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,

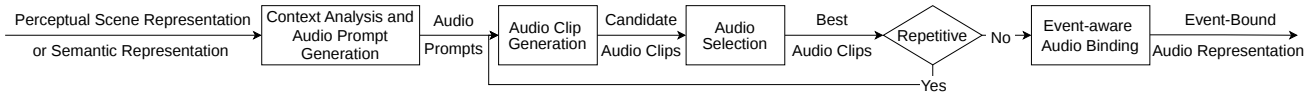


Figure 1. Proposed System Architecture.

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

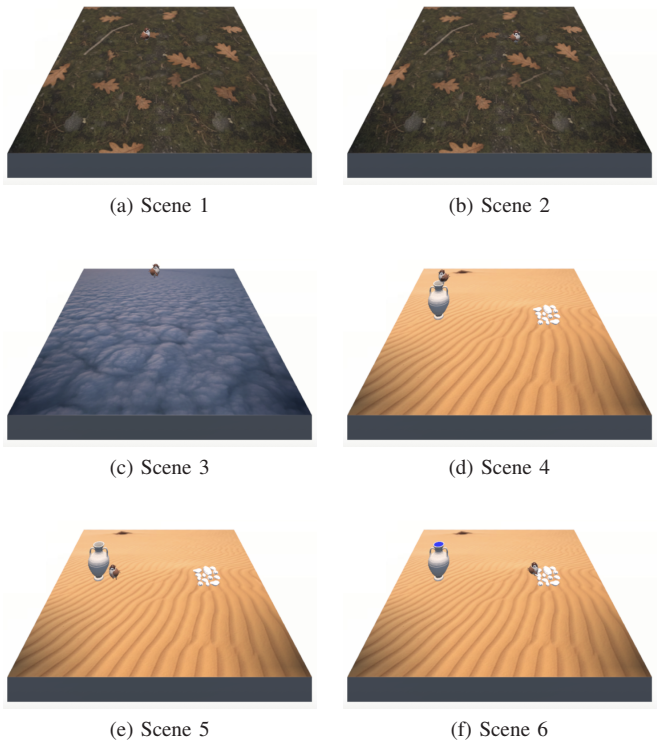
In this study, a Meta Quest headset was chosen as the primary interaction device due to its superior passthrough capabilities, high-quality visual fidelity, and its ability to operate as a standalone unit. A local workstation equipped with an NVIDIA RTX 5090 GPU was preferred as the generative AI pipeline 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/.