074 Würzburg, Germany
*Corresponding author
E-mail: lena.schubart@study.thws.de (L. Schubart), marie.herz@thws.de (M. Herz), susanna.goetz@thws.de (S. Götz), karsten.huffstadt@thws.de (K. Huffstadt), nicholas.mueller@thws.de (N. H. Müller)

***Abstract*—This study provides the comparison of prefrontal cortical activation during visual versus conditioned haptic emotion stimulation using functional Near-InfraRed Spectroscopy. Using a within-subject design with classical conditioning, 32 participants learned associations between facial expressions and vibrotactile patterns. Results showed no significant difference in prefrontal oxygenated hemoglobin concentration between modalities ($p = .640$), supporting modality-independent emotion processing. However, haptic processing required significantly higher cognitive effort as measured by the NASA Task Load Index (NASA-TLX) Mental Demand subscale ($p = .001$, $r = .57$). Recognition rates were comparable (94.8% visual, 93.8% haptic). These findings suggest that haptic emotion communication systems are neurally viable but may require extended training to reduce cognitive demands. The results have implications for the development of assistive technologies that could enable blind and visually impaired individuals to access non-verbal emotional information through the tactile channel.**

***Index Terms*—haptic emotion communication; cross-modal processing; prefrontal cortex; cognitive load; assistive system.**

I. INTRODUCTION

When people talk to each other, the spoken word is only a fraction of what is actually communicated. A raised eyebrow can signal skepticism, a fleeting smile can suggest agreement, an averted gaze can reveal disinterest [1]. This non-verbal communication, such as facial expressions, gestures, or eye contact, accounts for over 50% of emotional information exchange in conversations [1].

For the 2.2 billion people worldwide with visual impairments, including 36 million who are completely blind [2] [3], this visual dimension of human interaction remains systematically inaccessible. Without visual cues, affected individuals cannot recognize whether a statement is meant ironically, approvingly, or disapprovingly. Such information is typically derived from facial expressions and gestures [4], leaving them without a central source of information for understanding social situations. This situation is referred to as "conversational asymmetry," an imbalance in which sighted conversation partners have access to all communication channels, while visually impaired individuals can only receive verbal and auditory information [5]. This asymmetry affects not only individual conversations but the fundamental ability to build trust and maintain social relationships [5].

Existing assistive technologies for visually impaired people, such as screen readers, speech synthesis, or acoustic signals, primarily rely on auditory output [6]. In social interactions, these solutions reach their limits: Auditory cues interrupt con-versations, occupy the already heavily used auditory channel, and are perceptible to outsiders, preventing discretion [6]. The tactile channel offers a promising alternative: It can be used in parallel with auditory communication and enables discrete information transmission during ongoing conversations [7].

Research has demonstrated that emotions can be communicated through touch [4]. Liu et al. [7] transferred this principle to technical systems. Their programmable vibrotactile inter-face is based on elastomer actuators that generate vibrations in the frequency range of 50-450 Hz and enable four-dimensional haptic stimulation (time, position, amplitude, frequency) [7]. Users achieved emotion recognition rates of 64.6% without prior training; after systematic learning training, this increased to 95.8% [7]. Although these results demonstrate the behavioral effectiveness of haptic emotion transmission, there has been no neurophysiological study to date that clarifies whether haptic stimuli evoke the same cortical activation patterns as visual emotional stimuli [8].

The behavioral successes of Hertenstein et al. [4] and Liu et al. [7] demonstrate that humans can recognize emotions through haptic stimuli. However, high recognition rates only show that emotional information arrives, not whether the brain processes this information in the same way as visual stimuli [7]. It remains unclear whether haptically transmitted emotions activate the same neural networks as visual stimuli. Research on cross-modal emotion processing suggests that the brain may process emotional information independently of the sensory input modality [9]. Klinge et al. [10] also showed increased amygdala activation in blind participants during emotional auditory stimuli.

Even with comparable activation patterns, the question of cognitive costs arises. Cognitive Load Theory [11] postulates capacity-limited working memory resources; decoding unfamiliar stimulation modalities could require additional processing resources [12].

Despite established evidence for behavioral effectiveness of haptic emotion transmission and cross-modal emotion processing, a direct neurophysiological comparison of visual and haptic emotion stimulation is lacking [8].

Based on the identified research gap, this study addresses the following research question: Does emotional processing differ between haptic and visual modalities?

This study has several limitations, including the restriction of fNIRS to cortical surface activity, the relatively small sample size, and potential additional cognitive demands introduced by the conditioning paradigm.

The remainder of this paper is organized as follows. Section II reviews related work on haptic emotion communication, neural foundations of cross-modal processing, and functional Near-InfraRed Spectroscopy (fNIRS) in emotion research. Section III describes the methodology, including study design, stimuli, and measurement procedures. Section IV presents the results. Section V discusses the findings and their implications. Section VI concludes the paper and outlines future work.

II. RELATED WORK

This section reviews the theoretical and empirical foundations relevant to the present study.

A. Haptic Emotion Communication

Can humans understand emotions through touch alone? Hertenstein et al. [4] investigated this question in an experiment in which participants communicated eight different emotions exclusively through touching another person's arm, without words, facial expressions, or eye contact. Recipients recognized emotions with remarkable accuracy: Anger was correctly identified in 78% of cases, fear in 75%, disgust in 68%, and happiness in 83% (chance level: 12.5%). The researchers identified characteristic touch patterns: Anger was conveyed through short, intense touches; sadness through slow, gentle contact; joy through rhythmic, dynamic movements [4].

Liu et al. [7] transferred these findings to technical systems. Their flexible haptic interface uses elastomer actuators generating vibrations in the range of 50-450 Hz. In their study, users achieved emotion recognition rates of 64.6% for six basic emotions without prior training. After a systematic learning program, this increased to 95.8%, evidence that the brain can learn to associate vibration patterns with emotional meanings [7].

B. Neural Foundations of Cross-Modal Emotion Processing

For developing haptic emotion systems, it is crucial whether the brain can process emotions independently of the sensory channel. The meta-analysis by Lindquist et al. [9] provides important theoretical foundations: The absence of modality-specific emotion centers suggests that emotional information from different sensory channels could be processed in the same neural networks.

Klinge et al. [10] provided direct experimental evidence for this adaptability in a functional magnetic resonance imaging (fMRI) study. fMRI is an imaging technique that visualizes brain activity through changes in blood oxygenation and offers high spatial resolution [13]. The researchers compared brain activity of 12 congenitally blind and 12 sighted participants while both groups listened to emotional voices (fearful, angry, neutral). The central finding: Blind participants showed significantly stronger amygdala activation for emotional compared to neutral voices. The amygdala is an almond-shaped structure in the temporal lobe that plays a key role in evaluating the emotional relevance of stimuli [14]. Particularly revealing was the finding that the strength of amygdala activation correlated with individual recognition performance ($r = .54$). These results demonstrate that the emotional brain, when lacking a sensory channel, increasingly uses the remaining channels for emotional processing [10].

Theoretically, cross-modal emotion processing can be explained by Embodied Cognition, according to which bodily experiences, including haptic perception, are closely linked to cognitive and emotional processes [15]. Multiple Resource Theory [12] also postulates that processing new or unfamiliar stimuli requires more cognitive resources than automated processes.

C. fNIRS in Emotion Research

Compared to fMRI, fNIRS offers crucial advantages for studies with haptic stimulation. Participants can move, the device is portable, and measurement occurs under naturalistic conditions without the constraints of an MRI scanner [16]. The prefrontal cortex is particularly accessible for fNIRS and shows reliable emotion-related activation patterns [17].

Sánchez-Reolid et al. [8] demonstrated the validity of fNIRS for emotion recognition in a recent study. The researchers presented participants with emotion-inducing images from the IAPS database and classified the evoked emotions (happiness, sadness, fear, anger) based on valence and arousal according to Russell's Circumplex Model. With a 22-channel fNIRS setup over the prefrontal cortex and a Bagging-Trees algorithm, they achieved 64% classification accuracy. A value well above chance level that confirms the suitability of fNIRS for emotion research. While fNIRS has been successfully used to measure emotional responses to visual stimuli, a direct comparison with haptic emotion stimulation is lacking [8] [7].

Based on these findings, two hypotheses were formulated:

H1: When haptic emotion stimuli are tested, the prefrontal cortex is activated equally as with visual emotion stimuli.

H2: When haptic emotion stimuli are tested, the subjectively perceived cognitive load is higher than with visual emotion stimuli.

III. METHODOLOGY

This section describes the study design, stimuli used, measurement procedures, and experimental protocol.

A. Study Design and Sample

The study uses a within-subject design with classical conditioning paradigm. In a learning phase, associations between visual emotion stimuli and emotion-specific vibration patterns were established. Subsequently, it was tested whether haptic stimuli without visual information evoke comparable cortical activation. N = 32 participants were recruited following standard inclusion criteria (age > 18, normal or corrected-to-normal vision, no known neurological disorders). Participation was voluntary and conducted in accordance with institutional ethical guidelines, with informed consent obtained from all participants.

B. Visual Stimuli

Six emotion categories (happiness, sadness, fear, anger, disgust, neutral) from the FACES database were used [18]. The FACES database was chosen because it offers standardized, validated facial expressions with high recognition rates and is established in emotion research [18]. For each emotion, two images (one female, one male model) were presented alternately (Image1 → Image2 → Image1 → Image2), so that each emotion was shown four times. Using both genders increases the generalizability of results and controls for possible gender-specific differences in emotion perception.

C. Haptic Stimulation System

For haptic stimulation, a vibrotactile stimulation system was employed. The system consists of vibration motors attached to the fingers and wrist of the non-dominant hand. The non-dominant hand was chosen because it is less accustomed to fine motor tasks in everyday life, requiring increased conscious attention to tactile stimuli [19]. Each emotion was assigned to a specific finger, with assignment randomized, as emotional meaning is established through associative learning during the conditioning phase [14]. The haptic system was chosen because it enables discrete, non-intrusive information trans-mission and does not occupy the auditory channel, which are essential requirements for assistive technologies for visually impaired people [6].

D. fNIRS Measurement

Brain activity was measured using fNIRS. fNIRS is based on the principle that near-infrared light (760-850 nm wave-length) penetrates skull bone and brain tissue but is absorbed by hemoglobin. Active brain regions require more oxygen, causing blood to flow more strongly to these areas, changing the ratio of oxygenated (HbO) to deoxygenated hemoglobin (HbR). This change is measurable through light absorption [13].

For measurement, the Artinis Brite24 system with 27 channels was used with a frontal headband over the prefrontal cortex. The optodes covered Fp1, Fp2, F3, F4, F7, F8 (10-20 system). The mean HbO concentration across all channels served as an indicator of prefrontal activity. fNIRS was preferred over fMRI because, as already mentioned, participants can move around, the device is portable, and measurements are taken under natural conditions without the restrictions of an MRI scanner, which is an essential prerequisite for studies involving haptic stimulation [16]. The sampling rate was 25 Hz.

E. Assessment of Cognitive Load

The NASA Task Load Index was used to assess cognitive load [20], capturing subjectively perceived workload across six dimensions, including mental demand and frustration. The NASA-TLX was chosen because haptic emotion recognition may require more cognitive resources than familiar visual processing [11]. Assessing cognitive load is crucial for evaluating the everyday practicality of assistive systems, as excessive demands would limit practical usability in daily life [6] [11].

F. Experimental Procedure

Study duration was planned at 35 minutes per participant. The experimental protocol consisted of three consecutive phases.

In the conditioning phase, participants learned the assignment between FACES images and vibration patterns. The experimenter explicitly named each emotion (e.g., “Now follows happiness”) to support conscious linking between visual and haptic stimulus. A memory check at the end of the phase ensured that all assignments were correctly learned, with error free reproduction of all six assignments as the criterion [21]. Stimuli followed standardized timing. In the visual phase, after a 10s baseline, each emotion was shown four times for 4s, separated by 5s pauses. In the haptic phase, a 10s baseline preceded vibrations of approximately 200 ms, also separated by 5s pauses. The baseline served as a resting-state reference for all activations.

After successful conditioning, the visual test phase followed. Participants viewed FACES images without accompanying vibration and verbally reported the perceived emotion after each block. Responses were not corrected to avoid influencing the data. Brain activity was continuously recorded via fNIRS, followed by completion of the NASA-TLX to assess subjective cognitive load.

The haptic test phase concluded the experiment. With eyes closed, participants received vibration stimuli only to assess whether conditioned associations allowed correct emotion identification via haptic stimulation alone. Verbal emotion reports followed each block during continuous fNIRS recording, and the NASA-TLX was administered again to compare cognitive load between conditions.

IV. RESULTS

This section presents the results in four subsections: sample description, fNIRS findings, emotion recognition rates, and subjective cognitive load.

A. Sample

Thirty-two participants took part in the study (16 female, 16 male). Age ranged from 20 to 50 years (M = 26.38, SD = 5.70). Thirty-one participants were right-handed, one person was left-handed. All participants had normal or corrected vision. Participation was voluntary and required written consent.

B. fNIRS Results

fNIRS measures changes in the concentration of oxygenated hemoglobin in cortical blood vessels of superficial brain regions. An increase in HbO value indicates increased neural activity, as active brain areas are supplied with more oxygen-rich blood [16]. Values are given in micromoles per liter (µmol/L) and calculated relative to baseline (resting state).

The fNIRS data were processed using a Python script (version 3.9) with pandas, numpy, and scipy. Processing included artifact removal, event marker identification (S1 for visual, S2 for haptic phase), emotion block segmentation, and baseline correction using a 10-second rest period (250 data points at 25 Hz). As a block design with multiple stimuli per emotion block was employed, the mean HbO across each block served as the dependent measure, capturing the cumulative hemodynamic response rather than individual stimulus-evoked responses. Statistical analysis focused on the prefrontal cortex (PFC), as fNIRS primarily captures cortical surface activity and the PFC plays a central role in cognitive-emotional processing [22]. To avoid pseudo-replication, HbO values were first averaged across all prefrontal channels per emotion and condition then across the six emotions per participant, yielding one mean HbO value per condition and participant (N = 32 paired observations). The Shapiro-Wilk test revealed significant deviations from normal distribution for both conditions (visual: W = .722, p < .001; haptic: W = .463, p < .001); therefore, the nonparametric Wilcoxon signed-rank test was applied. Descriptive statistics are summarized in Table I.

TABLE I. DESCRIPTIVE STATISTICS AND WILCOXON TEST FOR PREFRONTAL HBO CHANGES

Stimulation	N	M (µmol/L)	SD	Z	p
Visual	32	−0.85	10.41	−0.47	.640
Haptic	32	5.93	15.74		

Note. N = number of participants (paired observations); M = mean; SD = standard deviation; Z = Wilcoxon signed-rank test statistic; p = significance value. Positive HbO values indicate an increase relative to baseline.

The Wilcoxon signed-rank test revealed no statistically significant difference between visual and haptic stimulation in the prefrontal region (Z = −0.47, p = .640). Descriptively, haptic stimulation was associated with a mean increase in HbO of 5.93 µmol/L, whereas visual stimulation showed a slight decrease of −0.85 µmol/L. Emotion-specific results are presented in Table II.

TABLE II. WILCOXON TESTS FOR VISUAL VS. HAPTIC PER EMOTION (PREFRONTAL REGION)

Emotion	Z	p	r
Sadness	−0.49	.627	.035
Anger	−0.08	.940	.006
Neutral	−0.66	.513	.048
Fear	−0.90	.369	.065
Happiness	−1.48	.140	.107
Disgust	−1.38	.166	.100

Note. All tests two-tailed. Z = Wilcoxon signed-rank test statistic; p = significance value; r = effect size ($r = Z/\sqrt{N}$).

Emotion-specific analyses revealed no significant differences between modalities (all p > .05). Small effect sizes were consistent with the absence of significant differences between modalities. In summary, the fNIRS results show that haptic emotional stimulation activates the prefrontal cortex at a level comparable to visual stimulation. The activation patterns do not differ significantly between the two modalities.

C. Emotion Recognition

Emotion recognition accuracy was captured through verbal responses after each emotion block. In the visual condition, 182 of 192 assignments were correct (94.8%). In the haptic condition, 180 of 192 were correct (93.8%), representing a difference of only 1.0 percentage points. Recognition rates are shown in Table III.

TABLE III. EMOTION RECOGNITION RATES BY PHASE AND EMOTION

	Visual		Haptic	
Emotion	correct/total	%	correct/total	%
Sadness	28/32	87.5	27/32	84.4
Anger	31/32	96.9	31/32	96.9
Neutral	32/32	100.0	32/32	100.0
Fear	28/32	87.5	28/32	87.5
Happiness	32/32	100.0	31/32	96.9
Disgust	31/32	96.9	31/32	96.9
Total	182/192	94.8	180/192	93.8

The highest recognition rates were observed for Neutral (100% in both phases) and Happiness (100% visual, 96.9% haptic). These results confirm successful conditioning and reliable stimulus identification in both modalities.

D. Subjective Cognitive Load (NASA-TLX)

Subjective cognitive load was captured using the NASA-TLX [20]. The NASA-TLX is a multidimensional instrument for assessing subjective workload and comprises six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Subjective workload results are summarized in Table IV.

TABLE IV. NASA-TLX RESULTS: COMPARISON OF VISUAL AND HAPTIC STIMULATION

Scale	M vis	SD vis	M hap	SD hap	Z	p
Total	173.81	56.36	187.28	71.16	1.81	.070
Mental Demand	23.44	15.83	38.13	19.19	3.21	.001

Note. Wilcoxon signed-rank test for paired samples. M = mean; SD = standard deviation; Z = Wilcoxon signed-rank test statistic; p = significance value.

The NASA-TLX total score showed no significant difference between conditions (Z = 1.81, p = .070). However, the Mental Demand subscale showed a highly significant difference (Z = 3.21, p = .001) with large effect size (r = .57) [23]. At the individual level, 24 of 32 participants (75%) reported higher mental demand during haptic stimulation.

These results show that processing haptic emotional stimuli is associated with significantly higher cognitive demand than processing visual stimuli.

V. DISCUSSION

Results support H1: No significant difference was found be-tween conditions (p = .640), indicating comparable prefrontal activation. Descriptively, haptic stimulation showed higher activation (M = 5.93 μmol/L) compared to visual stimulation (M = −0.85 μmol/L), though this did not reach significance due to high interindividual variability. H2 was confirmed: The Mental Demand subscale showed a highly significant difference (p = .001, r = .57). The high recognition rates (94.8% visual, 93.8% haptic) confirm successful conditioning and demonstrate that haptic emotion communication can achieve nearly equivalent accuracy to visual recognition.

The finding that prefrontal activation did not differ significantly has important theoretical implications. According to cognitive emotion regulation models [24], the prefrontal cortex evaluates emotional stimuli independently of sensory input modality. This aligns with cross-modal plasticity research [25]. The successful conditioning phase appears to have led to haptic stimuli activating similar prefrontal networks as visual stimuli, supporting modality-independent emotional representations at higher cortical levels. The emotion-specific analysis reinforces this: no significant differences for any emotion category, suggesting a general pattern consistent with Embodied Cognition theories [15].

The significant difference in Mental Demand reveals an important distinction: While neural activation is comparable, cognitive effort differs considerably. Processing haptic stimuli requires identifying vibration patterns and retrieving learned associations, while visual emotion recognition occurs automatically [26]. Visual recognition is automated through lifelong experience. Humans recognize emotional expressions from early childhood [27], whereas haptic coding is learned only during a brief conditioning phase. This aligns with Multiple Resource Theory [12]. The specificity of the effect is noteworthy: While the NASA-TLX total score did not differ significantly, the Mental Demand subscale showed a highly significant difference. This suggests that haptic processing selectively increases cognitive demand without affecting other aspects of workload, such as physical effort or frustration. One plausible explanation is that the haptic condition requires an additional layer of meaning-construction: participants must sustain a technically mediated representation and continuously map it onto socially learned emotional categories. Related work on technically mediated self-representation indicates that changes in representational format can affect subjective presence and increase interpretive processing demands, even when task accuracy remains high [35].

These findings are directly relevant for haptic assistive systems for blind and visually impaired people. Neural equivalence suggests haptic systems could provide functionally similar emotional information access at the level of prefrontal cognitive-emotional evaluation. However, increased cognitive load could be problematic in real-world applications where resources may be limited [28]. The high recognition rates (93.8% haptic) and potential reducibility of cognitive load through extended training support practical viability. Our haptic recognition rates (93.8%) exceed previous studies (50-78%) [4] [29], likely due to the explicit conditioning paradigm enabling more reliable retrieval. The comparable prefrontal activation extends previous fNIRS emotion research [16] to cross-modal communication.

Several limitations should be considered. fNIRS cannot capture deeper structures (amygdala, ACC, insula) central to emotion processing [30]; complementary fMRI studies would be necessary. Optode placement followed standardized landmarks, but anatomical correspondence may vary [31]. Findings from sighted participants may not generalize to blind individuals, who show different cortical organization [32] [33]; future studies should investigate whether they show increased efficiency in haptic processing. The FACES database excludes Surprise [34], and haptic patterns lacked empirical pre-validation [4] [29]. Manual stimulus triggering and fixed phase order prevent precise timing verification and counterbalancing. The increased prefrontal activation during haptic stimulation could reflect cognitive translation processes (retrieving learned associations) rather than emotion processing per se. A clear separation between emotional activation and cognitive demand is not possible with the present design.

VI. Conclusion and Future Work

This study provides fNIRS-based evidence for the assumption that conditioned haptic emotion stimulation activates the prefrontal cortex to the same extent as visual emotion stimulation. The absence of significant differences in prefrontal HbO concentration ($p = .640$) supports the hypothesis that emotion processing operates at a modality-independent level after successful conditioning.

However, haptic emotion processing requires significantly more cognitive effort, as demonstrated by the highly significant difference in NASA-TLX Mental Demand scores ($p = .001$, $r = .57$). This dissociation, characterized by comparable neural activation but increased cognitive load, has important implications for the design of assistive technologies. Haptic emotion communication systems appear to be neurally viable but may require structured training programs to achieve the level of automation typically associated with visual emotion recognition.

Recognition rates were comparable between modalities (94.8% visual, 93.8% haptic), demonstrating that haptic emotion communication can achieve nearly equivalent accuracy to visual recognition after appropriate conditioning.

Future research should extend this paradigm to blind and visually impaired populations to determine whether cross-modal plasticity enhances haptic emotion processing efficiency in this target group. Additional research directions include longitudinal studies on training effects on cognitive load, automated stimulus presentation for better temporal precision, and integration with fMRI to capture subcortical activation patterns.

Future work should also explore real-world applications and social use cases, such as supporting social interaction and everyday communication for visually impaired individuals.

Acknowledgment

The authors thank all participants for their voluntary participation in this study.

References

[1] Z. Shafique, M. A. Asghar, M. J. Khan, and W. Iqbal, "Nonverbal communication cue recognition: A deep multimodal transformer fusion approach," Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops, pp. 5666-5674, 2023.

[2] GBD 2019 Blindness and Vision Impairment Collaborators, "Causes of blindness and vision impairment in 2020 and trends over 30 years," Lancet Global Health, vol. 9, no. 2, pp. e144-e160, 2021.

[3] World Health Organization, "Blindness and vision impairment," 2023. [Online]. Available from: https://www.who.int/news-room/fact- sheets/detail/blindness-and-visual-impairment 2025.01.15

[4] M. J. Hertenstein, D. Keltner, B. App, B. A. Bulleit, and A. R. Jaskolka, "Touch communicates distinct emotions," Emotion, vol. 6, no. 3, pp. 528-533, 2006.

[5] K. Ghafoor, A. O. Alnoori, A. Munoz, and Z. Zhang, "Improving social interaction of the visually impaired through wireless technology," Int. J. Intelligent Computing and Cybernetics, vol. 17, no. 1, pp. 126-142, 2024.

[6] A. Kushnir and N. H. Müller, "Haptic feedback in everyday conversation situations," in HCI International 2020-Posters, C. Stephanidis and M. Antona, Eds. Springer, 2020, pp. 239-244.

[7] Y. Liu et al., "Emotional and directional enabled programmable flexible haptic interface for enhanced cognition in disabled community," Research, vol. 8, Article 0714, 2025.

[8] D. Sánchez-Reolid, R. Sánchez-Reolid, A. L. Borja, and M. T. López, "EEG and fNIRS signal-based emotion identification using machine learning," Electronics, vol. 13, no. 23, Article 4797, 2024.

[9] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: A meta-analytic review," Behavioral and Brain Sciences, vol. 35, no. 3, pp. 121-143, 2012.

[10] C. Klinge, B. Röder, and C. Büchel, "Increased amygdala activation to emotional auditory stimuli in the blind," Brain, vol. 133, no. 6, pp. 1729-1736, 2010.

[11] J. Sweller, "Cognitive load theory," in Psychology of Learning and Motivation, vol. 55, J. P. Mestre and B. H. Ross, Eds. Academic Press, 2011, pp. 37-76.

[12] C. D. Wickens, "Multiple resources and mental workload," Human Factors, vol. 50, no. 3, pp. 449-455, 2008.

[13] M. Ferrari and V. Quaresima, "A brief review on the history of human functional Near-InfraRed Spectroscopy (fNIRS) development and fields of application," NeuroImage, vol. 63, no. 2, pp. 921-935, 2012.

[14] J. E. LeDoux, The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon & Schuster, 1996.

[15] P. M. Niedenthal, "Embodying emotion," Science, vol. 316, no. 5827, pp. 1002-1005, 2007.

[16] P. Pinti et al., "The present and future use of functional Near-InfraRed Spectroscopy (fNIRS) for cognitive neuroscience," Ann. New York Acad. Sciences, vol. 1464, no. 1, pp. 5-29, 2020.

[17] A.-C. Ehlis, S. Schneider, T. Dresler, and A. J. Fallgatter, "Application of functional Near-InfraRed Spectroscopy in psychiatry," NeuroImage, vol. 85, pp. 478-488, 2014.

[18] N. C. Ebner, M. Riediger, and U. Lindenberger, "FACES - A database of facial expressions in young, middle-aged, and older women and men," Behavior Research Methods, vol. 42, no. 1, pp. 351-362, 2010.

[19] D. J. Goble and S. H. Brown, "The biological and behavioral basis of upper limb asymmetries in sensorimotor performance," Neuroscience & Biobehavioral Reviews, vol. 32, no. 3, pp. 598-610, 2008.

[20] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index)," in Human Mental Workload, P. A. Hancock and N. Meshkati, Eds. North-Holland, 1988, pp. 139-183.

[21] J. D. Karpicke and H. L. Roediger, "The critical importance of retrieval for learning," Science, vol. 319, no. 5865, pp. 966-968, 2008.

[22] H. Kober et al. "Functional grouping and cortical-subcortical interactions in emotion: A meta-analysis of neuroimaging studies," NeuroImage, vol. 42, no. 2, pp. 998-1031, 2008.

[23] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Lawrence Erlbaum Associates, 1988.

[24] K. N. Ochsner and J. J. Gross, “The cognitive control of emotion,” Trends in Cognitive Sciences, vol. 9, no. 5, pp. 242-249, 2005.

[25] D. Bavelier and H. J. Neville, “Cross-modal plasticity: Where and how?” Nature Reviews Neuroscience, vol. 3, no. 6, pp. 443-452, 2002.

[26] R. Adolphs, “Neural systems for recognizing emotion,” Current Opinion in Neurobiology, vol. 12, no. 2, pp. 169-177, 2002.

[27] C. A. Nelson, “The development and neural bases of face recognition,” Infant and Child Development, vol. 10, no. 1-2, pp. 3-18, 2001.

[28] C. Kilian, M. Ertl, and N. H. Müller, “Cognitive load in assistive technology use: A systematic review,” Assistive Technology, vol. 34, no. 6, pp. 721-733, 2022.

[29] H. Culbertson, S. B. Schorr, and A. M. Okamura, “Haptics: The present and future of artificial touch sensation,” Annual Review of Control, Robotics, and Autonomous Systems, vol. 1, pp. 385-409, 2018.

[30] K. L. Phan, T. Wager, S. F. Taylor, and I. Liberzon, “Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI,” NeuroImage, vol. 16, no. 2, pp. 331-348, 2002.

[31] D. Tsuzuki and I. Dan, “Spatial registration for functional Near-InfraRed Spectroscopy,” NeuroImage, vol. 85, no. 1, pp. 92-103, 2014.

[32] N. Sadato, T. Okada, M. Honda, and Y. Yonekura, “Critical period for cross-modal plasticity in blind humans,” NeuroImage, vol. 16, no. 2, pp. 389-400, 2001.

[33] D. Valente, A. Theurel, and E. Gentaz, “The role of visual experience in the production of emotional facial expressions by blind people,” Psychonomic Bulletin & Review, vol. 25, no. 5, pp. 1667-1680, 2017.

[34] P. Ekman, “An argument for basic emotions,” Cognition & Emotion, vol. 6, no. 3-4, pp. 169-200, 1992.

[35] L. Laskowitz and K. Huffstadt, “Between reality and fiction: Self-representation as an avatar and its effects on self-presence,” in Proc. 2023 Int. Joint Conf. Robotics and Artificial Intelligence (JCRAI), Shanghai, China, 2023, pp. 193-203. [Online]. Available from https://doi.org/10.1145/3632971.3633048 2026.02.16

Autonomous Mobile Robot Movement Algorithm with Human Collision Avoidance Perception

Kazuhisa Miwa*‡, Tomoki Osaki*, Yuki Ninomiya*§, Minoru Karasawa* and Hitoshi Terai†
*Graduate School of Informatics
Nagoya University, Nagoya 464-8601, Japan
Email: kazuhisa.miwa@gmail.com, osaki.tomoki.x5@s.mail.nagoya-u.ac.jp
ninomiya.yuki.t1@f.mail.nagoya-u.ac.jp, mkarasawa@nagoya-u.jp
†Graduate School of Humanity-Oriented Science and Engineering
Kindai University, Fukuoka 820-8555, Japan
Email: teraihitoshi@gmail.com
‡Faculty of Management
Nagoya University of Commerce & Business, Nisshin, Aichi 470-0193, Japan
§Center for Information and Communication Technology
Nanzan University, Nagoya 466-8673, Japan

***Abstract*—In shared spaces, Autonomous Mobile Robots (AMRs) must efficiently reach their goals while remaining physically safe and psychologically nonthreatening. Conventional Social Force Model (SFM) control uses distance-based repulsion and can underestimate future collision risk, which worsens the safety-efficiency tradeoff at higher speeds. We propose a collision-predictive SFM that weighs repulsion based on the anticipated risk of collision at the closest approach, as determined by distance and time. This is mapped through a braking-indicator probability. Simulations compared random, crossing, and straight flows across desired AMR speeds. The method kept close encounters low while shortening completion time in the Straight flow.**

Keywords-Autonomous mobile robot; Social force model; Collision predictive avoidance algorithm.

I. INTRODUCTION

In recent years, shared spaces, where pedestrians, bicyclists, and other users share the same space, have been expanding in urban areas [1][2]. These spaces have limited explicit traffic controls, such as signals and lanes, and order is maintained through mutual coordination among users, including yielding, adjusting speed, and changing paths [3]. This paper addresses the scenario of a single Autonomous Mobile Robot (AMR) operating within an environment of numerous moving pedestrians and performing tasks such as delivery and mobility assistance.

Transportation by AMR is promising due to its convenience. In this context, the AMR requires a navigation strategy that ensures safety while reaching its destination as efficiently as possible [2]. However, a trade-off is anticipated: increasing safety often reduces efficiency metrics, such as travel speed [2][4].

A key challenge when introducing AMRs into shared spaces is achieving "predictable and non-threatening" behavior within pedestrian groups [5][6]. Pedestrians avoid collisions by interpreting the intentions of others through their gaze, body orientation, and subtle changes in speed. If an AMR's behavior lacks human-like cues or consistency, however, it can trigger unnecessary avoidance actions and anxiety, which can disrupt the overall flow of the space [5].

The Social Force Model (SFM) is a well-known model that describes crowd behavior [7][8]. It describes crowd flow and avoidance by combining the driving force toward a destination with repulsive forces from others and obstacles. Applying the SFM to AMR movement could provide a unified approach to handling avoidance actions within pedestrian groups [5][9]. However, SFM's simple, distance-based avoidance is predicted to inadequately reflect differences in future risk based on congestion levels and crossing angles [10]. This makes the trade-off between efficiency and safety more apparent.

Safety in shared spaces must be considered from two perspectives: physical safety, which prevents collisions and excessive proximity, and psychological safety, which mitigates feelings of anxiety, surprise, and pressure experienced by pedestrians [1][2][11]. The widely used traditional risk metric, Time To Collision (TTC) [12], assumes one-dimensional movement. It is difficult to apply to two-dimensional movement, such as when an AMR's trajectory intersects with a pedestrian group's movement. Furthermore, the potential disconnect between the objective danger level and the subjective sense of safety is a unique design challenge of shared spaces [1][11].

Therefore, this study focuses on a braking indicator [13] that assigns a hazard level based on the probability that a person will apply the brakes. This probability is derived from the Distance of Closest Point Approach (DCPA) and the Time of Closest Point Approach (TCPA). We propose a collision-predictive avoidance algorithm that integrates this indicator into SFM. The proposed method aims to reduce travel time while avoiding sudden approaches to pedestrians. It accomplishes this by adjusting the evasion amount (reaction force) based on the "potential for future collision" rather than merely avoiding proximity. Furthermore, the evaluation criteria include the ripple effect on indirect groups of pedestrians (smoothing flow/eliminating congestion), as well as pedestrians who are directly interacting with the AMR. This allows for consideration

 ISBN: 978-1-68558-383-5

of safety and efficiency from the perspective of the entire shared space.

In Section 2, we describe the proposed method and the simulation setting, including the AMR control strategies and evaluation measures. In Section 3, we present the simulation results under the different pedestrian flow conditions. Finally, in Section 4, we discuss the implications of the findings, outline the limitations of the study, and suggest directions for future research.

II. METHOD

A. Overview

This study examines scenarios in which a single AMR enters an environment with numerous pedestrians moving within a shared space. The AMR traverses pedestrian flows while navigating to its destination. The evaluation aims to verify whether the proposed collision-predictive avoidance algorithm can efficiently reach the destination while ensuring safety by preventing excessive proximity to pedestrians. For comparison, a baseline (distance-based avoidance) was established using the SFM to describe the AMR's motion. The performance difference between the baseline and the proposed method was then compared under various pedestrian flow types and AMR desired speed conditions.

B. AMR Movement Control

Two control methods were set for the AMR:

- **Baseline (SFM-based avoidance):** The AMR performs avoidance actions based on the SFM, using repulsive forces proportional to its distance from pedestrians. The avoidance intensity primarily depends on distance and does not directly reflect differences in future collision risk.
- **Proposed (collision-predictive avoidance):** The proposed method retains the SFM framework, but it also estimates the potential for the AMR to make dangerously close approaches to pedestrians in the future. It then adjusts the avoidance intensity according to this risk level. The method uses the distance to closest point of approach (DCPA) and time to closest point of approach (TCPA) to estimate risk. These are calculated from the relative motion between the AMR and the pedestrian. These values are then input into a brake indicator [13], which assigns a risk level as a "probability of braking." The AMR's repulsive force is weighted by this risk level, strengthening avoidance for high-risk targets while mitigating it for low-risk ones. This enables avoidance based on "potential future collision risk" rather than mere proximity avoidance. The method simultaneously aims to maintain safety and suppress reduced mobility efficiency.

C. Simulation Environment

The simulation space was defined as a 20-meter-by-20-meter square field on a two-dimensional plane. Pedestrian agents and one AMR agent were placed within this space. Each agent moved from its starting point toward a destination (goal area). The simulation progressed in discrete time, updating acceleration, velocity, and position at each step. This study defined 1,000 steps as one trial and performed 500 trials. The final performance was calculated as the average of these trials.

D. Pedestrian Flow Scenarios

The SFM was used for the motion model of all pedestrians. The SFM default value of 1.34 m/s was set for all pedestrians to achieve the desired crowd speed. To reproduce typical pedestrian flows in shared spaces, three types of pedestrian flows were created by varying the placement of starting and destination positions: Random, Crossing, and Straight. "Random" represents mixed, multidirectional movement; "Crossing" represents intersecting flows; and "Straight" represents unidirectional, straight-line flow. These conditions include scenarios with different avoidance patterns, such as intersecting, overtaking, and oncoming traffic, enabling a cross-situational evaluation of the effectiveness of the proposed method.

E. Experimental Factors

In this simulation, the AMR's desired speed was the primary operational factor. It was set in steps within the range of 0.5–2.0 m/s to examine how the trade-off between increased efficiency from higher speeds and potential safety compromises could be mitigated. For each pedestrian flow scenario, all combinations of the AMR control method (baseline or proposed) and desired speed were executed.

F. Dependent Measures

Performance evaluation was conducted along two axes: AMR efficiency and safety.

- **Efficiency**: The completion time for the AMR to reach the goal area from its starting point was calculated.
- **Safety**: Events in which the distance between the AMR and a pedestrian fell below a certain threshold were defined as proximity events. The total number of these events (i.e., close encounters) was calculated (e.g., distance <0.1 m).

To examine the ripple effect of the AMR's presence on pedestrian crowd movement, pedestrians were classified into two groups: those that directly interacted with the AMR and those that did not. Direct interaction was defined as pedestrians whose field of view included the AMR during the simulation and who were included in the avoidance calculations. Similar to the AMR, completion time and number of close encounters were measured for both direct and indirect interaction groups.

III. RESULTS

Figures 1 to 3 illustrate the changes in (i) efficiency (completion time) and (ii) safety (number of close encounters) as the desired speed of the AMR is manipulated (0.5–2.0 m/s), categorized by pedestrian flow scenario. The figure shows the results for three groups—Mobility (AMR), Direct-interaction Pedestrians, and Indirect-interaction Pedestrians—and compares SFM (Baseline) with the collision predictive avoidance algorithm (Proposed). Error bars indicate standard errors.

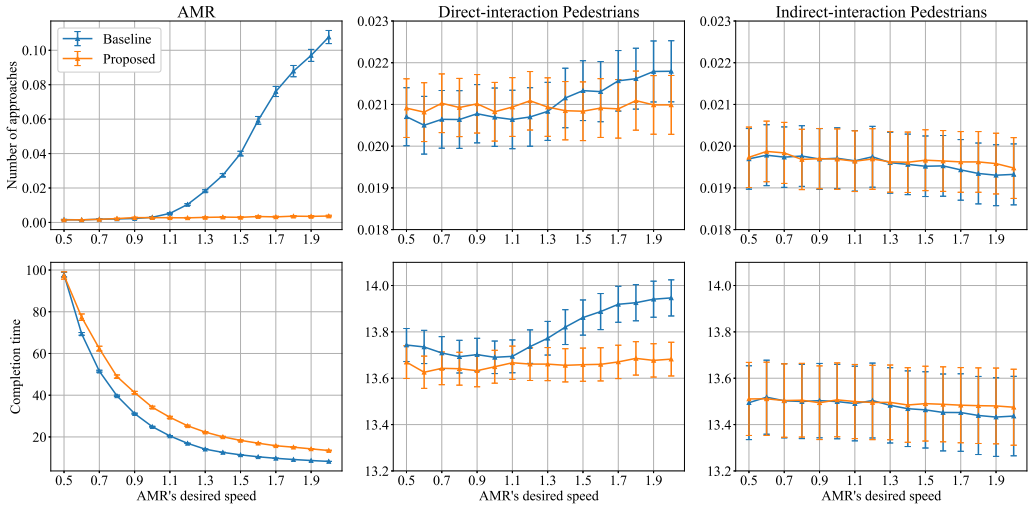


Figure 1. Trade-off between efficiency and safety in random situation.

A. Random Condition

Figure 1 illustrates the results in the random situation.

Focusing on AMR under the random condition, SFM showed an increase in the number of close encounters as the desired speed increased. In contrast, the collision predictive avoidance algorithm maintained a nearly constant number of close encounters regardless of changes in desired speed. Overall, however, SFM had slightly shorter completion times. Under random conditions, SFM exhibited a trade-off where "increased speed $\rightarrow$ shorter completion time" was accompanied by "increased number of close encounters." In contrast, the collision predictive avoidance algorithm exhibited a more desirable pattern of change: "shortening completion time without increasing the number of close encounters." Next, no clear difference in the number of close encounters was observed for Direct-Interaction Pedestrians based on desired speed or algorithm. Regarding completion time, however, under conditions with relatively high desired speeds (1.4 m/s or higher), the collision predictive avoidance algorithm had shorter completion times than SFM. For indirect-interaction pedestrians, no changes in the number of close encounters or completion time were observed due to differences in desired speed or algorithm.

B. Crossing Condition

Figure 2 illustrates the results in the crossing situation.

Mobility and pedestrian change patterns in the crossing condition were similar to those in the random condition. Specifically, SFM increased close encounters frequency with rising desired speed, while the collision predictive avoidance algorithm maintained a nearly constant close encounters frequency. Completion time shortened with increasing desired speed for both algorithms.

C. Straight Condition

Figure 3 illustrates the results in the straight situation.

The mobility change pattern in the straight condition was similar to that in the random and crossing conditions. However, different characteristics were observed in pedestrian movement performance.

For direct-interaction pedestrians, the collision predictive avoidance algorithm resulted in fewer close encounters than SFM under conditions with high desired speeds. Similarly, completion times tended to be shorter for the collision predictive avoidance algorithm. Furthermore, as with direct-interaction pedestrians, the collision predictive avoidance algorithm simultaneously reduced the number of close encounters and shortened completion times for indirect-interaction pedestrians.

IV. Discussion

This study examined the effectiveness of a collision-predictive avoidance algorithm that aims to balance safety and efficiency in a shared space environment with a single AMR and numerous moving pedestrians. A Baseline SFM performing distance-based avoidance was set as a point of comparison. Performance was evaluated by manipulating pedestrian flow scenarios and the desired speed of the AMR. The results showed that, under the Random, Crossing, and Straight conditions, the Baseline exhibited a clear trade-off between improved efficiency and reduced safety as the desired speed increased. In contrast, the proposed method tended to shorten completion time while maintaining a low number of close encounters. This supports the possibility of improving speed without sacrificing safety.

This difference is thought to stem from the disparity in the risk representation used for avoidance decision-making. The baseline method performs avoidance based on repulsive forces derived from distance (or instantaneous configuration). Consequently, as speed increases, situations arise where the heightened risk of future collisions is not sufficiently reflected in the avoidance strength. This results in more approach events at high speeds, making safety degradation more apparent. In contrast, the proposed method uses a braking indicator based

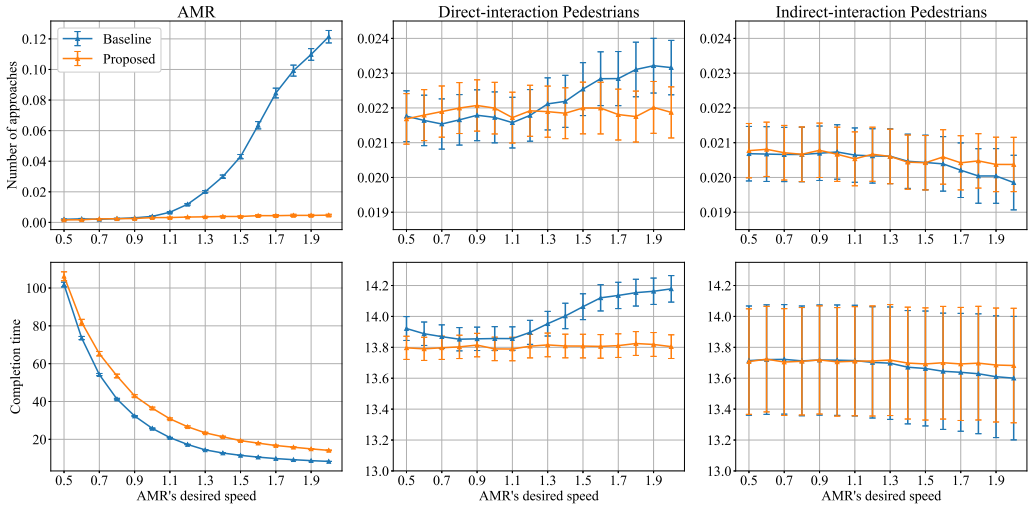


Figure 2. Trade-off between efficiency and safety in crossing situation.

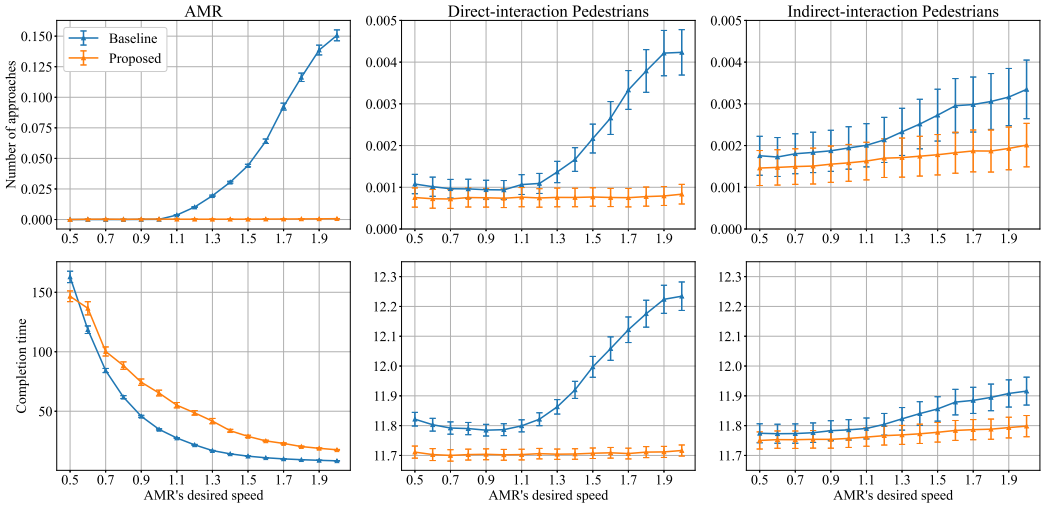


Figure 3. Trade-off between efficiency and safety in straight situation.

on DCPA/TCPA that adjusts the avoidance intensity according to the anticipated danger level of the future closest approach. This enables strong, early avoidance activation for “dangerous crossings” at high speeds while preventing overreaction to lower-risk targets. Therefore, it can be interpreted as achieving improved efficiency while maintaining safety.

The pedestrian-side results suggest that the effectiveness of the proposed method may extend beyond “the AMR’s own safe operation.” Under random/crossing conditions, significant differences were less apparent for indirect-interaction pedestrians, with effects primarily manifesting locally where AMRs interacted directly. This suggests that, in multidirectional, intersecting flows, localized avoidance actions are less likely to propagate order across the entire space. Thus, improvements for AMRs are less reflected in the travel time of indirectly interacting pedestrian groups. Conversely, under the straight condition, the proposed method reduced the number of approaches and shortened completion times for direct- and indirect-interaction pedestrians. This result indicates that, in situations with aligned flow directions, AMR behavior may suppress localized disruptions, such as congestion or chains of excessive avoidance, and propagate smoother flow throughout the entire pedestrian stream more readily.

This study has several limitations. First, the results reported here are from a homogeneous pedestrian group with specific properties. Robustness under realistic conditions, including variations among individuals (desired speed, reaction time, field of view, etc.), requires separate investigation. Second, safety metrics were defined based on the number of close encounters below a threshold; thus, they did not directly evaluate the “severity” of encounters or the abruptness of avoidance maneuvers (i.e., acceleration/jerk related to pedestrian discomfort). Third, only one AMR was tested. In scenarios with multiple AMRs, avoidance interactions could amplify nonlinearly. Future

 ISBN: 978-1-68558-383-5

work should use multifaceted metrics to evaluate performance, including approach distribution, avoidance smoothness, and pedestrian subjective evaluations. This work should also extend to multiple AMRs and real-world experiments.

As an important next step, real-world validation should be conducted to examine whether the proposed method also improves pedestrians' sense of psychological safety. In addition to field experiments in shared spaces, user studies could assess subjective reactions such as perceived safety, comfort, predictability, and threat while people interact with AMRs using different avoidance strategies. Such evaluations would help substantiate the claim that the proposed algorithm is not only physically safe, but also psychologically nonthreatening.

V. Conclusions

In this paper, we proposed a collision-predictive avoidance algorithm, showing promise in mitigating the safety-efficiency trade-off that is typically observed in distance-based AMR navigation in shared spaces. By incorporating anticipated collision risk through DCPA/TCPA-based braking indicators, the method reduced close encounters while maintaining or improving travel efficiency across multiple pedestrian flow conditions. The findings also suggest that its benefits may extend beyond the AMR itself to surrounding pedestrians, particularly in aligned flow environments. Although further validation under more realistic and complex conditions is necessary, the present results provide initial support for the usefulness of collision-predictive avoidance as a framework for achieving both physically safe and psychologically acceptable AMR behavior in shared spaces.

Acknowledgement

Support for this work was given by JSPS KAKENHI Grant Number 22H03912 and 22H00211, and by Toyota Motor Corporation (TMC). However, note that this article solely reflects the opinions and conclusions of its authors and not TMC or any other Toyota entity. There are no other conflicts of interest to declare.

References

[1] Y. Hasegawa, C. Dias, M. Iryo-Asano, and H. Nishiuchi, "Modeling pedestrians' subjective danger perception toward personal mobility vehicles", *Transportation research part F: traffic psychology and behaviour*, vol. 56, pp. 256–267, 2018.

[2] P. Salvini, D. Paez-Granados, and A. Billard, "Safety concerns emerging from robots navigating in crowded pedestrian areas", *International Journal of Social Robotics*, vol. 14, no. 2, pp. 441–462, 2022.

[3] B. Hamilton-Baillie, "Shared space: Reconciling people, places and traffic", *Built environment*, vol. 34, no. 2, pp. 161–181, 2008.

[4] A. Laureshyn, Å. Svensson, and C. Hydén, "Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation", *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1637–1646, 2010.

[5] M. Shiomi, F. Zanlungo, K. Hayashi, and T. Kanda, "Towards a socially acceptable collision avoidance for a mobile robot navigating among pedestrians using a pedestrian model", *International Journal of Social Robotics*, vol. 6, no. 3, pp. 443–455, 2014.

[6] Y. Wang, L. Hespanhol, S. Worrall, and M. Tomitsch, "Pedestrian-vehicle interaction in shared space: Insights for autonomous vehicles", in *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2022, pp. 330–339.

[7] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics", *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[8] D. Helbing, I. Farkas, and T. Vicsek, "Simulating dynamical features of escape panic", *Nature*, vol. 407, no. 6803, pp. 487–490, 2000.

[9] M. Prédhumeau, L. Mancheva, J. Dugdale, and A. Spalanzani, "Agent-based modeling for predicting pedestrian trajectories around an autonomous vehicle", *Journal of Artificial Intelligence Research*, vol. 73, pp. 1385–1433, 2022.

[10] F. Pascucci, N. Rinke, C. Schiermeyer, B. Friedrich, and V. Berkhahn, "Modeling of shared space with multi-modal traffic using a multi-layer social force approach", *Transportation Research Procedia*, vol. 10, pp. 316–326, 2015.

[11] A. Jafari and Y.-C. Liu, "Pedestrians' safety using projected time-to-collision to electric scooters", *Nature communications*, vol. 15, no. 1, p. 5701, 2024.

[12] J. C. Hayward, "Near miss determination through use of a scale of danger", *Highway Research Record*, no. 384, pp. 24–35, 1972.

[13] S. Matsubayashi, K. Miwa, H. Terai, and Y. Ninomiya, "Index of braking behaviour in two dimensions within risk perception", *Transportation research part F: traffic psychology and behaviour*, vol. 102, pp. 164–176, 2024.

Conversational Web Browsing: Voice-Only Navigation

Daniele Farriciello and Jing Hua Ye
Department of Computer Science
Munster Technological University
Cork City, Republic of Ireland
e-mail: daniele.farriciello@mymtu.ie | jinghua.ye@mtu.ie

***Abstract*—Web browsing shapes how we work and connect, yet it still relies heavily on the mouse and keyboard. For people with physical disabilities or professionals in hands-busy environments, this dependence makes accessing the web frustrating or impossible. This study presents a solution, a VoiceNav system that allows users to browse websites entirely using voice commands. The goal is to make this system work automatically on any website, bridging the gap by enabling users to browse and control websites entirely through voice commands. By processing natural speech, analyzing the underlying structure of web pages, and mapping spoken instructions to precise actions, the system transforms traditional navigation into a more intuitive and conversational experience. A critical feature is its ability to resolve ambiguity. When faced with multiple similar elements, such as two buttons labeled "Upload," the system asks the user to clarify their intent, mirroring real-life communication and ensuring that interactions remain accurate and dependable. In its current prototype, the system focuses primarily on voice interaction and real-time understanding of the page. Spoken requests are converted into structured commands by a parser, while the page is broken down through Document Object Model (DOM) analysis to identify the most relevant interactive elements. This enables support for common actions, such as opening websites, clicking buttons and links, scrolling, navigating menus, and typing into fields, without requiring any site-specific setup. Initial testing shows strong performance on navigation-centric browsing, with an overall success rate of around 80%, driven primarily by robust execution of general navigation and clicking actions (e.g., opening links, selecting buttons, and scrolling). The main weaknesses appear in form-heavy scenarios, especially when entering emails or other structured inputs, where transcription and formatting errors can reduce precision. These results highlight both the practicality of voice-first browsing today and the areas where more intelligent input handling and stricter validation would further enhance reliability. By demonstrating the capabilities of voice-based web navigation, this project showcases how accessibility and modern technology can collaborate to enhance everyday life. It offers not just a practical tool, but also a glimpse of a future where interacting with the web feels more natural, inclusive, and human.**

Keywords-web browsing, voice-based, personalized, conversational, hands-free

I. Introduction

Web browsing remains one of the most common digital activities, yet interaction with websites still depends heavily on the mouse and keyboard. This dependency limits the adoption of more natural human–computer interaction and creates barriers in hands-busy scenarios, where users cannot conveniently switch between physical tasks and manual browser control.

These barriers are more severe for people with motor impairments (e.g., limited hand mobility or conditions affecting fine control), for individuals recovering from injuries, and for older adults experiencing reduced mobility. While accessibility standards, such as the Web Content Accessibility Guidelines (WCAG) encourage inclusive design, many real-world websites are still difficult to operate without conventional input devices, leaving a gap between technical capability and practical inclusion.

This study addresses that gap by presenting VoiceNav, a browser-based system that enables users to navigate and operate websites using natural voice commands, without requiring site-specific customization. Users can issue direct instructions such as opening a website, scrolling, clicking links or buttons, and entering text into input fields, allowing common browsing tasks to be completed hands-free. The system is designed to work on arbitrary websites by analyzing the page structure and mapping spoken intent to concrete browser actions.

A key challenge in voice-driven interaction is ambiguity: a page may contain multiple similar interactive elements (for example, two buttons labeled "Upload"). To maintain reliable control, the system incorporates a clarification step that prompts the user when multiple plausible targets exist, mirroring how ambiguity is resolved in human conversation. By combining speech recognition, intent parsing, DOM-based element discovery, and action execution in a modular workflow, the approach aims to make voice-first browsing more usable, dependable, and scalable for accessibility and everyday hands-busy use cases.

The general relevant background on this field is discussed in Section II. The design of the VoiceNav system and the evaluation methodology are presented in Section III. The prototype of this system is narrated in Section IV. The evaluation of the performance of the VoiceNav system is articulated in Section V. The final remarks and the future extension of this system are presented in Section VI.

II. Background

Recent advances in web technologies and interaction paradigms have significantly transformed how users access and navigate online content. These developments have improved inclusivity by enabling alternative interaction modalities that reduce reliance on traditional input devices, such as keyboards, mice, and touchscreens. Among these, hands-free, voice-driven web browsing has emerged as a promising approach for enhancing accessibility and usability, particularly for users with motor impairments or in contexts where manual interaction is impractical.

 ISBN: 978-1-68558-383-5

Hands-free web browsing enables users to navigate websites, retrieve information, and perform complex tasks using spoken language. At its core, this interaction paradigm relies on Automatic Speech Recognition (ASR) systems, which convert spoken audio into textual representations by analyzing phonetic patterns and linguistic structures [1]. While early ASR systems were constrained by limited vocabularies and sensitivity to noise and accents, modern deep learning–based approaches have substantially improved robustness and accuracy across diverse speakers and environments [2].

Once speech is transcribed, natural language processing (NLP) techniques are employed to interpret user intent and map spoken commands to executable browser actions. These commands range from simple navigational requests (e.g., "scroll down") to complex multi-step instructions (e.g., "search for running shoes under $100 and add the first result to the cart"). NLP models analyze syntax, semantics, and contextual cues to infer user goals and translate them into appropriate web interactions [3]. This capability is central to enabling natural, flexible voice-based browsing experiences.

Voice-driven browsing is closely related to the evolution of voice assistants, such as Siri, Alexa, and Google Assistant, but places a stronger emphasis on direct interaction with web page elements, including forms, buttons, multimedia content, and dynamically generated interfaces. This requires precise grounding of language commands to the Document Object Model (DOM) and continuous adaptation to changing web states [4]. Human–Computer Interaction (HCI) research plays a crucial role in shaping these systems by guiding interface design, feedback mechanisms, and error-recovery strategies to support natural and trustworthy interaction [4].

Key advantages of voice-driven browsing include greater accessibility for users with motor disabilities, more natural and hands-free interaction for multitasking scenarios (e.g., driving or cooking), and enhanced usability for users with limited literacy or language skills. Furthermore, real-time continuous speech recognition enables always-on listening modes, reducing the need for repeated activation phrases and providing smoother user experiences.

Despite its advantages, hands-free browsing presents several challenges. ASR systems must handle background noise, accents, homophones, and disfluencies, while NLP components must resolve ambiguous or underspecified commands. Furthermore, dynamic and visually complex web layouts complicate the mapping between language and actionable elements, necessitating robust error detection and correction mechanisms to maintain user confidence [3].

Recent progress in large language models (LLMs) has significantly advanced voice-based web interaction by enabling deeper contextual understanding, multi-step reasoning, and iterative error correction. Frameworks, such as DexAssist demonstrate how dual-LLM architectures can separate planning and execution monitoring, leading to improved task success rates in complex browsing scenarios [3]. Similarly, multimodal approaches like WebVoyager incorporate visual perception to interact with real-world websites more effectively [5]. These developments represent a shift from rigid command-based systems toward intelligent, adaptive web agents capable of natural language interaction.

Understanding the evolution and technical foundations of hands-free web browsing is essential for situating current research within the broader fields of computer science and HCI. This work builds upon advances in ASR, NLP, accessibility-driven design, and LLM-based reasoning to contribute to the development of inclusive, intelligent web interaction systems.

A. Speech Recognition and Natural Language Processing

At the foundational level, hands-free browsing relies heavily on ASR and NLP technologies [1][2]. Speech recognition serves as the primary input mechanism, converting acoustic signals into digital text representations. This process involves complex signal processing, phonetic modeling, and language modeling components that must operate reliably across diverse speakers, environments, and linguistic variations.

Modern speech recognition systems employ deep learning architectures, including recurrent neural networks, transformer models, and connectionist temporal classification (CTC) algorithms to achieve robust performance [2]. The challenge extends beyond simple transcription to include understanding speaker intent, handling disfluencies, and adapting to domain-specific vocabularies relevant to web browsing tasks.

NLP complements speech recognition by interpreting the semantic content of transcribed utterances. NLP techniques enable systems to parse complex instructions, resolve ambiguities, and map user intentions to executable browser actions [3]. This semantic understanding is crucial for handling the varied and often imprecise nature of spoken commands in real-world browsing scenarios.

B. Human-Computer Interaction and Accessibility

The accessibility dimension is particularly significant, as voice-driven browsing serves as an assistive technology for users with motor disabilities, visual impairments, or other conditions that limit traditional input methods. This positions the research within the broader context of universal design and inclusive technology development, aligning with established accessibility standards and guidelines.

C. Speech Recognition Technologies

Speech recognition technology has evolved significantly from early template-matching approaches to sophisticated deep learning systems. Traditional ASR systems relied on hidden Markov models and Gaussian mixture models, which required extensive training data and performed poorly with speaker variations [2].

Contemporary ASR systems employ end-to-end neural architectures that can learn directly from raw audio to text mappings. These systems demonstrate improved robustness to noise, accents, and speaking styles, making them more suitable for real-world browsing applications [1]. However, challenges remain in handling domain-specific terminology, proper nouns, and the informal language patterns common in voice commands.

Recent research has focused on accent-specific adaptation techniques that improve recognition accuracy for diverse user populations [2]. This work is particularly relevant for voice browsing systems that must serve global user bases with varying linguistic backgrounds.

III. System Architecture and Evaluation Methodology

A. System Architecture

The proposed architecture for VoiceNav integrates both voice and eye-based interaction to achieve a fully hands-free web navigation experience. The system follows a modular, service-oriented design where each module is responsible for a specific functionality but communicates seamlessly with others to ensure consistency, accuracy, and responsiveness.

At a high level, the system comprises six core components: Voice Input, Speech Recognition, Eye Tracking, AI Processing, Action Execution, and Browser Interface. Figure 1 illustrates the overall structure and data flow.

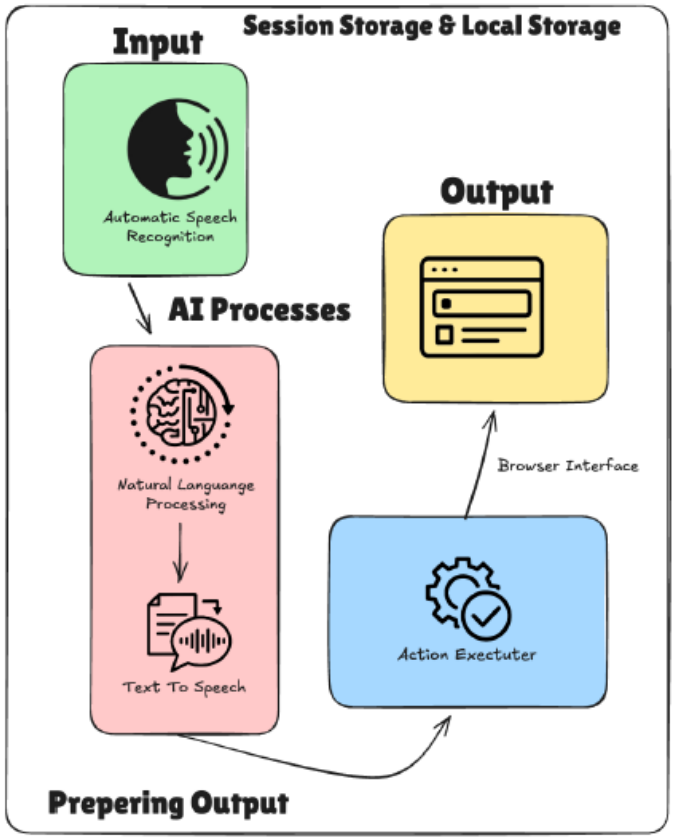


Figure 1. High-level architecture of the VoiceNav system showing the integration of voice

The voice input component captures the user's voice commands through the browser microphone interface. Using the Web Speech API, it continuously listens and converts spoken instructions into text. The system maintains a configurable confidence threshold (set to 0.3) to minimize misrecognition, automatically restarting recognition in noisy environments.

The AI processing module interprets both the text output from speech recognition. It uses natural language processing to determine user intent and to produce a specific action command. The module includes the following subcomponents:

- Intent Parser: Interprets command semantics (e.g., navigate, click, type).
- Context Resolver: Matches identified intent with elements or cross-checked elements in the DOM.
- Fallback Mechanism: Handles cases when voice input is unreliable, defaulting to the best available modality.

The action execution module receives the interpreted command and performs the corresponding operation directly on the web page. It interacts with the DOM to execute actions, such as clicking buttons, scrolling, typing input, or triggering events.

The interface feedback module provides visual or auditory feedback upon successful command recognition or action execution. Examples include highlighting the element being acted upon, confirming actions via voice response, or displaying temporary overlays.

B. Evaluation Methodology

The evaluation methodology for the VoiceNav system was designed to examine the effectiveness, robustness, and usability of hands-free, voice-first web navigation in realistic interaction scenarios. The methodology builds upon established evaluation practices in voice-driven browsing, multimodal interaction, and AI-powered web agents [3]–[5], emphasizing end-to-end task performance rather than isolated component accuracy.

1) Experimental Design: VoiceNav was evaluated using a task-based experimental design, in which the system was required to complete predefined browsing tasks on real, publicly accessible websites. This approach reflects common evaluation strategies used in prior studies on voice-controlled browsing systems and autonomous web agents [3][5], ensuring ecological validity and relevance to real-world use cases. The evaluation focused on voice-first interaction, meaning that all commands were issued verbally without reliance on traditional input devices, such as keyboards or mice. This constraint ensures that observed performance accurately represents hands-free usage conditions, which are central to the system's accessibility and HCI goals [4].

2) Task Selection and Websites: A set of representative browsing tasks was selected to cover a broad range of interaction types, including:

- Page navigation and scrolling
- Link and button selection
- Menu interaction
- Limited form interaction (e.g., entering text into input fields)

Tasks were executed across multiple websites with varying structural complexity, including content-oriented pages and form-heavy interfaces. This selection strategy allows performance comparison across different web layouts and interaction demands, a factor known to significantly influence voice-based system accuracy [3].

3) System Configuration: The VoiceNav system was deployed as a Chrome browser extension, leveraging the Web Speech API for speech recognition [1] and browser-level access to the DOM for action execution. Default system parameters were used throughout testing, including a fixed speech recognition confidence threshold and standard language settings. No user-specific calibration, adaptation, or personalization was applied, ensuring consistency across trials and enabling fair comparison across websites.

Speech input is processed in real time, and commands are interpreted using the system's AI processing module, which maps transcribed text to browser actions. This configuration mirrors the operational conditions of contemporary voice browsing systems and allows direct comparison with existing approaches described in the literature [3][4].

4) Evaluation Procedure: For each website, a sequence of tasks was executed sequentially. A task was considered successful if the system completed the intended action correctly without requiring repetition, clarification, or manual intervention. Failed tasks included misinterpreted commands, incorrect element selection, or incomplete execution.

To capture performance variability, task success rates were calculated per website, rather than aggregated across all tasks. This granular analysis enables identification of specific interaction contexts where performance degrades, particularly in scenarios involving structured data entry or complex page layouts [2][3].

5) Performance Metrics: The primary evaluation metric was task success rate, expressed as a percentage of correctly executed commands. This metric is widely used in evaluating voice-based and AI-driven web interaction systems, as it directly reflects practical usability [3]–[5].

In addition to the overall success rate, qualitative observations were recorded during testing to identify recurring error patterns, such as transcription errors, ambiguous command interpretation, and failures in form validation. These observations provide contextual insight into system limitations that are not fully captured by quantitative metrics alone [4].

6) Methodological Limitations: The methodology intentionally prioritizes system-level feasibility over large-scale statistical validation. As such, the evaluation does not include extensive user studies or long-term adaptation analysis. However, this approach is consistent with exploratory evaluations of emerging multimodal interaction systems and early-stage AI-assisted browsing frameworks [3]–[5].

Overall, this methodology provides a structured and realistic assessment of VoiceNav's current capabilities while clearly identifying directions for future refinement, including improved handling of structured input and enhanced error recovery strategies enabled by large language models [3].

IV. Prototype

In this section, the first version of the VoiceNav user interface is shown. These images (Figures 2, 3, 4, 5) give an idea of how users will interact with VoiceNav. This product is built as a Chrome extension. When loaded, the user sees the main bar showing that the project is ready (Figure 2). In the menu (Figure 3), users can switch languages (for example, English or Italian), mute the microphone, and choose whether the system is listening for commands. The green icon can be moved around the screen so users can place it where it is easiest to use. This makes the system flexible for everyone.

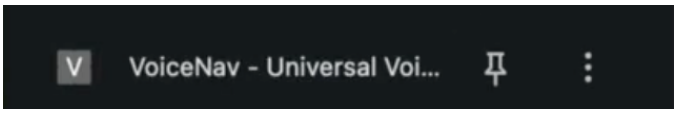


Figure 2. Extension loaded and ready to use (VoiceNav bar)

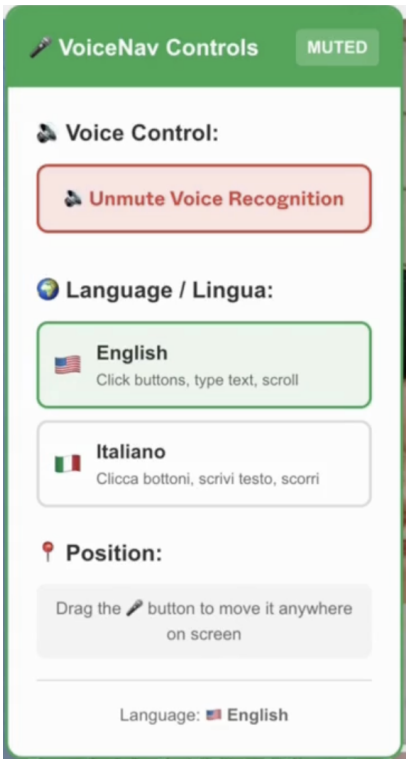


Figure 3. Menu interface lets users switch language and control the microphone. Icon can be moved anywhere

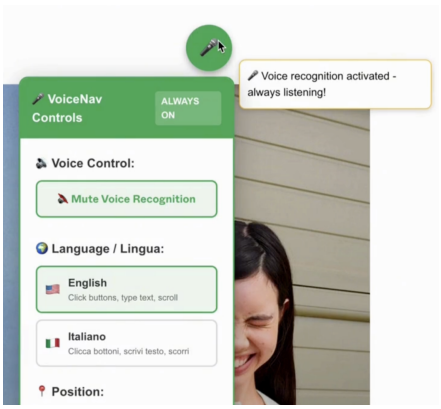


Figure 4. Voice recognition is always listening for commands and gives a quick response

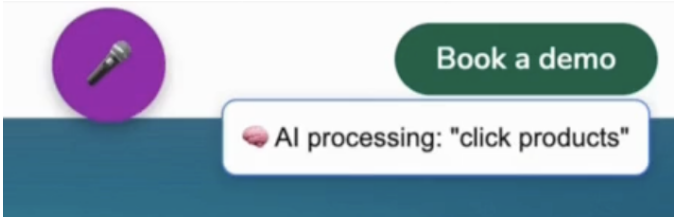


Figure 5. AI processes spoken commands and passes them as instructions to the command parser

When voice recognition is active, VoiceNav listens for spoken commands. A pop-up appears showing ”Voice recognition activated – always listening!”, letting users know they can speak and the system will react quickly (Figure 4). When the user says a command like ”click products,” an AI processes the instruction and shows a reminder that it is working on the user’s request (Figure 5). The spoken command is translated from speech to text, ensuring accuracy so actions match what users want.

V. Discussion | Evaluation

Scenario: Users want a drop-in paragraph added to their existing text that explicitly links the lower checkbox/radio accuracy to form-control ambiguity and UI implementation differences.

Initial testing shows strong performance on general navigation and clicking tasks, with an overall success rate of approximately 80% as illustrated in Figure 6. The figure summarizes command execution accuracy across the evaluated interaction categories and highlights the robustness of voice-driven navigation in typical browsing scenarios. The main weaknesses emerge in form-heavy interactions, particularly when entering email addresses or other structured inputs, where transcription and formatting errors reduce overall precision. These results demonstrate both the current practicality of voice-first browsing and the areas where smarter input handling and tighter validation mechanisms would further improve reliability.

Across navigation-centric websites, VoiceNav demonstrated consistent accuracy when executing commands, such as scrolling, opening links, navigating menus, and selecting visible interface elements, as reflected by the high success rates shown in Figure 6. These findings are consistent with prior research indicating that voice-driven interaction performs reliably when commands can be directly mapped to discrete DOM elements.

Performance degradation was observed on websites that require structured form input, particularly for email fields and text fields that enforce strict validation rules. As shown in Figure 6, these scenarios exhibit noticeably lower accuracy compared to navigation tasks. In such cases, minor transcription errors introduced by the speech recognition component resulted in failed submissions or incorrect input. This limitation reflects a broader challenge identified in voice-based interaction systems, where free-form speech must be translated into highly constrained input formats.

Form controls also contributed to reduced reliability, especially where the system had to resolve a spoken label to a specific option within a group. Figure 6 shows lower success rates for radio-button selection (60%) and checkbox selection (40%) compared to navigation tasks, indicating that VoiceNav is less accurate on state-based inputs than on discrete click targets. This reduction is partly explained by ambiguity in user phrasing (e.g., “enable/disable,” “tick/untick,” “the second option”), closely named alternatives, and inconsistent DOM/accessibility labeling on custom-styled inputs, all of which make correct grounding and state toggling harder than activating a single button or link.

Despite these limitations, the overall results indicate that voice-first browsing is already viable for a wide range of everyday web interactions. The contrast between high navigation accuracy and reduced performance on structured inputs, visible in Figure 6, underscores the importance of integrating intelligent input correction, validation-aware formatting, and iterative error-recovery mechanisms capabilities increasingly supported by large language model-based architectures.

VI. Conclusion and Future Work

This study demonstrates that voice-based hands-free web browsing is a practical and valuable step toward making everyday web access more inclusive for users who cannot reliably use a mouse and keyboard, as well as for hands-busy situations. The proposed approach emphasizes a modular, browser-native design in which continuous speech recognition, natural-language command parsing, DOM-based element discovery, and action execution operate as separate components that work together to deliver end-to-end voice control on arbitrary websites.

The main proposition supported by the prototype work is that voice control can be implemented effectively using standard browser technologies, provided the system includes robust intent parsing and reliable grounding of commands to page elements. This early voice-based system is evaluated using 40 test cases, with 10 per category across a single website. We also manually evaluated the system with 10 participants across mixed tasks, including tables, navigation, duplicate controls, and pop-up scenarios. Early prototype results indicate strong performance on common navigation operations (e.g., opening pages, clicking links/buttons, scrolling), while highlighting that form-heavy interactions and structured text entry remain the most error-prone areas due to transcription and formatting challenges. These findings suggest the system is already suitable for basic browsing assistance, with clear, well-scoped engineering work needed to improve reliability in noisy environments and to strengthen text-entry mechanisms (e.g., spelling modes, confirmation steps, and validation).

From a practical standpoint, the work implies that voice-first browsing can reduce friction for accessibility use cases without requiring site-specific configuration, which is critical for real-world adoption across diverse websites. To be dependable in daily use, the system must also prioritize user trust: clear feedback, safe error recovery, and transparent confirmations when actions are ambiguous are essential to prevent unintended clicks or submissions. Future development should focus on

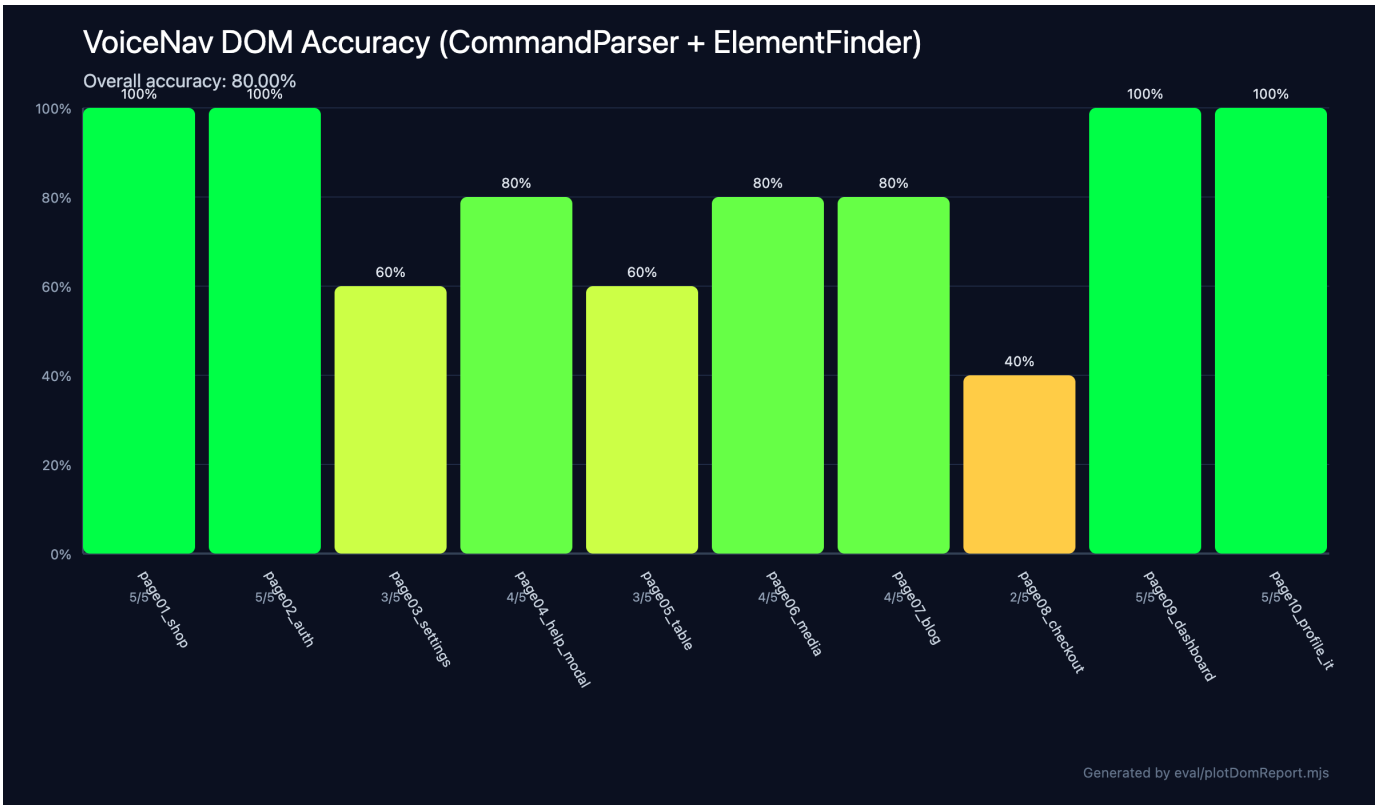


Figure 6. VoiceNav DOM command execution accuracy across evaluated interaction categories

expanding the command set, improving robustness and latency, and running structured user evaluations to quantify task success rates and usability across a wider range of websites and environmental conditions.

References

[1] J. Adorf, "Web Speech API", KTH Royal Institute of Technology, Technical Report, 2013.

[2] D. Prabhu, P. Jyothi, S. Ganapathy, and V. Unni, "Accented Speech Recognition With Accent-specific Codebooks", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore, Dec. 2023, pp.7175–7188. DOI: 10.18653/v1/2023.emnlp-main.444.

[3] S. Mehendale and A. Walishetti, "DexAssist: A Voice-Enabled Dual-LLM Framework for Accessible Web Navigation", in *Intelligent Human Computer Interaction: 16th International Conference, IHCI 2024*, D. Singh, J.-W. Van't Klooster, and U. S. Tiwary, Eds., Twente, The Netherlands: Springer-Verlag, May 2024, pp.171–177. DOI: 10.1007/978-3-031-88881-6_14.

[4] J. Cambre et al., "Firefox Voice: An Open and Extensible Voice Assistant Built Upon the Web", in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2021, pp.250:1–250:18. DOI: 10.1145/3411764.3445409.

[5] H. He et al., "WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models", in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp.6864–6890. DOI: 10.18653/v1/2024.acl-long.371.

Dynamic Diorama: Narrative-Driven Orientation Modeling and Object Placement for VR

Furkan Çelen
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: celen23f@itu.edu.tr

Meral Kuyucu
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: korkmazmer@itu.edu.tr

Bora Şenceylan
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: senceylan19@itu.edu.tr

Gökhan İnce
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: gokhan.ince@itu.edu.tr

***Abstract*—Spatial composition is a key factor in Virtual Reality storytelling, as object arrangement directly influences how users perceive meaning and emotion. However, converting text into 3D layouts is difficult because systems typically prioritize geometric rules over narrative context. Dynamic Diorama moves beyond simple geometry by analyzing text for structural details and emotional cues to shape the layout. Rather than just placing objects randomly or by rigid rules, our pipeline aligns the spatial relationships directly with the story's mood. This approach is benchmarked against standard baselines across four distinct narrative themes: happiness, fear, surprise, and sadness. Participant feedback indicated that the Large Language Model driven scenes offered significantly better narrative coherence and emotional alignment compared to the random and heuristic baselines. Eye-tracking data supported this finding, revealing that semantically informed scenes reduced the time-to-first-fixation on key narrative elements.**

***Keywords*—narrative-driven scene generation; VR storytelling; semantic layout reasoning; spatial placement.**

I. Introduction

Stories in Virtual Reality (VR) are shaped as much by spatial arrangement as by plot. Where an object sits, how characters face each other, and which items occupy the foreground all contribute to what users notice and how they feel. These decisions are not merely authoring challenges, but perceptual ones, as spatial decisions directly shape what users attend to and how narratives are experienced. Unlike traditional screen-based media, immersive Augmented Reality (AR) and VR environments amplify the perceptual impact of spatial decisions. Object placement, orientation, and proximity are not merely compositional choices but directly influence users' sense of presence, attention, and emotional engagement. Small spatial inconsistencies may disrupt narrative flow or reduce immersion, while subtle spatial cues can silently reinforce tension, intimacy, or anticipation.

Creating spatial narratives remains labor-intensive, as designers must constantly trade off physical plausibility, visual clarity, and affective intent—factors that do not always align. Many automated layout tools prioritize geometric or visibility constraints, producing scenes that are technically coherent but may fail to reflect the intended emotional cues of the story [1][2]. Narrative text contains both explicit relations (e.g., "the lamp sits on the table") and implicit affective cues, such as tension, intimacy, or distance, that can inform scene composition. Still, prior evaluations tend to emphasize per-object plausibility or geometry-focused metrics, with comparatively less attention given to human perception of narrative coherence in immersive environments. Recent language-conditioned systems demonstrate that free-text can guide object placement and reduce manual effort [3][4].

Advances in Artificial Intelligence (AI), particularly language-based models, offer an opportunity to bridge this gap between narrative intent and spatial realization. Narrative descriptions often encode affective and relational cues implicitly rather than explicitly, leaving room for interpretation that exceeds the expressiveness of rule-based systems. In this context, AI-driven reasoning is not introduced to maximize automation, but to explore whether semantic and affect-aware interpretations of narrative text can better support spatial storytelling in immersive environments. To study this interpretive space systematically, this study frames placement as an experimental variable of narrative-driven scene generation.

We introduce a controlled VR testbed, Dynamic Diorama, which exposes three placement paradigms (random baseline, rule-based heuristic, and Large Language Model (LLM) informed placement) and applies them to short vignettes designed to evoke distinct emotions. The testbed combines semantic parsing, heuristic validation, and placement ranking to maintain physical plausibility while enabling systematic comparisons across placement approaches. We investigated how variations in placement strategy shape viewer perception, attention, and spatial validity. This study addresses the following research questions:

- **RQ1:** How do layouts produced by different object placement strategies affect viewers' perceived narrative coherence and emotional alignment in VR scenes?
- **RQ2:** How do these layout strategies influence visual attention patterns, as measured through gaze-based metrics, during narrative scene exploration?
- **RQ3:** To what extent can layout strategies that incorporate higher-level semantic or affective reasoning improve perceived narrative quality without increasing spatial invalidity, such as collisions or physically implausible placements?

 ISBN: 978-1-68558-383-5

The contributions of this study are threefold: 1) a controlled experimental testbed for narrative scene layout in VR, where the object placement strategy is treated as an independent experimental variable, 2) an experimental setup with three random, heuristic, and LLM-based layout strategies designed to enable systematic comparison of their effects on viewer perception and attention and 3) an empirical VR study that examines the effects of different layout strategies on narrative coherence, emotional alignment, and visual attention using self-report and gaze-based measures.

The remainder of this paper is organized as follows: Section II reviews related work on text-to-scene generation. Section III details the Dynamic Diorama framework and placement strategies. Section IV describes the experimental design and methodology. Section V presents the quantitative and qualitative results. Finally, Section VI concludes the study and outlines future directions.

II. Related Work

A. Object Placement in Text-to-Scene Systems

Early attempts to generate scenes from text relied heavily on hand-crafted grammars and domain-specific rules, which proved brittle when applied to open-ended or creative narratives [2]. Recent work has shifted toward transformer-based language models, which are better suited to handling ambiguity and implicit structure. LLMs have been shown to extract elements, such as scene boundaries, characters, and relationships directly from unstructured natural language [5]. Systems like PlaceIt3D [3] and SceneTeller [4] illustrate how natural language can be used to guide 3D layout generation and reduce the need for manual scene construction.

In VR and Mixed Reality (MR) applications, object placement is typically constrained by geometric and semantic considerations. Common practices include managing occlusion, optimizing visibility, and respecting surface affordance, such as ensuring that objects are placed on appropriate supports rather than floating in space [1][3]. Work on AR label placement and occlusion-aware heuristics highlights the limits of geometry-first optimization when communicative or narrative goals are considered [6][7][8]. Scene-graph representations further formalize spatial relationships by linking object categories to likely locations and neighboring elements, demonstrating how semantic information supports functional placement [9]. These methods are effective at producing physically consistent environments.

Still, many of these approaches focus on identifying what should appear in a scene, paying less attention to how the emotional tone of a narrative should influence spatial composition. However, physical correctness alone does not guarantee that a scene supports a narrative. In many cases, layouts that are spatially sound fail to convey the emotional tension or intimacy implied by a story, resulting in environments that feel correct but emotionally unengaging [6].

B. Learning-Based Object Placement

Learning-based object placement models address some of the limitations of rule-driven systems by inferring spatial patterns from large collections of scenes [10]. This allows them to capture contextual relationships that are difficult to express through explicit heuristics. At the same time, unconstrained learning-based outputs can introduce practical issues, including object collisions or violations of physical affordance. Hybrid approaches attempt to balance these trade-offs by combining learned or language-derived proposals with rule-based validation mechanisms [11]. Such combinations are especially relevant in storytelling contexts, where a degree of spatial flexibility is needed, but basic physical coherence must still be preserved [12].

C. Interactive Layout Tools

Another line of work extends language-conditioned placement through open-vocabulary mappings, allowing free-form textual descriptions to be associated with 3D assets beyond fixed label sets [13]. This capability is particularly important for narrative scenes, where descriptions are often abstract or metaphorical. Interactive, chat-driven layout tools further point toward more fluid human–model workflows by enabling authors to iteratively refine layouts through dialogue rather than low-level parameter tuning, as demonstrated by systems, such as Chat2Layout [12]. Despite these advances, evaluating whether a generated scene actually aligns with an author's intent remains difficult. As a result, assessment often depends on a combination of automated measures and user-centered evaluation.

D. Emotion-Aware Spatial Design

Emotion-aware techniques are widely used in VR to drive reactive elements, such as lighting, sound, or avatar behavior [14]. Their influence on the spatial arrangement of objects, however, has received comparatively less attention. Frameworks like UniEmoX [15] suggest ways to model emotion perception in a general form, but there is still limited empirical evidence on how emotion-driven spatial composition affects user experience in immersive environments. This gap motivates our study that compares different placement strategies—ranging from random and heuristic methods to language-guided approaches—under controlled emotional narratives.

While existing text-to-scene systems primarily focus on geometric plausibility, our Dynamic Diorama framework introduces a novel approach by treating spatial placement as an experimental variable driven by affective cues. This explicitly bridges the gap between semantic LLM reasoning and emotional narrative alignment in VR.

III. Dynamic Diorama Framework

A. Framework Overview

In this study, we propose Dynamic Diorama as a research platform that supports systematic observation of how different object placement strategies influence narrative experience in

VR. This framework emphasizes comparability across conditions by using the same stories, assets, and physical constraints. This keeps the focus on placement behavior rather than content differences.

Figure 1 illustrates the high-level workflow of the Dynamic Diorama framework. In this workflow, the narrative text is first organized into scene-level representations that guide object placement. The resulting layouts are validated for spatial plausibility before being rendered as a VR story.

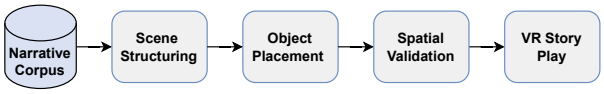


Figure 1. High-level workflow of the Dynamic Diorama framework.

B. Narrative Stimuli and Visual Scope

Four narratives were prepared to be used across all experiments to avoid variation. Each narrative was written to convey a single dominant emotion (happy, fearful, surprised, or sad). Narrative length ranged from approximately 60 to 100 words. To support temporal progression, each story is divided into five sequential scenes. Scene boundaries are derived using an LLM that produces a structured representation of narrative flow. Rather than generating geometry directly, this representation is later interpreted by the Unity runtime to drive scene transitions and placement logic. This step enables consistent story structure across placement strategies while keeping narrative content fixed.

All scenes are composed of a fixed pool of 30 3D assets. This collection comprises a diverse set of low-poly models, including environmental elements (e.g., furniture, foliage), human characters, and animals. These assets were selected for their narrative versatility, allowing the same objects to be recontextualized across different emotional scenarios (e.g., a dog functioning as a companion in a 'Happy' scene or a threat in a 'Fear' scene). The asset pool size was limited by hardware performance, development effort, and the requirements of standalone VR deployment. No placement approach is given access to additional assets or visual effects. Assets were reused across scenes, with differences arising only from their spatial arrangement, orientation, and relationships. Keeping the asset set fixed reduces the influence of visual richness and keeps spatial composition as the main experimental variable.

C. Scene Structuring

A narrative, N, can be represented as the following ordered sequence of scenes

$$N = \{s_1, s_2, \dots s_t \dots s_T\}, \tag{1}$$

where each scene s_t corresponds to a distinct segment of the story. For a given scene, the system operates over a fixed asset pool A, consisting of arbitrary assets (a_k) and is identical across all experimental conditions, as follows:

$$A = \{a_1, a_2, \dots a_k \dots a_K\}, \tag{2}$$

The outcome of scene composition is a spatial layout

$$L_t = \{(a_i, \mathbf{p}_i, \theta_i) \mid a_i \in A_t \subseteq A\}, \tag{3}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ denotes the 3D position vector of asset a_i, and $\theta_i \in [0, 2\pi)$ represents its orientation around the vertical axis. Differences between placement approaches arise from how these layouts are produced.

The placement process is defined as mapping a scene description and asset pool to a spatial layout:

$$f : (s_t, A) \rightarrow L_t, \tag{4}$$

In this study, all placement approaches implement the same mapping interface f, but differ in the information used to guide it as: random placement ignores narrative semantics, heuristic-based placement relies on predefined spatial rules, and the LLM placement incorporates narrative and affective cues extracted from the text.

D. Object Placement

Three placement approaches are implemented within the same framework, each operating on identical narrative input and spatial limits but differing in how placement decisions are produced.

1) Constrained Random Placement: The random baseline is intentionally simple, while still enforcing spatial constraints. Object locations are generated without regard to narrative meaning or emotional tone. Basic checks are applied to avoid collisions and invalid placements. This ensures that scenes remain navigable and visually acceptable, despite lacking semantic or emotion-aware structure. This condition serves as a baseline reference for evaluation. The underlying logic for this approach is detailed in Figure 2.

Algorithm 1 Random Placement Strategy

Require: Asset pool A, valid spatial regions R
Ensure: Scene layout L
1: **for** each asset $a \in A$ **do**
2: Random position $p \sim R$
3: Random orientation $\theta \sim [0, 2\pi)$
4: Add (a, p, θ) to L
5: **end for**
6: **return** L

Figure 2. Pseudo-code for the random placement strategy.

2) Heuristic-based Placement: In the heuristic condition, object placement follows a fixed set of designer-defined rules specified in a JSON configuration file. These rules determine which assets may appear in each scene and constrain their allowable regions, orientations, and basic spatial relationships. The rules reflect common assumptions about plausible object

arrangements but are not sensitive to narrative emotion. As a result, the layouts are consistent and easy to interpret, but they do not change in response to the emotional tone of the story. This limitation is clearly illustrated in Figure 5, where the heuristic agent correctly orients towards the target but fails to exhibit the bodily expression required for the 'Surprised' emotion. In practice, the heuristics are implemented through spatial constraints, such as raycast-based surface checks, predefined anchor regions, and simple orientation rules. The execution flow of these rules is outlined in Figure 3.

Algorithm 2 Heuristic-Based Placement Strategy

Require: Asset pool A, rule set $\mathcal{H}$
Ensure: Scene layout L
1: **for** each asset $a \in A$ **do**
2: Retrieve rule $h \in \mathcal{H}$ corresponding to asset type of a
3: Determine position $p \leftarrow h_{\text{pos}}(a)$
4: Determine orientation $\theta \leftarrow h_{\text{ori}}(a)$
5: Add (a, p, θ) to L
6: **end for**
7: **return** L

Figure 3. Pseudo-code for the heuristic-based placement strategy.

3) LLM-Informed Placement: In the LLM-informed condition, the narrative text is interpreted by Google's Gemini 3 Pro model. We selected this model for its advanced reasoning capabilities in spatial context understanding and multimodal processing [16]. The model was accessed via API with a temperature setting of 0.7 to balance structural adherence with creative interpretation. It produces a structured description used by the Unity pipeline. This description encodes relational cues, such as relative distance, spatial priority, and orientation (e.g., near a window, facing the viewer), which are then translated into concrete 3D placements by the engine. Unlike the other placement approaches, this condition allows cues from the narrative to influence spatial decisions. These cues can affect object proximity, character orientation, facial expressions, and gaze direction, including how elements are oriented relative to the viewer. To maintain a clear separation between conditions, affect-driven adjustments are applied only in the LLM-informed placement strategy. For example, the structural adjustments made to visually emphasize the antagonist in the 'Fear' scenario are demonstrated in Figure 6. The LLM pipeline relies on a zero-shot prompting strategy where the scene boundaries and object relations are parsed into json objects. These objects dictate the exact spatial parameters (e.g., proxemics, gaze vectors) before the Unity engine renders the final layout. The integration of this semantic parsing into the pipeline is summarized in Figure 4.

E. Spatial Validation

All scenes undergo the same validation checks before rendering, regardless of placement strategy. These checks include collision detection, surface support verification, and spacing constraints. Applying the same validation across conditions prevents physically implausible layouts while still allowing differences in spatial composition. Validation is implemented using Unity's PhysX engine through collider intersection checks and surface support tests.

Algorithm 3 LLM-Informed Placement Strategy

Require: Narrative text N, asset pool A
Ensure: Relational placement specification R
1: Extract spatial relations $S \leftarrow \text{Analyze}(N)$
2: **for** each relation $r \in S$ **do**
3: Infer spatial parameters (distance, orientation, salience) for r
4: Associate parameters with relevant assets in A
5: **end for**
6: **return** R

Figure 4. Pseudo-code for the LLM-Informed placement strategy.

IV. Evaluation

A. Hardware and Software

The study was conducted using the HTC Vive Focus Vision headset. All scenes were rendered in real time using Unity 6, and the same application build was used across all experiments. Audio was delivered through the headset's built-in speakers.

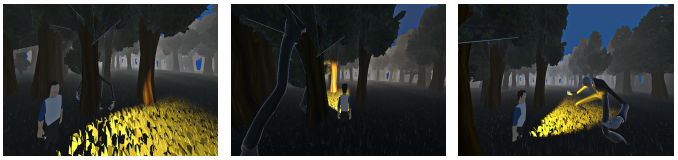
Interaction was limited to natural head movement, keeping attention on the narrative and the surrounding scene rather than on controls. All sessions were conducted in the same physical environment, with room layout, ambient lighting, and verbal instructions kept consistent to minimize external variation. To capture granular attention metrics, we leveraged the HTC Vive Focus Vision's built-in eye-tracking capabilities via the OpenXR interface. Gaze origin and direction vectors were accessed in real-time, synchronized with the application's update loop. We implemented a custom raycasting system that continuously mapped these vectors to the 3D scene geometry, allowing the system to log specific object fixations, gaze duration, and scan paths with high precision throughout the narrative experience [17].

B. Experimental Design

Participants joined the experiment voluntarily and were recruited informally. No specific experience with VR, games, or interactive storytelling was required. Each participant completed a single session in which the same narrative was presented three times under different scene configurations. The narrative text, asset pool, and runtime constraints remained unchanged, so differences between scene versions were limited to object placement and spatial relationships. Scene configurations were labeled neutrally (Scene A, B, and C), and participants were not informed of the underlying placement strategies to avoid expectation bias. The order of the three scene versions was random between participants to reduce sequence effects.

(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 5. 'Surprised' scenario (Diorama 1). (a) **Random-Based:** Character gazes at an arbitrary point, lacking context. (b) **Heuristic-Based:** Character correctly looks at the target (bird) based on rules, but the body orientation fails to convey the emotional state to the viewer. (c) **LLM-Based:** Character maintains gaze on the target while orienting the body towards the camera, effectively displaying the 'Surprised' emotion and maximizing user immersion.



(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 6. 'Fear' scenario (Diorama 2). (a) **Random-Based:** Character gazes aimlessly into the dark forest, failing to acknowledge the nearby antagonist. (b) **Heuristic-Based:** Character adheres to the gaze rule by facing the target coordinates, but the composition fails due to occlusion, leaving the antagonist visually obstructed by a tree. (c) **LLM-Based:** The scene is semantically restructured to reveal the antagonist clearly, while the character's body orientation and gaze align to vividly portray the 'Fear' state to the viewer.

(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 7. 'Sadness' scenario (Diorama 3). (a) **Random-Based:** Character stands in the background facing away from the focal point (photograph), completely missing the narrative beat. (b) **Heuristic-Based:** Character satisfies the gaze constraint by facing the target, but the substantial distance and rigid posture fail to evoke the intended intimate atmosphere. (c) **LLM-Based:** The agent demonstrates semantic understanding of proxemics by positioning the character intimately close to the memento and adjusting the head tilt downward, effectively embodying the emotion of mourning through non-verbal cues.

(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 8. 'Happiness' scenario (Diorama 4). (a) **Random-Based:** Spatial layout is disjointed; the secondary character (dog) is obscured by foliage in the background, severing the narrative link. (b) **Heuristic-Based:** The agent aligns orientation to the target, but the excessive physical distance results in a detached observation rather than a shared emotional experience. (c) **LLM-Based:** The model interprets 'Happiness' as active companionship, significantly reducing the spatial distance to foster a sense of intimacy and interaction between the character and the animal.

Each session began with a brief verbal introduction. Participants were informed that they would experience a VR-based story and later answer questions about their impressions and observations. Each participant was assigned one of four narratives, each centered on a single dominant emotional tone. The selected story was presented three times, once for each scene configuration, with short pauses between viewings to allow participants to rest without disrupting the overall narrative context.

After each scene viewing, participants completed a short questionnaire assessing narrative coherence, emotional alignment, and presence for the scene they had just experienced. Narrative coherence was evaluated through questions focusing on the relationship between object placement and story understanding, specifically how logical the arrangement of objects felt within the context of the story and whether placement helped direct attention to important elements. Presence was assessed by asking participants whether they felt physically present in the environment, became engrossed in the virtual world, and felt like part of the story rather than an external observer. Emotional alignment was measured by asking participants to identify the dominant emotion they felt in the scene and to rate how strongly that emotion was conveyed. After all three scene versions had been viewed, participants completed an additional comparative questionnaire. These questions asked participants to directly select which scene best conveyed the story's emotion, supported narrative understanding, guided visual attention most naturally, elicited the strongest sense of immersion, and which version they would choose to show to another person.

C. Measures and Data Collection

A total of 20 participants (16 male, 4 female) ranging in age from 21 to 45 participated in the study. The cohort possessed diverse educational backgrounds, predominantly holding or pursuing undergraduate (50%) and graduate (45%) degrees. The majority of the participants (90%) reported having beginner-level experience with VR technologies, while 10% possessed intermediate familiarity. Participant background information also included field of study (e.g., Computer Engineering, Electronics) to represent a wider range of user perspectives.

Responses were collected using a combination of short-answer prompts and 7-point Likert-scale ratings. For constructs measured using multiple Likert-scale items, responses were averaged to form composite scores for analysis.

V. Results

A. Results on Narrative Perception

To address RQ1, we examined how different object placement strategies influenced participants' perceptions of narrative coherence, emotional alignment, and presence. A one-way Analysis of Variance (ANOVA) was conducted to evaluate the statistical significance of the perception scores. As shown in Table I, the *LLM-based* placement strategy consistently received higher ratings in semantic categories. It achieved significantly higher perceived narrative coherence ($\mu = 5.7, \sigma = 0.9$) and stronger emotional alignment ($\mu = 5.4, \sigma = 0.9$) compared to the heuristic-based and random conditions ($p < 0.001$).

Interestingly, the sense of presence was comparable between the *Heuristic* ($\mu = 5.1, \sigma = 0.8$) and *LLM-based*

($\mu = 5.0, \sigma = 0.8$) conditions ($p > 0.05$). This suggests that while rule-based layouts can achieve physical plausibility, semantic reasoning is essential for conveying the narrative's emotional context and improving the storytelling capability of the scene.

TABLE I
PERCEPTION SCORES ACROSS PLACEMENT APPROACHES (N=20).

Metric	Random	Heuristic	LLM-based
Narrative Coherence	3.0 ± 1.7	5.1 ± 1.3	$\mathbf{5.7 \pm 0.9}$
Emotional Alignment	3.2 ± 1.5	4.4 ± 1.3	$\mathbf{5.4 \pm 0.9}$
Sense of Presence	4.4 ± 1.1	$\mathbf{5.1 \pm 0.8}$	5.0 ± 0.8

B. Results on Visual Attention

To address RQ2, we examined how object placement strategies influenced visual attention during scene viewing. As shown in Table II, layouts generated using the LLM-informed strategy consistently guided attention more effectively than the other conditions. Participants oriented to relevant objects more quickly in the LLM-based scenes. This is reflected in a significantly shorter *Time to First Fixation* ($\mu = 2.1\,\text{s}, \sigma = 0.7$) compared to the heuristic-based ($\mu = 3.6\,\text{s}, \sigma = 1.0$) and random layouts ($\mu = 4.8\,\text{s}, \sigma = 1.2$). Once fixated, participants also spent more time attending to key story elements in the LLM-informed condition, exhibiting longer *Dwell Times* ($\mu = 5.4\,\text{s}$) than in the other two conditions.

TABLE II
GAZE-BASED ATTENTION RESULTS ACROSS PLACEMENT APPROACHES.

Metric	Random	Heuristic	LLM-based
Time to First Fixation (s)	4.8 ± 1.2	3.6 ± 1.0	$\mathbf{2.1 \pm 0.7}$
Dwell Time (s)	2.3 ± 0.9	3.1 ± 1.1	$\mathbf{5.4 \pm 1.3}$

The differences in both Time to First Fixation and Dwell Time across the three conditions were found to be statistically significant ($p < 0.01$) using a one-way ANOVA.

C. Results on Physical Plausibility

We analyzed the physical plausibility metrics summarized in Table III. The *Random* baseline exhibited the highest instability, with an 11.2% initial collision rate and 7 visible artifacts in the final scenes, demonstrating the necessity of constraints. The *Heuristic* approach remained the most stable (2.1% collision) due to rigid rules.

The *LLM-based* strategy showed a moderate initial collision rate (3.8%), primarily driven by the model's semantic attempts to create intimate object proximity. However, the Spatial Validation layer effectively mitigated these risks. Consequently, the LLM-based approach achieved a highly plausible final result with minimal artifacts (2 instances), significantly outperforming the Random baseline and approaching the stability of handcrafted heuristics.

Since the physical plausibility metrics primarily consist of frequency counts, a Chi-square test was utilized, confirming that the reduction in invalid placements compared to the random baseline was statistically significant ($p < 0.05$).

TABLE III
PHYSICAL VALIDITY AND SYSTEM PERFORMANCE METRICS.

Metric	Random	Heuristic	LLM-based
Initial Collision Rate (%)	11.2	2.1	3.8
Validation Rejection Rate (%)	14.5	1.5	2.4
Avg. Generation Retries	1.8	0.2	0.8
Final Invalid Placements (Count)	**7**	**1**	**2**

*Refers to minor artifacts visible across all 20 experimental sessions.

D. Qualitative Results

Participants were asked to select the diorama version they felt best represented the story. As summarized in Table IV, the majority of participants (60%) explicitly preferred the *LLM-based* layouts. Qualitative feedback indicated that users found these scenes "more alive" and "narratively accurate," particularly praising the meaningful interactions between characters, such as the intimate proximity to the dog in the 'Happiness' scenario in Figure 8 or the mourning posture in the 'Sadness' scenario in Figure 7.

The *Heuristic* approach was preferred by 35% of users, primarily for its "clean and organized" structure, though some described it as "emotionally distant." The *Random* baseline was largely rejected (5%), with participants citing that the chaotic placement often "broke the immersion" and made the story difficult to follow. These preference rates align with the quantitative gains in coherence and emotional alignment reported in RQ1, confirming that users prioritize semantic depth over simple geometric order.

TABLE IV
USER PREFERENCES ACROSS PLACEMENT APPROACHES (N=20).

Preferred Scene Version	Percentage (%)
Random	5.0
Heuristic	35.0
LLM-based	**60.0**

In summary, the qualitative feedback and user preferences strongly corroborate the quantitative gaze and perception metrics, confirming that semantic layout reasoning significantly enhances the immersive storytelling experience.

VI. CONCLUSION AND FUTURE WORK

This study presented a framework that reimagines spatial layout not merely as a geometric puzzle, but as a narrative medium. By incorporating LLMs into the VR pipeline, we explored how spatial composition can move beyond static rules to reflect emotional context—translating abstract sentiments like "happiness" or "fear" into concrete proximity and orientation adjustments.

Our evaluation suggests that semantically informed layouts offer a tangible advantage in narrative coherence and emotional alignment over traditional baselines. Although the generative approach introduced high initial collision rates due to its ambitious placement strategies, our findings confirm that a validation layer can effectively mitigate these risks, balancing semantic expressiveness with physical plausibility.

For future work, we plan to integrate procedural mesh deformation and inverse kinematics to resolve physical conflicts dynamically, ensuring that intimate character interactions remain both semantically powerful and geometrically seamless without relying on expensive regeneration cycles. The evaluations will involve a larger and more diverse participant sample to further validate these findings.

REFERENCES

[1] M. Billinghurst, A. Clark, and G. Lee, "A survey of augmented reality," *Foundations and Trends in Human-Computer Interaction*, vol. 8, no. 2–3, pp. 73–272, 2015.

[2] M. Cavazza, F. Charles, and S. J. Mead, "Character-based interactive storytelling," *IEEE Intelligent Systems*, vol. 17, no. 4, pp. 17–24, 2002.

[3] A. Abdelreheem *et al.*, "PlaceIt3D: Language-guided object placement in real 3D scenes," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2025.

[4] B. M. Öcal, M. Tatarchenko, S. Karaoglu, and T. Gevers, "SceneTeller: Language-to-3D scene generation," in *Proc. European Conference on Computer Vision (ECCV)*, 2024, pp. 362–378.

[5] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.

[6] Y. Zhou, I. Nuriddinov, A. E. Rhesa, W.-T. Lo, and T.-Y. Li, "RL-LABEL: Reinforcement learning-based label placement for dynamic AR scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 2540–2552, 2023.

[7] M. Fiala, "ARTag: A fiducial marker system using digital techniques," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 590–596.

[8] C. Lee and A. Varshney, "Human-centric label placement in augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 11, pp. 3660–3675, 2022.

[9] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5410–5419.

[10] U. Parihar *et al.*, "MonoPlace3D: Learning 3d-aware object placement for 3d monocular detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[11] W. Feng *et al.*, "LayoutGPT: Compositional visual planning and generation with large language models," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] C. Wang *et al.*, "Chat2Layout: Interactive 3D furniture layout with multimodal large language models," *arXiv preprint arXiv:2407.21333*, 2024.

[13] S. Peng *et al.*, "OpenScene: 3D scene understanding with open vocabularies," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 815–824.

[14] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[15] C. Chen *et al.*, "UniEmoX: Cross-modal semantic-guided large-scale pretraining for universal scene emotion perception," *IEEE Transactions on Image Processing*, 2025.

[16] G. Team and G. DeepMind, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[17] M. Kuyucu *et al.*, "Emotion recognition in virtual reality using sensor fusion with eye tracking," *Computers in Biology and Medicine*, vol. 197, p. 111070, 2025.

It Could Literally Change My Life: Exploring the Potential of Conversational Interaction for Indoor Wayfinding Among People with Visual Impairments

Segun J. Samuel, Mohammad Adnaan, Jeremy R. Cooperstock
Department of Electrical and Computer Engineering
McGill University
Montreal, Canada
e-mail: {segun.samuel | mohammad.adnaan}@mail.mcgill.ca, jer@cim.mcgill.ca

Ahmed Farooq
School of Information Sciences
Tampere University
Tampere, Finland
e-mail: {ahmed.farooq}@tuni.fi

***Abstract*—Navigating unfamiliar and complex indoor environments independently is challenging for People with visual Impairments (PVI). As a result, PVI rely on a variety of Assistive Technologies (AT) to preplan and execute their journeys indoors. Among these AT are the emerging Conversation Interaction (COI) systems that leverage Large Language Models (LLMs) to deliver engaging experiences. Specifically, for indoor location assistance, there is an increasing focus on the remote exploration of Points of Interest (POIs), such as shopping malls, before planning a visit. However, there are few studies on the use of COI systems for indoor wayfinding and related aspects, such as route rehearsals—a process of learning the required sequence of actions from one location to another. Consequently, it is not clear what kind of spatial information PVI would expect from conversational interaction agents, how they would use these aids, and the challenges that might arise from their use. To explore the potential of conversational interaction as a modality for indoor wayfinding assistance for PVI, we developed 'GeoChatre," an interactive mobile COI app that enables contextual learning of unfamiliar indoor routes. Our study reveals the specific types of spatial information that PVI expect from conversational agents for indoor wayfinding, including step-based distance estimates, directional guidance, landmarks, obstacle awareness, and shows that contextual, progressive disclosure of information supports route recall. Additionally, our results show that voice-based interactions in shared indoor spaces raise privacy concerns, particularly on destination disclosure, highlighting the need for coded interactions and discreet input alternatives.**

Keywords-AI and Accessibility; Assistive Technologies; Navigation; Conversational Interaction; Visual Impairments; Privacy.

I. INTRODUCTION

Buildings with complex layouts present significant navigation barriers for People with Visual Impairments (PVI), particularly when visiting for the first time [1][2]. While a range of Assistive Technologies (AT) have been developed to support indoor mobility [3], the integration of Large Language Models (LLMs) into COI systems (e.g., ChatGPT) has opened a new avenue for delivering spatial information through natural dialogue [4][5], since LLMs provide a more accessible knowledge source than conventional Question and Answering (Q&A) systems [6].

For example, there is an increasing use of COI for indoor wayfinding assistance to support exploration of points of interest [7], such as in shopping malls [8], to provide users with information based on personal interests before making travel decisions. Conversational interaction is also being used to select or specify destinations [9]–[12], or for human localization in indoor environments through intelligent conversation between users and an agent [13]. While some studies have investigated COI for indoor wayfinding [13]–[16], there is limited work targeting indoor route rehearsals for PVI—the process of learning the required sequence of actions to reach destinations or return to their origin. Instead, blind individuals often acquire such route knowledge through the use of tactile maps [17] or with the assistance of Orientation and Mobility (O&M) specialists, especially when changing environments [18].

However, tactile maps assume tactile literacy on the part of the users and lack the capacity to provide detailed information [19], while O&M experts might not always be available. To overcome these limitations, this research explores the use of COI agents as wayfinding assistants to provide users with advance knowledge of POIs and facilitate their learning of indoor routes [5]. Our research considers the following questions:

RQ 1: What kind of spatial information do PVI expect from conversational interaction for indoor route learning?

RQ 2: How would PVI like to use conversational interaction for indoor route learning?

RQ 3: What challenges/reservations might PVI have with the use of conversational interaction for indoor wayfinding assistance?

The contributions of our study are twofold: First, we identify the specific types of spatial information that PVI expect from a conversational agent for indoor wayfinding assistance, including step-based distance estimates, directional guidance, landmarks, and obstacle awareness. Second, we uncover that voice-based interactions in public indoor spaces raise privacy

concerns that might have often been overlooked in AT designs, with participants distinguishing between socially neutral and sensitive locations, thereby contributing new insights into the social dynamics of using COI systems in shared indoor environments.

The rest of the paper is structured as follows: Section II provides an overview of related work, followed by a brief explanation of system design in Section III. User recruitment, demographics, and study procedures are discussed in Section IV. Sections V and VI present the results and the implications of findings, respectively, while the limitations of the study are highlighted in Section VII. Section VIII concludes the paper, summarizes the central findings, and outlines directions for future work.

II. Related Work

Independent wayfinding in unfamiliar environments is challenging for most people, especially for PVI [11]. Accordingly, AT for navigation usually provide guidance or facilitate knowledge of the surroundings for PVI [20][21]. Previous research has shown the potential of traditional tactile maps to convey spatial environmental knowledge [22]. There are also digital interactive maps that provide dynamic spatial information and auditory feedback for spatial learning [17][23][24], but they impose hardware requirements [20] and are limited in their adaptability to different user needs or contexts without major modifications [17]. Other attempted solutions use commodity smartphones to provide virtual navigation via a sequence of turn-by-turn instructions, and render relevant POI information [11][12][20][25], but as these are passive systems, they do not support interactive Q&A as might be needed by PVI [8][17].

Since natural language dialogue is often the preferred interaction method by PVI [26], more AT are integrating COI agents to improve user experience [7][8][27][28]. Research on COI systems, particularly for indoor location assistance, has primarily focused on exploration of POIs. Such approaches allow remotely "probing" POIs for information based on individual interests [8][27][29]. There is also research on general scene understanding, mostly for outdoor settings [30], and on obstacle avoidance [31]. However, to the knowledge of the authors, only a few studies have been carried out on COI for wayfinding [8][13], and related aspects, such as preplanning indoor routes. Addressing these gaps in the literature is of value to PVI, since acquiring cognitive maps of both outdoor and indoor environments, before independent navigation, is crucial [5][17]. Moreover, there is no guidance available as to the kind of spatial information that PVI would expect from conversational interaction agents for indoor wayfinding, how such agents would be used, and the challenges that might arise in their use.

III. System Design of GeoChatre

To address the gaps in the literature identified above, we developed "GeoChatre," an interactive mobile app that enables contextual learning of unfamiliar routes in complex indoor environments, following the principles recommended for designing conversation agents for navigation [32]. GeoChatre is intended for a number of situations: 1) where PVI have a specific destination in mind for independent navigation, 2) scenarios in which users are temporarily disoriented, e.g., in complex building environments, and seek to regain orientation, and 3) where individuals simply want to know what POIs exist along a given route, either for the sake of their spatial awareness or to plan journeys there.

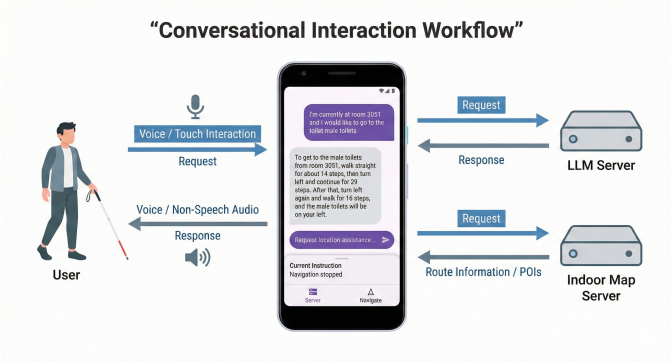


Figure 1. Conversational interaction architecture: GeoChatre.

The GeoChatre system, as shown in Figure 1, was implemented on Android, and tested on a Google Pixel 9 Pro smartphone. The app uses the "Take Me Out of Error" (TACME) indoor localization framework based on pedestrian dead reckoning that we developed as part of a larger study. TACME runs on a server built on indoor maps from floor plans to render routes and POIs information. The architecture comprises three principal components: a client-side, an LLM server, and an indoor map server. Communication between these components follows a request-response pattern over the REST Application Programming Interfaces (APIs).

A. Input Processing and User Intent Classification

Users interact with GeoChatre through voice or text input. For voice interaction, users double-tap the screen, triggering distinct audio cues (speech and non-speech) to signal the start and end of voice registration. Voice input is captured using Android's built-in SpeechRecognizer. The system also supports textual input. Once a verbal request is transcribed, it is sent to the LLM server, which runs Llama 3.2 3B. The LLM is prompted to perform intent classification to determine whether the input represents a new route request, a follow-up inquiry, or an ambiguous query requiring clarification. For new route requests, the mobile application sends the extracted source and destination to the indoor map server via a REST API call. The server provides spatial data, including relevant safety cues for each route segment.

B. Data Retrieval and Response Generation

The route data retrieved from the map server is combined with the user request and passed to the LLM. The LLM, by prompts, synthesizes route information into a natural-language response tailored for PVI. For initial direction requests, the

LLM generates concise summaries by combining consecutive similar actions (e.g., aggregating multiple short segments into a concise instruction such as *Walk straight for 15 steps, then turn right at the water fountain and walk for an additional 6 steps. Main restroom will be on your left*. Users can request more detailed information through follow-up queries. Following this approach to information rendering, our information design can be situated within some established theoretical frameworks: the Landmark-Route-Survey model of spatial knowledge acquisition [33][34] and the information-seeking mantra of "overview first, zoom and filter, then details on demand" [35]. In other words, when a user first makes a direction request, GeoChatre delivers a concise route summary that aggregates consecutive navigation actions into an overview. Users can then progressively access richer spatial information through follow-up questions—effectively achieving details on demand through natural dialogue. The generated response is delivered to the user through Android's text-to-speech engine, with audio feedback beeps providing interaction cues for different user request and response states. The LLM is also customized through prompting to provide multilingual indoor navigation assistance required by PVI.

IV. User Study

We secured the approval of the McGill University Research Ethics Board (REB #19-10-041) for our study. Seven participants, ($N = 7$; 2 female, 5 male, mean age of $M = 41.9$ years ($SD = 10.2$) took part in the experiment, and were compensated CA $15 per hour for their time. One completely blind person participated in the study. The remaining six participants had corrected-to-normal vision. The study was conducted on the third floor of a multi-story university building. The indoor map server is developed from the floor plan. The map indexed 28 POIs across the navigable area, organized into categories, including classrooms, offices, and amenities such as a printing station, a lounge area, study areas, emergency exits, and vertical circulation points. The building features an L-shaped main corridor spanning approximately 200 meters, with multiple perpendicular branches extending to classrooms, offices, and amenities.

We collected demographic information from the subjects, including their vision level, time of blindness onset, use of wayfinding aids, and whether they have any hearing impairments. A short training session was conducted to familiarize them with the control settings for the app, such as for the loudness and speed of text-to-speech. Also, during this phase, participants received names of different indoor locations and learned to request direction guidance from their current position. They were also informed that they can ask for additional details, such as nearby landmarks, estimated walking time, safety cues, and any other information they deemed helpful for constructing mental representations. In the experiment, subjects were requested to choose from indoor locations that they would visit and use the COI tool to explore and familiarize themselves with the route information prior to undertaking the journey. The goal was to explore how COI systems would support the development of spatial knowledge for real-world navigation. After the participants carried out these tasks, they completed a post-questionnaire survey. We also conducted semi-structured interviews to gain deeper insights into the experiences of the participants. The sessions were audio recorded, and each study session lasted approximately 60 minutes.

V. Results

A. RQ1: What kind of spatial information do PVI expect from conversational interaction for indoor route learning?

The analysis reveals that subjects expected a rich set of spatial information from GeoChatre. Their expectations clustered around four distinct categories of spatial information, each serving a complementary role for independent wayfinding indoors.

1) Step-count Information: Step-based distance estimation provided by GeoChatre emerged as one of the most widely discussed types of spatial information. Step counts were valued as actionable cues that support self-location awareness. As some participants explained: *It is able to calculate the distance between where I am and where I am going to by steps. That means I am able to count my steps and get to where I am going to.* (P001); *... it can also tell me how many steps to take to get there ...* (P002). Some expressed that the step information can help to determine how much effort is required to complete wayfinding tasks. For example, *it can also help me to know how much effort it will take me to get there ...* (P002). Importantly, participants demonstrated an understanding of the limitations of step-count cues. One acknowledged that *the number of steps might not be accurate, since the number of paces for each person may be different.* (P002). However, they still valued it as *a rough idea of what the distance is.* (P007). This suggests that PVI may not require absolute precision, but rather a reliable distance estimate that supports spatial knowledge of the journey ahead.

2) Landmark Identification and Sequential Ordering: Landmarks emerged as prominent spatial features expected with their sequential presentation along the routes. There were several expressions, such as, *it is able to tell me the different landmarks I will encounter on my path... the water fountain, study area, the fire extinguisher, the elevator, then the female toilet..., the female restroom before...* (P001). Some highlighted the value of the order of information presented: *the aspect I find interesting is the sequence of landmarks starting from where you are leaving to... the way it arranged it sequentially... it is very interesting.* (P004). It appears the sequential ordering of landmarks creates a narrative structure of the journey that supports the construction of mental maps, just as expressed by one participant that *... being able to map the indoor area is quite interesting.* (P006)·

3) Turn-by-Turn Directional Information: Participants valued explicit Turn-by-Turn instructions (TbT) as directional cues. Most of them successfully recalled the routes. For example, a participant recounted one of the route instructions saying, *It said from here, then I go... take a right 14 steps,*

take left 29 steps, left again 16 steps, and my destination will be at the left. (P001). This demonstrates the degree to which these instructions were internalized and retained. Another account reveals how TbT and landmarks work in concert: *when they say you go straight, you turn left and you continue a bit forward and I think they gave indications also you're going to see... bins, you're going to see fountain and you're going to pass one of the labs.* (P007). The integration of directional, landmark, and metric information constitutes the comprehensive spatial information expected by PVI.

4) Obstacle Awareness and Safety Cues: Obstacle and safety information emerged as an important expectation, particularly regarding the dynamic nature of indoor spaces. Participants valued information about obstacles and how to avoid them. One expressed satisfaction for being told to *be careful of the obstacles on the way* (P002) while some raised concerns about how the system would *adapt quickly to understanding when...* objects are moved around. This suggests that PVI expect spatial information from COI agents to reflect the current state of the environment.

B. RQ2: How would people with visual impairments like to use conversational interaction for indoor route learning?

1) Journey Preplanning: A striking finding is the strong emphasis participants placed on using the COI app before undertaking a journey. Being able to plan before traveling will enable PVI to build cognitive maps of the environment. This was clearly articulated by one of the participants who said: *It will help before one embarks on a journey, you can have an idea of the path, and how to get there.* (P002). This remote environmental awareness capability with conversational agents extends to exploration of POIs and querying for their existence in an interactive manner: *even without standing up from my seat, I can know if a particular object or location exists, and if it exists, it can also help me to know how much effort to take me to get there* (P002). Some of the participants expressed satisfaction with knowing what to expect by simply interacting with the COI agent, much more like with human assistance: *knowing what to expect when going... before going to a place* (P003). They also commented on the advantage of being aware of certain information from COI aids that would not have otherwise been available: *Even for people..., who can see, you may not notice some of the information it has given you, you may not even have noticed them or even pay attention to them.* (P001). The COI agent offers a fundamentally different experience from conventional prejourney tools, which lack adaptability and flexibility to user requests [8].

2) Duality Modes, Active-Journey Error Correction and Integrated Navigation: Subjects articulated a clear desire for the COI agent to serve both planning and real-time goals, and critically, to integrate with existing mobility aids. Participants (P002, P003, P005, and P007) expressed satisfaction with the dual role—the possibility of using the app for remote planning and in-situ navigation: *It can give information beforehand and when you are in the process of carrying out the action.* (P005). A real-time use case could even be for when individuals at certain indoor locations (e.g., at the reception) wanting to find directions to a destination of interest. They would simply interact with the agent in such situations and receive similar assistance as that from humans: *I have gone to a number of places where you don't know where certain things are, where certain things are arranged, or even where the restroom is; you will be asking people around, but with this, it will guide you.* (P001).

Some individuals specifically noted the value of the app for getting directions: *It is also good for a shopping mall; you want to shop and you are looking for the directions ...* (P004); *It would be nice in navigating new places, new environment like malls, indoor marketplaces, new spaces... you have never been before, you are looking for a particular location inside that building.* (P006) and *I like to use this application in a complex environment.* (P005). They also want to use the COI app for wayfinding error corrections as part of real-time navigation support. This reflects a practical case when people are lost or temporarily disoriented in buildings and would like to regain their orientation: *I also find the part that you enter a wrong direction and it guides you back... so it doesn't waste your time.* (P001).

Most of the subjects explained the need to integrate the COI app with existing mobility systems: *it can not be standalone. It will need other mobility aids. Integrating this with a mobility aid will make navigation seamless.* (P002). One participant said *I will use this to plan and then use the mobility aid for the real-time movement.* (P003). Another PVI further explained that *if you can combine it with..., it would just be like okay, it gives you an overview of the route first, and you start walking, and it kind of updates you as you work. So I think it would be very good.* (P007). All these suggest that PVI see the conversational agent as a builder of a cognitive map that should complement other mobility tools.

3) Diverse Indoor Environments and Application Scenarios: All seven participants discussed other potential use cases, indicating a remarkably broad applicability. Although the most consistently recurring scenario was navigating unfamiliar environments (P001, P002, P003, P005, P006, P007), participants also identified specific indoor venues including shopping malls, subway stations (P004, P006, P007), workspaces (P006), university campuses (P007), or finding location of washrooms, cardio stations, cafetaria, gyms and fitness centres (P001, P007) within buildings. Based on their prior working experience, one participant offered a particularly "rare" application in maintenance and inventory management for finding missing items: *I have seen where some maintenance organizations are unable to use some spare parts because they are unable to find their location, and because they are unable to find them, they won't be able to use them until those spare parts are expired.* (P004). Some also attested to the value of the COI agent in familiar environments: *even places where you have been, but you need direction..., to the same place.* (P006) and even for the sighted: *That information will help both somebody with visual impairments and somebody who can really see.* (P001).

C. RQ3: What challenges/reservations might PVI have with the use of conversational interaction for indoor wayfinding assistance?

Despite an overwhelmingly positive reception from participants, particularly on the perceived usefulness and enjoyment of COI for indoor navigation experience as shown in Figure 2, participants identified several key issues that could impede the effective use of COI for indoor navigation.

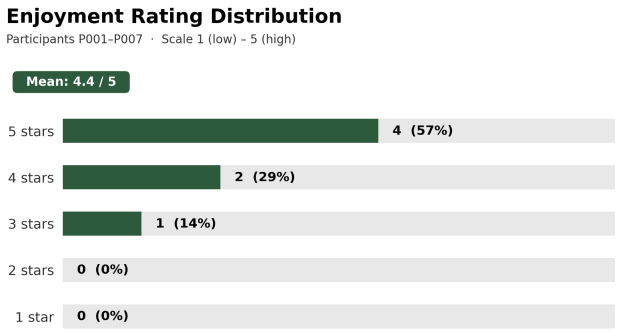


Figure 2. Participant enjoyment rating with conversational interaction: GeoChatre.

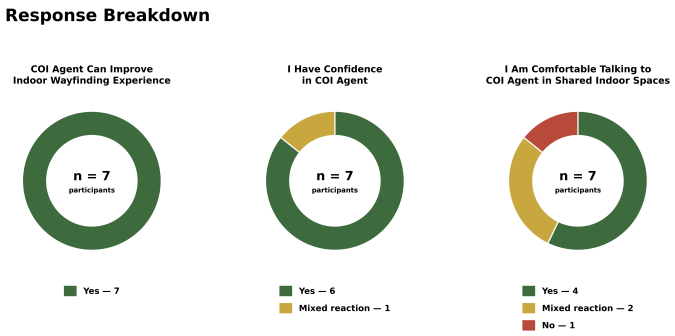


Figure 3. User perception of conversational wayfinding agent.

1) Accent and Speech Recognition Barriers: A few subjects had challenges with the speech recognition. This likely results from accents, dialectal variation, or low-resource language influences that affected their interaction with the app. One participant framed this as an equity issue and said *it should be accessible for all people irrespective of their accents.* (P005). There appears to be a sort of semantic substitution: *It's misplacing some words for other words, like, restroom for restaurants.* (P005). This finding carries significant weight, given that visual impairments may necessitate voice-based interactions. As depicted in Figure 3, the "accent-bias" elicited mixed reactions, affecting user confidence in the agent.

2) Privacy and Social Concerns in Indoor Public Spaces: A nuanced challenge emerged around the social dynamics of speaking to a COI agent in public. While several participants were generally comfortable interacting with GeoChatre, some provided insights into the situational nature during such interactions, distinguishing between socially neutral destinations: *I don't mind if people hear that I need to go to the cafeteria...* and socially sensitive ones: *But ... stuff like the washroom or even like, yeah, like even I need the bin, well, nobody needs to know that I'm going to throw my dirty t-shirt in the bin.* (P007). Some proposed solutions, such as coded destination options: *it would be nice to just use your phone and be able to click. Or option one, washroom; ...* (P007, P006) to allow navigation requests without explicitly disclosing destinations. In other words, participants want COI systems to automatically prompt with number-coded options for destinations or support the customization of location labeling when users explicitly request such interactions. These findings reveal the desire and challenge to control the information that becomes publicly audible through voice-based interactions for wayfinding in shared indoor environments.

3) Noise and Environmental Interference: Some participants (P003, P007) raised concerns about performance in noisy environments. They expressed discomfort using the COI agent in large indoor spaces *because there could be noise, like, interference, and it might not be able to, like, pick it.* (P003). Another participant drawing on experience with voice assistants: *if you're like Siri or all the vocal assistants, they tend to glitch when there are many noises.* (P007). This challenge is particularly critical because the environments where PVI might most need wayfinding assistance—shopping malls, subway stations, university buildings—are often among the noisiest. In those situations, PVI might be required to speak louder, which is often undesirable: *if you yell out loud and there's a lot of people.* (P007) while planning navigation to socially sensitive areas (e.g., washrooms).

VI. Discussion

A. Multi-Layered Spatial Information Requirements

Our findings from this study suggest that COI systems for indoor wayfinding should deliver spatial information in an integrated manner, combining steps, landmarks, turn-by-turn directions, and obstacle warnings. By leveraging the reasoning capabilities of LLMs, this information should be delivered progressively. In one of the accounts from the participants, they like that GeoChatre *did not output all those details* (P006) at once but provided information based on contexts and as requested.

There is also a strong desire for conversational interaction agents to support both journey planning and in-situ navigation. Planning would allow users to remotely explore the environments, assess effort, and build mental models of routes before traveling. This knowledge can be further "rebuilt" at the sites (e.g., within the building) just before navigation, or even during wayfinding, to self-correct from disorientation. Additionally, the consistent framing of conversational agents as complementary aids by PVI means they should be designed and used alongside existing mobility tools.

B. Contextual and Progressive Information Delivery

Results revealed that participants value the COI agent for its mode of information delivery, which was both contextual—

tailored to the request—and progressive—structured from overview to detail. This finding aligns with prior research on the standard principles of presenting information—overview-first, and detail-on-demand [35][36]. Participants captured both dimensions of information presentation: *in the first go, it did not put all those details, which might not be necessary until if I have to ask, and when I want details, it gives me details* ... (P006); *when I asked for landmarks, it provided landmarks* ... (P007). This combined delivery mode appears to support route recall as subjects were able to recount spatial elements and describe the routes after they interacted with the agent.

C. Accessibility and Inclusivity Implications

For PVI who rely on voice as one of their primary input modalities, speech recognition failure is technically equivalent to an inaccessible interface. One explanation for this recognition issue lies in the composition of COI training data [37]. Although many AI models are trained on vast datasets, these often lack linguistic and cultural diversity because they are disproportionately composed of Western-centric data [16]. COI systems should therefore be trained on diverse accents from the outset, with strong consideration for potential influences from low-resource languages [37].

The findings around social comfort suggest that systems should incorporate privacy-aware features such as coded POIs. The distinction between socially neutral and sensitive locations should inform interaction designs that preserve the privacy of use in shared spaces. In addition, the recurrent feedback that COI can benefit sighted individuals as well as PVI promotes a universal design that could enlarge the user base and reduce stigma associated with such assistive technologies. In deed, participants expressed strong willingness to adopt the technology, with one stating *I would just use it all the time. Because it's super practical* ... and describing it as something that *could literally change my life.* (P007).

VII. Limitations

While this study provided insights into the expectations of PVI, desired use scenarios, and reservations regarding "conversational" indoor wayfinding, several limitations should be acknowledged. First, the work drew on interviews with seven participants. This might limit the generalization of findings to the broader community of PVI. Future studies should recruit a larger sample. In addition, the present study evaluates spatial knowledge acquisition rather than actual navigation performance. While our results demonstrate that participants can form cognitive maps of the indoor environment using COI, we did not measure how well this knowledge translates to successful real-world indoor wayfinding. This will be addressed in a planned follow-up study involving in-situ navigation tasks with PVI participants. Furthermore, analysis of results relies on self-reported data, so there might be differences between the reports of participants and their actual behaviour in real-world settings.

VIII. Conclusion and Future Work

Our study on conversational interaction for indoor wayfinding assistance shows that people with visual impairments have multi-dimensional expectations. PVI desire contextualized and multilayered spatial information and the protection of personal privacy when interacting with COI agents. They prefer using COI systems flexibly across planning and in-situ navigation, in diverse indoor environments, and as a complement to existing mobility aids. Generally, "conversational wayfinding" is perceived as promoting independence while offering cross-ability benefits. However, there remain significant problems, including "linguistic-bias", privacy concerns, and environmental interference in crowded environments. Future research should pursue privacy-preserving interaction designs and scalable deployment strategies to advance this emerging technology toward real-world impact for indoor navigation.

Acknowledgment

This research was supported by NSERC Discovery Grant. The authors thank the participants for their time and valuable insights.

References

[1] C. Engel et al., "Travelling more independently: A requirements analysis for accessible journeys to unknown buildings for people with visual impairments," in *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–11.

[2] W. Jeamwatthanachai, M. Wald, and G. Wills, "Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings," *British Journal of Visual Impairment*, vol. 37, no. 2, pp. 140–153, 2019.

[3] A. T. Parker et al., "Wayfinding tools for people with visual impairments in real-world settings: A literature review of recent studies," in *Frontiers in Education*, Frontiers Media SA, vol. 6, 2021, p. 723 816.

[4] X. Tang, A. Abdolrahmani, D. Gergle, and A. M. Piper, "Everyday uncertainty: How blind people use genai tools for information access," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–17.

[5] R. Adnin and M. Das, "I look at it as the king of knowledge: How Blind People Use and Understand Generative AI Tools," in *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, 2024, pp. 1–14.

[6] J. Choi, D. Choi, S. Jeong, H. Hong, and J. Seering, "How far i'll go: Imagining futures of conversational ai with people with visual impairments through design fiction," *arXiv preprint arXiv:2510.12268*, 2025.

[7] P. Karmaker, D. Korre, M. H. u. Rehman, and M. Khodadadzadeh, "AI-Enhanced Landmark Recognition For Self-Guided Tour Application Using Large Language Models," in *Adjunct Proceedings of the 27th International Conference on Mobile Human-Computer Interaction*, 2025, pp. 1–5.

[8] Y. Kaniwa et al., "Chitchatguide: Conversational interaction using large language models for assisting people with visual impairments to explore a shopping mall," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. MHCI, pp. 1–25, 2024.

[9] M. Matei, L. Alboaie, and A. Iftene, "Safety navigation using a conversational user interface for visually impaired people," *Procedia Computer Science*, vol. 207, pp. 1164–1173, 2022.

[10] L. Ran, S. Helal, and S. Moore, "Drishti: An integrated indoor/outdoor blind navigation system and service," in *Second IEEE annual conference on pervasive computing and communications, 2004. Proceedings of the*, IEEE, 2004, pp. 23–30.

[11] D. Sato et al., "Navcog3: An evaluation of a smartphone-based blind indoor navigation assistant with semantic features in a large-scale environment," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 2017, pp. 270–279.

[12] D. Sato et al., "Navcog3 in the wild: Large-scale blind indoor navigation assistant with semantic features," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 12, no. 3, pp. 1–30, 2019.

[13] A. Bokolo Jnr, "Examining the use of intelligent conversational voice-assistants for improved mobility behavior of older adults in smart cities," *International Journal of Human–Computer Interaction*, vol. 41, no. 7, pp. 3867–3888, 2025.

[14] H. Cuayáhuitl, N. Dethlefs, K.-F. Richter, T. Tenbrink, and J. A. Bateman, "A dialogue system for indoor wayfinding using text-based natural language.," *Int. J. Comput. Linguistics Appl.*, vol. 1, no. 1-2, pp. 285–304, 2010.

[15] P. Dang, J. Zhu, W. Li, and J. Lai, "A large language model-based agent for wayfinding: Simulation of spatial perception and memory," *Cartography and Geographic Information Science*, vol. 52, no. 4, pp. 350–369, 2025.

[16] K. Rahimi, M. W. Haque, S. Dasgupta, and M. Rahman, "Vision-based localization and llm-based navigation for indoor environments," *arXiv preprint arXiv:2508.08120*, 2025.

[17] M. Manzoni, S. Mascetti, D. Ahmetovic, R. Crabb, and J. M. Coughlan, "Mapio: Embodied interaction for the accessibility of tactile maps through augmented touch exploration and conversation," *arXiv preprint arXiv:2412.00946*, 2024.

[18] R. G. Long and E. Hill, "Establishing and maintaining orientation for mobility," *Foundations of orientation and mobility*, vol. 1, p. 45, 1997.

[19] J. Ducasse, A. M. Brock, and C. Jouffrais, "Accessible interactive maps for visually impaired users," in *Mobility of Visually Impaired People: Fundamentals and ICT Assistive Technologies*, Springer, 2017, pp. 537–584.

[20] J. Guerreiro et al., "Virtual navigation for blind people: Transferring route knowledge to the real-world," *International Journal of Human-Computer Studies*, vol. 135, p. 102 369, 2020.

[21] J. R. Blum, M. Bouchard, and J. R. Cooperstock, "What's around me? spatialized audio augmented reality for blind users with a smartphone," in *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Springer, 2011, pp. 49–62.

[22] M. Bleau, C. van Acker, N. Martiniello, J. P. Nemargut, and M. Ptito, "Cognitive map formation in the blind is enhanced by three-dimensional tactile information," *Scientific Reports*, vol. 13, no. 1, p. 9736, 2023.

[23] A. Brock, "Interactive maps for visually impaired people: Design, usability and spatial cognition," Ph.D. dissertation, Université Toulouse 3 Paul Sabatier, 2013.

[24] K. Papadopoulos, E. Koustriava, and M. Barouti, "Cognitive maps of individuals with blindness for familiar and unfamiliar spaces: Construction through audio-tactile maps and walked experience," *Computers in Human Behavior*, vol. 75, pp. 376–384, 2017.

[25] D.-R. Chebat, S. Maidenbaum, and A. Amedi, "The transfer of non-visual spatial knowledge between real and virtual mazes via sensory substitution," in *2017 International Conference on Virtual Rehabilitation (ICVR)*, IEEE, 2017, pp. 1–7.

[26] C.-H. Chen, M.-F. Shiu, and S.-H. Chen, "Use learnable knowledge graph in dialogue system for visually impaired macro navigation," *Applied Sciences*, vol. 11, no. 13, p. 6057, 2021.

[27] B. Gamage, T.-T. Do, N. S. C. Price, A. Lowery, and K. Marriott, "What do blind and low-vision people really want from assistive smart devices? comparison of the literature with a focus study," in *Proceedings of the 25th international ACM SIGACCESS conference on computers and accessibility*, 2023, pp. 1–21.

[28] L. Clark et al., "What makes a good conversation? challenges in designing truly conversational agents," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–12.

[29] G. Jain et al., "I want to figure things out: Supporting exploration in navigation for people with visual impairments," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–28, 2023.

[30] G. Jain, L. Findlater, and C. Gleason, "Scenescout: Towards ai agent-driven access to street view imagery for blind users," *arXiv preprint arXiv:2504.09227*, 2025.

[31] J.-E. Kim, G. Sahas, and M. Bessho, "Toward assisting blind individuals in exploring unfamiliar indoor environments using multimodal llm and smartphone lidar," in *2025 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, 2025, pp. 1–6.

[32] J. Balata, Z. Mikovec, and P. Slavik, "Conversational agents for physical world navigation," in *Studies in Conversational UX Design*, Springer, 2018, pp. 61–83.

[33] A. W. Siegel and S. H. White, "The development of spatial representations of large-scale environments," in *Advances in child development and behavior*, vol. 10, Elsevier, 1975, pp. 9–55.

[34] S. Werner, B. Krieg-Brückner, H. A. Mallot, K. Schweizer, and C. Freksa, "Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation1," in *Informatik'97 Informatik als Innovationsmotor: 27. Jahrestagung der Gesellschaft für Informatik Aachen, 24.–26. September 1997*, Springer, 1997, pp. 41–50.

[35] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The craft of information visualization*, Elsevier, 2003, pp. 364–371.

[36] B. Craft and P. Cairns, "Beyond guidelines: What can we learn from the visual information seeking mantra?" In *Ninth International Conference on Information Visualisation (IV'05)*, IEEE, 2005, pp. 110–118.

[37] D. I. Adelani et al., "Irokobench: A new benchmark for african languages in the age of large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 2732–2757.

Improving Acceptability of Energy Efficiency Recommender Systems Through HCI Design

Hayet Hammami and Yacine Ghamri-Doudane
L3I Lab
University of La Rochelle
La Rochelle, France
e-mail: {hayet.hammami | yacine.ghamri}@univ-lr.fr

***Abstract*—Energy efficiency recommender systems are increasingly introduced in workplace environments, yet their adoption remains challenging due to acceptability issues, such as lack of trust, intrusiveness, and limited user control. While Human-Computer Interaction (HCI) design principles for recommender systems are well established, their application in shared professional contexts remains underexplored. This paper reports an empirical investigation through two successive studies: (1) an exploratory co-design workshop with corporate employees, and (2) a two-phase evaluation in a university setting, including an acceptability study using medium-fidelity and video prototypes followed by a hands-on evaluation of a functional mobile application (SmartLR). Across these studies, we examine how HCI design elements, such as notification strategies, transparency mechanisms, and personalization features, shape user trust, perceived control, and short-term acceptability in professional settings. The results highlight recurring design trade-offs related to notification intrusiveness, automation, transparency, and cognitive load. Based on these findings, we propose five practical HCI design guidelines for energy-efficiency recommender systems in workplace environments.**

Keywords-HCI design; Recommender Systems; Energy Efficiency; Workplace Environments; User Centered Design; User Acceptability.

I. INTRODUCTION

Recommender Systems (RSs) are increasingly used to support energy efficiency in buildings and workplace environments by providing occupants with personalized recommendations based on sensor data and contextual information [1][2]. Unlike consumer-facing recommender systems, workplace energy RSs primarily operate in shared spaces, rely on continuous environmental sensing, and often offer limited direct personal benefit to individual users. Beyond algorithmic accuracy, adoption also depends on user acceptability, trust, and perceived disruption to everyday work activities.

Prior research in Human-Computer Interaction (HCI) has identified several factors that influence the acceptability and use of recommender systems, including perceived usefulness, ease of use, transparency, and user control [3]–[5]. These dimensions have been widely studied in domains such as e-commerce, media recommendation, and smart home systems [6]–[8]. However, their applicability to workplace energy recommender systems remains insufficiently understood. In professional contexts, users must reconcile energy-saving recommendations with productivity demands, privacy concerns related to environmental sensing, and the collective nature of decision-making in shared spaces, which fundamentally shapes system acceptability [9][10]. Indeed, most existing work on energy-related recommender systems has focused on algorithmic approaches, system architectures, or quantified energy savings [11]–[13]. While some studies examine user engagement and feedback mechanisms [14][15], fewer investigate how concrete interface design choices influence acceptability during early and medium-term adoption phases, particularly as systems evolve from conceptual prototypes to deployed applications. Understanding how user perceptions change across increasing levels of system fidelity is therefore critical for designing recommender systems that can be meaningfully integrated into workplace environments.

This paper presents an empirical investigation of a workplace energy efficiency recommender system conducted through two successive studies. First, an exploratory co-design workshop to surface user expectations, concerns, and perceived adoption barriers. Second, a two-phase study: the first phase employed medium-fidelity and video-based prototypes to examine initial reactions to alternative interface designs following comparative evaluation approaches [16][17]. The second phase examined user experience with a functional mobile application in a workplace setting, focusing on usability, trust, and perceived effort.

Rather than proposing novel interaction techniques, in this work we synthesize five design guidelines grounded in our findings in a setting where recommender systems are often ignored. Our findings reveal recurring design trade-offs specific to workplace environments, such as balancing notification intrusiveness with awareness, automation with perceived user control, and transparency with cognitive load. By articulating these tensions, the paper provides practical insights for researchers and practitioners designing recommender systems in organizational contexts, particularly those addressing sustainability goals.

This paper is structured as follows: Section II reviews related work on HCI design principles and recommender systems. Sections III, III-C, and III-D present our empirical investigation, results and discussion. Finally, Section IV concludes with a summary of contributions and defined limitations of our work.

II. Related Work

In this section, we review related work on HCI design in recommender systems, with a focus on energy efficiency RSs. We structure the review in two parts: (1) a general overview of recommender systems; and (2) HCI design in recommender systems, which are central to our case study. We synthesize gaps, and position our work within the intersection of these two domains.

A. Overview of Energy Efficiency Recommender Systems

According to [2], Recommender Systems (RSs) are *classified as information systems that primarily focus on information retrieval tasks*. They are initially designed for content personalization based on explicit or implicit user preferences. RSs for energy management have emerged as a critical domain, especially in the context of smart buildings [2][11][13]. These systems typically leverage Internet of Things (IoT) sensor networks to monitor environmental conditions (temperature, humidity, occupancy, lighting), analyze user behavior and occupancy patterns, and generate recommendations for energy-saving actions [18]. For example, such systems may recommend opening a window when humidity becomes too high, turning off lights in unoccupied rooms, adjusting blinds according to sunlight, or modifying heating and cooling settings when temperature exceeds comfort thresholds. While much of the literature focuses on the functioning and algorithmic performance of RSs, comparatively little attention has been paid to their interface design. Indeed, HCI considerations remain secondary in this literature, most studies focus on prediction accuracy, optimization efficiency, or energy outcomes, with interface design treated as an implementation detail rather than a research variable.

In this work, our focus lies on the HCI aspect, specifically, how HCI design influences trust, acceptability and acceptance of these systems, key factors of their adoption and success. A comprehensive taxonomy of energy saving RSs can be found in [2]. This gap persists in recent work, e.g., while DigiGuide [19], and the user-centered recommendation system based on a Pareto-efficient optimization algorithm presented by Cipollone et all. [20] represent technical sophistication, neither reports on user interface design, usability testing, or acceptability factors.

B. HCI Design in Recommender Systems

Because RSs are largely unidirectional delivering recommendations rather than soliciting input, user interaction is often limited. As a result, maintaining engagement and fostering trust requires thoughtful interface design. In this context, different researchers explored how reflective interfaces can influence user behavior in energy consumption contexts [9][21]. Different factors influencing user's acceptability and acceptance have been examined, including perceived usefulness, ease of use, trust, and social influence. Each of these can be shaped through HCI strategies. For example, T. Schwartz et al. demonstrate how interfaces that encourage users to reflect on their personal data can lead to more conscious and sustainable energy use [10]. Their findings highlight the importance of providing users with meaningful feedback and actionable insights. Indeed, users are more receptive when systems provide data that encourages reflection. The impact of interface design is further supported by Swearingen and Sinha, who show that navigation and layout significantly influence user satisfaction, more so than aesthetics or algorithm quality [6]. Hammami et al. reach similar conclusions in HCI evaluation studies, reinforcing the need to prioritize interface clarity over aesthetics [17]. Tintarev et al. emphasize the importance of explanation and transparency in RSs [8]. Their work shows that users are more likely to trust and follow recommendations when they are provided with clear, contextualized justifications. Similarly, Falconnet et al. [22] investigate how message framing affects user beliefs, system attitudes, and behavioral intention, demonstrating that well-framed and justified messages enhance engagement and decision making speed. Indeed, tailored messaging within RSs can significantly enhance user engagement and satisfaction. This study demonstrates the critical role of transparent and contextually appropriate communication in improving user experience. In a broader HCI perspective, Wang et al. [23] use the UTAUT model to analyze acceptance factors such as performance expectancy, effort expectancy, and social influence. They underline the importance of contextual adaptation and community perception in fostering adoption, while also highlighting the need for further research into how specific HCI elements affect these variables.

Finally, HCI for sustainability (also called Sustainable HCI) has long investigated how interface design can promote pro-environmental behavior [14]. Sustainable HCI has produced extensive knowledge on eco-feedback, but workplace energy recommender systems, which combine recommendation algorithms with feedback interfaces, remains underexplored. Empirical studies of deployed systems in shared professional environments are a few. Our work responds to this gap by exploring the HCI design of energy efficiency focused RSs in workplace environments.

1) HCI and Trust in Recommender Systems: Trust is a well-established determinant of recommender system success [9]. Research has identified multiple trust antecedents: transparency and explainability (users trust systems when they understand recommendation rationale) [8]; perceived competence (recommendation accuracy and relevance) [24]; privacy assurance (clear data handling policies); and user control (ability to override or customize) [7]. Explainability must be personalized and context-adaptive. Chromik and Butz [25] articulate design principles for explainable user interfaces (XIA), emphasizing adaptation to users' evolving cognitive states and expertise levels. However, these advances remain concentrated in domains such as e-commerce, healthcare, and generic AI decision-making; application to energy efficiency or workplace sustainability contexts is minimal. Little work examines how transparency, control, and privacy perceptions operate in shared workplace environments where users receive recommendations but derive no direct personal benefit.

2) Workplace Specific HCI Challenges: Workplace environments impose distinct HCI challenges for recommender system adoption [26][27]. Most relevant challenge is notification fatigue, recent work on novel notification modalities (e.g., pneumatic shape-changing smartwatch backs) demonstrates efforts to combat alert fatigue, but remains in early prototyping stages and untested in workplace contexts. Also, privacy and surveillance sensitivity is a very important challenge. In shared spaces, users are acutely aware of data collection and monitoring [7]. Unlike e-commerce where data sharing is implicitly traded for personalization, workplace sensing raises concerns about performance evaluation and autonomy infringement. These concerns are amplified when occupants are passive data providers rather than active participants in system governance [3]. While workplace HCI challenges are increasingly recognized, e.g., new ISO standards [28] and Industry 5.0 frameworks [29], empirical research validating design solutions through deployed systems and standardized evaluation frameworks is lacking.

To summarize, while RSs for energy efficiency have made significant progress, their success depends heavily on HCI design. Despite considerable advances, several critical gaps remain in the literature. First, while algorithmic performance has been extensively studied, the impact of specific HCI design elements on long-term acceptability and engagement is less understood, particularly in workplace contexts where motivations differ from consumer applications. Second, though the benefits of transparency and explanation have been established, optimal approaches for delivering this information remain unclear. Our study builds on this literature by examining how HCI design choices could impact acceptability and acceptance of RSs in workplace environments.

III. Empirical Investigation

Building on the research gaps identified in the literature, this section presents our empirical studies conducted across two workplace environments. Our studies directly address the need to better understand how specific HCI design elements influence user acceptability and engagement with energy efficiency recommender systems. In particular, we explore how interface design, notification strategies, and interaction patterns shape users' willingness to accept and adopt such systems in professional contexts.

Our research was conducted as part of two ongoing projects with similar goals but distinct implementation contexts: BuildOn and Smart Campus, presented below. The first study, exploratory in nature, was conducted in a corporate setting through the BuildOn project [30] using a co-design focus group approach. The second study took place in our university through the Smart Campus project [31] , involving iterative user-centered design and evaluation of a functional mobile recommender system. Together, these studies offer comprehensive insights into user expectations, interaction patterns, and trust factors that shape system acceptability and engagement across different professional environments.

TABLE I. Overview of the empirical studies and project phases

Study	Project	Purpose
Study #1	BuildOn	Exploratory co-design workshop to identify user expectations, preferences, and adoption barriers.
Study #2	Smart Campus - Phase 1	Acceptability evaluation using medium-fidelity and video prototypes.
Study #2	Smart Campus - Phase 2	Hands-on evaluation of the functional SmartLR mobile application.

We structure this section as follows: subsection III-A presents the first study, subsections III-B, III-C, and III-D present the SmartLR system used in the Smart Campus project and the two studies related to this project.

A. Study #1: Co-Design Focus Group (BuildOn Project)

To ground our research in real user needs, we conducted an exploratory study as part of an ongoing European project (BuildOn), which aims to design recommender systems for indoor environmental quality optimization in office buildings. This initial study employed a co-design approach involving employees from a corporate office environment (EDF R&D Paris) to understand user expectations and concerns before developing functional prototypes.

1) Methodology and Participants: The co-design focus group involved 12 occupants of the EDF R&D office building in Paris, including engineers, researchers, and project managers. An open invitation email was distributed to employees working in the building, and the 12 participants who volunteered to take part were included in the workshop. No participant selection or sampling procedure was applied as the objective was to capture diverse workplace perspectives related to comfort, energy use, and daily interaction practices in office environments. Participation was unpaid and based entirely on voluntary involvement.

The workshop was facilitated by the author as part of the BuildOn research project. Before the session, an EDF R&D executive introduced the BuildOn project objectives and remained present during the workshop. The co-design activities were then moderated by the author, who guided discussions, asked follow-up questions to better understand participants' design choices and reasoning, and coordinated the different collaborative activities.

The session followed a structured co-design and design thinking format lasting approximately three hours. Activities included presentation of usage scenarios, collaborative brainstorming, group discussions, and interaction blueprinting exercises centered on future energy recommendation services. Participants worked in small groups to propose interface ideas, discuss positive and negative aspects of potential solutions, and present their final concepts to the other participants. The

workshop relied on presentation slides, whiteboards, handwritten notes, and participant-generated mockups.

Qualitative data collection combined direct observation, handwritten notes, photographs of participant-produced mockups and collaborative boards, and spontaneous verbal feedback shared during discussions and presentations. During group presentations, participants explained and justified their proposed interaction designs, while the author documented recurring concerns, expectations, and design rationales.

Qualitative feedback was analyzed using an exploratory thematic approach. Notes, participant comments, and workshop artifacts were reviewed iteratively to identify recurring expectations, usability concerns, interaction preferences, and perceived adoption barriers. Emerging observations were progressively grouped into broader themes related to system acceptability, user control, personalization, and communication preferences. This exploratory analysis aimed to identify key interaction principles and acceptance factors to guide subsequent design phases of the recommender system.

2) Key Findings: Workshop notes and participant discussions were systematically documented and then analyzed using an exploratory thematic approach to identify recurring expectations, preferences, and adoption concerns related to workplace energy recommender systems. Several recurring themes emerged from participant feedback, which fell into two main categories: essential features for acceptability ("must have") and potential adoption barriers ("must avoid").

a) Essential Features:

- **Personalization:** Ability to customize thresholds based on personal comfort preferences.
- **Contextual Information:** Connection with external weather conditions and forecasts.
- **Positive Reinforcement:** Messages that congratulate users for energy-friendly behaviors, periodic challenges, or gamified comparisons between offices.
- **Advisory Tone:** Recommendations framed as suggestions rather than commands.
- **Comparative Feedback:** Options to compare performance with others or against historical data.
- **Bi-directional Communication:** Ability to report discomfort or override recommendations.

b) Adoption Barriers:

- **Mobile Notifications:** Users strongly rejected smartphone alerts, anticipating they would be quickly ignored or disabled.
- **Lack of Control:** Systems that offer recommendations without allowing user adjustments.
- **Absence of Visual Signals:** Participants preferred ambient visual indicators (such as LED-based signals) over text-only interfaces or push notifications.

Participants also emphasized the importance of bi-directionality in interactions. Being able to report discomfort, signal absences, or override incorrect recommendations when context is missing (e.g., "*I am feeling cold despite the system saying the temperature is optimal*") was seen as essential for maintaining a sense of agency. These insights underline a recurring tension in workplace recommender systems: users may accept passive systems only if they retain control and can understand or challenge the system's reasoning.

In summary, this initial study highlighted interaction principles essential for designing acceptable recommender systems in workplace environments. It confirmed that transparency, personalization, and perceived control are critical to foster trust and engagement.

B. The Smart Campus Project

Smart Campus is an ongoing initiative on our university campus aiming to reduce energy consumption through context-aware, sensor-driven recommendations delivered to building occupants via a mobile application (named SmartLR). The SmartLR system continuously monitors indoor conditions using IoT sensors installed in staff offices and shared spaces. These sensors collect real-time data on temperature, humidity, presence, and lighting, which are processed to generate energy-saving recommendations. The system targets a wide range of university users—including faculty, administrative staff, and PhD students—encouraging behavior change without disrupting daily routines.

We conducted a two-phase study within this project to investigate how different HCI design elements influence users' willingness to engage with such a system. To ensure a user-centered approach, we adopted an iterative evaluation process spanning two complementary phases:

- **Phase 1 – Early acceptability evaluation:** Participants were introduced to the system through medium-fidelity prototypes and video scenarios. We aimed to explore their initial reactions, concerns, and expectations, and to identify HCI elements likely to affect acceptability (e.g., interface clarity, feedback granularity, perceived control, notification strategies).
- **Phase 2 – Acceptance evaluation:** Based on Phase 1 findings, we developed a functional mobile application and tested it with a new group of participants. This phase focused on real-time interaction, mobile usability, and the role of trust and personalization in shaping user engagement and long-term adoption.

Both studies were conducted on campus and involved 32 participants in total (15 for the first study and 17 on the second study), covering a diversity of roles in the university. The next two sections detail the methodology and findings from each study.

C. Study #2 - Phase #1: Early Acceptability Evaluation

This study evaluates the acceptability of our recommender system (SmartLR) prior to deploying the full app. We focused on how different HCI design approaches influence user perceptions and willingness to engage with energy-saving recommendations in workplace settings. By examining initial reactions to different interface designs, we aimed to identify key factors that enhance or inhibit system acceptability.

1) Protocol: The evaluation session began with a brief introduction to SmartLR, explaining its functionality, the role of IoT sensors installed throughout university facilities, and its mobile application as the primary interface for user interaction. This introduction aimed to familiarize participants with how the mobile app delivers energy-saving recommendations, and its intended benefits.

Participants were presented with three different prototype designs of the mobile application, each illustrating a unique design approach. Following this, they viewed two short video demonstrations simulating real-world interactions with the system. These videos highlighted typical user interactions, showcasing how SmartLR offers recommendations and how users respond to notifications for energy-saving actions through their smartphones. This experimental approach builds on prior HCI evaluation research, which supports the use of multiple design alternatives and video-based evaluation to enhance feedback quality and user engagement. Research has shown that *"testing many is better than testing one"*, as comparative evaluations help users express their preferences and usability concerns more effectively [16][17]. Moreover, using video alongside paper prototypes, as noted by Hammami et al. [32], allows participants to visualize system interactions more clearly, leading to more detailed feedback.

2) Participants: We recruited 15 voluntary participants through internal laboratory messaging and direct email invitations sent to colleagues and administrative staff within the university. No compensation was provided. Participants were drawn from diverse professional backgrounds, including researchers (PhD students, postdocs, and research engineers) and administrative staff. While demographic data such as age were not collected, the participant pool ensured representation across different roles within the university.

3) Prototype Elaboration: To evaluate the impact of different HCI design elements, we developed different design prototypes of the SmartLR mobile application. These prototypes were designed to explore variations in interface layout, interaction flow, and visual hierarchy.

a) Paper Prototype: The paper prototypes were medium-fidelity mockups created in Figma and printed for the session. We created three different interface designs of the mobile application, each with a distinct layout while maintaining the same core functionalities. The three prototype designs are presented in Figure 1.

The design prototypes included key system features: weather information to contextualize recommendations; room identification; room status displaying sensor readings for temperature, humidity, and lighting conditions; and a menu icon providing access to additional system functions (contents not detailed at this stage).

Each prototype contained two different UIs, representing two distinct system states: normal and alert. The normal state shows the system's standard operation when environmental conditions are stable: temperature and humidity within optimal ranges, and lighting conditions aligned with occupancy. The alert state is triggered when conditions deviate from optimal ranges, such as excessive temperature/humidity or lights left on in unoccupied rooms, prompting corrective action recommendations.

Figure 1. Three alternative interface designs tested for SmartLR in the first phase of Study #2

b) Video Prototype: In addition to the paper prototypes, we created two short video simulations to demonstrate user interaction with SmartLR using the mobile application. The first video lasted 52 seconds, while the second lasted 1 minute and 20 seconds. Each scenario illustrated a different context in which SmartLR provides recommendations and how users respond to system notifications through their mobile phone.

*c) **Scenario 1: High Humidity Alert**:* In the first scenario, a person works in their office with the blinds open, door and windows closed, and the light on, maintaining an ideal temperature. They receive a notification on their phone that the room's humidity is high and a prompt is given to ventilate it. The individual then opens the windows and the door to ventilate the space, following the system's recommendation, and later resumes work. After a short while, they recheck the humidity status on their phone using the application.

*d) **Scenario 2: Unnecessary Lighting**:* The second scenario depicts a person working in their office under normal conditions: blinds open, door and windows closed, and light on. Upon preparing to leave, the individual turns off their computer, gathers their belongings, and exits the office forgetting light on. After leaving, they receive a phone notification advising them to turn off the light to conserve energy. This scenario emphasizes the system's ability to prompt users to save energy even if they forget to perform actions themselves.

The combination of paper and video prototypes allowed participants to evaluate both static interface designs and interactive system behavior, ensuring a comprehensive assessment of usability, clarity, and effectiveness in real-world contexts.

4) Data Collection: To collect user feedback, we employed two complementary methods: semi-structured interviews and a questionnaire based evaluation. Initially, participants engaged in an interactive feedback session (the interview). We used a whiteboard to visually categorize feedback into three columns: *Likes*, *Dislikes*, and *Suggestions*. Participants shared their thoughts directly on the whiteboard using pens or post-it notes. This method captured both qualitative and quantitative feedback, providing diverse insights into user preferences and responses. To enhance feedback depth, we asked participants to think aloud during the evaluation, capturing real-time reactions. Sessions were audio recorded to ensure comprehensive documentation of all feedback.

After the interview, participants completed a structured questionnaire based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model [4]. To clarify the relationship between the questionnaire and our research questions, the questionnaire items were grouped into four constructs derived from the UTAUT model. Items related to perceived usefulness of SmartLR for supporting energy efficiency and workplace comfort were associated with Performance Expectancy (RQ1). Items related to ease of use, clarity of interaction, and usability were associated with Effort Expectancy (RQ2). Items reflecting the perceived influence of colleagues and the organizational context were associated with Social Influence (RQ3). Items related to confidence in the system's recommendations and data handling were associated with Trust (RQ4). In addition, separate items captured usage intention and overall evaluation of the application.

Qualitative feedback collected during these individual sessions was analyzed using an exploratory thematic approach. Feedback from think-aloud observations, interview responses, and whiteboard annotations was reviewed and grouped into recurring themes related to usefulness, usability, aesthetics, and privacy concerns.

5) Research Questions: Our study explores how different HCI design elements influence user experience with the SmartLR mobile application. We explored four specific Research Questions:

- **RQ1 (Performance Expectancy):** Users who perceive the system as beneficial for their work and energy efficiency will be more likely to adopt it.
- **RQ2 (Effort Expectancy):** Users who find the system easy to use will be more likely to adopt it.
- **RQ3 (Social Influence):** Users who perceive a positive social climate around SmartLR will be more likely to integrate it into their routines.
- **RQ4 (Trust):** Users who trust the system's accuracy, privacy, and reliability will be more likely to adopt it.

6) Results:

a) Qualitative Results: We collected 149 feedback during interviews: 63 Likes (42%), 22 Dislikes (15%), and 64 Suggestions (43%). The high proportion of suggestions indicates constructive engagement, with users actively contributing to system improvement rather than merely criticizing existing features.

Feedback was categorized into four dimensions:

- **Utility:** 85 comments (57%) – Users appreciated the ecological focus and real-time environmental measurements. Suggestions focused on personalization (customizable thresholds) and visualization of environmental impact.
- **Usability:** 37 comments (25%) – Concerns emerged regarding notification frequency, with users expressing preference for non-intrusive alerts that would not disrupt work.
- **Aesthetics:** 19 comments (13%) – Minor feedback on visual design elements.
- **Privacy and Security:** 8 comments (5%) – Exclusively categorized as Dislikes or Suggestions, indicating potential trust barriers. Users raised questions about data storage duration, location, and access permissions.

Users expressed strong interest in gamification elements (badges, weekly reports, office competitions) to make energy-saving behaviors more tangible and engaging. This aligns with research by Mendez et al. [15] and Chatzigeorgiou et al. [14], which demonstrates how gamification strategies promote long-term engagement with energy-saving systems.

Notably, participants did not express strong preferences for any single interface design among the three prototypes. Instead, they identified useful or problematic elements across all versions, providing valuable guidance for our final design decisions.

b) Quantitative Analysis: The questionnaire contained 16 items rated on a 5-point Likert scale from "Strongly Disagree" (1) to "Strongly Agree" (5). We analyzed participants' responses by calculating mean scores for each construct and interpreting them with respect to the corresponding Research Questions:

- **Performance Expectancy (RQ1):** 4.0 – Users recognized the system's potential benefits for energy efficiency.
- **Effort Expectancy (RQ2):** 4.4 – Perceived ease of use was rated highest.
- **Social Influence (RQ3):** 3.4 – Workplace norms showed limited influence on adoption intention.
- **Trust (RQ4):** 3.8 – Moderate confidence in data handling and recommendation accuracy.
- **Usage Intention:** 3.9 – Positive willingness to adopt the system.
- **Overall Application Evaluation:** 4.2 – Generally positive perceptions.

These findings suggest that perceived usefulness (RQ1), ease of use (RQ2), and trust (RQ4) were all positively perceived by participants and may play an important role in shaping acceptance of the system.

7) Discussion: This study provides critical insights into factors influencing early acceptability of energy efficiency recommender systems in workplace environments. Key implications for the interface design in such systems include:

1) **Prioritize intuitive, low cognitive load interfaces:** The high Effort Expectancy score (4.4) suggests that usability

is an important factor, though not sufficient alone to guarantee adoption.

2) **Implement transparent data handling practices:** Privacy concerns, while representing only 5% of feedback, were exclusively negative, indicating that clear explanations and GDPR compliance are necessary to maintain trust.
3) **Demonstrate concrete benefits through visualization:** Users requested features showing energy savings and environmental impact, confirming that tangible feedback motivates sustained engagement.
4) **Balance notification frequency:** Concerns about disruptive alerts underscore the need for well-timed, non intrusive notifications in workplace contexts.
5) **Offer personalization options:** The strong demand for customizable thresholds and display preferences reinforces that perceived control is critical for user acceptance.

D. Study #2 - Phase #2: Acceptance Evaluation

Building on insights from phase 1, we developed a fully functional version of SmartLR to conduct a hands-on evaluation of the system in a controlled workplace-like setting. This second phase shifted from hypothetical scenarios toward direct interaction with a working application, allowing us to examine how the HCI design principles identified during the acceptability evaluation influenced usability, trust, perceived effort, and user engagement during realistic usage situations. While phase 1 focused on anticipated acceptability based on prototypes, phase 2 explored user experience through supervised interaction with the functional application.

1) Protocol: User tests were conducted using a fully functional mobile application on a provided Android smartphone. The session began with a brief introduction to SmartLR, after which participants were given time to freely explore the application's features and interface. During this exploration period, we simulated environmental changes to trigger alert conditions, generating notifications that participants could respond to in real time. This controlled evaluation approach allowed us to observe reactions to system alerts and assess how intuitively users navigated the interface when prompted to take action. Although the evaluation took place in a supervised setting, the interaction scenarios were designed to reproduce realistic workplace situations, with participants seated at office desks while interacting with the system. The main SmartLR application interfaces used during testing are shown in Figures 2, 3, and 4. The home screen features several key elements: Current weather conditions displayed prominently. University branding for institutional context. Profile access for personalization. Real-time sensor readings (temperature, humidity, presence, lighting). Alert indicators when measurements exceed thresholds. And navigation menu for accessing additional features.

Based on phase 1 of the study, we implemented several enhancements:

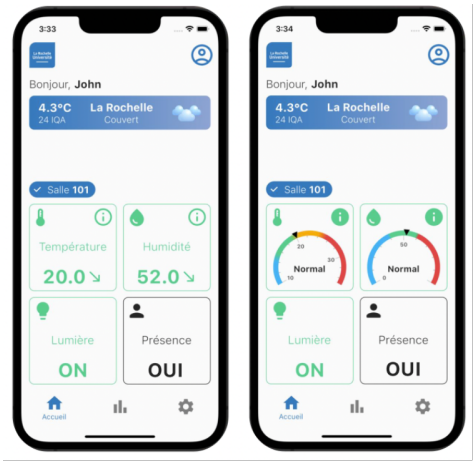


Figure 2. SmartLR home screen showing normal conditions with sensor readings

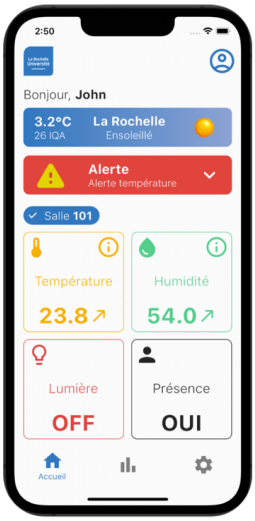


Figure 3. SmartLR home screen with alert notification for high humidity

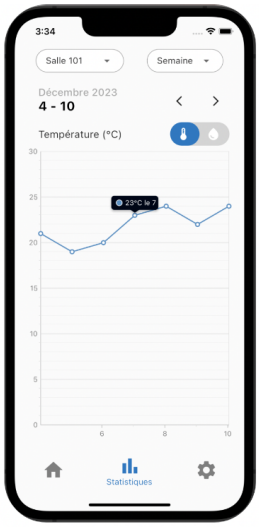


Figure 4. SmartLR statistics page displaying historical environmental data

- **Dynamic indicators** showing temperature and humidity trends (increasing, stable, and decreasing arrows), providing users with proactive awareness of changing conditions. These visual cues allow users to anticipate changes and take preventive actions, such as opening a window or adjusting heating, before reaching critical thresholds.
- **Visual gauges** indicating "normal" ranges for environmental measurements, making it easier for users to interpret data at a glance, a feature particularly appreciated for humidity levels.
- A dedicated **statistics page** (see figure 4) offering historical measurement data organized chronologically, enabling users to track patterns over time and better understand environmental fluctuations.
- **Enhanced privacy protection** through comprehensive GDPR compliance documentation, developed in collaboration with the university's specialized legal services. This included the integration of a formal data protection charter within the application, directly addressing privacy concerns raised during Study #1.

These design decisions directly addressed feedback from the first evaluation, particularly the desire for more personalization options, clearer visualization of environmental data, and transparent privacy controls. The interface prioritized intuitive navigation and minimal cognitive load while providing the detailed information users requested.

2) Participants: We recruited 17 voluntary participants from the same university population as Phase 1 through internal laboratory messaging and direct email invitations sent to colleagues and administrative staff. No compensation was provided. Participants came from similar profiles, including researchers, faculty members, and administrative staff.

3) Data Collection: As in phase 1, we employed a mixed-methods approach combining qualitative feedback through interactive sessions and quantitative assessment through a standardized questionnaire. During the interview, we asked participants to categorize their feedback as likes, dislikes, or suggestions, providing consistent comparison with our first study results.

The questionnaire was modified slightly from Study #1 to include additional items related to expected effort and expanded questions about design aesthetics and privacy perceptions. These additions were motivated by a specific focus on how interface design influences perceived effort and usability. The revised questionnaire maintained the 5-point Likert scale structure from the previous study.

As in phase 1, qualitative feedback collected during these individual sessions was analyzed using an exploratory thematic approach. Participant comments and suggestions were reviewed and grouped into recurring themes related to utility, usability, aesthetics, and personalization.

4) Results:

a) Interactive Feedback Analysis.: We collected 132 feedback which we categorized and quantified. The distribution showed 32% Likes, 30% Dislikes, and 38% Suggestions. Compared to phase 1, we observed a relative increase in Dislikes (from 15% to 30%) and a decrease in Likes (from 42% to 32%). This shift likely reflects the transition from conceptual evaluation to practical usage, where users encountered actual limitations rather than hypothetical capabilities.

We further classified feedback based on the system aspects users commented on most directly. The analysis revealed three main categories:

- **Utility:** 54.3% of feedback. Representative suggestions included adding comparative consumption data between dates, language selection options, and visibility of threshold values on the statistics page. These suggestions reveal users' desire for more personalized and informative features that enhance the practical value of the system.
- **Usability:** 37.4% of feedback. Participants requested highlighting relevant measurements during alerts and praised the overall simplicity and ergonomics of the interface. Comments such as "easy to handle" and "fairly simple and ergonomic" indicate generally positive reception.
- **Aesthetics:** 8.3% of feedback. Only minor comments about color choices, indicating that participants were primarily concerned with functionality and ease of use rather than visual appearance.

A notable pattern in the feedback was the emphasis on personalization. Multiple users requested the ability to customize which measurements appeared on the home screen, highlighting the importance of adaptability in interface design.

b) Questionnaire Analysis.: The questionnaire analysis revealed consistently positive perceptions across all dimensions, with improvements in almost all categories compared to phase 1.

- **Performance Expectancy:** Increased from 4.0 to 4.2, suggesting that hands-on experience with the functional system strengthened users' belief in its utility. This suggests that participants perceived the system as useful and relevant for supporting energy-efficiency practices.
- **Effort Expectancy:** Remained high (4.5 compared to 4.4 in Study #1), confirming that the implemented interface successfully preserved the intuitive usability identified as crucial in our first phase of the study. This result suggests that ease of use was positively perceived and likely contributed to overall system acceptance.
- **Trust:** Showed a slight increase from 3.8 to 4.1, indicating that the privacy controls and transparency features incorporated into the functional application helped address users' data protection concerns.
- **Design:** The addition of a specific Design score (4.5) suggests that visual and interactive elements contributed positively to user satisfaction and perceived usability.
- **Social Influence:** Increased substantially from 3.4 to 3.9, approaching the thresholds of the other factors. This suggests that as users gain concrete experience with energy-saving recommendations, they become more aware of the social and organizational context of their

energy consumption behaviors.

5) Discussion: this second phase of Study #2 provided valuable insights into how users interact with and perceive SmartLR in a functional context, revealing both strengths in our implementation and opportunities for future refinement. The questionnaire results provide encouraging indications regarding the acceptability of the proposed design approach and the relevance of the explored Research Questions. The high Effort Expectancy score (4.5) suggests the importance of usability and interface design in participants' perceptions of the system, demonstrating that our user-friendly interface contributed positively to system acceptability. The increase in Performance Expectancy (4.2) from phase 1 supports RQ1, suggesting that users who recognize the application's practical value may be more willing to adopt it Similarly, improvements in Trust (4.1) and Social Influence (3.9) indicate more positive participant perceptions regarding confidence in the system and its integration within workplace practices.

A recurring theme in user feedback was the desire for personalization. Multiple participants requested the ability to customize displayed measurements on the home screen, highlighting the importance of adaptability in interface design. This preference reveals that even when users appreciate an interface's overall design, they still value the ability to tailor it to their specific needs and preferences.

The functional aspects of the system emerged as particularly crucial to users, with utility related feedback dominating the comments (54.3%). This suggests that while aesthetic considerations contribute to overall satisfaction, the core capabilities and practical benefits remain primary drivers of user engagement. Features that provide tangible evidence of impact, such as comparative energy consumption data, appear especially valuable in reinforcing performance expectations and sustaining motivation.

Based on these findings, we identify several implications for future development:

1) **Enhance personalization options:** Allow users to configure their information displays according to individual preferences and priorities (linked to RQ2).
2) **Provide more concrete visualization of energy savings:** Reinforce performance expectations through tangible evidence of impact (linked to RQ1).
3) **Maintain interface clarity and usability:** Continue to evolve the design while preserving the high standards identified in this study.
4) **Further strengthen trust:** Maintain transparent communication about data handling practices and regular updates on security measures (linked to RQ4).
5) **Develop social features:** Encourage users to share their experiences, leveraging the increased importance of social influence observed in this study (linked to RQ3).

IV. Conclusion and Future Work

This paper investigated how Human-Computer Interaction (HCI) design influences the acceptability of energy-efficiency recommender systems in workplace environments. Through two complementary studies, a co-design workshop in the BuildOn project and a two-phase SmartLR evaluation in the Smart Campus project, we examined how users perceive and react to different interface design choices across early design and hands-on evaluation stages.

The results consistently indicate that short-term acceptability is shaped by perceived usefulness, ease of use, trust, perceived control, and the way recommendations are presented through the interface. In particular, participants responded positively to personalization options, contextualized feedback, and clear interaction mechanisms, while raising concerns about intrusive notifications, insufficient transparency, and limited control over recommendations.

Based on these findings, this work contributes five practical design guidelines for workplace energy-efficiency recommender systems:

1) **Balance notification intrusiveness:** Recommendations should be delivered in ways that support awareness without unnecessarily interrupting work activities.
2) **Enable personalization:** Users should be able to adapt thresholds, displayed information, and notification preferences to their needs and workplace situations.
3) **Provide layered information access:** Interfaces should present essential information first while allowing access to more detailed explanations, sensor values, and historical data when needed.
4) **Visualize impact:** Systems should make the consequences of recommendations more concrete through clear and meaningful feedback on environmental conditions and energy-related effects.
5) **Build trust through transparency:** Interfaces should clearly communicate recommendation logic, privacy-related information, and data-handling practices in order to support user confidence.

Rather than claiming long-term adoption effects, these guidelines should be understood as design recommendations grounded in iterative empirical evaluation of initial user perceptions and reactions. Although these guidelines were derived from the context of energy-efficiency recommender systems, some of the identified HCI considerations, such as transparency, personalization, feedback clarity, and user engagement, may also be relevant to other domains involving persuasive or behavior-change technologies. However, additional studies would be required to validate their applicability beyond the energy domain.

This work has several limitations. First, the studies relied on relatively small participant samples in specific workplace contexts, which limits the generalizability and replicability of the findings. Second, the evaluations primarily captured short-term acceptability and supervised interaction rather than long-term adoption in everyday practice. In particular, the second phase of the SmartLR study involved controlled hands-on sessions with simulated alert conditions rather than a fully naturalistic deployment. Third, the qualitative analyses were exploratory and intended to identify recurring expectations, concerns, and

design implications rather than establish predictive or causal relationships.

Future work should extend this research through longitudinal studies in real workplace environments in order to examine sustained use, behavioral change, and long-term acceptance. Additional work is also needed to explore adaptive notification strategies, richer personalization mechanisms, and stronger integration with building management systems. Future investigations could also examine the role of gamification elements, which emerged as a user interest in our studies, while considering the risk of engagement fatigue. Finally, evaluating these design principles in other organizational settings would help assess their robustness beyond the specific contexts studied here.

REFERENCES

[1] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Techniques, applications, and challenges," *Recommender systems handbook*, pp. 1–35, 2021.

[2] Y. Himeur *et al.*, "A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects," *Information Fusion*, vol. 72, pp. 1–21, 2021.

[3] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[4] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[5] V. Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions," *Decision sciences*, vol. 39, no. 2, pp. 273–315, 2008.

[6] K. Swearingen and R. Sinha, "Beyond algorithms: An hci perspective on recommender systems," in *ACM SIGIR 2001 workshop on recommender systems*, vol. 13, 2001, pp. 1–11.

[7] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4–5, pp. 441–504, 2012.

[8] N. Tintarev and J. Masthoff, "Explaining recommendations: Design and evaluation," in *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys)*, ACM, 2015, pp. 147–150. DOI: 10.1145/2792838.2792841.

[9] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of management annals*, vol. 14, no. 2, pp. 627–660, 2020.

[10] T. Schwartz, G. Stevens, L. Ramirez, and V. Wulf, "Uncovering practices of making energy consumption accountable: A phenomenological inquiry," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 20, no. 2, pp. 1–30, 2013.

[11] T. Pinto *et al.*, "Multi-agent-based cbr recommender system for intelligent energy management in buildings," *IEEE Systems Journal*, vol. 13, no. 1, pp. 1084–1095, 2018.

[12] F. Taghvaei and R. Safa, "Efficient energy consumption in smart buildings using personalized nilm-based recommender system," *Big data and computing visions*, vol. 1, no. 3, pp. 161–169, 2021.

[13] R. O. Panizza, S. M. M. Mohammadi, M. Anbia, and M. NikBakht, "Design recommender system for building energy performance," in *Canadian Society of Civil Engineering Annual Conference*, Springer, 2023, pp. 223–236.

[14] I. M. Chatzigeorgiou and G. T. Andreou, "A systematic review on feedback research for residential energy behavior change through mobile and web interfaces," *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110 189, 2021. DOI: 10.1016/j.rser.2020.110189. [Online]. Available: https://www.researchgate.net/publication/343563334_A_systematic_review_on_feedback_research_for_residential_energy_behavior_change_through_mobile_and_web_interfaces.

[15] J. I. Méndez, T. Peffer, P. Ponce, A. Meier, and A. Molina, "Empowering saving energy at home through serious games on thermostat interfaces," *Energy and Buildings*, vol. 263, p. 112 015, 2022. DOI: 10.1016/j.enbuild.2022.112015. [Online]. Available: https://escholarship.org/uc/item/4959v9tf.

[16] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen, "Getting the right design and the design right," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 1243–1252.

[17] H. Hammami, G. Calvary, M. Riahi, F. Moussa, and S. Bouzit, "Comparative evaluation? yes, but with which alternative ui?" In *Electronic Visualisation and the Arts (EVA 2017)*, 2017, pp. 32–1.

[18] M. T. Siddique, P. Koukaras, D. Ioannidis, and C. Tjortjis, "Smartbuild recsys: A recommendation system based on the smart readiness indicator for energy efficiency in buildings," *Algorithms*, vol. 16, no. 10, p. 482, 2023.

[19] J. Ma, R. Yus, and G. Bouloukakis, "Digiguide: A dt-based occupant guiding system for optimizing comfort and energy consumption," in *2025 IEEE International Conference on Smart Computing (SMARTCOMP)*, IEEE, 2025, pp. 9–17.

[20] V. Cipollone *et al.*, "User-centric comfort measurement and energy optimization: A pareto-efficient approach for a personalized recommendation strategy," in *2025 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv)*, IEEE, 2025, pp. 312–316.

[21] B. Xiao and I. Benbasat, "E-commerce product recommendation agents: Use, characteristics, and impact," *MIS quarterly*, pp. 137–209, 2007.

[22] M. Wutz, M. Hermes, V. Winter, and J. Köberlein-Neu, "Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: Integrative review," *J Med Internet Res*, vol. 25, e46548, 2023, ISSN: 1438-8871.

[23] Y.-Y. Wang, A. Townsend, A. Luse, and B. Mennecke, "The determinants of acceptance of recommender systems: Applying the utaut model," *18th Americas Conference on Information Systems 2012, AMCIS 2012*, vol. 3, pp. 2238–2246, Jan. 2012.

[24] K. Acharya and J. Mikkonen, "Energy usage responsive space and personal mobile devices," in *Proceedings of the 12th international conference on human computer interaction with mobile devices and services*, 2010, pp. 407–408.

[25] M. Chromik and A. Butz, "Human-xai interaction: A review and design principles for explanation user interfaces," in *IFIP Conference on Human-Computer Interaction*, Springer, 2021, pp. 619–640.

[26] M. Chen *et al.*, "Values of user exploration in recommender systems," in *Proceedings of the 15th acm Conference on recommender systems*, 2021, pp. 85–95.

[27] Y. Wang *et al.*, "Surrogate for long-term user experience in recommender systems," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 4100–4109.

[28] International Organization for Standardization, "Iso standard 88373," 2026, [Online]. Available: https://www.iso.org/standard/88373.html (visited on 05/2026).

[29] R. Hamdani and I. Chihi, "Adaptive human-computer interaction for industry 5.0: A novel concept, with comprehensive review and empirical validation," *Computers in Industry*,

vol. 168, p. 104 268, 2025, ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2025.104268. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361525000338.

[30] BuildON Project, "Buildon project," 2026, [Online]. Available: https://buildon-project.eu/ (visited on 05/2026).

[31] University of La Rochelle, "Smart campus," 2026, [Online]. Available: https://www.univ-larochelle.fr/luniversite/notre-vision/mettre-en-place-un-campus-intelligent-durable-et-responsable/smart-campus/ (visited on 05/2026).

[32] H. Hammami, F. Camara, G. Calvary, M. Riahi, and F. Moussa, "The benefits of combining paper-and video-based prototypes for user interface evaluation," Apr. 2020.

Event-Aware Audio Generation for LLM-Driven Storytelling in Extended Reality

Mehmet Karaaslan, Meral Kuyucu, Bora Şenceylan, Gökhan İnce
Department of Computer Engineering
Istanbul Technical University
Istanbul, Türkiye
e-mail: {karaaslan18 | korkmazmer | senceylan19 | gokhan.ince}@itu.edu.tr

***Abstract*—Sound is essential for immersion in Extended Reality (XR), yet audio design is often manual and disconnected from narrative context. This paper presents a Large Language Model (LLM) driven pipeline for event-aware audio generation in XR storytelling. The system extracts sound-inducing events from scene inputs, generates context-specific audio using a diffusion model, and selects high-quality samples through an LLM-based judging mechanism. Generated sounds are bound to narrative events using timing and repetition cues. We evaluated the system using a six scene story displayed in XR with 20 participants. Results show improved narrative-audio alignment and immersion when using Semantic Representation.**

Keywords-extended reality; generative audio; LLM-based systems; event-aware audio; immersive storytelling.

I. Introduction

Sound is one of the most fundamental modalities through which humans perceive and interpret their surroundings. It provides essential information about spatial orientation, material properties, and the cause of events, and often reaches our consciousness faster than visual stimuli. Auditory feedback is not supplementary, but a vital component in how humans build a mental model of the world and feel present within it.

As it is transitioned from physical to digital environments through Virtual Reality (VR) and Extended Reality (XR), establishing a strong sense of presence becomes a significant challenge. VR aims to transport users into entirely synthetic worlds, while XR seeks to integrate digital content seamlessly into the user's physical environment. In both domains, the goal is to support immersion and presence, enabling users to suspend disbelief and meaningfully engage with the experience. Although visual technologies in immersive environments have reached high levels of fidelity, the auditory experience has often lagged behind. Without synchronized, contextually accurate sound, even the most visually stunning immersive experience feels hollow and artificial. This creates a sensory mismatch that disrupts the user's immersion.

Extensive research has been conducted in an attempt to bridge this gap. Initial efforts centered on spatial audio and pre-recorded sound libraries activated by specific user actions. Even though recent studies have moved towards more sophisticated context-aware approaches [1][2][3]. Most existing approaches primarily focus on generating individual sound effects or scene-level audio, without addressing narrative continuity across scenes or automatic binding of sounds to evolving events. While recent generative models demonstrate impressive audio quality in isolation, they are typically evaluated at the model level rather than within interactive experiences. As a result, little attention is given to how generated audio supports story progression, maintains consistency over time, or impacts user perception during actual XR use. These limitations become especially critical in narrative settings, where sounds must be context-aware, temporally aligned, and perceptually coherent to sustain immersion.

To address these limitations, this study proposes a Large Language Model (LLM)-integrated pipeline designed to enhance storytelling in XR through generative audio. The proposed sonification pipeline automatically generates audio content and binds it to XR scenes based on semantic structure. This study investigates the following research questions:

- **RQ1:** Does using structured scene descriptions improve narrative-audio alignment and immersion compared to video-based input with a brief textual summary?
- **RQ2:** Do technically literate users perceive the proposed framework as a viable tool for automated audio authoring in XR?

Building on these research questions, we present an LLM-driven pipeline that analyzes narrative context, object identities, and environmental interactions to extract sound-producing events and generate context-specific audio. The generated sounds are integrated back into the experience through event-aware binding, supporting temporal alignment and continuity across scenes. We evaluate the proposed system through a within-subject VR user study with 20 participants using a six-scene narrative experience, comparing input representations and examining the automated audio selection mechanism. The contributions of this paper are four-fold: 1) a modular pipeline for automatic, event-aware sound generation in XR storytelling, 2) an LLM-based judging mechanism for automated ranking and selection of diffusion-generated audio, 3) an event-aware audio binding strategy that supports temporal alignment and narrative continuity across scenes, and 4) an XR user study evaluating narrative-audio alignment, synchronization, and perceived immersion.

The remainder of the paper is organized as follows: Section II reviews related work in XR audio and generative models. Section III details the proposed Event-Aware Audio Generation System. Section IV describes the experimental setup and user study methodology. Section V presents the results, and Section VI concludes the paper and outlines future work.

II. Related Work

The role of sound in XR has evolved from simple triggering of pre-recorded assets to systems that attempt to adapt audio dynamically to user interactions and scene context. Early approaches emphasized spatial accuracy and realism, while more recent work explores generative methods that produce sound based on semantic or visual inputs. As

immersive experiences become increasingly narrative-driven and multi-scene in structure, the need for audio systems that support temporal coherence, event continuity, and contextual reasoning has grown.

A. Audio in XR and Interactive Systems

Early work in immersive audio primarily emphasized spatial fidelity and the triggering of pre-recorded assets. These systems typically relied on rule-based or retrieval-based mappings, where specific user actions were manually associated with fixed sound libraries. While effective for basic interactions, this approach required substantial authoring effort and offered limited flexibility.

More recent approaches aim to reduce this overhead through in situ generation. For example, SonifyAR [1][2] employs a pipeline called Programming by Demonstration to capture physical interactions (such as a ceramic cup sliding on wood) as text, which is then processed by an LLM to retrieve or generate sound. SandTouch [3] demonstrates that gesture-responsive audio feedback can improve presence in virtual art experiences. Similarly, Sonify Anything [4] uses computer vision to infer material properties and generate physically plausible interaction sounds in real time.

Despite these advances, most interactive audio systems remain focused on short-term physical interactions or isolated object-level feedback. They do not reason over narrative structure, temporal continuity, or evolving scene context. As a result, they are limited in their ability to support story-driven experiences in XR. Recent exploratory studies on generative Artificial Intelligence (AI) for immersive storytelling [5][6] also highlight the absence of mechanisms for maintaining coherent audio behavior across multi-scene narratives. This gap motivates the need for systems that move beyond local interaction cues toward event-aware, narrative-level audio generation.

B. Generative Audio Models

Recent advances in generative audio have been largely driven by latent diffusion models. Systems, such as AudioLDM [7] and AudioLDM 2 [8] generate high-quality sound from natural language prompts. These models support zero-shot generation, style transfer, and audio inpainting, significantly improving output realism. PicoAudio [9] further introduces timestamp-aware generation and frequency controllability, enabling finer temporal alignment between audio and visual content.

Despite their strong generative capabilities, these models are primarily designed for standalone audio production. They do not inherently account for scene structure, object behavior, or narrative progression within XR environments. In immersive applications, audio must be aligned not only with visual timing but also with semantic context and continuity across scenes. Diffusion models alone do not provide this linkage.

As a result, deploying generative audio in XR requires an intermediate reasoning layer that determines which sounds should be generated, when they should occur, and how they should persist over time. This layer must translate structured scene information into generation prompts and bind the synthesized outputs back into the runtime environment. Our work addresses this integration gap by combining generative audio models with LLM-driven event extraction and event-aware audio binding.

C. Multimodal Reasoning with LLMs

LLMs are increasingly used as reasoning components in multimodal systems. They enable the interpretation of narrative text, visual inputs, and structured environmental data. Prior work has shown that LLMs can decompose stories into object descriptions and scene layouts, as demonstrated in Metabook [10] and Stepping Into Stories [6]. DreamFoley [11] jointly models video, text, and audio to generate foley sounds aligned with visual motion, while Scene2Hap [12] uses LLMs to infer physical properties, such as material density for haptic feedback.

Beyond content generation, recent studies have explored the use of LLMs for automated evaluation. Audio-aware LLM judges [13] and LLM-as-a-judge protocols [14] demonstrate that LLMs can approximate human preferences when ranking generated outputs. These approaches enable scalable quality assessment without manual annotation.

While these systems highlight the potential of LLMs for multimodal reasoning and evaluation, they largely operate at the level of individual scenes or isolated interactions. They do not address how narrative events evolve over time, nor how generated content should be persistently bound to objects and actions within immersive environments. Our work builds on this line of research by using LLMs not only for content generation and evaluation, but also for extracting event-level structure that supports continuity and audio binding across multi-scene XR experiences.

D. Audio-Visual Alignment and Object-Level Sonification

Aligning audio with specific visual elements is essential in immersive systems, particularly when multiple sound sources coexist within a scene. Sounding That Object [15] generates audio from user-selected visual regions within images, linking sound directly to object-level input. Scene-to-Audio [16] converts complex visual scenes into representative audio renderings, primarily for accessibility purposes. Similarly, SEE-2-SOUND [17] produces spatial audio by identifying and localizing multiple sound sources in visual content without requiring explicit training data. ImmerseDiffusion [18] further extends this direction by conditioning diffusion models on spatial and environmental parameters to generate immersive 3D soundscapes.

Together, these approaches demonstrate a shift from scene-wide sonification toward object-aware audio generation. However, most systems operate at the level of isolated frames or single-scene inputs. They focus on spatial alignment or object-level correspondence, but do not explicitly model how sounds should persist, evolve, or repeat across a sequence of narrative events. In story-driven XR experiences, continuity goes beyond recognizing objects; sounds must remain consistent over time, align with unfolding actions, and reflect changes across scenes.

Taken together, prior work advances interaction-level sonification, generative audio synthesis, multimodal reasoning, and spatial alignment. However, these efforts remain largely fragmented. Most systems focus on individual components,

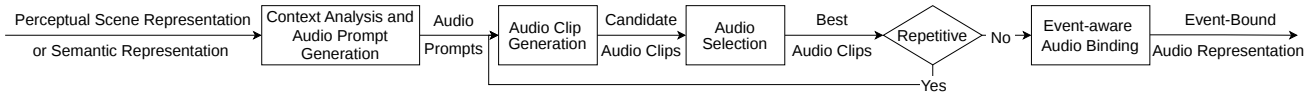


Figure 1. Proposed System Architecture.

such as object-level sound generation or visual-audio correspondence, rather than treating audio as part of a continuous narrative experience. As a result, narrative reasoning, temporal consistency, and runtime audio binding are rarely addressed within a single framework. In contrast, our work introduces an end-to-end pipeline that extracts sound-producing events from structured scene representations and maintains consistent audio behavior across multi-scene XR narratives.

III. Event-Aware Audio Generation System

A. Framework Overview

The proposed system is designed as a modular framework for event-aware audio generation in XR. As shown in Figure 1, the framework operates in four components: 1) context analysis and audio prompt generation, 2) audio generation, 3) audio selection, and 4) event-aware audio binding. Each component is intentionally decoupled, allowing different LLMs, audio generators, or runtime engines to be substituted without altering the overall pipeline.

The framework accepts two forms of input: 1) a Perceptual Scene Representation (PSR) consisting of a scene video and its short summary, and 2) a Semantic Representation (SR) that encodes objects, actions, and temporal information extracted from the XR environment. Both inputs can be transformed into a shared intermediate representation in the form of structured audio events as proposed in this paper. These events define when sounds occur, how they evolve over time, and how they relate to narrative progression.

Each audio event is converted into *(N)* audio clips using a generative audio model. An LLM-based selection mechanism then ranks these candidates based on semantic relevance and perceptual quality. Finally, the selected audio clips are bound to the XR runtime using timing, repetition, and transition parameters, enabling consistent and context-aware sound behavior across multiple scenes. The following subsections describe each component of the framework in detail.

B. Context Analysis and Audio Prompt Generation

The framework supports two input representations for different types of XR data. These are processed by an LLM to bridge the gap between environmental data and audio synthesis. The first is a perceptual scene representation, consisting of a scene video paired with its short summary. This setting approximates vision-based approaches commonly used in prior work, while preserving minimal narrative context required for audio generation. The second is a semantic representation, which encodes contextual information, such as objects, actions, and temporal relationships extracted from the XR environment.

The LLM acts as a digital sound designer. First, it processes the given input to identify every sound-generating event within the scene. Then, for each event, it generates a structured audio specification that includes:

- Event type: *new*, *continue* (persisting from a previous scene), or *copy* (reusing an existing sound).
- A text prompt for the audio generation model.
- Start and end times aligned with scene animations.
- A repetition flag for recurring actions (e.g., footsteps), enabling variation across repeated sounds.
- A volume level scaled from 1 to 10.
- Fade-in and fade-out durations to support smooth transitions.

To generate these specifications, the LLM follows a list of rules and constraints. The first rule is event type management, which ensures temporal consistency across the narrative. The LLM determines if a sound is being created for the first time, flows uninterrupted from a previous scene, or is a specific recurring sound effect that must remain identical.

The second rule focuses on the audio prompt. To ensure the generated audio clips integrate realistically into the XR environment, the LLM is told to specify physical attributes for each event, including material composition and specific acoustics. This descriptive detail guides the generation model to produce audio that respects the context of the scene.

The third rule establishes a hierarchical volume scaling logic. On a scale of 1 to 10, the LLM assigns volume levels based on the event's narrative priority; background ambiances are restricted to a lower range of 2–4, while discrete action events are prioritized with higher values between 5–8. This automated mixing ensures that key interactions remain audible and clear without being masked by environmental soundscapes.

The final rule defines timing and transitions. The LLM assigns start and end times in a (mm:ss) format to synchronize each sound with scene animations. It also specifies fade-in and fade-out durations to ensure smooth transitions and prevent abrupt auditory cuts.

For recurring interactions, such as walking or wing flapping, events are marked as repetitive. In these cases, the framework requests multiple audio samples for the same prompt, allowing variation during playback rather than relying on a single repeated clip. This distinction enables the system to handle both continuous ambient soundscapes and discrete interaction events within a unified representation.

C. Audio Generation and Selection

After audio effect descriptions are produced, the framework performs a generate-and-select cycle for each sound event. For every prompt, the audio generation model produces N number of candidate audio clips. This allows the system to explore different versions of the same prompt and reduces the impact of randomness in diffusion-based generation. Candidate selection is performed using an LLM-as-a-Judge mechanism. Each generated clip is scored on a scale from 0 to 100 according to three criteria:

- **Thematic accuracy:** how well the sound matches the textual prompt.

- **Technical quality:** clarity and absence of artifacts or digital noise.
- **Atmospheric coherence:** perceived realism and contribution to immersion.

The highest-scoring clip is selected as the final output for non-repetitive events. For repetitive actions, such as footsteps, this process is repeated three times to produce a small pool of variations. During playback, clips are sampled from this pool to avoid perceptual repetition and auditory fatigue.

This generate-and-select strategy reduces randomness in diffusion models, where a single output may not match the intended sound. Evaluating multiple candidates helps maintain consistency between audio and scene context.

D. Event-Aware Audio Binding

The final component of the framework integrates the generated audio files into the XR environment. In this context, event-aware binding associates generated sounds with specific animation events (e.g., synchronizing wing-flap audio with the bird's wing-flap motion). This module follows the event parameters produced during the Context Analysis (Section III-B) to determine when and how each sound should be played. Audio sources are instantiated dynamically at runtime and attached to their corresponding scene objects. For each event, the system applies the specified start and end times, volume level, and transition parameters. Fade-in and fade-out durations are used to avoid abrupt onsets or cutoffs. This ensures smooth integration with ongoing scene dynamics and animation timelines. Repetitive events, such as footsteps or wing flaps, are handled differently. Instead of replaying a single clip, the system draws from a small pool of generated variations. During playback, clips are selected in sequence to cover the required animation duration. This reduces perceptual repetition and creates a more natural auditory effect.

IV. Experiments

A. Hardware and Software

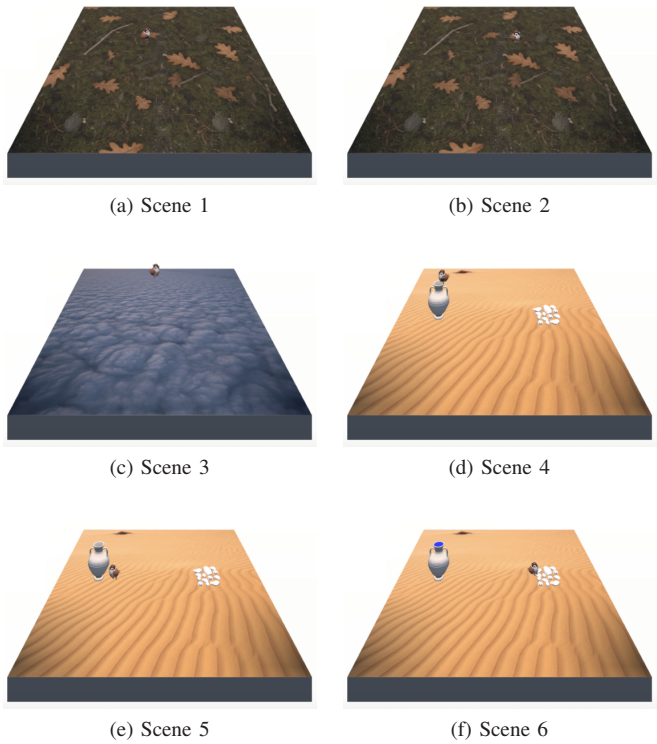
In this study, a Meta Quest headset was chosen as the primary interaction device due to its superior passthrough capabilities, high-quality visual fidelity, and its ability to operate as a standalone unit. A local workstation equipped with an NVIDIA RTX 5090 GPU was preferred as the generative AI pipeline host. The proposed audio generation workflow was implemented in Python. The latest Gemini Pro API [19] was used to analyze inputs and generate audio effect descriptions, which were then used by the AudioLDM 2, specifically the cvssp/audioldm2-large checkpoint [20], latent diffusion model, to generate audio clips. Additionally, the Gemini Flash API [19] functioned as a judge to evaluate and select the optimal audio samples. The Unity game engine was used to develop the XR environment.

B. Narrative and Experimental Environment

The framework was evaluated using a six-scene XR narrative inspired by the Thirsty Crow fable [21], implemented in Unity. The narrative was designed to cover a range of acoustic conditions, including ambient environments, object interactions, and repetitive motion events. This allowed the system to be tested across both continuous soundscapes and discrete action-driven audio.

The six scenes were structured as follows:

- The bird walking through a forest environment (Figure 2a).
- The bird flying within the forest (Figure 2b).
- The bird flying above clouds (Figure 2c).
- The bird standing on the edge of a pitcher and looking inside (Figure 2d).
- The bird pecking the pitcher (Figure 2e).
- The bird collecting stones and dropping them into the pitcher (Figure 2f).



(a) Scene 1 (b) Scene 2

(c) Scene 3 (d) Scene 4

(e) Scene 5 (f) Scene 6

Figure 2. Narrative scenes used in the XR evaluation.

The full narrative experience lasted approximately two minutes, with individual scenes designed with a fixed duration of eleven seconds to maintain experimental consistency. The Thirsty Crow narrative was selected due to its clear sequence of actions and physical interactions.

The study was conducted in a quiet indoor setting rather than a controlled laboratory. Participants experienced the narrative using the headset under consistent lighting and audio conditions. This setup was chosen to approximate realistic usage scenarios while maintaining basic experimental consistency.

C. Experimental Design

To evaluate the proposed framework, we conducted a user study with 20 participants recruited through purposive sampling. The primary selection criteria required participants to have a baseline familiarity with XR systems and generative AI tools to ensure the evaluation focused on the quality of the generated audio rather than the novelty of the hardware. Additionally, all participants were required to have normal or corrected-to-normal vision and hearing. The group consisted of 5 female and 15 male participants, aged between 20 and 27 years. All participants were either university students or

 ISBN: 978-1-68558-383-5

TABLE I. MEAN (± SD) LIKERT RATINGS FOR PERCEPTUAL SCENE REPRESENTATION AND SEMANTIC REPRESENTATION WITH AND WITHOUT REPETITION.

	Perceptual Scene Representation		**Semantic Representation**	
	wo. Repetition	**w. Repetition**	**wo. Repetition**	**w. Repetition**
Audio Narrative	2.8 ± 1.41	2.35 ± 1.29	3.75 ± 1.33	3.15 ± 1.08
Audio Visual Synchronization	2.56 ± 1.28	2.26 ± 1.19	4.3 ± 0.78	3.09 ± 0.83
Immersion / Presence Contribution	2.94 ± 1.36	2.56 ± 1.32	3.8 ± 1.43	2.95 ± 1.27

graduates from diverse academic backgrounds and reported moderate familiarity with XR systems and contemporary generative tools. Participation was voluntary, and informed consent was obtained prior to the experiment. No personally identifiable data was collected.

The study followed a within-subjects design, where each participant experienced automatically generated auditory scenes using two different input methods: perceptual scene representation and semantic representation. Additional system components, including repetitive audio handling, were evaluated as part of the overall framework. Dependent measures included perceived narrative–audio alignment, audio–visual synchronization, immersion and presence, and subjective preference between conditions. In this study, the number of candidate audio files *(N)* is selected as three.

D. Procedure

Each participant completed one session lasting approximately 15 minutes. Upon arrival, participants received a brief introduction to the narrative experience and the headset. After the device was fitted, a short calibration was performed to ensure visual and auditory clarity. Participants experienced the narrative twice, generated with both PSR and SR input types. The order of auditory scenes generated was randomized across participants to eliminate order effects. After each exposure to the audiovisual content, participants completed a short questionnaire, in which participants rated their experience using a 5-point Likert scale (1: Strongly Disagree, 5: Strongly Agree) across three dimensions:

- **Audio Narrative:** whether the sounds matched the scenes.
- **Audio–Visual Synchronization:** whether sounds occurred at appropriate moments.
- **Immersion and Presence:** whether audio contributed to a sense of immersion.

To assess the framework's potential as a developer-facing tool, participants also rated the system's usefulness for automatic sound generation across the following dimensions:

- **Utility and Adoption:** whether the tool is perceived as useful for development and likely to be used in future projects.
- **Audio Quality:** whether generated audio clip's quality is good
- **Workflow Efficiency:** whether the system increases efficiency compared to traditional methods.
- **Preference for Automation:** whether automated sound creation is preferred over manual searching and curation.

V. RESULTS

A. Results on Narrative–Audio Alignment and Immersion

Our analysis focuses on the comparative performance of the input representations (PSR vs. SR), the impact of the repetition mechanism, and the perceived utility of the system as an automated authoring tool.

The data indicate a clear preference for the Semantic Representation (SR) over the Perceptual Scene Representation (PSR) across all measured dimensions. As shown in Table I, participants rated the SR condition (without repetition) significantly higher in terms of Audio Narrative alignment (M=3.75, SD=1.33) compared to the PSR condition (M=2.8, SD=1.41). This suggests that providing the LLM with structured data allows for more contextually accurate audio prompt generation than video-based inputs.

The most substantial difference was observed in Audio-Visual Synchronization. The SR condition achieved a mean score of 4.3±0.78, while the PSR condition fell to 2.56±1.28. Feedback from participants suggested that the PSR-based pipeline occasionally generated sounds that were thematically relevant but failed to play at the correct moment. For example, in the second scene, the sound of the bird flapping its wings started late, failing to sync with the moment the bird began to fly. In contrast, the SR input allowed the event-aware binding module to anchor audio clips to specific animation timestamps with higher precision.

As expected, adding the repetition mechanism led to a decrease in Likert ratings across both input representations. This outcome was mainly caused by the quality of the generated audio files. It aligns with the observations made during the development phase.

B. Result on Perceived Usefulness as an Automated Audio Authoring Tool

The secondary objective of our study was to evaluate the system from a developer's perspective, specifically targeting its potential as an automated authoring tool. As summarized in Table II, the framework received positive marks for its potential integration into XR development workflows.

TABLE II. AUTHORING UTILITY QUESTIONNAIRE RESULTS (5-POINT LIKERT, MEAN ± STD).

Measure	**Mean ± Std**
Usefulness	3.8 ± 1.75
Output quality	3 ± 1.33
Workflow efficiency	3.7 ± 1.33
Adoption intent	3.9 ± 1.44

Participants expressed a strong **Adoption Intent** (M=3.9,SD=1.44), viewing the system as a practical solution for immersive sound design. This is supported by **Workflow Efficiency**, as they highlighted that automated generation could reduce manual search effort. However, **Output Quality** scored lowest; while the system managed long-form ambient sounds successfully, the quality of short-duration clips for

repetitive actions was perceived as lower. Overall, these results show that participants viewed the system as both practical and effective for supporting automated audio authoring.

VI. Conclusion and Future Work

This paper presented a modular framework for event-aware generative audio in XR. The system combines LLM-based scene reasoning with diffusion-based sound synthesis and automated quality selection. Structured audio events are extracted from narrative context and bound directly to animation timelines. This enables sounds to follow story intent while remaining synchronized with runtime behavior. User study results showed consistent trends favoring structured scene descriptions over video-based input. Participants viewed the framework as a promising tool for automated audio authoring.

Evaluation identified specific performance variances between audio types; while long ambient samples were well-received, short-duration clips occasionally lacked the perceptual precision required to sonify discrete actions. Scalability also remains a consideration for high-density environments. The current generate-and-select strategy introduces latent processing overhead for each event. These results suggest that as scene complexity grows, the computational demands of LLM-driven reasoning and the *(N)* generation ratio will require further optimization to sustain real-time performance.

Future work will focus on practical deployment. We plan to develop a custom runtime introspection mechanism to automatically derive semantic representations from the XR scene by capturing narrative events directly from the running environment, and an author-facing interface for previewing and refining generated sounds. Larger and more diverse studies will be conducted to validate the findings. We also aim to explore real-time constraints and alternative audio models. Our long-term goal is to support scalable audio pipelines that reduce manual effort while preserving creative control in immersive storytelling.

References

[1] X. Su, J. E. Froehlich, E. Koh, and C. Xiao, "Sonifyar: Context-aware sound generation in augmented reality", in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '24, Pittsburgh, PA, USA: Association for Computing Machinery, 2024, pp. 1–13, ISBN: 9798400706288. DOI: 10.1145/3654777.3676406.

[2] X. Su, E. Koh, and C. Xiao, "Sonifyar: Context-aware sound effect generation in augmented reality", in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24, Honolulu, HI, USA: Association for Computing Machinery, 2024, pp. 1–7, ISBN: 9798400703317. DOI: 10.1145/3613905.3650927.

[3] L. Liu et al., "Sandtouch: Empowering virtual sand art in vr with ai guidance and emotional relief", in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25, Association for Computing Machinery, 2025, pp. 1–21, ISBN: 9798400713941. DOI: 10.1145/3706598.3714275.

[4] L. Schütz, S. Matinfar, U. Eck, D. Roth, and N. Navab, *Sonify anything: Towards context-aware sonic interactions in ar*, 2025. arXiv: 2508.01789 [cs.HC].

[5] H. Doh, J. Shi, R. Jain, H. Kim, and K. Ramani, *An exploratory study on multi-modal generative ai in ar storytelling*, 2025. arXiv: 2505.15973 [cs.HC].

[6] A. Vitali, C. Schneegass, and T. Dingler, *Stepping into stories: Envisioning a generative ai pipeline to create story-based vr reading environments*, Mensch und Computer 2025 - Workshopband, 2025. DOI: 10.18420/muc2025-mci-ws09-144.

[7] H. Liu et al., *Audioldm: Text-to-audio generation with latent diffusion models*, 2023. arXiv: 2301.12503 [cs.SD].

[8] H. Liu et al., "Audioldm 2: Learning holistic audio generation with self-supervised pretraining", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024. DOI: 10.1109/TASLP.2024.3399607.

[9] Z. Xie, X. Xu, Z. Wu, and M. Wu, *Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation*, 2024. arXiv: 2407.02869 [cs.SD].

[10] Y. Wang et al., *Metabook: A mobile-to-headset pipeline for 3d story book creation in augmented reality*, 2025. arXiv: 2405.13701 [cs.HC].

[11] F. Li, W. Zhao, Y. Li, Z. Zhou, and D. He, *Dreamfoley: Scalable vlms for high-fidelity video-to-audio generation*, 2025. arXiv: 2512.06022 [cs.SD].

[12] A. Jingu, E. AliAbbasi, P. Strohmeier, and J. Steimle, *Scene2hap: Combining llms and physical modeling for automatically generating vibrotactile signals for full vr scenes*, 2025. arXiv: 2504.19611 [cs.HC].

[13] C.-H. Chiang et al., *Audio-aware large language models as judges for speaking styles*, 2025. arXiv: 2506.05984 [eess.AS].

[14] L. Zheng et al., *Judging llm-as-a-judge with mt-bench and chatbot arena*, 2023. arXiv: 2306.05685 [cs.CL].

[15] T. Li et al., *Sounding that object: Interactive object-aware image to audio generation*, 2025. arXiv: 2506.04214 [cs.CV].

[16] C. Gupta, A. Ram, S. Sridhar, C. Jouffrais, and S. Nanayakkara, "Scene-to-audio: Distant scene sonification for blind and low vision people", in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '25, Association for Computing Machinery, 2025, pp. 1–9, ISBN: 9798400713958. DOI: 10.1145/3706599.3719849.

[17] R. Dagli, S. Prakash, R. Wu, and H. Khosravani, *See-2-sound: Zero-shot spatial environment-to-spatial sound*, 2025. arXiv: 2406.06612 [cs.CV].

[18] M. Heydari, M. Souden, B. Conejo, and J. Atkins, *Immersed-iffusion: A generative spatial audio latent diffusion model*, 2025. arXiv: 2410.14945 [cs.SD].

[19] "Google deepmind gemini", Accessed: 2026-02-24. [Online]. Available: https://deepmind.google/models/gemini/.

[20] "Audioldm2 large checkpoint", Accessed: 2026-02-24. [Online]. Available: https://huggingface.co/cvssp/audioldm2-large.

[21] "The crow and the pitcher", Accessed: 2026-02-24. [Online]. Available: https://en.wikipedia.org/wiki/The_Crow_and_the_Pitcher/.

Castillo de San Marcos AR: Spatial Augmented Reality Interactive Learning System for Cultural Heritage Education

Markus Santoso, David Ramtulla,
HuaGuo Tian, Jonah Matousek, Yixin Hou
Digital Worlds Institute, University of Florida
Gainesville, United States
Email: markus.santoso@ufl.edu, ramtulladavid@ufl.edu,
h.tian@ufl.edu, jonahmatousek@ufl.edu, yixinhou@ufl.edu

***Abstract*—This paper presents a spatial Augmented Reality (AR) interactive learning system designed to enhance cultural heritage education and immersive human-computer interaction using Castillo de San Marcos in St. Augustine, Florida. The system employs Vuforia Ground Plane spatial tracking to anchor reconstructed historical content directly onto the physical environment of the fort, enabling situated and contextual learning experiences. Users explore architectural structures, defensive layouts, and historical narratives through interactive spatial visualization and AR-based storytelling. The proposed system contributes to research in interactive educational technologies and spatial computing by integrating markerless AR interaction with cultural heritage interpretation in outdoor environments. Preliminary observations indicate improved user engagement, spatial comprehension, and experiential learning potential. Future work includes formal user evaluation, collaborative learning features, and expanded historical reconstruction.**

Keywords—augmented reality; cultural heritage; immersive learning; interactive education; spatial AR.

I. INTRODUCTION

Cultural heritage sites provide opportunities for experiential and contextual learning. Traditional interpretation methods often rely on static signage and narration, which may limit user engagement and spatial understanding. Augmented Reality (AR), originally defined by Azuma [1] as the integration of virtual and physical environments in real time, enables digital content to be embedded within real-world environments to support situated learning. Billinghurst et al. [2] further demonstrated the growing role of AR in Human-Computer Interaction (HCI) and interactive visualization. Castillo de San Marcos presents a historically significant masonry fort whose architectural and defensive structures benefit from spatial visualization. This work presents a spatial AR system that overlays reconstructed historical elements onto the physical site to enhance immersive learning, interactive exploration, and contextual engagement.

Augmented Reality has emerged as an effective medium for cultural heritage education by enabling immersive and situated learning experiences. Unlike traditional interpretation methods, AR allows digital historical reconstructions and contextual information to be directly integrated with the physical environment, improving spatial understanding and learner engagement. Prior research has shown that AR enhances conceptual comprehension, supports experiential learning, and increases motivation by transforming passive observation into interactive exploration [3][4]. These capabilities make AR particularly suitable for historical sites where spatial context and environmental immersion are essential for meaningful learning.

This work is relevant to the conference themes of advanced Human-Computer Interaction, immersive educational technologies, contextual interaction, and interactive digital heritage systems. By combining spatial Augmented Reality, situated learning, and interactive visualization, the proposed system demonstrates how immersive interfaces can enhance user engagement and experiential learning within real-world cultural heritage environments.

The rest of the paper is structured as follows. Section II reviews related work in Augmented Reality and cultural heritage education. Section III presents the system design and implementation of the spatial AR learning platform. Finally, Section IV concludes the paper and discusses future work.

II. RELATED WORK

Augmented Reality has been widely applied in cultural heritage education to enhance engagement and contextual understanding. AR enables digital reconstructions to be overlaid onto physical environments, supporting experiential learning and improved knowledge retention. Fonseca et al. [5] explored mixed reality applications for cultural heritage visualization and interactive learning environments. Jantke et al. [6] demonstrated an AR system for Castle Scharfenstein integrating gamification and visitor interaction to enhance exploration and engagement. Situated learning theory emphasizes learning within real-world contexts, which AR effectively supports. While many AR heritage systems rely on markers or indoor deployment, fewer studies explore spatial AR in indoor and outdoor environments. Situated learning theory emphasizes learning within real-world context, which AR effectively supports. While many AR heritage systems rely on markers or indoor deployment, fewer studies explore spatial AR in indoor/outdoor environments.

Numerous studies have demonstrated the educational benefits of Augmented Reality in cultural heritage contexts. AR has been shown to improve knowledge retention and engagement by enabling users to interact with reconstructed historical environments and contextual visualizations [3]. Furthermore, AR-based heritage systems promote experiential and inquiry-based learning, allowing visitors to actively explore historical narratives rather than passively consume information [4]. These findings highlight the potential of AR as an effective tool for enhancing cultural heritage interpretation and educational outcomes.

Situated learning theory suggests that learning is more effective when knowledge is acquired within real-world contexts. AR supports this approach by embedding digital information directly within the physical environment. While many existing systems focus on indoor museum environments or marker-based AR, fewer works explore outdoor heritage interpretation using spatial ground-plane tracking. This work contributes by applying an AR to a masonry fort environment with spatially contextualized interaction.

III. System Design & Implementation

The system was developed in Unity using Vuforia Ground Plane technology. Ground Plane enables the detection of horizontal surfaces and allows digital models to be anchored without predefined visual markers. Reconstructed 3D architectural elements are overlaid onto the physical site to support spatial visualization and contextual learning. Figure 1 shows the 3D reconstruction model of Castillo de San Marcos.

The application was implemented in Unity and deployed on a mobile handheld platform using Vuforia Ground Plane tracking for markerless spatial anchoring. Reconstructed 3D models were optimized for mobile rendering performance to support stable real-time interaction in outdoor environments. Users interact with the virtual content through touchscreen-based translation, rotation, and scaling controls integrated within the Graphical User Interface (GUI).

Figure 1. 3D Model of Castillo de San Marcos.

Spatial AR does not rely on predefined image targets or fiducial markers to align virtual content with the real environment. Instead, it detects characteristic features and planar surfaces of a scene in real time, enabling flexible deployment in unprepared environments and supporting natural user interaction. This approach reduces setup complexity and allows AR content to be anchored directly onto real-world terrain, which is particularly suitable for outdoor cultural heritage environments such as Castillo de San Marcos. Advances in computer vision and mobile sensing technologies further enable spatial systems to utilize the physical environment itself as a tracking reference, improving usability and immersion [7].

Figure 2. 3D Reconstruction of Castillo de San Marcos in a Spatial AR application.

As shown in Figure 2, users initialize the system by scanning the environment to detect a ground plane for spatial anchoring. Once the anchor is established, AR overlays are aligned with the physical structure of the fort. Contextual hotspots provide access to historical narratives, reconstructed visualizations, and explanatory content. On-site deployment demonstrates the feasibility and robustness of spatial AR under varying indoor/outdoor lighting conditions. The application further enables users to manipulate the reconstructed castle model through rotation and scaling interactions via the graphical user interface (GUI) on the mobile device screen.

IV. Conclusion and Future Works

Spatial AR enhances spatial comprehension and engagement in outdoor heritage environments, although challenges remain in tracking stability, lighting variability, and reconstruction accuracy. Optimization for mobile performance is essential to ensure a consistent user experience. Augmented Reality also offers clear pedagogical benefits for cultural heritage education by embedding contextual information directly within the physical

environment, thereby supporting situated learning and improving spatial cognition. Interactive visualization and user-driven exploration promote deeper understanding and engagement compared to traditional static interpretation methods, consistent with prior findings on AR's positive impact on learner motivation and comprehension [3][4].

Several limitations remain in the current implementation. Outdoor spatial tracking performance may vary under inconsistent lighting conditions and complex environmental surfaces. Mobile hardware constraints can also affect rendering performance and tracking stability in large-scale outdoor environments. In addition, the current system currently focuses primarily on single-user interaction and has not yet undergone formal user evaluation for learning effectiveness and usability assessment.

This paper presented a Spatial AR system for cultural heritage education at Castillo de San Marcos. The platform demonstrates the effectiveness of spatially anchored visualization for immersive historical learning. Future work will focus on controlled user evaluation, multi-user collaboration, and expansion of historical reconstruction and educational content.

REFERENCES

[1] R. Azuma, "A Survey of Augmented Reality," Presence, 1997.

[2] J. Billinghurst, A. Clark, and G. Lee, "A Survey of Augmented Reality," Foundations and Trends in Human–Computer Interaction, vol. 8, no. 2–3, pp. 73–272, 2015.

[3] M. Wojciechowski and W. Cellary, "Evaluation of learners' attitude toward learning in AR systems," Computers & Education, 2013.

[4] M. Dunleavy and C. Dede, "Augmented reality teaching and learning," Journal of Technology and Teacher Education, vol. 22, no. 1, pp. 7–22, 2014.

[5] D. Fonseca, E. Martí, A. Redondo, I. Navarro, and S. Sánchez, "Mixed reality applications in cultural heritage," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-5, pp. 231–238, 2014

[6] K. P. Jantke, J. Krebs, and M. Santoso, "Game Amusement and CRM: Castle Scharfenstein AR Case Study," IEEE GCCE, 2014, pp. 488–491

[7] Z. Oufqir, A. El Abderrahmani, and K. Satori, "From Marker to Markerless in Augmented Reality," Springer, 2020, doi:10.1007/978-981-15-0947-6_57

Function Discoverability and Perceptual Accessibility in Interfaces for Adults Aged 60+: Task-Based UX Study

Julia Manikowska
Institute of Computing Sciences, Faculty of Computing and Telecommunications,
Poznan University of Technology
Poznan, Poland
e-mail: julia.manikowska@student.put.poznan.pl

Julia Samp
Institute of Computing Sciences, Faculty of Computing and Telecommunications,
Poznan University of Technology
Poznan, Poland
e-mail: julia.samp@student.put.poznan.pl

Piotr Lukasiak
Institute of Computing Sciences, Faculty of Computing and Telecommunications,
Poznan University of Technology
Poznan, Poland
e-mail: piotr.lukasiak@put.poznan.pl

***Abstract*— Ageing populations increasingly rely on digital services, yet older adults often face usability and accessibility barriers, especially when interfaces depend on hidden actions, icon-only controls, and visually demanding layouts. This study assessed whether four different application prototypes (Shopping List, Messenger, Login/Registration, Shop) and alternative User Interface (UI) and Cascading Style Sheets (CSS) variants support function discoverability and perceptual accessibility for adults aged 60+. Overall task difficulty was moderate (mean M = 2.22, median Me = 2, standard deviation SD = 1.09), but only 48.3% of task attempts were completed without moderator help. Messenger was the most challenging module (M = 2.73; 29.7% independent), with "add member to group" performing worst (M = 3.12; 16.2% independent). Shop was the easiest (M = 1.82; 64.0% independent). Participants aged 60+ were less independent and slower than those < 60 (p < 0.001). Barriers clustered around discoverability, navigation, icon semantics, form feedback, contrast, and target size. Design improvements should prioritise explicitly labelled actions, stronger feedback, higher contrast, and larger interactive targets; visual styling alone was insufficient without structural interaction changes.**

Keywords*— *usability testing; function discoverability; web accessibility; user interface design.

I. INTRODUCTION

In the context of ongoing population ageing and the continued migration of services to digital channels, the accessibility and usability of interfaces for adults aged 60+ has become an issue of growing social and economic relevance [1]. From the perspective of older users, the central challenge is not merely "access to technology" but also the cost of interaction: increased cognitive effort, uncertainty about the consequences of actions, heightened risk of error, and frustration, which may ultimately result in disengagement from online services [2]. Research on technology acceptance in this group emphasises that sustained use of digital solutions depends largely on a sense of control and ease of use rather than on the system's objective functionality alone [3]. Against this background, the present study addressed a practical question: to what extent the designed application modules and User Interface (UI) and Cascading Style Sheets (CSS) variants are comprehensible, discoverable, and executable independently by older adults, who often experience functional limitations.

Cognitive mechanisms help explain why identical interface solutions may operate differently for younger and older users. Processing-speed theory indicates that with age the cost of operations requiring rapid attentional switching and the maintenance of multiple elements in working memory increases, thereby elevating susceptibility to pauses and errors in sequential tasks [4]. In parallel, age-related changes in the visual system (including reduced contrast sensitivity and poorer performance under low luminance) directly affect typographic legibility and the recognisability of elements with weak visual separation [5].

Within these constraints, the literature identifies recurring challenges in interface design for older adults: limited discoverability of functions, navigational disorientation, ambiguity of icons, difficulties with forms (login/registration), and visual and ergonomic barriers associated with small clickable targets. Empirical studies show that age significantly differentiates both perceived website usability and task performance, and that these differences intensify as information architectures become more complex and the number of potential action paths increases [6]. Importantly, older adults rely more heavily on structural cues and layout consistency, and, as informational queries grow more complex, they more frequently commit errors attributable to the costs of search strategies (e.g., goal loss and ineffective exploration) [7]. From a navigation perspective, solutions such as vertical menus have been shown to support effectiveness and user preference in tasks of increasing complexity, reinforcing the importance of explicit and predictable option presentation [8]. Moreover, research on navigation style indicates that the organisation of transitions and the manner of presenting options exert a measurable influence on older adults' efficiency in web-based environments [9].

 ISBN: 978-1-68558-383-5

Studies have demonstrated that, in older adults, icon comprehension in navigation tasks depends on the semantic distance between a symbol and its intended meaning, as well as on the presence of a textual label; labelling reduces interpretive uncertainty and may limit selection errors [10]. Related work on mobile-device icons suggests that age-related differences are particularly pronounced during initial use, before users acquire interface conventions [11]. Practically, this supports recommendations for selective redundancy (e.g., icon plus label) in task-critical locations, while avoiding unnecessary stimulus density in secondary areas. Research on redundancy further indicates that such solutions can improve performance and confidence among older users, even if younger users sometimes perceive them as superfluous [12]. In communication modules (e.g., a messenger), this body of evidence justifies efforts to minimise unlabeled icons and to increase the clarity of control intent.

Another major category includes forms, particularly login and registration, which combine security requirements with memory demands and error risk. Evidence suggests that older adults more often encounter barriers in authentication, and that the critical factors become unambiguous error messages, support for account recovery, and mechanisms that prevent mistakes (e.g., the ability to reveal a password during entry) [13]. Accordingly, form design for older adults should prioritise clarity, proximity of feedback to the relevant field, reduction of unnecessary steps, and minimisation of the need to infer system rules.

Visual accessibility is likewise an area in which seemingly minor design choices can substantially increase interaction costs for older adults. Research in the context of medical devices indicates that font size and button position influence task execution in older users, underscoring the importance of information hierarchy and the placement of primary actions [14]. Systematic reviews focusing on typography for older adults emphasise that legibility parameters (size, typeface, line spacing) should be considered in relation to task type (reading versus scanning) and users' perceptual conditions [15].

Interaction ergonomics (e.g., the size and spacing of touch targets) constitutes another central axis of design for older adults, particularly given reduced motor precision. Empirical work has shown that button size and spacing affect touch characteristics in older adults, translating into error rates and user comfort [16]. Similar conclusions have been reported for smartphone pointing performance, where older users exhibited poorer outcomes with small targets and dense layouts [17].

Existing studies synthesising design recommendations indicate that effective support for older adults typically requires a coherent set of interventions rather than a single change. For example, heuristic accessibility checklists for smartphone interfaces have been proposed, addressing legibility, consistency, visibility of system status, and the reduction of cognitive load [18]. Case studies of redesign for older adults demonstrate that usability improvements usually result from bundles of changes: simplifying information architecture, reducing visual clutter, and increasing the explicitness of controls [19]. Systematic reviews of mobile-app guidelines for older adults highlight recurrent recommendations, including large interactive elements, consistent navigation, predictable flows, and redundant information coding in critical locations [20]. Relatedly, calls for consolidating standards for the "ageing web" reinforce the need to connect design practice with accessibility requirements and established guidelines [21].

More recent empirical work on interface element characteristics suggests that specific design decisions can influence task performance in older adults, encouraging task-based prototyping and evaluation rather than reliance on design intuition alone [22]. In parallel, a strand of research has advanced in-action user support (e.g., real-time interactive guides), which may enhance older adults' web accessibility and reduce reliance on external assistance [23]. The present study therefore evaluated User Experience (UX) and the perception of key User Interface (UI) elements in an older adult cohort, with particular attention to how typographic legibility, contrast, colour scheme, and clickable-target size shape the completion of typical tasks in web applications. An additional objective was to identify interaction barriers and formulate design recommendations supporting inclusive solutions for older users.

The rest of this paper is organized as follows. Section II describes the methodology of the study. Section III presents the results. Section IV discusses the findings. Section V concludes the paper and outlines directions for future work.

II. METHODS AND METHODOLOGY

A mixed-methods design was employed, integrating quantitative and qualitative components. The quantitative component comprised (i) participants' subjective ratings of task difficulty on a five-point scale (1-5), (ii) a dichotomous record of whether moderator assistance was required (Yes/No), and (iii) dichotomous (Yes/No) evaluations of the legibility of key user interface elements. The qualitative component was based on participants' open-ended statements concerning experienced difficulties, accompanying feelings and emotions, and suggestions for interface improvements. Sampling was non-probabilistic: the core sample consisted of older adults (60+), complemented by a small group of younger participants included as a comparative reference in selected analyses. Data collection was conducted between 16 July 2025 and 3 February 2026.

A. *Procedure*

The procedure was structured and consisted of four stages. All sessions were conducted individually in the presence of a moderator, who observed task performance, recorded difficulty ratings and completion status, and provided clarification only when necessary. This moderated setting ensured consistent task administration across participants. First, participants completed an introductory questionnaire including demographic items as well as questions regarding technology use and physical health. Second, a task-based usability test was conducted across four modules/prototypes, described in detail in Section F (*Prototypes and tasks*); after each task, the moderator

recorded the perceived difficulty rating (1-5), whether assistance or clarification was necessary, and the task completion time, based on the recorded end-time stamp, enabling subsequent analysis of task efficiency. Third, after completing each module, participants filled in a post-module evaluation including Yes/No questions on font legibility, perceived colour scheme, and button size, supplemented with open-ended questions about difficulties and emotions. Finally, a closing questionnaire was administered, in which participants provided, among other responses, a subjective assessment of workload (e.g., tiring/demanding) and indicated the elements they perceived as easiest and most difficult.

B. *Variables and measures*

The study included a set of variables and measures capturing both task outcomes and subjective experience. The primary outcome was task independence, operationalised as task completion without assistance versus completion with moderator support. Additional measures comprised subjective task difficulty (1-5) and task duration, derived from recorded task end-time stamps. Perceptual evaluation of the interface (font legibility, colour scheme, and button size) was captured in a Yes/No format and enriched with participant comments. A separate category included qualitative data, i.e., the content of open-ended responses describing usability barriers and design recommendations.

C. *Characteristics of cohorts*

A total of 74 individuals participated in the study (N = 74), enabling a description of the sample in terms of basic demographic and social characteristics. The gender distribution indicated a predominance of women: 53 female participants (71.6%) and 21 male participants (28.4%). The age structure was dominated by older adults aged 60+ (65 participants; 87.8%), with the largest subgroup comprising participants aged 70-79 years (31 participants; 41.9%). Participants younger than 60 years constituted 9 individuals (12.2%) and served as a comparison group in the analyses. With respect to educational attainment, higher education was reported most frequently (38 participants; 51.4%), followed by secondary education (26 participants; 35.1%), indicating a relatively high level of educational capital in the study group.

Residence was characterised along two territorial dimensions: settlement size and participants' self-reported location descriptors. The sample structure indicated a predominance of urban residents, particularly from large urban centres, which should be taken into account when interpreting the results and generalising them to older populations with different territorial characteristics. Consequently, while the collected material supported inferences about typical barriers to interaction and interface perception among older adults, the scope of generalisation should consider potential differences arising from everyday-life context, the availability of digital services, and place-related variation in technological competence.

D. *Technology experience and functional limitations*

Most participants reported regular contact with technology: 64 individuals (86.5%) used technology daily. The most frequently reported device was a smartphone (55 participants; 74.3%), alongside a small group of non-users (4 participants; 5.4%). For interpretation, the high prevalence of self-reported health-related limitations was also salient: 50 participants (67.6%) indicated conditions or disabilities that could affect interaction with the user interface. Notably, visual difficulties predominated among the declared limitations (46 participants; 62.2% of the total sample). As limitation categories could overlap (e.g., concurrent visual impairment and motor difficulties), percentages do not sum to 100%.

E. *Tools*

The questionnaire and data-recording protocol were implemented using Google Forms. The research materials comprised: (1) the introductory questionnaire, (2) task scenarios for the four prototypes, (3) post-module evaluation questions, and (4) the closing questionnaire.

F. *Prototypes and tasks*

The test covered four web-application prototypes reflecting typical user activities:

1. Shopping list/form (4 tasks): adding, deleting, marking as purchased and filtering, and editing an item.

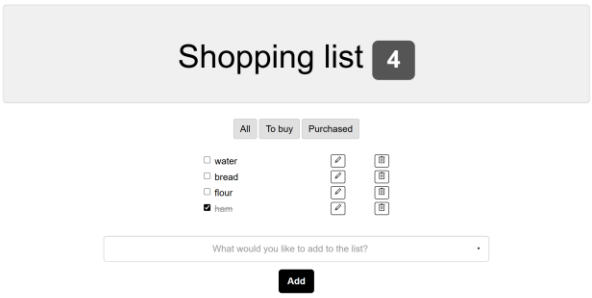


Figure 1. Shopping list interface.

2. Messenger (3 tasks): sending a message, adding a user to a group, and sending a photo with a caption and an emoji.

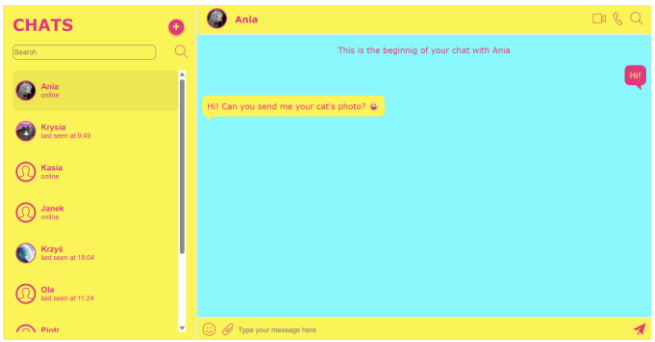


Figure 2. Messenger interface.

3. Login/password recovery/registration (3 tasks): logging in, password recovery via a code-based procedure, and account registration.

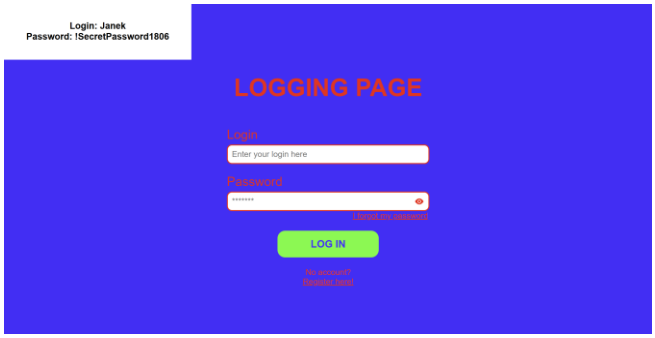


Figure 3. Login page interface.

4. Online shop (3 tasks): searching for a product, selecting a variant and adding it to the cart, and completing the purchase flow (delivery and payment).

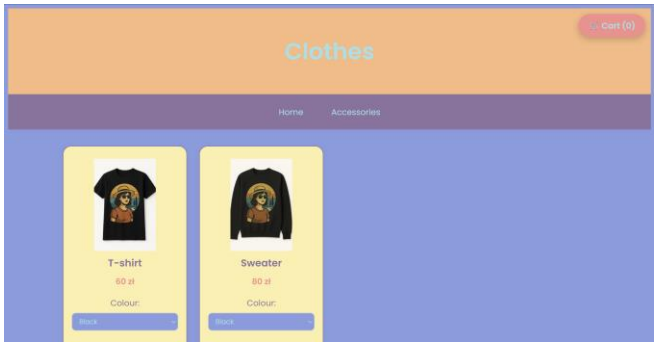


Figure 4. Online shop interface.

The prototypes were made available online (hosted, inter alia, on Render) and were presented in pairs of visual UI variants (e.g., versions differing in colour scheme and contrast). Within each module, participants worked with one of two variants: Shopping List (49 vs 24; 1 missing), Messenger (52 vs 22), Login (56 vs 18), and Shop (38 vs 36). This design enabled comparisons of UI perception and task-performance patterns (Fig. 1-4).

III. RESULTS

The analysis included N = 74 participants (women: 53; men: 21). Most participants were aged 60+ (n = 65), although a smaller group of participants aged <60 also took part (n = 9). Most participants lived in urban areas (n = 64), particularly in cities with more than 100,000 inhabitants (n = 40), whereas fewer lived in rural areas (n = 10). Self-reports regarding health-related limitations indicated that 50/74 participants reported at least one condition/limitation, including 46/74 reporting visual problems (e.g., “vision impairment”).

At the global level (all tasks combined), the mean difficulty rating on the 1-5 scale was M = 2.22, the median was Me = 2, and the standard deviation was SD = 1.09 (number of ratings = 958). Overall, 48.3% of task attempts were completed without moderator assistance (463/958), 41.5% required moderator assistance (398/958), and 10.1% involved the use of a scenario (97/958). A detailed breakdown of difficulty and the proportion of independent completions for each task is provided in Table I, while aggregated module-level results (including time) are presented in Table II.

For the time analysis (based on task completion timestamps, excluding records of 00:00:00 treated as missing), the median total session time among participants with complete time measurements (n = 66) was 29.0 minutes, with interquartile range IQR = 18.8 minutes (Q1 = 22.0; Q3 = 40.8). At the module level, tasks in the Login/Registration module took the longest (Me = 6.0 minutes; IQR = 7.0), whereas the shortest durations were observed for the Shopping List module (Me = 3.0 minutes; IQR = 4.0) and the Shop module (Me = 3.5 minutes; IQR = 4.8) (Table II).

TABLE I. TASK DIFFICULTY AND PERCENTAGE OF COMPLETION WITHOUT ASSISTANCE (PER TASK)

#	Module	Task	n (results)	M	Me	SD	No help (%)
1	List	List: add "cheese"	73	1.93	2	1.03	63.0
2	List	List: del"cheese"	73	2.00	2	1.08	52.1
3	List	List: mark"bread"	73	2.32	2	0.94	41.1
4	List	List: edit "water" → "juice"	73	2.48	2	1.20	38.4
5	Messenger	Send: "hello"	74	2.30	2	1.02	43.2
6	Messenger	add Jan to group	74	3.12	3	1.06	16.2
7	Messenger	Send pict	74	2.78	3	1.12	29.7
8	Login/ Registration	Login	74	2.04	2	0.94	43.2
9	Login/ Registration	Login pass recov	74	2.05	2	1.00	61.0
10	Login/ Registration	Login account reg	74	2.34	2	0.99	47.3
11	Shop	Find item	74	1.55	1	0.88	78.4
12	Shop	Shop: add item to basket	74	1.86	2	0.96	52.7
13	Shop	Shop: buy and pay	74	2.06	2	1.04	60.8

TABLE II. AGGREGATED RESULTS PER MODULE (DIFFICULTY, INDEPENDENCE, TIME)

Module	M	Me	SD	No help (%)	Time Me (min)	Time IQR (min)
Messenger	2.73	3	1.12	29.7	5.0	3.0
List	2.18	2	1.09	48.6	3.0	4.0
Login/Registration	2.14	2	0.98	50.9	6.0	7.0
Shop	1.82	2	0.99	64.0	3.5	4.8

A. *Results by module/prototype (Shopping List, Messenger, Login/Registration, Shop)*

In the Shopping List module, a moderate level of difficulty was observed (M = 2.18; Me = 2), and nearly half of task attempts were completed without assistance (48.6%). The greatest cognitive burden concerned tasks requiring

interpretation of controls related to filtering and content editing (tasks 3-4; M = 2.32-2.48), which co-occurred with a lower proportion of independent completions (38.4-41.1%).

The Messenger module yielded the highest mean difficulty ratings among all modules (M = 2.73; Me = 3) and the lowest share of independent completions (29.7%). Adding a user to a group was particularly problematic (task 6; M = 3.12; 16.2% without assistance), as was a compound action involving sending a photo with a caption and an emoji (task 7; M = 2.78; 29.7% without assistance). This pattern suggested that barriers in this prototype were more strongly related to locating functions and understanding interaction logic than to entering short text messages per se (task 5; M = 2.30).

In the Login/Registration module, difficulty was similar to that of the Shopping List module (M = 2.14; Me = 2), whereas completion time was the highest (Me = 6.0 minutes). The account creation task was the most demanding (task 10; M = 2.34). Notably, password recovery using a code was associated with relatively higher independence (task 9: 61.0% without assistance) than login (task 8: 43.2%), which may have resulted from a more unambiguous step-by-step sequence.

The Shop module was rated as the easiest (M = 1.82; Me = 2) and exhibited the highest independence (64.0% without assistance). The easiest task in the entire study was finding a product (task 11; M = 1.55; 78.4% without assistance). Difficulty increased for actions requiring multiple decisions (product variant selection and order finalisation; tasks 12-13), but ratings remained moderate overall (M = 1.86-2.06).

B. UI evaluation (font legibility, colours, button size; emojis in the messenger)

Evaluations of interface elements were predominantly positive, although clear differences emerged across modules. Overall (aggregated across all modules), font legibility was rated positively in 223/286 responses (78.0%), colour selection in 220/287 responses (76.7%), and button size in 232/287 responses (80.8%). In the Messenger module, emojis received positive ratings in 60/74 responses (81.1%).

The lowest proportion of "Yes" responses was observed for colour scheme in the Shop module (39/71; 54.9%), and a relatively low proportion was also found for colour scheme in the Messenger module (48/72; 66.7%). Negative comments in open-response fields related to UI questions included, among others, overly intense colours or insufficient contrast, as well as excessively small text and controls.

IV. DISCUSSION

Interpretation suggested that the observed difficulties resulted primarily from limited discoverability of functions and ambiguity of controls, and only secondarily from the intrinsic "content complexity" of the tasks. Across the entire study, difficulty was moderate (M = 2.22; Me = 2; SD = 1.09; n = 958 ratings), yet the proportion of completions without assistance was only 48.3% (463/958), indicating that many barriers emerged at the stage of selecting the appropriate action in the interface rather than during execution itself.

The Messenger module exhibited the highest difficulty among all modules (M = 2.73; Me = 3; SD = 1.12; n = 222 ratings) alongside the lowest independence (29.7% without assistance). The most difficult task was adding a group member (task 6), with M = 3.12 (Me = 3; SD = 1.06) and the lowest proportion of independent completions (16.2%). This profile was consistent with the qualitative material: an illustrative statement directly noted: "Finding the option to add to the group." The qualitative responses were consistent with the quantitative findings and helped clarify that the main barriers were not task goals themselves, but rather the discoverability of controls, uncertainty about next steps, and the perceptual accessibility of key interface elements.

A second source of difficulty was the increased number of steps and context switches in the task involving an attachment, caption, and emoji (task 7): M = 2.78 (Me = 3; SD = 1.12) with 29.7% independent completions. By comparison, the relatively simpler operation of sending a short text message (task 5) had lower difficulty (M = 2.30; Me = 2; SD = 1.02) and a higher proportion of independent completions (43.2%). This gradient (text → multi-step action) suggested that the key issue in the messenger was not the concept of "communication" itself, but rather navigation across functions, symbolism, and interface states. The Human-Computer Interaction (HCI) literature has noted that older adults are particularly sensitive to hidden actions and the cognitive demands of interface exploration, which translates into slower and less confident performance in navigation tasks.

Shop results indicated an apparent paradox: this module was simultaneously the least difficult (M = 1.82; Me = 2; SD = 0.99; n = 222) and the most independent (64.0% without assistance). The easiest task overall was product search (task 11): M = 1.55 (Me = 1; SD = 0.88) with 78.4% independent completions. In e-commerce contexts, users often achieve task success despite mediocre aesthetics due to well-established interaction schemas, which may explain high effectiveness and independence. By contrast, colour evaluation reflected perceptual comfort rather than "ability to complete," and qualitative comments included remarks about tiring colour combinations (e.g., "Too sharp a colour; in the long run it is tiring…").

Accordingly, the findings were interpreted as reflecting two separable UX dimensions: efficiency/feasibility of task completion (relatively high here) and perceptual accessibility (relatively lower). Accessibility requirements indicate that legibility depends, among other factors, on minimum contrast thresholds (e.g., 4.5:1 for standard text).

At the perceptual and ergonomic level, efforts should focus on modules with the weakest UI evaluations. Accessibility standards, including criteria for text contrast (e.g., 4.5:1 for standard text), provide clear directions and measurable thresholds for auditing such changes. Additionally, given the comments concerning ergonomics and clickability (22 mentions in reported difficulties and 22 in improvement suggestions), increasing target sizes and/or spacing was justified, consistent with accessibility-oriented approaches to minimum target size (including 24×24 CSS px in Web Content Accessibility Guidelines 2.2 for certain

cases). Finally, considering the substantial age-related differences (independence 38.5% vs 92.3% and time 32.0 vs 17.0 minutes; $p < 0.001$), personalisation features supporting readability (larger text, high contrast) and explicit labelling should constitute a permanent interface component, given their potential benefit for the most sensitive user group.

A. Study limitations

Several variables were self-reported (e.g., health-related limitations), which may have reduced the validity of classifying participants as "with limitations" versus "without". UI/CSS variants were not fully randomised and balanced; therefore, differences in UI ratings may have been partly modulated by expectations and aesthetic preferences. Furthermore, time measurement based on task end-time stamps was sensitive to missing values, and the presence of a moderator may have reduced variability in outcomes, particularly in tasks where assistance was frequent. The study did not include a formal cognitive screening measure (e.g., the Montreal Cognitive Assessment or Mini-Mental State Examination), which limits the precision of interpreting whether some observed difficulties were primarily perceptual, functional, or partly cognitive in nature. Participant characterization relied on self-reported information regarding physical health, digital skills, and everyday technology use. The study design and sample size were sufficient to identify the main interaction difficulties associated with the perceptual and functional demands of the tasks. Future studies should include standardized cognitive assessment to strengthen interpretability.

V. CONCLUSION AND FUTURE WORK

As part of the completed work, a full set of research materials was developed (task scenarios, a UX/UI evaluation questionnaire, and visual interface variants), and usability tests were conducted for four prototypes: Shopping List, Messenger, Login/Registration, and Shop. The study provided a coherent account of how older adults perform tasks in web-based interfaces and which UI elements constitute genuine barriers to independent task completion. Quantitatively, overall task difficulty was moderate ($M = 2.22$; $Me = 2$; $SD = 1.09$), yet independence proved limited. Only 48.3% of task attempts were completed without assistance. This indicated that the primary obstacle was not the substantive content of the tasks but rather the stage of "locating the appropriate function" and understanding how to execute the next step within the interface.

The most consequential practical insight was that the weakest outcomes were observed in the Messenger prototype (module-level $M = 2.73$; independence 29.7%), particularly for the task of adding a user to a group ($M = 3.12$ and only 16.2% completed without assistance). The qualitative data consistently pointed to a discoverability issue ("Finding the option to add to the group."), which mapped directly onto Jakob Nielsen's heuristics (recognition rather than recall; consistency and standards; visibility of system status). At the same time, the Shop prototype achieved the best performance parameters (module-level $M = 1.82$; independence 64.0%) while receiving the weakest evaluation of colour scheme (54.9% positive responses). This pattern suggested a separation of two dimensions: participants were able to accomplish task goals by relying on familiar e-commerce conventions, yet they experienced reduced perceptual comfort. This, in turn, provided a direct rationale for adjustments aligned with the Web Content Accessibility Guidelines (WCAG) developed by the W3C (notably with respect to contrast and the legibility of components).

Overall, the findings indicated that, when designing interfaces for older adults, priority should be given to solutions that increase discoverability and the unambiguity of key actions. Accordingly, the following should be treated as critical: (a) exposing primary actions as visible, clearly labelled buttons (rather than unlabeled icons or hidden functions), (b) guiding users through multi-step tasks (stepwise structure, clear process state, and confirmations), (c) reducing the cost of errors through intelligible messages and timely feedback, and (d) improving interaction ergonomics (larger clickable targets and more salient component states). In form-based contexts (login/registration/checkout), mechanisms that support data entry (e.g., "show password" and real-time validation) were practically important, as they reduced hesitations and the need for assistance. From an accessibility perspective, removing solutions that reduced legibility (mainly in the Shop module) was essential, because even when tasks could be completed successfully, such design choices lowered comfort and increased the risk of errors during extended interaction.

In subsequent iterations, development should proceed along two parallel tracks: refining interaction structure (navigation, labels, and function visibility) while simultaneously improving the perceptual layer (contrast, visual hierarchy, and element sizing). In practice, this entails starting with "high-impact" fixes in areas with the lowest independence (notably within the messenger), and only then refining aesthetics which (as the results demonstrated) may improve UI ratings but cannot substitute for structural changes. Future research should extend usability testing with more fine-grained process measures (errors, backtracking, misclicks, and step-level time), and randomise assignment to UI variants. Additionally, one should focus on a deeper qualitative analysis of participants' open-ended responses and combine user studies with a formal WCAG conformance audit. In addition, including a formal cognitive screening measure (e.g., the Montreal Cognitive Assessment) would strengthen the study by providing a standardized measure of cognitive abilities that may affect user interface interaction. It is also advisable to test across devices (smartphone vs desktop) and in more naturalistic conditions, as certain barriers for older adults may become more pronounced in real contexts of use (visual fatigue, distractions, and time pressure). Ultimately, the completed work underscored that the decisive success factor was not a one-off "prototype assessment" but an iterative, evidence-based design cycle: identifying points of breakdown, implementing improvements, retesting, and verifying whether independence increases where it previously declined.

REFERENCES

[1] J. R. Beard and D. E. Bloom, "Towards a comprehensive public health response to population ageing," The Lancet, vol. 385, no. 9968, pp. 658–661, 2015.

[2] T. L. Mitzner et al., "Older adults talk technology: Technology usage and attitudes," Computers in Human Behavior, vol. 26, no. 6, pp. 1710–1721, 2010.

[3] Q. Ma, K. Chen, A. H. S. Chan, and P. L. Teh, "The influence of technology acceptance factors on older adults' adoption of smart home services," International Journal of Human-Computer Interaction, pp. 1–12, 2021.

[4] T. A. Salthouse, "The processing-speed theory of adult age differences in cognition," Psychological Review, vol. 103, no. 3, pp. 403–428, 1996.

[5] C. Owsley, "Aging and vision," Vision Research, vol. 51, no. 13, pp. 1610–1622, 2011.

[6] N. Wagner, K. Hassanein, and M. Head, "The impact of age on website usability," Computers in Human Behavior, vol. 37, pp. 270–282, 2014.

[7] A. Chevalier, A. Dommes, and J.-C. Marquié, "Strategy and accuracy during information search on the Web: Effects of age and complexity of the search questions," Ageing & Society, pp. 1–25, 2013.

[8] S. Leuthold, P. Schmutz, J. A. Bargas-Avila, A. N. Tuch, and K. Opwis, "Vertical versus dynamic menus on the world wide web: Eye tracking study measuring the influence of menu design and task complexity on user performance and subjective preference," Computers in Human Behavior, vol. 27, no. 1, pp. 459–472, 2011.

[9] D. Castilla et al., "Effect of Web navigation style in elderly users," Computers in Human Behavior, pp. 1–10, 2016.

[10] J. A. Dosso and A. Chevalier, "How do older adults process icons? The influence of semantic distance and text labels," Educational Gerontology, pp. 1–12, 2021.

[11] R. Leung, J. McGrenere, and P. Graf, "Age-related differences in the initial usability of mobile device icons," Behaviour & Information Technology, pp. 1–12, 2011.

[12] S. Reddy, A. Chattopadhyay, and K. Moffatt, "The effects of redundancy in user-interface design on older users," International Journal of Human-Computer Studies, pp. 1–12, 2020.

[13] K. Renaud and J. Ramsay, "Now what was that password again? A more flexible way of identifying and authenticating our seniors," Behaviour & Information Technology, pp. 1–12, 2007.

[14] Y.-Y. Yeh, Y.-J. Chang, and W.-C. Li, "Impact of button position and touchscreen font size on the healthcare device operation by older adults," Heliyon, vol. 6, no. 6, pp. e04147–e04147, 2020.

[15] W. Hou, A. Li, C. Lin, and S. Nie, "How to design font size for older adults: A systematic literature review," Frontiers in Psychology, vol. 13, pp. 931646–931646, 2022.

[16] M. E. Sesto, C. B. Irwin, K.-B. Chen, A. O. Chourasia, and D. A. Wiegmann, "Effect of touch screen button size and spacing on touch characteristics of older adults," Human Factors, vol. 54, no. 3, pp. 425–436, 2012.

[17] H. Hwangbo, S. H. Yoon, B. S. Jin, Y. S. Han, and Y. G. Ji, "A study of pointing performance of elderly users on smartphones," International Journal of Human-Computer Interaction, pp. 1–12, 2013.

[18] N. Mi, L. A. Cavuoto, K. Benson, T. Smith-Jackson, and M. A. Nussbaum, "A heuristic checklist for an accessible smartphone interface design," Universal Access in the Information Society, pp. 1–12, 2014.

[19] C. Patsoule and P. Koutsabasis, "Redesigning websites for older adults: A case study," Behaviour & Information Technology, pp. 1–12, 2014.

[20] L. Gomez-Hernandez, I. Garcia-Magariño, and S. Alcaraz, "Design guidelines of mobile apps for older adults: Systematic review and thematic analysis," JMIR mHealth and uHealth, vol. 11, pp. e43186–e43186, 2023.

[21] A. C. Cavender and J. P. Bigham, "Towards standards for the aging web," Universal Access in the Information Society, pp. 1–12, 2011.

[22] J. Zhou, H. Yuan, M. Huang, S. Zhang, and I. K. Kaner, "The impact of interface design element features on task performance in older adults," International Journal of Environmental Research and Public Health, vol. 19, no. 15, pp. 9251–9251, 2022.

[23] M. Taieb-Maimon, A. Cohen, and S. Shahar, "Improving older adults' accessibility to the web using real-time online interactive guides," International Journal of Human-Computer Studies, pp. 1–12, 2022.