Connotation and 3D Modeling from Limited, Raw Textual Descriptions

Ella Berman
Computer Science
Grinnell College
Grinnell, USA
email: bermanel@grinnell.edu

Mahiro Noda
Biological Chemistry
Grinnell College
Grinnell, USA
email: nodamahi@grinnell.edu

Kailee Shermak
Sociology
Grinnell College
Grinnell, USA
email: shermakk@grinnell.edu

Zi Ye
Computer Science
Grinnell College
Grinnell, USA
email: yezi@grinnell.edu

David Rothfusz
Computer Science
Grinnell College
Grinnell, USA
email: rothfusz2@grinnell.edu

Jiayi Chen
Risk Management
Pennsylvania State University
State College, USA
email: jjc7655@psu.edu

Thammik Leungpathomaram
Computer Science
Grinnell College
Grinnell, USA
email: leungpat@grinnell.edu

Shuta Shibue
Computer Science
Grinnell College
Grinnell, USA
email: shibuesh@grinnell.edu

Chenxing Liu
Computer Science
Grinnell College
Grinnell, USA
email: liutommy@grinnell.edu

Fernanda Eliott
Computer Science
Grinnell College
Grinnell, USA
email: eliottfe@grinnell.edu

***Abstract*—In emotion-rich contexts, how do you comprehend the meaning behind your perception? This exploratory multi-phase project seeks to gather insights into how abstraction and emotions travel different spaces. The investigated spaces are: images, human-made textual descriptions, mental models, and 3D (three-dimensional) scenes. In previous work, we described our project idea; here, we detail our pilot for Project Phases 1 and 2, in which a team first creates *raw descriptions* of memes (in addition to creating detailed descriptions and the Observer-Centered Dataset Attributes) so that the Phase 2 team, so-called modelers, read the raw descriptions and build a 3D scene as accurately and faithfully as possible to the meaning behind their perception of the description. Raw descriptions are created by "unsaying" (*i.e.*, by identifying and removing the *unsaid elements* from a detailed description); and therefore, are more vague than alt-text since raw description purposefully leave details out (to see if the modelers "got" the message in spite of gaps). We designed a diagram to illustrate how modelers decided a 3D scene was complete, called "*Accuracy* and *Faithfulness* Gateways Diagram", detailed here. We launched this project as a pilot to inform our methods to ensure objectivity and replicability. A key challenge in identifying the *unsaid elements* comes from making the implicit explicit, and our approach to accomplishing that can inspire frameworks for detecting biases and microaggressions in visual content and help to create cultural sensitivity awareness. We pinpoint our work's social impact applications, which will be detailed in future work. Finally, investigating abstraction within and across spaces is notably relevant right now. In fact, as more people interact with generative AI platforms (such as AutoGen or Vertex AI), prompt designers deal with and add abstraction into a prompt as they instruct an AI-powered model to behave in certain ways.**

***Keywords-abstraction; connotation; memes; 3D-modeling; textual descriptions*.**

I. Introduction

Connotation offers versatile approaches to communication. It can support argot languages or even hidden messages and expression against oppression, such as in Brazilian songs known for their response to dictatorship, e.g., "Sinal Fechado" (Paulinho da Viola, 1969), "Comportamento Geral" (Gonzaguinha, 1973), "Mosca na Sopa" (Raul Seixas, 1973), and "Cálice" (Chico Buarque and Gilberto Gil, 1978). These songs illustrate how abstraction, emotions, and connotation blended together can help to create complex messages.

Besides songs, poems, and others, memes widely shared online rely on connotation to deliver a message. Here, we build on our previous work [1] on memes [2], which are a "form of media communicating a thought or idea through some shared understanding" [3]. Memes "often hide complex, abstract reasoning mechanisms behind their humorous front" [3]. Connotative meanings refer to the "associations, overtones, and feel that a concept has, rather than what it refers to explicitly (or denotes, hence denotative meaning). Two words with the same reference or definition may have different connotations" [4]. "In writing, you can choose a word that has a clear denotation and few connotations—a word like tall or quiet—or you can choose a word that connotes something more—like statuesque or tranquil" [5].

But how do we *comprehend* the overtones and overall meaning behind our perception? More importantly, how do we play with connotation to create hidden messages that others can understand? Emotion knowledge enables children to identify emotions in themselves and others and facilitates emotion recognition in complex social situations. Thus, social-cognitive processes, such as theory of mind (ToM), may contribute to developing emotion knowledge by helping children comprehend the emotion expression's variability across individuals and situations [6]. Theory of mind can be defined as the "human ability to ascribe mental states, intentions, and feelings to other human agents and to oneself" [7].

When telling a story, we do not provide every single detail; we expect others to fill in the gaps and evoke mental models consistent with the story. E.g., if the story involves a library, it may be associated with a quiet place filled with books and other associated behaviors/rules/objects (if those clues correspond to one's cultural experiences). Mental models are "internal representations of the external world consisting of causal beliefs that help individuals deduce what will happen in a particular situation" [8]. Meanwhile, emotional mental models cover emotions and feelings connected to mental models: "Mental models cause certain expectations/thoughts of how things should look like/work and connect certain emotions with this. Consequently, a mental model is a cognitive and an emotional framework in the brain, influenced by person's personality (genes) and the environment including social variables" [9] (see [10] for a theoretical review on the role of shared mental models in human-AI teams).

Yet, what if you wanted to somehow architect similar skills into a machine? Suppose you wanted to model an *emotion-driven Artificial Intelligence* (AI) system able to cooperate purposefully with humans and other biological creatures. You may ask: "How to encode abstract and emotion-rich contexts into an AI system's mental models and assist its decision-making process?" That is one of our main research goals, although we wonder: 1. Would that enable a more holistic contextual evaluation and better-informed AI's decision-making process? and 2. If one is to architect an AI system modeled after emotions and feelings, should it be influenced in any way by task-irrelevant emotion stimuli [11]? If so, what does that look like?

We use the term *emotion-rich* to convey emotional messages that the human senses can perceive (which could be translated, in robotics, for example, *via* the robot's sensors [12]). Our lab builds cognitively inspired computational models, and we are designing a computational architecture that uses traditional Reinforcement Learning techniques (RL) [13] and models emotions and moral processes [14], [15], [16]. RL is "learning what to do—how to map situations to actions – so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning" [13].

To tackle that goal, we decided to examine humans first and narrowed our questions to "How do abstraction and emotions travel different spaces?" In [1], we present our multi-phase multi-team project idea to investigate that question, which explores distinct spaces: images of memes, human-made textual descriptions, mental models, and 3D scenes (see figs. 1 and 3).

A short overview of the project's Phases 0-2 is as follows: **Phase 0.** Manually collect images of memes. **Phase 1.** Analyze the images and create a database with 1. raw textual image descriptions, 2. detailed textual image descriptions, and 3. a set of attributes to analyze the memes, resulting in 4. the Observer-Centered Dataset. **Phase 2.** Create 3D scenes from the raw descriptions. We launched this project as a pilot, enabling us to create methods and gateways across and within phases using an *ad-hoc* and data-driven approach. Therefore, we will rerun the project once all phases are consolidated.

We hypothesize that by investigating how abstraction and emotions travel different spaces, we will gather insights into key elements for producing a consistent and holistic understanding of complex, abstract messages and connotations – finally, getting a better picture of how and what to model in an *emotion-driven* AI system that uses traditional RL techniques.

"Abstraction enables humans to distill a cascade of sensory experiences into a useful format for making sense of the world and generalizing to new contexts" [17]. Highlighting that knowledge exists at multiple levels of abstraction, Reed [18] provides a taxonomic analysis of abstraction that examines three senses of abstraction: "(a) an abstract entity is a concept that has no material referent, (b) abstraction focuses on only some attributes of multicomponent stimuli, and (c) an abstract idea applies to many particular instances of a category." Forward, Ho et al. [19] illustrate the importance of abstraction for AI and RL frameworks: abstractions are important for adaptive decision-making, e.g., abstractions guide exploration and generalization, facilitate efficient trade-offs, and simplify computation. Note that providing AI modeling details and identifying different uses of abstraction (e.g., visual abstraction, relational abstraction, temporal abstraction) falls out of our scope; the same goes for disassociating abstraction from emotions and connotation, but we provide definitions in the Glossary, see Appendix A.

We use "abstraction" as a blanket term for *something untied from concrete elements*, which covers both abstract words (e.g., *honor* and *freedom*) and/or dealing with abstraction (abstract problem-solving)."There is reason to assume that abstract concepts are more sensitive to contextual constraints than concrete concepts" [20] and "Statistically, abstract words are more emotionally valenced than are concrete words" [21]. Finally, challenges from interpreting an image that poses multiple emotional mental models drove us to *networked emotions* [22], helping us deal with the messy layers of emotions in meme comprehension (see Section VI). Still, we understand that humor "is a universal phenomenon but is also culturally tinted", and "some humor coping strategies may have different connotations under different cultural backgrounds, which would directly impact how humor is used in different cultural backgrounds" [23].

Hence, in our research project, abstraction intercepts emotions to the extent that, similar to the *Telephone Game*, people form different mental/emotional models as a message travels through them – and aspects of comprehending a message from a raw description are abstract and open-ended. (The Telephone Game starts with a line, or circle, of people; the first person in line privately receives or creates a message and whispers it to the next person in line. The process repeats until the end of the line; finally, the last person shares the message out loud to check if the original message accurately made its way through the whispers.) We acknowledge this research's challenges and limitations: abstraction and emotions are multifaceted topics, whose combination with technologies brings even more layers. Still, in spite of the challenges, our research outcomes motivate directions for social impact tools (see Section V), in addition to insights for cognitively-inspired AI modeling and dealing with networked emotions.

A note on Phase 0 image collection. We identified the digital space as a good fit for our purposes given the emotions' social nature and their central place in digital cultures: "The socially mediated communication of emotion is intricately linked to the social textures of networking technologies" [24]. This led us to images that convey jokes or metaphors characteristic of memes since they often hide abstract reasoning mechanisms and given their ability to be either easily understood or learned through examples, making them a viable format for idea transfer [3]. Memes can be used for various purposes, e.g., to entertain, instruct, or express political views and expose others to political content [25]; for simplicity, we target their use within humor or entertainment.

Our contributions are to:

1) Detail a methodological breakdown to textually describe memes (or similar images). Albeit the raw descriptions' purpose is to check what message will be encoded by the 3D modelers, our methods are still relevant to others working with textual description tools. We provide two kinds of image descriptions: detailed and raw (the latter is created *via* the identification and removal of *unsaid elements* from detailed descriptions).
2) Visually organize and contextualize a set of attributes to inform the analysis of memes. The attributes take into account the observer's perspective and networked emotions; we call those the Observer-Centered Dataset attributes.
3) Illustrate how a 3D modeler deals with limited, raw textual descriptions and decides whether a 3D scene is complete. Two gateways (*accuracy* and *faithfulness*) are identified, and they serve as a checkpoint for evaluating and inspecting the 3D model before it is complete, becoming a so-called 3D scene – modelers decide whether the scene sufficiently reflects the observable visual features of the mental models they created based on the raw description.
4) Provide a glossary and a multidisciplinary literature review as we situate our research.

Although it is not our claim that our Gateways Diagram covers the 3D modeling process in general, we do hope this work can a) benefit the decision-making process of similar 3D modeling initiatives and b) inform the creation of richer alt-text tools or even other assistive technologies, such as 3D modeling tools for the visually impaired – ultimately assisting in creating 3D printing blueprints; see resources such as Round Table on Information Access for People with Print Disabilities [26], See3D [27], and the Accessible Graphics hub [28]. More specifically, we hope to inform richer assistive technology tools' creation as we call attention to a tension between the *unsaid elements* (in a description or explanation, for example) and the audience's *assumed elements*. Therefore, in how abstraction and emotions make their way through different spaces, especially when a message is heavy on connotation.

That tension, worked through our proposed descriptions' breakdown, dataset, and diagram, can **inform the development of AI tools better equipped to deal with abstraction** (e.g., using Generative AI tools for creating a 3D scene from vague prompts), connotations, and cultural elements, aiming for culturally sensitive human-machine interaction and output/outcomes. Deviating from memes, one may ponder upon the everyday news on AI achievements, which play with connotations and anthropomorphism [29].

This work is organized as follows: introduction in Section I, followed by our methods, which split into two: Project's Phase 1 details in Section II, followed by Phase 2 in Section III. We show our research outcomes, such as the Observer-Centered Dataset attributes and a sample of 3D scenes' static images in Section IV, discussion in Section V, followed by related literature in Section VI, and conclusion in Section VII.

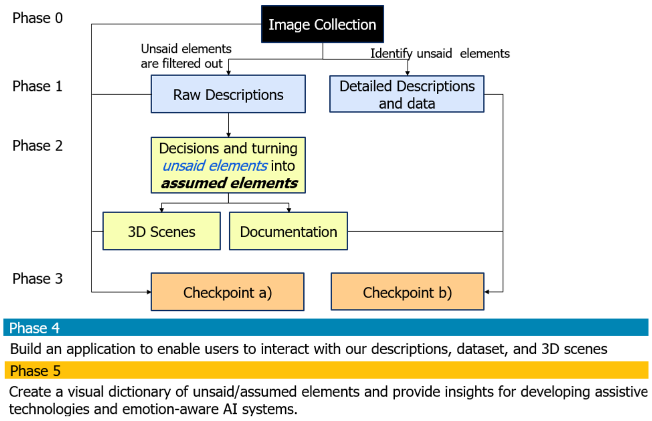

Phase 0
Image Collection
Unsaid elements are filtered out
Identify unsaid elements
Phase 1
Raw Descriptions
Detailed Descriptions and data
Phase 2
Decisions and turning *unsaid elements* into ***assumed elements***
3D Scenes
Documentation
Phase 3
Checkpoint a)
Checkpoint b)
Phase 4
Build an application to enable users to interact with our descriptions, dataset, and 3D scenes
Phase 5
Create a visual dictionary of unsaid/assumed elements and provide insights for developing assistive technologies and emotion-aware AI systems.

Fig. 1. Project Overview. Project phases 1 and 2 are this manuscript's focus: producing and encoding raw textual descriptions into a 3D scene (and documenting the decision-making process).

II. Background and Methods

A summary of all project phases is given next, and an overview in Figure 1. As Figure 2 shows, Phase 1 covers human-made textual descriptions followed by the memes' attributes identification; then, Figure 3 illustrates, for Phases

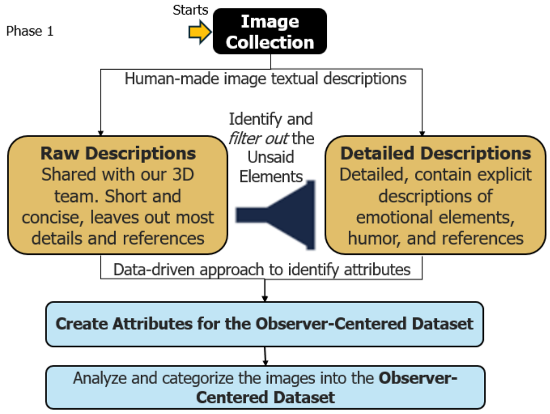

Phase 1
Starts
Image Collection
Human-made image textual descriptions
Raw Descriptions
Shared with our 3D team. Short and concise, leaves out most details and references
Identify and *filter out* the Unsaid Elements
Detailed Descriptions
Detailed, contain explicit descriptions of emotional elements, humor, and references
Data-driven approach to identify attributes
Create Attributes for the Observer-Centered Dataset
Analyze and categorize the images into the **Observer-Centered Dataset**

Fig. 2. Phase 1 starts by accessing Phase 0 meme collection to create detailed and raw descriptions and the Observer-Centered Dataset attributes. Finally, the memes are categorized into the dataset.

0 – 3, in what ways we envision abstraction and emotions traveling spaces.

Note that **each phase has a unique team with a strict non-sharing policy**: everything a team produces is kept within the team only, unless at the Phase's cycle conclusion when we share our results with the community – for consistency, we use "we" across this manuscript as we integrate the teams' results. We launched this project as a *pilot* to establish procedures and methods to ensure objectivity and replicability. To that end, we investigated related literature combined with a data-driven approach (see Section IV-C).

Project Phases in a Nutshell:

- Phase 0, **Image (Meme) Collection**. Manually collect images that hide complex, abstract reasoning mechanisms characteristic in memes – any political or hateful content is forbidden. The $\approx$ 400 memes were collected from social platforms such as Instagram and WhatsApp, and they cover two countries (the USA and Brazil), as we sought to investigate more than one country. As expected, humorous memes are abundant online; therefore, our collection sits within humor, with just a few exceptions (see Figure 7). Still, some memes seek to evoke humor from negative tones, given the use of self-deprecating humor.
- Phase 1, **Database**. Write raw and detailed image descriptions in English and categorize the memes in a dataset called the Observer-Centered Dataset. Feed the *Phase 2* team with raw descriptions – i.e., leaving details out, by which we named *unsaid elements*. Example of a raw description: *A* soaking wet cat sits inside a sink with open eyes that pop out. There is a leading text: "I leave the bathroom shaking cold, and the person asks:" follow-up text: "Are you cold?' Nope, a ghost is entering me."
- Phase 2, **3D Scenes and Decisions**. Without access to the memes, interpret and encode the raw image descriptions into a 3D scene using a tool such as Blender and document the decision-making process. *Unsaid elements* can either be on the a) concrete side, e.g., it mentions a cat on a sink but no details about the fur's color or the sink's shape, size, and material/color, or the environment; or b) more abstract and emotionally-tinted, e.g., 3D modelers may reflect: "this seems to imply discomfort; is it supposed to be humorous?" Hence, 3D modelers have to fill in the gaps and make decisions to build a 3D scene, by which we call *assumed elements*. Therefore, *unsaid elements* from Phase 1 become *missing elements* in Phase 2, as modelers identify that something is missing in the description and subsequently make assumptions of how to fill in the gaps, resulting in the *assumed elements* – see Figure 3.
- Phase 3, **Checkpoint**. Compare: a) raw descriptions, memes, and 3D scenes (focus on the 3D scenes' canonical view, which should coincide with the front view), and b) unsaid with *assumed elements* and documentation. Examine how/if those differ, analyze our dataset, and document what we learned about abstraction/emotions across spaces.
- Phase 4, **Software application**. We will apply human-centered design (HCD) practices to develop a software application, e.g., Shiny app [30], to enable people to interact with our project's data and outcomes.
- In Phase 5, we will investigate in what ways our findings can inform the development of a **Visual Dictionary** that refers back to emotions, abstraction, and connotative meanings – we hypothesize this work will provide valuable insights for fostering assistive technologies and modeling *emotion-driven* AI systems.

To recap, our overall goal is to architect an *emotion-driven* AI system; to inform our processes, we are investigating how abstraction and emotions travel through spaces. The comparison *Unsaid Elements* (from Phase 1) with the *Missing* and *Assumed Elements* (from Phase 2) will be key in investigating/mapping the different elements people may combine to interpret abstract messages. Likewise, our dataset will enable us to filter and group meme details in various ways (such as comparing memes from Brazil and the US), contributing insights for AI modeling, such as dealing with humor and connotation in different cultures. We detail Phase 1 next.

A. *Textual Descriptions and a Narrative Approach*

Phase 1's overall goal is to devise a method for creating descriptions that are *as raw as possible* but still retain the meme's overall meaning. Abstract and emotionally-tinted elements are a) included in a detailed description but b) filtered out from a raw description. Therefore, it involves dealing with both connotative and denotative meanings. According to Schnotz [31], "text comprehension includes the formation of at least three kinds of mental representations: a text surface representation, a propositional representation, and a mental model" and "Inferences are an integral component of text comprehension, because the author of a text omits information which can be easily completed by the reader." More than that, through the *unsaid elements*, we seek to identify and omit more information than one typically would to convey a message.

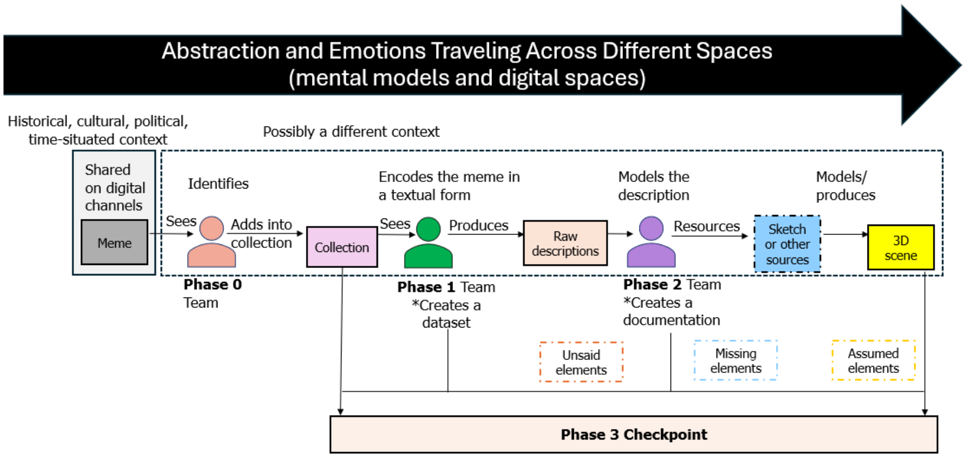

Abstraction and Emotions Traveling Across Different Spaces
(mental models and digital spaces)
Historical, cultural, political, time-situated context
Possibly a different context
Shared on digital channels
Meme
Identifies
Sees
Adds into collection
Collection
Encodes the meme in a textual form
Sees
Produces
Raw descriptions
Models the description
Resources
Sketch or other sources
Models/ produces
3D scene
Phase 0 Team
Phase 1 Team *Creates a dataset
Phase 2 Team *Creates a documentation
Unsaid elements
Missing elements
Assumed elements
Phase 3 Checkpoint

Fig. 3. An illustration of how abstraction and emotions travel different spaces within the project Phases 0-3. *Unsaid*, *missing*, and *assumed elements* are identified within dashed boxes.

Phase 1 includes creating the Observer-Centered Dataset attributes (see Section IV-C), which are split into three dimensions: 1. Concrete Design, 2. Blend, and 3. Emotional Design. The Concrete Design dimension covers objective elements that facilitate an image identification; the Blend focuses on the image's observer, whereas the Emotional Design on an image's messy layers of emotions (see Section VI). As we place the observer as an image's target, the Blend dimension blends together the three dimensions, see Figure 9. Finally, our dataset's **focus on *the observer* and *networked emotions* is its most distinguishing feature** – we will analyze our dataset in future work.

We noticed key linkages with assistive technologies during our investigation of image description methods. Hence, our work is inspired by the *Accessible Publishing* [32] advice on how to write accessible image descriptions for people with *print disability*, "which includes individuals who are blind or visually impaired, people with cognitive and comprehension disabilities, and persons who have physical mobility challenges" (see in Section II-C our parallel with alt-text).

In Figure 4, we illustrate our process for a human interacting with an image aiming to describe it: we depict the process as a ladder, starting from the initial viewing of the image and collecting information as we move up to finally reach a total understanding at the top. We follow a narrative approach to describing the images as we combine the guidance from three resources, all explained below:

1) Methods and advice from [32].
2) Heuristics from [33] to capture the whole image's meaning in an image description.
3) Advice from [34] on how to describe memes.

Our preliminary image descriptions' version was similar to the *Accessible Publishing's* [32] long descriptions, which are detailed textual descriptions that can be "several paragraphs long and/or may contain other elements such as Tables and lists. This technique is generally used for complex images where spatial information needs to be conveyed to the reader such as maps, graphs, and diagrams. Sometimes called extended description, these descriptions are too long and complex for alt-text."

However, as our process evolved, we felt the need to create two kinds of descriptions: raw and detailed – giving rise to the *unsaid elements*, which are the elements to be removed/filtered out of a detailed description to create a raw description. In other words, the elements to be "unsaid". Finally, instead of describing complex maps/diagrams, our detailed descriptions aim to describe images that rely on abstract reasoning mechanisms characteristic of memes.

Nganji and colleagues [33] propose heuristics to capture the whole meaning and description of the image. It addresses the "who", "what", "when", "where", and "how" of the image: "*who* asks the questions relating to the people in the image, while *what* relates to other non-human objects including buildings, trees, automobile, etc. including their descriptions such as colour. *When* on the other hand asks questions related to time such as when the picture was taken (time, season, etc.) while *where* seeks to find out the location such as where the image was taken, the positions of various objects in the image, etc. *How* relates to actions, emotions, etc". The authors propose an Image Description Assessment Tool, which is a Java-based tool for assessing how well an image description matches the actual content of the image on the web (it also provides a speech interface so that people can listen to the description of an uploaded image); thus, weights are applied to the heuristics categories to determine how close a description is to the original image.

To conclude, we are inspired by Lewis [34] specific strategies for describing memes, summarized as follows: 1. write any text that precedes the image; 2. describe the subject briefly (who or what is depicted) and 3. note any alterations in the subject's appearance, if relevant. 4. explain what the subject is doing; 5. Be explicit about the punchline.

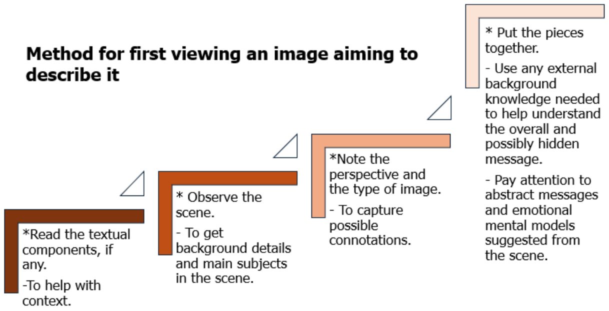

Method for first viewing an image aiming to describe it
*Read the textual components, if any.
-To help with context.
* Observe the scene.
- To get background details and main subjects in the scene.
*Note the perspective and the type of image.
- To capture possible connotations.
* Put the pieces together.
- Use any external background knowledge needed to help understand the overall and possibly hidden message.
- Pay attention to abstract messages and emotional mental models suggested from the scene.

Fig. 4. We depict, as a ladder, our process for interacting with an image aiming to fully describe it. The process starts from the lower step and finishes at the top.

B. Detailed Textual Descriptions

Our detailed descriptions are created according to the instructions below (it would be interesting to run human studies and investigate if these instructions can assist in creating accessible image description tools):

General Instructions. Prioritize describing an image simultaneously with viewing it for the first time to ensure a fresh perspective and avoid leaving details out. Write a description as a "building up" process, ultimately leading to an overall understanding of the image, as Figure 4 shows. Be as detailed and precise as possible with your visual and emotional explanations, but be sensitive to cultural differences. There is no need to describe things we assume to be common knowledge (as long as the visual details match those), such as the shape of objects (or even color, as in the case of a polar bear, which frequently *appears* to be white). Then, the specific steps are:

1) Provide a general overview of the type of image and the main subjects, e.g., "This image is a (photograph/drawing/etc.) that depicts a (cat/person/plate of food/etc.)"
2) If present, describe the location of any text in reference to the image, write the text verbatim, and note the original language. E.g., (above/below/etc.) in the image, there is text (originally in English/Portuguese/etc.) that reads "...".
3) Provide the subjects' and the scene's detailed description. E.g., position, actions being performed, colors, materials, etc. Note: some details are left out, such as the color of the ocean, as long as details match what is assumed to be common knowledge.
4) Optional: if the image's perspective is necessary for overall understanding, include that. E.g., in the case of food on a plate, a top view is most likely essential to see all the elements clearly. So, one would describe, "The perspective of the image is directly above the plate of food." – We initially assumed the image's perspective was unnecessary; however, while it may not be the most critical aspect of interpreting an image, it may support connotative meanings in memes.
5) Connect everything together, explaining the punchline. Important: provide the context and set up the scene first to allow readers to discover the punchline by themselves before reading this part of the description. Explain the humor/emotional elements/meaning of the image and provide any additional details/context/pop culture knowledge/background information if necessary. E.g., "There is an urban legend in Brazil about a ghost who haunts bathrooms, and there was the pandemic shutdown. Thus, the joke is that the ghost is upset by the absence of students in the school's bathroom to scare." Without these two pieces of context, the corresponding meme does not make sense. This step is the most open-ended and complicated, as you need to analyze the emotional layers in an image, as well as any crucial outside information; also, it varies based on the image observer and interpretation – one person might find an image funny, and another might not.

To conclude, detailed image descriptions must both a) **Accurately** capture the image's explicit and concrete elements, and b) **Reliably** capture the image's abstraction, coherence, and contextual bridges needed to convey its meaning. Then, as a message travels through different spaces (see the box below), it should remain **consistent** with the image's concrete elements and abstraction:

Meme → Phase 1 teams' mental models → detailed description → reader

On the other hand, raw image descriptions must a) Encode the bare minimum amount of the meme's explicit and concrete elements in a way that only just allows a reader to get the message's meaning. b) Lack of any reference to how abstraction, emotions, and connotation are used to build a message. The purpose is to check how well one recreates the message's overall meaning from only raw pieces of it. Therefore, checking how abstraction and emotions travel different spaces, as illustrated in Figure 3.

Hence, a message should remain **consistent** in spite of traveling through spaces (see the box below). We hypothesize this will help us investigate how humans get connotative meanings, abstraction, and emotional messages from missing information.

Meme → Phase 1 teams' mental models → raw description → 3D teams' mental models → 3D scene

C. Raw Descriptions and the "Unsaid Elements"

A raw description is created by "breaking" or "tearing down" a detailed description through filtering out what we call by *unsaid elements*. One could think of using **alt-text**, as there are "no hard rules on how long alt-text should be, but they are usually a short phrase or at the most, a couple of sentences" [32]. Although that is somewhat similar to raw descriptions, as they are short image descriptions, they are not the same: raw descriptions are more vague and lean than alt-text since they are used to check if/how a reader (the 3D team) catches and portrays the *unsaid elements*.

Thus, elements such as colors, shapes, descriptions of what things look like, affordances [35], [36], and explanations of the emotional/humorous components are all removed. Following [7] take on [36], we define an affordance as a "relation between an agent's abilities and the physical states of its environment" [7]. As we improve our methods, we will add the following step: **check/rewrite the raw description to make sure the used words have a clear denotation and as few as possible connotations**.

The instructions below detail how to write a raw description – note that we are moving down from the top of the ladder depicted in Figure 4:

1) Remove any explanation of humor/emotional elements/image meaning, as well as any explanation of background information/context (*i.e.*, exclude everything written for the detailed description's step 5).
2) Determine whether to remove perspective, if present. E.g., there is an image of a snake in a banana peel, and since perspective is unnecessary for interpretation, it should be removed. However, it should be kept otherwise; e.g., in another image, a fancier car must appear in the front to seem like it is the image's focus and "trick" the observer; another example: there is an image of eggs on a plate, in which perspective helps to understand that the eggs are supposed to look like two people holding hands. Still, perspective is somewhat open to interpretation, and justification should be documented for the choice, whether perspective is included or not.
3) Remove any visual details that do not interfere with understanding the image, such as color, type of material, and any non-objective descriptions. Frequently, the shape or color of an object, such as a chair or a Table, does not interfere with the image's interpretation. Therefore, those only remain if absolutely necessary to convey the image's meaning; e.g., in an image, the orange color of a cat's fur provides visual cues for it to look similar to a croissant, and the same comparison would not be as evident with a different color – for example, see the "croissant-cat image" in [37].
   - Remove mentions of the number of objects (as long as they are not necessary to convey the image's meaning). In the "machine learning" meme (see Section IV), the specific number of computers in the classroom is considered an *unsaid element* and therefore replaced with "rows of computers". In this case, we are checking whether the 3D modeler understood that the computers are meant to imitate students in a classroom. Also, details about the specific type of computers are removed, as most computers would still convey the appropriate message.
   - Remove/replace unnecessary details about image subjects. E.g., an image shows a little girl wearing glasses; however, since that is not needed to convey the message, we used the word "child" instead.
4) Text location and wording are objective and should thus remain as-is in the raw description.
5) An image's type (e.g., photograph, drawing) is *almost* always removed since the image's recreation is 3D modeled. However, there are exceptions, such as when the medium helps to drive the image's meaning. For example, if an image's element was clearly drawn with simple black and white lines, and such detail is needed for understanding, it should be kept in the raw description.
6) Finally, clean up and observe the used language to not give details away. For instance, in the eggs on a plate image, we made sure to keep the raw description as objective as possible. Instead of writing the yolk has a "smiley face", we wrote that there are "two dots next to each other, with an upward curved line underneath". It is vital to identify and filter out these "micro interpretations" and leave it up to the 3D modeling team to realize that the yolk has a face. Make sure to avoid repetitions: analyze the raw description and remove any general introductory descriptions if they repeat information unnecessarily.

"When are you done writing a raw description? What determines that it is finalized and ready to be sent to the 3D modeling team?" That is not a trivial question, as sometimes the team felt the need to keep cleaning up raw descriptions within multiple iterations. The team engages with explicit and implicit knowledge as they identify the *unsaid elements* – Zheng *e*t al. [38] summarize [39]: when "knowledge has been articulated, then it is explicit knowledge. Otherwise, another question is raised: Can it be articulated? If the answer is yes, then it is implicit knowledge. If the answer is no, then it is tacit knowledge".

Still, we consider the task to be complete once a raw description does not include any unnecessary elements. Therefore, it basically has the subjects of the image, any text if present, and the bare minimum for other details. It does not over-describe visual elements, does not hint at the image's meaning/humor, and does not emphasize an element as more important than the others.

Preparation for Phase 3, checkpoint. In parallel to identifying the *unsaid elements*, we document a "checklist" of things to look for in the 3D scenes and check if/how the 3D modelers depicted the messages' overall meaning. Once we finish processing the Phase 0 images, we will have a collection of *unsaid elements*, which we hypothesize will help to create a visual Dictionary that refers back to emotions, abstraction, and connotative meanings (Phase 5).

Finally, we will process all memes within our collection but model 3D scenes from a subset only. Then, in Phase 3, we will check how the 3D team filled in the gaps from missing information and interpreted both the meme's main visual components (concrete elements) and the abstract and emotional components. Two potential lines of inquiry for Phase 3 are as below:

1) Llorens-Gámez *e*t al. [40] show that components, such

as form and geometry, space distribution and context, color and texture, among others, influence memory and/or attention, and can be assessed objectively. The verbal description of a sink may bring up very different mental models based on each individual's background, as architecture differs across countries and cultures. It would be interesting to investigate to what extent familiar shapes or contexts populate a 3D modeler's *assumed elements*. If a modeler is used to seeing wood-made and square-like sinks, will those occupy the *assumed elements*? (Of course, other players are in place, such as how easy it is to design that shape and texture in the chosen 3D modeling tool.)

2) Leshin et al. [41] provide preliminary evidence that brain representations of emotional facial expressions are influenced by two sources of conceptual knowledge: a person's access to emotion category words and their cultural background. Their findings support evidence that conceptual knowledge activated in the minds of perceivers influences emotion perception. If an emotional context is related to disgust in the modeler's culture but anger in the original image's culture, will the 3D scene still be consistent with the original image? Images that are meant to be humorous to some may not be to others because humor shifts in different cultural contexts – see [42] for a view on how cultures create emotions or [43] for findings suggesting that emotion depends on context, culture, and their interaction.

III. Methods: 3D Scenes from Raw Descriptions

In this Section, we describe our processes for modeling a 3D scene from raw descriptions. For clarity purposes, in this Section only, we use interchangeably 'description' and 'raw description.'

We examined several 3D modeling tools before selecting Blender Version 3.5.1 for its flexibility and learning curve – see [44] for a review on Blender's versions and interfaces, and [45] for an application built on top of Blender. As stated earlier, modelers do not have access to the images that originated descriptions. To prevent influencing each other's style and approach, they individually worked on the 3D scenes. Finally, the glossary terms (Appendix A) are key to interpreting the Gateways diagram, shown in Figure 6.

We ask questions such as: "How to model an AI system that *gets* abstract and emotional messages from spatial communication? What does that even mean?" Hence, memes are a key resource, given their use of spatial communication to convey a message. According to Tversky [46], by using position, form, and movement in space, gestures, and actions convey meanings. In that sense, differently from solely symbolic words, visual communication can directly convey content and structure (both literally and metaphorically). Although it may lack the rigorous definitions words can offer, visual communication delivers flexibility and suggestions for meanings. Such flexibility, in turn, requires context and experience to interpret conveyed meanings [46]. At the same time, "the layout of the physical environment, including the apparent steepness of a hill and the distance to the ground from a balcony can both be affected by emotional states" [47].

Cohen and colleagues [48] detail four technological affordances represented in research on emotion: interactivity, personalization, accessibility, visibility, and social cues; finally, the authors discuss how technological affordances relate to emotional regulation via media use. *Social cues* are particularly important in our project since they are nonverbal signals that "infuse meaning into messages, including information about a sender's emotional state"; and "Technologies vary in terms of the type and number of the social cues they afford to users for emotional expression" [48]. That context helps to answer a question such as below.

Why the use of 3D scenes? We chose a 3D format since it provides different perspectives and enables people to play with the objects on the screen, enabling a richer experience (this interactivity is unique to 3D spaces compared to 2D images, while there are mixed results on its advantage for learning [49]). In addition, we can take screenshots of a 3D scene if needed (as in Figure 8). Finally, we sought to investigate how the modelers translate a raw description into a dynamic encoding (dealing with spatial organization and hierarchy), which opens avenues for applications within spatial thinking skills.

Instructions to create a 3D Scene are shown below:

1) Read the description and create a 3D scene as close as possible to your comprehension of the description.
2) You are free to sketch your ideas and to search online for reference images if that helps the modeling process (e.g., to model an airplane or some other unfamiliar object).
3) Do not search for memes and do not observe the other modelers working on their models – so that you do not influence each other's style.
4) Focus on developing your own shapes and avoid, as much as possible, importing shapes and libraries into the 3D modeling tool.
5) Engage with your peers to share tips on the modeling tool.
6) Finally, focus on portraying what you interpreted the message to be. Do not focus (or spend your time) on creating fancy-looking 3D scenes.

Often, modelers felt the need to use sketches either to make sense of the description, fill in the gaps, and/or visualize objects' details in different dimensions and perspectives (more in Section III-A). In that case, the description and sketch are revisited during the modeling process.

"How do you decide that a 3D scene is complete?" We investigated that question and concluded modelers were following two main gateways to evaluate and decide if a 3D scene was complete. We named those "accuracy" and "faithfulness" gateways, see Section III-B. Our focus relies on the description's message but not on creating fancy-looking 3D scenes. Hence, striking a balance between time and detail in the scenes was crucial. Also, evaluating each model's *accuracy* and *faithfulness* helped determine when the modeling process was complete.

Accuracy is assessed by revisiting the description (and sketch if used) to check if the description's explicitly stated elements have been included in the model. *Faithfulness* is assessed more complexly by evaluating the emotions in the final scene and checking if these align with the interpreted emotions from the description.

Since cultural background and experiences play a key role in how individuals make sense of a description and encode it into a 3D scene, we briefly provide our cultural context. Our research is being conducted within the USA Liberal Arts institution's cultural context, and we are a multicultural team: in addition to the US, some people lived or are originally from countries such as Brazil, China, Japan, and Thailand, and so far, six modelers (undergraduate students) have worked on this project.

A. Sketches

A sketch is a drawing draft that helps the modeler brainstorm and navigate through mental models triggered by the description, usually made before building the actual 3D scene. A sketch is an external representation [50], a visual-spatial display that augments cognition [51].

"When people read text, they construct representations of several levels, including" text-based representation (a representation of the text itself, the propositional content of the text) and "a representation of the situation or object described in the text" [52]. Depending on different descriptions, the modeler's mental models may be easily formed as a whole scene, or they may initially appear as separate objects or parts and need to be joined together after considering their relationships with each other – all these considerations are recorded in the modelers' written documentations. Modelers also reflect on any elements that feel to be missing from the description for the scene to make sense. Hence, modelers can combine potential *missing elements* into their mental models.

In addition, they may need to search for certain objects or parts to draw or model details successfully. Reference images help to visualize objects that are unfamiliar or hard to imagine (e.g., an armadillo body). Then, they can draw the sketches according to their mental models and the reference. The drawing process may be done with pen and paper or with digital drawing apps. The sketch will usually be a simple line drawing with black drawings and a white background.

Modelers review the description and check with their sketches, and they may continue to identify what is missing in their sketches and adjust it accordingly. They can also add annotations to help them better model their drawing and document their decisions. The primary purpose of having a sketch is to facilitate the process of building a 3D scene. Specifically, a sketch can help the modelers in three ways:

1) **Augments Cognition.** As the modelers navigate through mental models triggered by the description, the initial product can be vague and blurry at first glance. However, the process of drawing can help to consolidate ideas and make them clearer. Modelers can externally visualize the scene they are considering, enabling them to review their mental models. It also helps them think about what is missing from the description and elaborate on this.
2) **Reference.** It provides a standard reference for the 3D scene that boosts the modeling process efficiency. A modeler may find that transitioning directly from mental models into the actual 3D model can be difficult, so a sketch acts as a bridge. For instance, Blender allows modelers to import reference images; thus, they can import sketches into the tool and build the 3D scenes according to the sketch. A sketch also provides a standard for the size of the objects, the proportion and layout of the whole scene, and how the model parts relate.
3) **Consistency.** It may be easier to maintain consistency between mental models and the 3D model if there is a sketch to compare with. The 3D scenes must be consistent with the modeler's initial mental models of the scene. Thus, the 3D scene reflects the modeler's interpretation. However, many factors may decrease this consistency, such as technical issues with the modeling tool and the difficulty of different models. In this case, a sketch helps to record a modeler's initial mental models after reading the description, as the drawing process tends to be more flexible than 3D modeling.

In Figure 5, we show a sketch with a scene's different perspectives. Before drawing it, the modeler read the description of a polar bear standing on an iceberg in the middle of the ocean and formed mental models of this scene. The description mentions a reflection of the polar bear's skeleton on the ocean's surface, so the modeler considered the different perspectives and what should be seen under each perspective according to the physical properties. The modeler searched for images of polar bears and icebergs to observe details to draw them better and draw the scene by combining mental models with details from reference, real images. First, the modeler drew the scene from the front view, where the reflection of the skeleton cannot be seen. Then, from the top front view, the skeleton can be seen on the surface of the ocean. Additionally, there is a sketch of the skeleton itself to show its details – in Figure 8, we show the modeler's corresponding 3D scene (the image that inspired the corresponding raw description is shown in Figure 7).

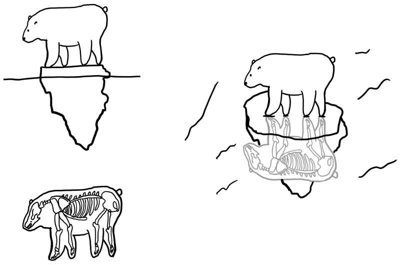


Fig. 5. An example of a sketch that contains a scene's different perspectives. The modeler sketched those to inform the 3D modeling process.

B. 3D Modeling Outcomes and the Gateways Diagram

We first detail the Gateways diagram and then show, in Section IV, our 3D scenes' static images. Modelers engage with networked emotions and emotional mental models as they switch between and across mental models to guide a description's sense-making and decision-making that leads to creating a 3D scene. Although we understand that "cognitive and emotional mental models are activated at the same time" [9], we bridge the modeling task with a dual-process account of decision-making [53], and each process has its own gateway. We decided to do so to account for both the way modelers' described their processes and highlight the importance of emotional mental models. Similarly, modelers transit between explicit, implicit, and tacit knowledge in both gateways.

Inspired by work on diagrams and cognition such as [53], [54], and [55], we designed the Gateways diagram (Figure 6) to understand the modelers' decision-making process and how they navigated the 'layers' or dimensions of emotional processing during the 3D modeling process. In the end, our diagram was informative not only in understanding the modeling process but also in checking for consistency across modelers – we will add the diagram to the 3D modeling instructions in future work.

As the diagram shows, modelers start by reading a description, and their purpose is to pass the *accuracy* and *faithfulness* gateways to complete a 3D model, producing a 3D scene. Generally, the beginning process (diagram's top/first half) tends to focus on *accuracy*, and *faithfulness* is prioritized towards the end of the modeling process (diagram's second half). We designed the diagram to allow for flexibility in the 3D modeling process, as modelers seek to create accurate and faithful 3D scenes that capture both visual and abstract details. The gateways are not completely divisible, and the process of addressing each gateway is open-ended. Therefore, achieving accurate and faithful 3D scenes can look different for distinct modelers or even for the same modeler on different days. However, despite the open-endedness, a scene must be consistent with the description.

Frequently, modelers started from the explicit and concrete elements and launched a Raw 3D model. By this point, models can be checked for the *accuracy* gateway. They may search for clues, such as image references, or sketch a few ideas to help identify *missing elements*. Modelers decide whether the model sufficiently reflects the observable visual features of the mental model they created based on the description. By 'passing' the *accuracy* gateway, they ensure that the model is accurate with the description.

At some point, modelers make assumptions to turn *missing elements* into *assumed elements* they can incorporate into the model (diagram's 2nd half). Likely starting from implicit knowledge to identify *missing elements* but then engaging more with explicit knowledge to instantiate *assumed elements* and incorporate them into the 3D model. Modelers use imagination (see [56] for a detailed view of human imagination), mental models, and knowledge/experiences to make assumptions, turning *missing elements* into *assumed elements* that can be added to the Raw 3D model to complete the model. They may sketch to reflect on different facial expressions or search online for a clue (for example, to investigate: "how does a happy turtle look like?"). Finally, the modeler documents assumptions that guided the specific *assumed elements* and other notable details about their decisions throughout the modeling process – we will investigate that documentation in phase 3.

As the model nears completion, modelers frequently focus more on emotional mental models and networked emotions. That helps examine the *faithfulness* gateway: the modeler reflects whether the 3D model sufficiently reflects the description's abstraction and emotional tone. Once the modeler decides that the model passes both gateways, the modeling process is complete, producing the 3D scene: a completed 3D model that is accurate and faithful to the description and to the modeler's mental models resulting from the description. In Section IV we show a sample of our 3D scenes and sketches.

In [1], we illustrate possible questions modelers may ask themselves while modeling emotions. The questions refer to 'layers' or dimensions of emotional processing during the 3D modeling process: modeler's, 3D model's encoding, observer/audience, and image's *via* raw descriptions. Once an observer views and interacts with a 3D scene, if the observer's overall response matches the *Observer's Intended Emotional Response* (see below), abstraction likely made it successfully through spaces and gateways. It would be interesting, in future work, to run human studies in that direction.

Modelers reported that some descriptions were harder to navigate since they evoked multiple emotional mental models to make sense of - particularly when there were conflicting or unaligned (*messy*) emotional layers. Given that challenge, we list below some of the guiding questions that helped ground the modeling process:

1) **Source.** From the raw description, what can I assume about the image?
2) **Mediated Communication.** Given my experience with popular culture, social media, and memes, what does it seem to mean? Is this supposed to be humorous? How do I feel about that?
3) **Characters.** Are there multiple characters? Do they have aligned, neutral, or unaligned emotions? Who is the main subject?
4) **Modeler's Emotional Response.** Emotions triggered in the modeler as part of making sense of the description and finalizing the 3D scene.
5) **Observer Intended Emotional Response.** Emotions the observers are supposed to have when viewing the 3D scene for the first time. How is an observer supposed to feel? (Should that be similar to how I felt when I read the description?) Should that be aligned with the characters in the scene?
6) **Emotional Mental Models.** Given the raw description and *assumed elements*, it is time to put on multiple "emotional

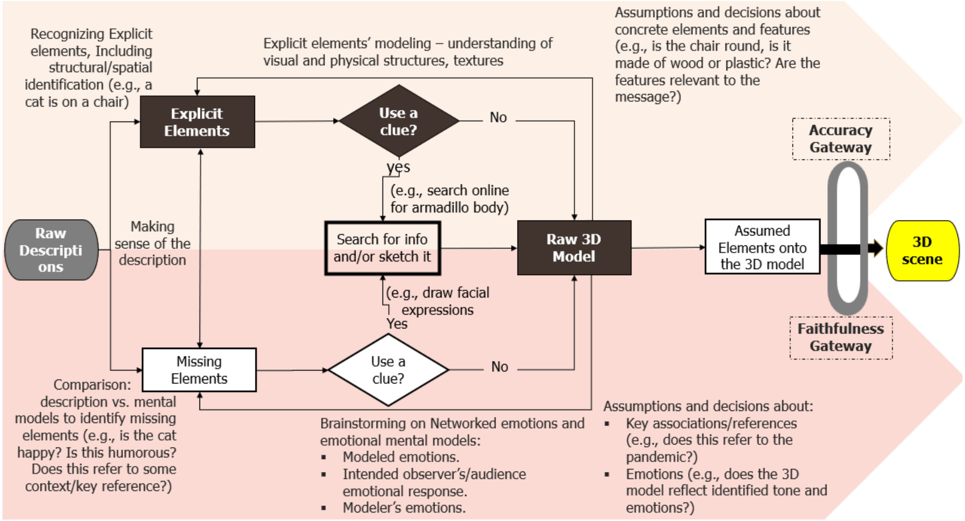

Recognizing Explicit elements, Including structural/spatial identification (e.g., a cat is on a chair)
Explicit elements' modeling – understanding of visual and physical structures, textures
Assumptions and decisions about concrete elements and features (e.g., is the chair round, is it made of wood or plastic? Are the features relevant to the message?)
Explicit Elements
Use a clue?
No
yes
(e.g., search online for armadillo body)
Accuracy Gateway
Raw Descripti ons
Making sense of the description
Search for info and/or sketch it
Raw 3D Model
Assumed Elements onto the 3D model
3D scene
(e.g., draw facial expressions
Yes
Faithfulness Gateway
Missing Elements
Use a clue?
No
Comparison: description vs. mental models to identify missing elements (e.g., is the cat happy? Is this humorous? Does this refer to some context/key reference?)
Brainstorming on Networked emotions and emotional mental models:
▪ Modeled emotions.
▪ Intended observer's/audience emotional response.
▪ Modeler's emotions.
Assumptions and decisions about:
▪ Key associations/references (e.g., does this refer to the pandemic?)
▪ Emotions (e.g., does the 3D model reflect identified tone and emotions?)

Fig. 6. Gateways diagram - Process to create and decide that a 3D scene is complete: it has to pass the Accuracy and Faithfulness gateways (right).

hats", deal with the messy layers, and model the scene and its components.

IV. A Glance at the Project's Outcomes

We show three examples of our detailed textual descriptions along with *unsaid elements*. In Figure 8, we depict the corresponding 3D scenes' static images and raw descriptions in Table I. We named the examples for reading purposes – the Phase 2 team receives ID numbers only. Finally, we present the Observer-Centered Dataset in Section IV-C.

A. Raw, Detailed Descriptions, and 3D Scenes

As the three examples below show, the identification of *unsaid elements* includes a reflection of what needs to be removed or reframed to create a raw description. We consider that reflection insightful for identifying clues people may use (potentially without realizing) to comprehend a message and ultimately key in reflecting on human-AI interaction.

Finally, as mentioned in Section II-C, we will add another step in our process to ensure that raw descriptions' words have a clear denotation and as few as possible connotations.

Example 1: "Machine Learning", ID 13.

A "machine learning" meme can be found, for example, in a Reddit post [57].

Detailed Description. This image is a photograph that depicts a white room with upright computer screens (no keyboards). There are three rows of computers, with three computers per row, on a wooden floor facing the front of the room. In the front of the room, there is a larger screen facing the rows of computers. On the large computer screen, it shows text that says "machine learning" implying the joke that the computers are learning by being in a classroom setting like humans. The perspective of the image is low to the ground, behind, and to the left of the rows of computers.

Reflection: Unsaid Elements. How many computers are needed to convey the sense of a classroom? What types of computers come to mind? Elements: room's color, floor, type of computers, number of computers, the fact that there are no keyboards and an explanation of the joke/background knowledge. To understand the joke, one would need to know what a standard classroom setup looks like, and a basic understanding of what machine learning is. **Is perspective important?** Yes, to ground the metaphor.

Example 2: "Polar bear", ID 18.

The "Future we all face" cartoon, or "Polar bear" meme [58], see Figure 5.

Detailed Description. This image is a drawing of a polar bear balancing on a small iceberg in the ocean, with the sky as the background. The image conveys a sad message. The polar bear's four feet can barely fit on the iceberg. Its bottom is pointing towards the top right of the image, and its nose touches the water. The water shows a reflection of the polar bear on the iceberg, but as a skeleton. We believe this image is intended to provide dark commentary on the state of global warming and the polar ice cap melting; we think the reflection is meant to be a window into the future extinction of the polar bear population and perhaps ours. **Is perspective important?** Yes, because both the bear and reflection must be seen to understand the message.

Reflection: Unsaid Elements. Orientation of the bear's body, shape, and color; the same goes for the sky, ocean, and what a reflection is. Climate change understanding and how it relates to a polar bear, the connection between body and skeleton. The bear is "facing" a skeletal version of itself, highlighting a possible reality, which conveys another layer to the image's connotations.

Example 3: "Butterfly", ID 22.

This description is based on a pun that generates the word "butterfly", see [59].

Detailed Description. This image is a photograph of illustrated elements and real insects on a paper. On the left of a paper, there is a simple black line drawing of a person's behind, starting at the waist and ending at the top middle of the thighs. In the middle, the letters "ER" are written. To the right of those letters, there are two flies placed on the paper. The perspective of the image is directly above the paper. The joke is that it is intended to represent the word "butterflies". **Is perspective important?** Yes, because the image includes drawings that can't be seen from a side view of the paper.

Reflection: Unsaid Elements. The number of flies (how many flies are needed to convey the message?) and their appearance. How the objects in the scene visually spell out the word "butterflies" and a person's behind creates the beginning of the word. How different elements merge/blend to create a single word.

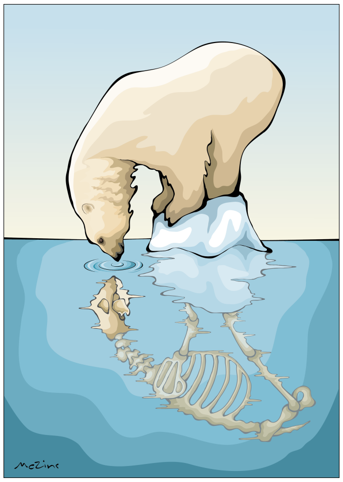


Fig. 7. The future we all face, by Mary Zins [58] (used with permission).

B. A sample of 3D scenes along with corresponding raw descriptions

In Figure 8, we show a sample of our 3D scenes, sketches and corresponding raw descriptions in Table I. Although the modeling task may seem simple, modelers faced insightful challenges along the way. For example, some descriptions mention cultural references the modelers did not recognize, and another interesting barrier came from a description centered around veganism, with a sarcastic tone. Since veganism was not a common reference to the modeler, it was challenging to understand the atmosphere the description created. Therefore, in addition to cultural background, contextual knowledge (e.g., context brought by COVID-19), popular culture, and the media were often needed to make sense of the descriptions – which was expected, given our focus on memes.

Some descriptions contained references to popular movies that must be understood for the description to make sense – there is a description that combines the context of the pandemic with the 2000 movie Castaway *via* the ball the main character bonds with. Background knowledge of the movie not only informed the modeler of what the face should look like but also what it meant in the context of the text (see Figure 8, lower left).

When we launched the project's phase 1, we were still examining what a "raw description" should look like. As we experimented with continuously removing details, we finally decided on a final method. However, we decided to keep older versions to record our trajectory. For instance, notice Table's I first row, which is derived from an older method of creating raw descriptions. To conclude, an observation on the memes that inspired #12 and #20: a) We were unable to retrieve an online source for the Brazilian version of #12. Thus, we provide an English version found within other COVID memes [60]; and b) Multiple versions of #20 are described in [61] and [62].

C. The Observer-Centered Dataset

Phase 1's goal includes the identification of attributes to examine many memes at once. Here, we present our dataset dimensions (see the dataset attributes in Appendix B). We conducted a data-driven approach to identify ad-hoc categories and image attributes, similarly to [63], whose work provides methodological directions for the study of memes.

As we kept cataloging new attributes and writing descriptions, we saw the need to better organize them, leading us to group the categories within three dimensions. Therefore, each dimension covers a set of categories, and each category has a set of attributes. That organization assisted us in capturing the images' observer experience and the interplay between concrete and abstract elements, and networked emotions.

Our approach to creating the dataset is similar to [63], which asks two questions: "Which meme formats are currently circulating online?" and "How do popular meme formats convey their message?" to then propose a methodological toolkit to analyze Internet memes. Giorgi [63] conducts a data-driven approach to create eight ad hoc categories to examine a sample from a dataset of static images collected on Instagram within the Italian cultural context. Although similar, our work presents important distinctions (in addition to the languages explored), such as our focus on abstraction and emotions, leading us to consider the observer's experience.

Similarly, Cochrane et al. [64] create a dual classification system for meme categorization: meme composition and multimodal quality. Meme composition focuses on a meme's structure, i.e., on how memes recontextualize images and text to create new meanings, whereas multimodal quality on the ways that text interacts with the image. Although the authors also consider an image's structure to get into a meme's meaning, our approach is different, given our focus on how a message travels through spaces.

Currently, the Observer-Centered Dataset has 26 attributes distributed within three dimensions. The dimensions are: Concrete Design, Blend, and Emotional Design. The Blend dimension centers around an image's observer, while the Emotional design on networked emotions. These two dimensions share two categories (Emotional Alignment and Humorous Intent), as Figure 9 shows, and cover categories that are human-interpreted and more flexible than the Concrete Design dimension (shown at the top of the Figure).

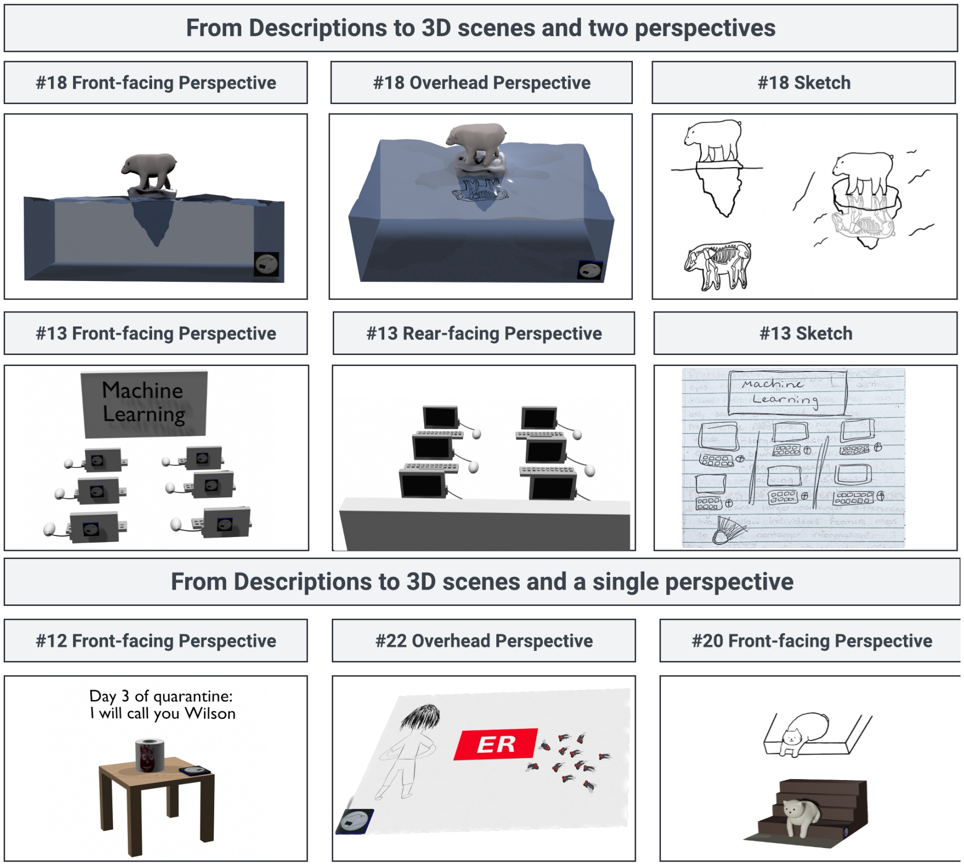

From Descriptions to 3D scenes and two perspectives
#18 Front-facing Perspective
#18 Overhead Perspective
#18 Sketch
#13 Front-facing Perspective
#13 Rear-facing Perspective
#13 Sketch
Machine Learning
Machine Learning
From Descriptions to 3D scenes and a single perspective
#12 Front-facing Perspective
#22 Overhead Perspective
#20 Front-facing Perspective
Day 3 of quarantine:
I will call you Wilson
ER

Fig. 8. Phase 2 and a sample of 3D scenes' and sketches' from a raw description. Numbers refer to the corresponding raw description ID.

Process and attribute labels.

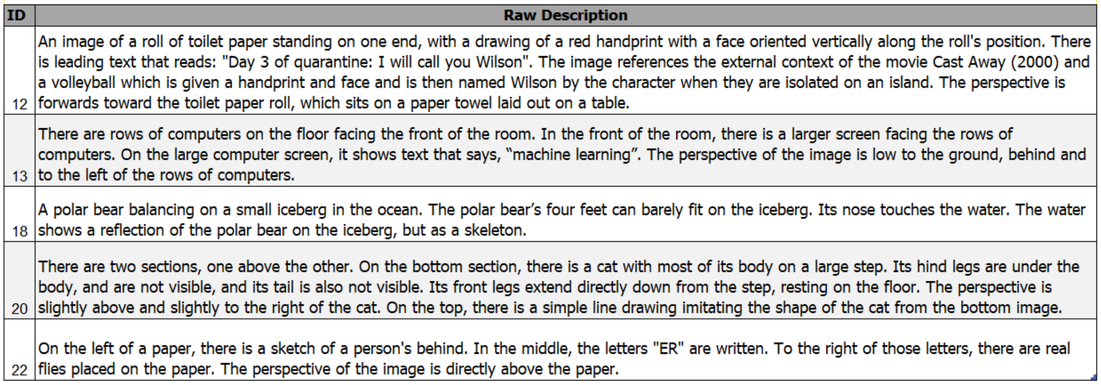
As we added memes' descriptions and identified new attributes, we faced challenges in defining concrete labels and definitions. It was difficult to create a dataset that covered the meaning of any meme-like image (within our collection scope) without excluding important details of some images or including attributes that are irrelevant to others. We kept adding new attributes as additional images were processed and reshaping older attributes, but the attributes were not always relevant for all of the images, and some were too ambiguous.

For instance, it was difficult to name the image attributes in a concise way that reflects their meaning. This challenge is explored in [65], which presents the idea that words have different meanings depending on the individual. Their results show that "at least ten to thirty quantifiably different variants of word meanings exist for even common nouns. Further, people are unaware of this variation, and exhibit a strong bias to erroneously believe that other people share their semantics. This highlights conceptual factors that likely interfere with productive political and social discourse". Their findings support our hunch that categorizing abstraction and emotions using attributes containing one or two words is challenging, as different interpretations of words can hinder understanding, especially in the context of dimensions meant to convey abstract/interpretive attributes.

A note on irony. Lozano-Palacio et al. [66] provide a broad cognitive-pragmatic perspective on the irony that interprets "ironic meaning" as a result of complex inferential activity that arises from conflicting conceptual scenarios. They distinguish basic and re-adapted uses of irony; basic uses are: Socratic irony, rhetorical irony, satirical irony, tragic irony, dramatic irony, and metafictional irony. Irony is then "determined by the attitudinal element arising from the clash between an epistemic and an observable scenario". We follow the authors' approach and consider verbal and situational irony as different materialization of *the same phenomenon*: "In both cases, the epistemic scenario is drawn from the speaker's certainty about a state of affairs (be it formed through an echo or not), and the observable scenario from the situation that is evident to the speaker" [66].

Especially if focusing on humor, the layers of emotion do not always align with the image observer and the characters in the scene. For example, an image was clearly conveyed through the detailed description "Photograph with text at the top stating 'My cat isn't paying enough attention so I improvised.' We see the back of an orange/brown cat's head with its ears up and half of its body facing away from the camera. The cat's head is to the left of the image, and its body is to the right of the image. It appears to be sitting on a couch, with the background showing part of a door and some

TABLE I
RAW DESCRIPTIONS PROVIDED TO THE 3D MODELING TEAM CREATE THE 3D SCENES DEPICTED IN FIGURE 8.



ID	Raw Description
12	An image of a roll of toilet paper standing on one end, with a drawing of a red handprint with a face oriented vertically along the roll's position. There is leading text that reads: "Day 3 of quarantine: I will call you Wilson". The image references the external context of the movie Cast Away (2000) and a volleyball which is given a handprint and face and is then named Wilson by the character when they are isolated on an island. The perspective is forwards toward the toilet paper roll, which sits on a paper towel laid out on a table.
13	There are rows of computers on the floor facing the front of the room. In the front of the room, there is a larger screen facing the rows of computers. On the large computer screen, it shows text that says, "machine learning". The perspective of the image is low to the ground, behind and to the left of the rows of computers.
18	A polar bear balancing on a small iceberg in the ocean. The polar bear's four feet can barely fit on the iceberg. Its nose touches the water. The water shows a reflection of the polar bear on the iceberg, but as a skeleton.
20	There are two sections, one above the other. On the bottom section, there is a cat with most of its body on a large step. Its hind legs are under the body, and are not visible, and its tail is also not visible. Its front legs extend directly down from the step, resting on the floor. The perspective is slightly above and slightly to the right of the cat. On the top, there is a simple line drawing imitating the shape of the cat from the bottom image.
22	On the left of a paper, there is a sketch of a person's behind. In the middle, the letters "ER" are written. To the right of those letters, there are real flies placed on the paper. The perspective of the image is directly above the paper.

shelves. On the back of the cat's head, there are two googly eyes facing the camera. The joke is that the human had to put googly eyes on the cat to pretend that the cat was looking at/paying attention to them."

The image can also be successfully conveyed through the raw description: "Text at the top stating 'My cat isn't paying enough attention, so I improvised.' We see the back of a cat's head. The cat appears to be sitting on a couch. There are two googly eyes placed on the back of the cat's head." When it came to the dataset coverage of this image, interpretive challenges presented themselves, especially for humor alignment. To label the type of **Emotional-Alignment**, the emotion of the image's observer and the emotion of the image's subject must be determined, so they can be compared. But in the image, who is the subject? Is the subject the cat, or is it the human? This is a matter of opinion, so the image cannot easily/justifiably fit into the category of "aligned" or "unaligned"; therefore, we used the "ambiguous" data entry.

Also, **there are memes that call for an Outward (ad hoc) participant/observer**: they expand their scope as they incorporate us, outside observers, as if we were part of the image/meaning (as an illustration, consider the "Hand with Reflecting Sphere" by Maurits C. Escher). These kinds of memes informed us to center the Blend dimension around the image's observer. This dimension raises an interesting reflection: how to model an AI system that "sees itself" within a context and uses that to produce a holistic interpretation and successful predictive processing?

V. DISCUSSION

Project similarities with the Telephone Game come with caveats, such as the flexibility and open-endedness in modeling a 3D scene from a textual description. Still, that is exactly it: we seek to investigate how abstraction and emotions make their way through people (calling attention to a tension between *unsaid elements* and the audience's assumed elements) and foster ideas on developing *emotion-driven* AI systems and assistive technologies.

In Phase 3, we will compare the *unsaid elements* from phase 1 with the "missing" and "assumed elements" from phase 2. The amount to which they match will help us to reflect back into abstraction and emotions. Moreover, we are setting metrics to ensure objectivity, such as keeping consistent terminologies and processes across phases and building gateways; besides, concrete elements are easy to track across spaces (for example, check if a cat "made its way through spaces").

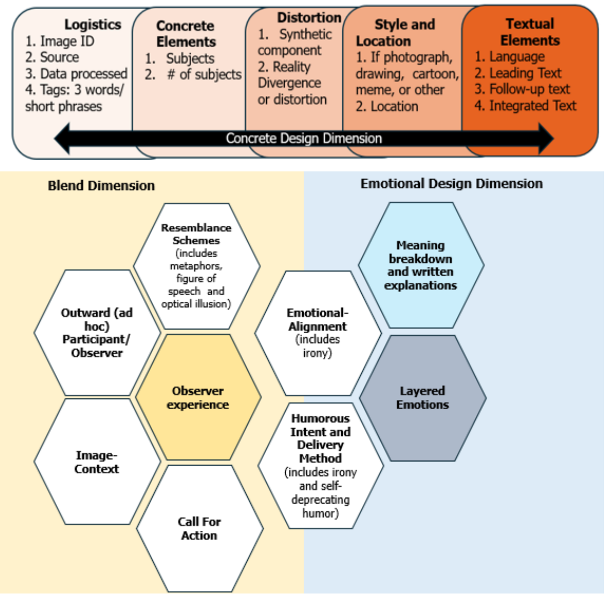

Logistics
1. Image ID
2. Source
3. Data processed
4. Tags: 3 words/ short phrases
Concrete Elements
1. Subjects
2. # of subjects
Distortion
1. Synthetic component
2. Reality Divergence or distortion
Style and Location
1. If photograph, drawing, cartoon, meme, or other
2. Location
Textual Elements
1. Language
2. Leading Text
3. Follow-up text
4. Integrated Text
Concrete Design Dimension
Blend Dimension
Emotional Design Dimension
Resemblance Schemes (includes metaphors, figure of speech and optical illusion)
Outward (ad hoc) Participant/ Observer
Observer experience
Image-Context
Call For Action
Emotional-Alignment (includes irony)
Humorous Intent and Delivery Method (includes irony and self-deprecating humor)
Meaning breakdown and written explanations
Layered Emotions

Fig. 9. The dataset attributes are categorized within three dimensions: Concrete Design, Blend, and Emotional Design. The categories *Emotional Alignment* and *Humorous Intent* belong to two dimensions: Blend and Emotional Design.

By translating an image's message into different presentation modalities (e.g., detailed and raw descriptions, 3D scenes), we are changing the conditions of comprehension [31]. Furthermore, with the removal of *unsaid elements*, a raw description becomes more abstract (it detaches as much as possible from the source image), making the message's comprehension task more abstract and open-ended (see the *concreteness effect*, in Section VI). "The advantages of concrete materials are that they can activate real-world knowledge during learning, induce physical or imagined action, enable learners to create their own knowledge of abstract concepts, and activate brain regions associated with perceptual processing. The advantages of abstract materials are that they can focus attention on more useful functional features rather than superficial features and increase generalization across multiple contexts" [18].

Raw Descriptions and Local Coherence. As Schnotz [31] points out, texts are not carriers of meaning; instead, "they trigger processes whereby multiple coherent mental representations are constructed through an interplay between text-driven bottom-up and knowledge-driven top-down activation of cognitive schemata". Then, in a "text with only local coherence, successive sentences are semantically related but without an overarching thematic connection. In a locally and globally coherent text, successive sentences are connected and there is an overarching thematic connection." Detailed descriptions offer locally and globally coherent text, whereas **raw descriptions potentially offer local coherence but a weaker global coherence**, and readers "have to reconstruct the local and global text coherence in their minds" [31]. How raw can an image description be to still allow the image's message to be transmitted? Detailed descriptions make a task clearer: since the text is rich in details and coherence, the reader's task is *to comprehend* the abstract and emotional elements in context. On the other hand, although raw descriptions are shorter, they make a reader's task more abstract and open-ended, as a reader has to fill in the gaps using their background knowledge and experiences.

Generative AI. There is incredible work being done under the generative AI tools' umbrella, such as a foundation world model [67], adding voice and image capabilities [68], text-to-3D content creation [69], text-to-image [70], [71], and text-to-video [72], among many others. Our research substantially differs from those: we are not using artificial neural networks (or any other computational approach) to identify patterns and structures within data to generate content. Rather, our agents are humans, and we seek to investigate how a message's abstraction and emotional tone, among others, travel through spaces. Also, generative AI tools are not allowed in our project – except perhaps in later phases to compare our outcomes with a tool's output such as *genz 4 meme* [73], a tool that receives a meme as input, and outputs an explanation of inside jokes and hidden meanings (although, as of now, a tool such as GPT-4 has yet to develop robust abstraction abilities at human levels [74]). Likewise, although humans are our only agents and in Phase 2 we investigate what makes a modeler decide that a 3D scene is complete, our focus relies on the *message*, not on humans themselves - hence, for simplicity, we use 'mental models' as an umbrella term, and the Gateways diagram has a focus on the task rather than on the modeler.

Sentiment Analysis. Something distinctive about our work, in particular in relation to sentiment analysis (or opinion mining; see [75] for a review in computational sentiment analysis), is that we are not trying to determine if a message' emotional tone is positive, negative, or neutral. We are seeking to investigate if a message is kept consistent within spaces without necessarily classifying its emotional tone. In fact, we take into account 1) the messy layers of emotion [22], and 2) that humor shifts in different cultural contexts (e.g., images that are meant to be humorous to some may not be to others) – we acknowledge that cultures create emotions [42] and findings suggesting that emotion depends on context, culture, and their interaction [43].

Meme Sentiment Classification. While research is being done to classify hateful memes targeted at particular audiences [76], meme sentiment classification is still an area to be explored [77]. Perhaps our methods to break down images to write descriptions and our dataset attributes could be explored in that direction - for example, our dataset's focus on the observer could help to investigate if an observer is somehow being attacked through the message's tone or connotation (e.g., one could build on the *Outward participant/observer* attribute). Although we do not include typeface data in the Observer-Centered Dataset, **typeface effects are often used to convey strong connotation messages, and it would be interesting to investigate that further**.

Our dataset covers memes from Brazil and the US, imposing a unified view of both, which, although not ideal, allows us to compare them; we include information about the language of origin in the dataset as it provides additional context. When creating and reading the textual descriptions, we must consider the historical, cultural, political, and time-situated context of when an image was created (political memes offer a key example since they can become obsolete quickly). That potentially "interferes" with how a message makes its way through different spaces, and we are using the documentation created within project Phases to help us navigate that.

We launched this project with the aim of getting insights into the modeling of *emotion-driven* AI systems; still, our work offers applications for social impact and assistive technologies. Below, we provide a few ideas for further examination.

- **Learning from Meme templates → 3D Scenes.** We described the Gateways diagram, which serves as a guideline for the modelers to create 3D scenes out of missing information. What if we could combine meme templates with raw descriptions and our Gateways diagram's process to automate the generation of 3D scenes from memes? The 3D scenes could elaborate on the unsaid elements and have a focus on teaching a domain (e.g., computer science [3]) or assist with meme humor comprehension in adolescents with language disorder or hearing loss [78], emotion regulation in depression [79], spatial thinking skills, or help individuals with aphantasia [80] create various visualizations and explore them from various distinct.
- **3D Blueprints.** Inspired by [81]'s investigation of using text-to-image generators to create concept art for the 3D-modeling process of a character, it would be interesting to check if our processes to create a 3D scene can help to embed emotions and abstract concepts to design accessible interactive 3D blueprints for blind and low-vision people [82].
- **Humor Comprehension.** Similarly to [83], our work can inform the development of intervention resources to remedi-

ate humor comprehension deficit. In that direction, we would use as inspiration: a) Dr. Temple Grandin's strategies for creating concrete exemplifications of abstract concepts [84], [85], and b) Buxbaum et al [78], which moves from "old humor" (e.g., jokes, videos, and cartoons) to web-based humor (memes), and c) Dr. Spector's work on abstract language and cognition, which informed the creation of resources such as [86], [87] [88] and [89].

- **Descriptions and Cultural Sensitivity.** In this research, we propose a systematic approach to deconstructing memes into their fundamental elements and unsaid elements reflection. The breakdown helps identify an image's references/connotations and highlight key cultural and contextual knowledge, ultimately helping a description writer to 1) notice any gaps in the description and 2) ensure cultural sensitivity. Finally, as mentioned earlier, a key challenge in identifying the *unsaid elements* originates from making the implicit explicit, and we hope to inspire frameworks for identifying biases and microaggressions in visual content.
- **Diagnostic Images.** We wonder if the Phase 1 process of creating detailed, raw descriptions (in particular, the unsaid elements) and a dataset would help to inform the identification of diagnostic images [90].
- **Strategic Decision-Making and External Representations.** According to Csaszar and colleagues [91], there is work to be done to understand external representations' central role (visuals, more specifically) in the search for new strategies. Their research "highlights that the design and use of external representations — much like navigation tools — hold consequences for decision-making quality." The authors propose a few directions for study, and computer-aided representations are among them. In that regard, the sketches created by our 3D modelers (to make sense of raw descriptions and connotations) could provide insights for further examination.

VI. Related Literature

Here, we provide a more in-depth literature review of concepts relevant to this research.

Concreteness Effect and Emotion Words. The concreteness effect "refers to the observation that concrete nouns are processed faster and more accurately than abstract nouns in a variety of cognitive tasks" [92]. There are two well-known theories for explaining the effect's neuronal basis: the dual-coding theory, and the context availability theory. Jessen et al. [92] studies suggest a combination of both theories [92]. To account for experimental findings, both theories should link abstract words with experiential information [21]: Kousta et al. [21] study emotional content (a type of experiential information) to demonstrate that it plays a vital role in the processing and representation of abstract concepts.

Starting from the question "Are the concepts represented by emotion words different from abstract in memory?", Altarriba and Bauer [93] examine emotion concepts in three experiments. According to the authors, "although emotion words have often been included in the abstract stimuli in the literature, when rated on concreteness, imageability, and context availability they are different from abstract and concrete words". Altarriba and Bauer [93] results indicate that emotion words are more memorable and readily recalled than concrete and abstract words, and that concepts represented by emotion words are more imageable and are easier to find a context for than abstract words, although they are less concrete than abstract words. **Although we acknowledge these studies, we make a loose distinction between emotions and abstraction for simplicity since a deeper analysis falls out of the scope of our project.**

Text Comprehension. Research suggests that language comprehension involves sensorimotor representations; thus, Zwaan [20] reviews the literature on mental models focusing on how mental representations are constrained by linguistic and situational factors, which are then extended to include sensorimotor representations. Text Comprehension "is equivalent to the construction of multiple mental representations in working memory. (...) Mental representations include a text surface representation, a propositional representation, and a mental model. They are characterized by different forgetting rates. As speakers and authors omit information which can be easily completed by listeners and readers, text comprehension always includes inferences" [31]. With respect to mental models of the text content, "text comprehension can be characterized as the construction, evaluation, and (if needed) revision of a mental model of the subject matter described in the text" [31]. According to Schnotz [31], text meanings are constructed by the individual through an interaction between external information received through the text and internal information from the individual's prior knowledge" [31]

According to Butterfuss, Kim and Kendeou [94], reading involves three interrelated elements, all situated into a broader sociocultural context: 1) the reader, 2) text, and 3) the reading task. The authors provide considerations on individual differences in readering comprehension, and the importance of a readers' prior knowledge. They also consider the role of emotions in reading comprehension (information may elicit emotional responses). Within emotions, they point us to two key dimensions: valence and arousal. "Valence refers to whether the subjective experience of emotions is pleasant or unpleasant. Arousal refers to the level of physiological arousal and intent to engage in activity. These two dimensions of emotions may independently influence reading comprehension via attention, working memory, motivation, learning strategies, memory processes, and self-regulation" [94].

Networked Emotions and Mental Models. The term Networked Emotions (or "Messy Layers") takes into account the social nature of emotions and the messy layers of emotion and emotion regulation; it refers to the view of "emotions as multi-layered processes in which intraindividual processes are tightly coupled and often cannot be separated from interindividual processes" [22]. There are many instances where "regulation and elicitation can best be described by nested layers of feedback loops (...) Dealing with nested layers is messy because all layers can potentially influence emotional components" [22]. Finally, according to Giaxoglou, Döveling, and Pitsillides [24], it "involves the mobilization of affect in online emotional cultures as a transmittable, spreadable, and self-contained resource, bringing out formerly privately shared emotions into online spaces and collective experience".

Nissenbaum and Shifman [95] present a cross-lingual study of memes to trace global and local expressive repertoires; and Flecha Ortiz and colleagues [96] investigate memes and collective coping theory, while discussing how memes can help to reinterpret a problematic situation. Continuing on coping theories, Schramm and Cohen [97] discuss emotion regulation and coping via media use.

Culture and Cognition. For Hutto et al. [98], sociocultural influences operate with respect to our explicitly formed and expressed beliefs and values but can additionally inform and infuse what we see and feel. Then, the authors [99] provide an interesting reflection on the production of the self and how continuous interaction with local cultural niches amplifies its scope through engagements with social media, ending up contributing to new ecologies of human existence. The authors [7] argue that we learn the shared habits and expectations of our culture through immersive participation in cultural practices that selectively shape attention and behaviour, a process by which the authors call "thinking through other minds" – finally, see [100] for more details in neuroscience research and culture.

Human Perception and 3D scenes. The computer modeling literature is active in producing insightful work on human perception (e.g., initiatives such as the Emotion Recognition Challenge [101]); as a review in computational sentiment analysis [75], and a survey on computational methods for modeling human perception of 3D scenes [102] show. The authors cover visual attention, 3D object quality perception, and material recognition. Forward, [103] review advances seeking to capture human efficiency in real-world scene and object perception, and [104] proposes a 3D modeling framework that uses visual attention characteristics to obtain compact models more adapted to human visual capabilities. Then, [105] offer insights into the development of applications in 3D knowledge of the scene, ranging from early stages of the 3D acquisition process to the higher-level tasks over 3D data. Finally, [106] provides a biologically constrained model of visual attention (with the capability of object recognition and localization) against large object variations of a visual search task in virtual reality.

Interestingly, [107] investigates the question of how to develop common sense in AI systems. Moving to semantic modeling, it could be used, for example, for large-scale scenes, automatically generating complex environments or supporting intelligent behavior on the virtual scenes, semantic rendering, and adaptive visualization of complex 3D objects [108]. Switching gears to narratives, Ong et al. [109] review time-series emotion recognition and time-series approaches in affective computing; finally, they introduce the Stanford Emotional Narratives Dataset (SENDv1), a set of rich, multimodal videos of self-paced, unscripted, emotional narratives, annotated for emotional valence over time. Finally, see [110] for a context-aware emotion recognition framework that combines four contexts: multimodal emotion recognition based on facial expression, facial landmarks, gesture, and gait.

Skurka and Nabi [111] discuss four traditions of emotion theory and highlight how digital spaces can contribute to emotional arousal and impact. Then, [112] focuses on the cognitive science of human variation in the field of spatial navigation since studies either using the real world or virtual reality show that there are significant individual differences in navigation competencies. Aiming to help researchers and designers develop emotionally interactive devices or designs, [113] examine emotional interactions between humans and deformable objects; they investigate how the design of a flexible display (depicted in 3D images in which an object is bent at different axes) interacts with emotion. Thinking of spatial skills and objects in 3D, Munns et al. [114] present an approach for developing computer-based tests of spatial skills and illustrate it by creating a test of the ability to visualize cross-Sections of 3D objects.

Nissenbaum and Shifman [95] present a cross-lingual study of memes to trace global and local expressive repertoires; and Flecha Ortiz and colleagues [96] investigate memes and collective coping theory, while discussing how memes can help to reinterpret a problematic situation. Continuing on coping theories, Schramm and Cohen [97] discuss emotion regulation and coping via media use. Finally, we conclude with considerations on empathy: Zaki et al. [115] reflect on a lack of a consistent demonstration of a correspondence between affective empathy (perceivers' experience of social targets' emotions) and empathic accuracy (perceivers' ability to accurately assess targets' emotions) – important since theories suggest that affective empathy should contribute to empathic accuracy. Their findings suggest that perceivers' self-reported affective empathy can predict their empathic accuracy, but only when targets' expressivity allows their thoughts and feelings to be read.

VII. Conclusion

In Active Threat Response Training, you are likely urged to comprehend the meaning behind your perception. What if you wanted to somehow architect these skills into a machine? Seeking to investigate what a holistic encoding of abstract and emotion-rich contexts could look like for an emotion-driven AI system, we created a multi-phase project to examine how abstraction and emotions travel different spaces.

We detailed our project's phases 1 and 2: the Phase 1 team provides raw textual image descriptions to the Phase 2 team, which is responsible for turning a description into a 3D scene. We presented our methods for creating textual descriptions from memes and a dataset that focuses on the image's observer. We also show a sample of 3D scenes' images along with corresponding raw descriptions and the Gateways diagram, designed to help us understand the 3D modelers' decision-making. We identified applications for social impact but will expand on that in future work. We hope our dataset, raw descriptions, and Gateways diagram can provide insights into exploring the concreteness effect in connection with sensorimotor representations.

Alt-text and long descriptions are essential for conveying the visual aspects of images to individuals with print disabilities, as we have previously discussed. Memes and other humorous images are key components of digital culture, fostering connections among people. If a framework can be provided that methodically and thoughtfully takes into account the *assumed elements* and guides the creation of image descriptions in reference to images with emotional/hidden meanings, such as memes, it will better include those who cannot see the visuals. Our collected images focus on humorous/entertaining aspects, but we hope our methods can inform approaches to expand on other themes.

Given emotions' investigation challenges, we are identifying processes to ensure objectivity and map how a message travels through spaces. As we do so, challenging questions emerge, and we hypothesize they will bring insights into research in emotions and how to build emotion-aware AI systems and assistive technologies. For example, how informative would it be if an *emotion-driven* AI system outputted its decision log on missing and *assumed elements* in a narrative-like sequence of pictures (or 3D scenes) and text?

To conclude, the work [116] describes a computational model that uses multiple representations in problem-solving. The model's behavior is illustrated by simulating the "cognitive and perceptual processes of an economics expert as he teaches some well-learned economics principles while drawing a graph on a black-board". It would be interesting to combine [116] with our methods and the Gateways Diagram to simulate a 3D modeler creating 3D scenes from raw descriptions.

Acknowledgments

This work would not have been possible without the support from Grinnell College's Harris Faculty Fellowship and the Mentored Advanced Project (MAP) program.

Finally, we would like to thank our reviewers for their thoughtful comments.

References

[1] J. Chen, E. Berman, M. Noda, K. Shermak, Z. Ye, D. Rothfusz, and F. Eliott, "How do abstraction and emotions travel different spaces?" in *Proc. of The Tenth International Conference on Human and Social Analytics HUSO*. IARIA, 2024.

[2] R. Dawkins, "The selfish gene," in *The selfish gene*, 1976, pp. 224–p.

[3] B. Bettin, A. Sarabia, M. C. Gonzalez, I. Gatti, C. Magnan, N. Murav, R. Vanden Heuvel, D. McBride, and S. Abraham, "Say what you meme: Exploring memetic comprehension among students and potential value of memes for cs education contexts," in *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, 2023, pp. 416–429.

[4] J. Scott, *A Dictionary of Sociology (Connotative versus denotative meaning entry)*. Oxford University Press, 2015.

[5] Merriam-Webster, "Connotation vs. denotation: Literally, what do you mean?" 2024. [Online]. Available: https://www.merriam-webster.com/grammar/connotation-vs-denotation-literally-what-do-you-mean

[6] A. M. Seidenfeld, S. R. Johnson, E. W. Cavadel, and C. E. Izard, "Theory of mind predicts emotion knowledge development in head start children," *Early Education and Development*, vol. 25, no. 7, pp. 933–948, 2014.

[7] S. P. L. Veissière, A. Constant, M. J. D. Ramstead, K. J. Friston, and L. J. Kirmayer, "Thinking through other minds: A variational approach to cognition and culture," *Behavioral and Brain Sciences*, vol. 43, p. e90, 2020.

[8] K. L. van den Broek, J. Luomba, J. van den Broek, and H. Fischer, "Evaluating the application of the mental model mapping tool (m-tool)," *Frontiers in Psychology*, vol. 12, p. 761882, 2021.

[9] B. Stangl, "Emotional mental models," in *Encyclopedia of the Sciences of Learning*. Springer, 2012.

[10] D. S. Robert W. Andrews, J. Mason Lilly and K. M. Feigh, "The role of shared mental models in human-ai teams: a theoretical review," *Theoretical Issues in Ergonomics Science*, vol. 24, no. 2, pp. 129–175, 2023.

[11] S. Lodha and R. Gupta, "Irrelevant angry, but not happy, faces facilitate response inhibition in mindfulness meditators," *Current Psychology*, vol. 43, no. 1, pp. 811–826, 2024.

[12] S. Gadanho, "Reinforcement learning in autonomous robots: an empirical investigation of the role of emotions," Ph.D. dissertation, U. of Edinburgh. College of Science and Engineering. School of Informatics., 1999.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction (2nd ed.)*. MIT press, (1992), 2018.

[14] X. Yu, R. Morri, and F. Eliott, "Eda, an empathy-driven computational architecture," in *Proceedings of the Ninth Goal Reasoning Workshop at ACS.*, 2021.

[15] F. Eliott and C. Ribeiro, "Moral behavior and empathy modeling through the premise of reciprocity," in *Proc. of The First International Conference on Human and Social Analytics HUSO*. IARIA, 2015.

[16] ——, "Emergence of cooperation through simulation of moral behavior," in *Hybrid Artificial Intelligent Systems. HAIS 2015: 10th I. Conf. on Hybrid Artificial Intelligence Systems, Bilbao, Spain. Lecture Notes in Artificial Intelligence*, vol. 9121. Springer International Pub., 2015, pp. 200–212.

[17] M. G. Mattar, J. E. Fan, W. K. Vong, and L. Wong, "How does the mind discover useful abstractions?" in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, no. 45, 2023.

[18] S. K. Reed, "A taxonomic analysis of abstraction," *Perspectives on Psychological Science*, vol. 11, no. 6, pp. 817–837, 2016. [Online]. Available: http://www.jstor.org/stable/26358684

[19] M. K. Ho, D. Abel, T. L. Griffiths, and M. L. Littman, "The value of abstraction," *Current opinion in behavioral sciences*, vol. 29, pp. 111–116, 2019.

[20] R. A. Zwaan, "Situation models, mental simulations, and abstract concepts in discourse comprehension," *Psychonomic bulletin & review*, vol. 23, pp. 1028–1034, 2016.

[21] S.-T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, and E. Del Campo, "The representation of abstract words: why emotion matters." *Journal of Experimental Psychology: General*, vol. 140, no. 1, p. 14, 2011.

[22] A. Kappas, "Social regulation of emotion: messy layers," *Frontiers in psychology*, vol. 4, p. 51, 2013.

[23] T. Jiang, H. Li, and Y. Hou, "Cultural differences in humor perception, usage, and implications," *Frontiers in psychology*, vol. 10, p. 438919, 2019.

[24] K. Giaxoglou, K. Döveling, and S. Pitsillides, "Networked emotions: Interdisciplinary perspectives on sharing loss online," pp. 1–10, 2017.

[25] A. Halversen and B. E. Weeks, "Memeing politics: Understanding political meme creators, audiences, and consequences on social media," *Social Media + Society*, vol. 9, no. 4, p. 20563051231205588, 2023.

[26] Round Table team, "Round table on information access for people with print disabilities," 2024. [Online]. Available: https://printdisability.org/

[27] C. F. Karbowski, "See3d: 3d printing for people who are blind." *Journal of Science Education for Students with Disabilities*, vol. 23, no. 1, p. n1, 2020.

[28] M. University, "Accessible graphics hub," 2024. [Online]. Available: https://accessiblegraphics.org/

[29] E. Swaim and F. Eliott, "Complex behavior vs. design-interpreting ai: Reminders from synthetic psychology," in *Proc. of The 9th International Conference on Human and Social Analytics HUSO*. IARIA, 2023.

[30] Posit, "R shiny application." [Online]. Available: https://shiny.posit.co/

[31] W. Schnotz, *Comprehension of Text*. Cambridge University Press, 2023, p. 63–86.

[32] Accessible Publishing Contributers, ""accessible publishing. an online portal featuring information and resources for the advancement and development of accessible publishing in canada and beyond."." [Online]. Available: https://www.accessiblepublishing.ca/a-guide-to-image-description/#terms

[33] J. T. Nganji, M. Brayshaw, and B. Tompsett, "Describing and assessing image descriptions for visually impaired web users with idat," in *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011*. Springer, 2012, pp. 27–37.

[34] V. Lewis, ""veronica with four eyes"," 2024. [Online]. Available: https://veroniiiica.com/how-to-write-alt-text-for-memes/

[35] J. J. Gibson, "The senses considered as perceptual systems." 1966.

[36] ——, *The ecological approach to visual perception: classic edition*. Psychology press, [1979], 2014.

[37] K. Zmanovskaia, "Cats photoshopped into food are so cute you could just eat them up, by emma taggart." [Online]. Available: https://mymodernmet.com/cats-in-food-photoshop-funny/

[38] J. Zheng, M. Zhou, J. Mo, and A. Tharumarajah, "Background and foreground knowledge in knowledge management," in *International Working Conference on the Design of Information Infrastructure Systems for Manufacturing*. Springer, 2000, pp. 332–339.

[39] F. Nickols, "The tacit and explicit nature of knowledge: The knowledge in knowledge management," in *The knowledge management yearbook 2000-2001*. Routledge, 2013, pp. 12–21.

[40] M. Llorens-Gámez, J. L. Higuera-Trujillo, C. S. Omarrementeria, and C. Llinares, "The impact of the design of learning spaces on attention and memory from a neuroarchitectural approach: A systematic review," *Frontiers of Architectural Research*, vol. 11, no. 3, pp. 542–560, 2022.

[41] J. Leshin, M. J. Carter, C. M. Doyle, and K. A. Lindquist, "Language access differentially alters functional connectivity during emotion perception across cultures," *Frontiers in Psychology*, vol. 14, p. 1084059, 2024.

[42] B. Mesquita, *Between us: How cultures create emotions*. WW Norton & Company, 2022.

[43] Z. H. Pugh, S. Choo, J. C. Leshin, K. A. Lindquist, and C. S. Nam, "Emotion depends on context, culture and their interaction: evidence from effective connectivity," *Social Cognitive and Affective Neuroscience*, vol. 17, no. 2, pp. 206–217, 07 2021. [Online]. Available: https://doi.org/10.1093/scan/nsab092

[44] L. Soni, A. Kaur, and A. Sharma, "A review on different versions and interfaces of blender software," in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2023, pp. 882–887.

[45] T. M. Takala, M. Mäkäräinen, and P. Hämäläinen, "Immersive 3d modeling with blender and off-the-shelf hardware," in *2013 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2013, pp. 191–192.

[46] B. Tversky, "Visualizing thought," in *Handbook of human centric visualization*. Springer, 2014, pp. 3–40.

[47] J. R. Zadra and G. L. Clore, "Emotion and perception: The role of affective information," *Wiley interdisciplinary reviews: cognitive science*, vol. 2, no. 6, pp. 676–685, 2011.

[48] E. L. Cohen and J. G. Myrick, "Emotions and technological affordances," *Emotions in the Digital World: Exploring Affective Experience and Expression in Online Interactions*, p. 32, 2023.

[49] S. Kriz and M. Hegarty, "Top-down and bottom-up influences on learning from animations," *International Journal of Human-Computer Studies*, vol. 65, no. 11, pp. 911–930, 2007.

[50] M. Hegarty, "Diagrams in the mind and in the world: Relations between internal and external visualizations," in *Diagrammatic Representation and Inference: Third International Conference, Diagrams 2004, Cambridge, UK, March 22-24, 2004. Proceedings 3*. Springer, 2004, pp. 1–13.

[51] ——, "The cognitive science of visual-spatial displays: Implications for design," *Topics in cognitive science*, vol. 3, no. 3, pp. 446–474, 2011.

[52] M. Hegarty and M.-A. Just, "Constructing mental models of machines from text and diagrams," *Journal of memory and language*, vol. 32, no. 6, pp. 717–742, 1993.

[53] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, "Decision making with visualizations: a cognitive framework across disciplines," *Cognitive research: principles and implications*, vol. 3, no. 1, pp. 1–25, 2018.

[54] S. Pinker, "A theory of graph comprehension," in *Artificial intelligence and the future of testing*. Psychology Press, 2014, pp. 73–126.

[55] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5. McLean, VA, USA, 2005, pp. 2–4.

[56] J. Pearson, "The human imagination: the cognitive neuroscience of visual mental imagery," *Nature reviews neuroscience*, vol. 20, no. 10, pp. 624–634, 2019.

[57] "Machine learning meme." [Online]. Available: https://www.reddit.com/r/ProgrammerHumor/comments/jxyawv/machine_learning/

[58] M. Zins, "The future we all face." [Online]. Available: https://www.cartoonmovement.com/cartoon/future-we-all-face

[59] "Butterfly meme, by i love killing flies." [Online]. Available: https://www.facebook.com/ILoveKillingFlies/

[60] "Covid memes: Why we're using laughter to get us through a pandemic, by karine bengualid." [Online]. Available: https://copyhackers.com/2020/06/covid-memes/

[61] "Awkward half-cat loafing on the stairs sparks photoshop battle no one expected, by andrea romano." [Online]. Available: https://mashable.com/article/awkward-half-cat-photoshop-battle

[62] "These cute illustrations prove cats are just funny little shapes." [Online]. Available: https://www.buzzfeed.com/pablovaldivia/silly-cat-drawings

[63] G. Giorgi, "Methodological directions for the study of memes," in *Handbook of research on advanced research methodologies for a digital society*. IGI Global, 2022, pp. 627–663.

[64] L. Cochrane, A. Johnson, A. Lay, and G. Helmandollar, ""one does not simply categorize a meme": A dual classification system for visual-textual internet memes," *Proceedings of the Linguistic Society of America*, vol. 7, no. 1, pp. 5260–5260, 2022.

[65] L. Marti, S. Wu, S. T. Piantadosi, and C. Kidd, "Latent Diversity in Human Concepts," *Open Mind*, vol. 7, pp. 79–92, 03 2023.

[66] I. Lozano-Palacio and F. J. R. de Mendoza Ibáñez, *Modeling Irony : A Cognitive-Pragmatic Account*. John Benjamins, 2022.

[67] Genie team, "Genie: Generative interactive environments," 2024. [Online]. Available: https://sites.google.com/view/genie-2024/

[68] OpenAI team, "Chatgpt can now see, hear and speak," 2024. [Online]. Available: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak

[69] C. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M. Liu, and T. Lin, "Magic3d: High-resolution text-to-3d content creation," in *IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR)*, 2023.

[70] G. Chechik, R. Gal, and Y. Atzmon, "Generative ai research spotlight: Personalizing text-to-image models," 2024. [Online]. Available: https://developer.nvidia.com/blog/generative-ai-research-spotlight-personalizing-text-to-image-models/

[71] Adobe team, "Adobe firefly," 2024. [Online]. Available: https://www.adobe.com/products/firefly.html

[72] OpenAI team, "Creating video from text," 2024. [Online]. Available: https://openai.com/sora

[73] "Genz 4 meme," https://chatgpt.com/g/g-OCOyXYJjW-genz-4-meme, 2024, hosted on ChatGPT by OpenAI.

[74] M. Mitchell, A. B. Palmarini, and A. Moskvichev, "Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks," *arXiv preprint arXiv:2311.09247*, 2023.

[75] Y. Ophir and D. Walter, "Computational sentiment analysis," *Emotions in the Digital World: Exploring Affective Experience and Expression in Online Interactions*, p. 114, 2023.

[76] A. Aggarwal, V. Sharma, A. Trivedi, M. Yadav, C. Agrawal, D. Singh, V. Mishra, and H. Gritli, "Two-way feature extraction using sequential and multimodal approach for hateful meme classification," *Complexity*, vol. 2021, pp. 1–7, 2021.

[77] P. Behera, A. Ekbal *et al.*, "Only text? only image? or both? predicting sentiment of internet memes," in *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 2020, pp. 444–452.

[78] M. Buxbaum, H. F. Pedersen Ed D *et al.*, "What do you meme? meme humor comprehension in adolescents with language disorder or hearing loss," *The Journal of Special Education Apprenticeship*, vol. 11, no. 1, p. 6, 2022.

[79] U. Akram, J. Drabble, G. Cau, F. Hershaw, A. Rajenthran, M. Lowe, C. Trommelen, and J. G. Ellis, "Exploratory study on the role of

emotion regulation in perceived valence, humour, and beneficial use of depressive internet memes in depression," *Scientific reports*, vol. 10, no. 1, p. 899, 2020.

[80] A. Zeman, "Aphantasia and hyperphantasia: exploring imagery vividness extremes," *Trends in Cognitive Sciences*, 2024.

[81] S. O. Mathiesen and A. Canossa, "If you can't beat them, join them: How text-to-image tools can be leveraged in the 3d modelling process," in *HCI International 2023 – Late Breaking Papers*. Cham: Springer Nature Switzerland, 2023, pp. 162–181.

[82] S. Reinders, "Accessible interactive 3d models for blind and low-vision people," *ACM SIGACCESS Accessibility and Computing*, no. 129, pp. 1–7, 2021.

[83] C. C. Spector, "Remediating humor comprehension deficits in language-impaired students," *Language, speech, and hearing services in schools*, vol. 23, no. 1, pp. 20–27, 1992.

[84] T. Grandin, "Emergence: Labeled autistic (with margaret scariano)," *Arena Press*, 1986.

[85] ——, *Thinking in pictures: My life with autism*. Vintage, 1995.

[86] C. C. Spector, *As far as words go: activities for understanding ambiguous language and humor*. Brookes Publishing, 2009.

[87] ——, "Just for laughs: Understanding multiple meanings in jokes," 1995.

[88] ——, "Saying one thing, meaning another: Activities for clarifying ambiguous language," 1997.

[89] ——, "Between the lines enhancing inferencing skills," 2006.

[90] Y. Bai and W. Bainbridge, "Diagnostic images for alzheimer's disease show distinctions in biomarker status and scene-related functional activity between patients and healthy controls," *Journal of Vision*, vol. 23, no. 9, pp. 5600–5600, 2023.

[91] F. A. Csaszar, N. Hinrichs, and M. Heshmati, "External representations in strategic decision-making: Understanding strategy's reliance on visuals," *Strategic Management Journal*, vol. n/a, no. n/a, 2024.

[92] F. Jessen, R. Heun, M. Erb, D.-O. Granath, U. Klose, A. Papassotiropoulos, and W. Grodd, "The concreteness effect: Evidence for dual coding and context availability," *Brain and language*, vol. 74, no. 1, pp. 103–112, 2000.

[93] J. Altarriba and L. M. Bauer, "The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words," *The American journal of psychology*, pp. 389–410, 2004.

[94] R. Butterfuss, J. Kim, and P. Kendeou, "Reading comprehension." *Grantee Submission*, 2020.

[95] A. Nissenbaum and L. Shifmancohen2023emotions, "Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study," *Journal of Computer-Mediated Communication*, vol. 23, no. 5, pp. 294–310, 2018.

[96] J. A. Flecha Ortiz, M. A. Santos Corrada, E. Lopez, and V. Dones, "Analysis of the use of memes as an exponent of collective coping during covid-19 in puerto rico," *Media International Australia*, vol. 178, no. 1, pp. 168–181, 2021.

[97] H. Schramm and E. L. Cohen, "Emotion regulation and coping via media use," *The international encyclopedia of media effects*, pp. 1–9, 2017.

[98] D. D. Hutto, S. Gallagher, J. Ilundáin-Agurruza, and I. Hipólito, *Culture in Mind – An Enactivist Account: Not Cognitive Penetration but Cultural Permeation*, ser. Current Perspectives in Social and Behavioral Sciences. Cambridge University Press, 2020, p. 163–187.

[99] *The Situated Brain: Introduction*, ser. Current Perspectives in Social and Behavioral Sciences. Cambridge University Press, 2020, p. 159–272.

[100] S. Han and G. Northoff, *Cultural Priming Effects and the Human Brain*, ser. Current Perspectives in Social and Behavioral Sciences. Cambridge University Press, 2020, p. 223–243.

[101] Odyssey, "Emotion recognition challenge," 2024. [Online]. Available: https://www.odyssey2024.org/emotion-recognition-challenge

[102] Z. C. Yildiz, A. Bulbul, and T. Capin, "Modeling human perception of 3d scenes," in *Intelligent Scene Modeling and Human-Computer Interaction*. Springer, 2021, pp. 67–88.

[103] T. Lauer and M. L.-H. Võ, "The ingredients of scenes that affect object search and perception," *Human perception of visual information: Psychological and computational perspectives*, pp. 1–32, 2022.

[104] M. Chagnon-Forget, G. Rouhafzay, A.-M. Cretu, and S. Bouchard, "Enhanced visual-attention model for perceptually improved 3d object modeling in virtual environments," *3D Research*, vol. 7, pp. 1–18, 2016.

[105] M. Poggi and T. B. Moeslund, "Computer vision for 3d perception and applications," p. 3944, 2021.

[106] A. Jamalian, F. Beuth, and F. H. Hamker, "The performance of a biologically plausible model of visual attention to localize objects in a virtual reality," in *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 2016, pp. 447–454.

[107] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu *et al.*, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[108] J. Divya Udayan and H. Kim, "Semantic modeling and rendering," in *Intelligent Scene Modeling and Human-Computer Interaction*. Springer, 2021, pp. 105–127.

[109] D. C. Ong, Z. Wu, Z.-X. Tan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling emotion in complex stories: the stanford emotional narratives dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 579–594, 2019.

[110] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, and L. Zhang, "Emotion recognition for multiple context awareness," in *European Conference on Computer Vision*. Springer, 2022, pp. 144–162.

[111] C. Skurka and R. L. Nabi, "Perspectives on emotion in the digital age." 2023.

[112] N. S. Newcombe, M. Hegarty, and D. Uttal, "Building a cognitive science of human variation: Individual differences in spatial navigation," pp. 6–14, 2023.

[113] J. M. Lee, J. Baek, and D. Y. Ju, "Anthropomorphic design: emotional perception for deformable object," *Frontiers in psychology*, vol. 9, p. 1829, 2018.

[114] M. E. Munns, C. He, A. Topete, and M. Hegarty, "Visualizing cross-sections of 3d objects: Developing efficient measures using item response theory," *Journal of Intelligence*, vol. 11, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2079-3200/11/11/205

[115] J. Zaki, N. Bolger, and K. Ochsner, "It takes two: The interpersonal nature of empathic accuracy," *Psychological science*, vol. 19, no. 4, pp. 399–404, 2008.

[116] H. J. Tabachneck-Schijf, A. M. Leonardo, and H. A. Simon, "Camera: A computational model of multiple representations," *Cognitive science*, vol. 21, no. 3, pp. 305–350, 1997.

[117] Merriam-Webster, "Accuracy," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/accuracy

[118] ——, "Reliable," 2024. [Online]. Available: https://www.merriam-webster.com/dictionary/reliable

[119] ——, "Consistent," 2024. [Online]. Available: https://www.merriam-webster.com/dictionary/consistent

[120] ——, "Object," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/object

[121] ——, "Faithful," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/faithfulness

[122] ——, "Abstract," 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/abstract

[123] A. Damasio and G. B. Carvalho, "The nature of feelings: evolutionary and neurobiological origins," *Nature reviews neuroscience*, vol. 14, no. 2, pp. 143–152, 2013.

[124] X. Hu and M. Twidale, "A scoping review of mental model research in hci from 2010 to 2021," in *HCI International 2023 – Late Breaking Papers*. Cham: Springer Nature Switzerland, 2023, pp. 101–125.

[125] B. Tversky, "Cognitive maps, cognitive collages, and spatial mental models," in *European conference on spatial information theory*. Springer, 1993, pp. 14–24.

[126] M. Hegarty, "Mechanical reasoning by mental simulation," *Trends in cognitive sciences*, vol. 8, no. 6, pp. 280–285, 2004.

[127] M. Hegarty and D. Waller, "A dissociation between mental rotation and perspective-taking spatial abilities," *Intelligence*, vol. 32, no. 2, pp. 175–191, 2004.

[128] M. Hegarty, "Chapter 7 - components of spatial intelligence," in *The Psychology of Learning and Motivation*, ser. Psychology of Learning and Motivation. Academic Press, 2010, vol. 52, pp. 265–297.

Appendix A: Glossary

We created a glossary to ensure consistency in our communication and processes. Overall, terms are sorted for better understanding instead of alphabetically.

We identified at least three concepts for communicating connotative, emotion-rich messages in digital spaces, which are the extent to which a message's material representation is:

1) **Accurate/Accuracy.** A "conformity to truth or to a standard or model" [117]. In our context, if something captures the source's explicit and concrete elements.
2) **Reliable.** "suitable or fit to be relied on." [118] Here, if a detailed textual description reliably captures the images' abstraction and contextual linkages needed to retrieve its meaning (see "Faithfulness" below).
3) **Consistent.** "marked by harmony, regularity, or steady continuity: free from variation or contradiction" [119]. Here, if a message remains *consistent* with the original source, in spite of traveling through various spaces.

Terms More Related to the task of building the 3D Scenes:

- **3D Models.** Refer to the 3D modeling process. Once the model passes the two gateways (figure 6), it is complete/finished, and we call it a **3D Scene**. A modeler's goal is not to make fancy 3D scenes; they stop modeling once they determine a model matches their mental models triggered by the raw description.
- **Concrete Elements.** Elements that add specific and objective visual elements to an object, e.g., modeling a cat sitting on a chair.
- **Object.** "Something material that may be perceived by the senses" [120], e.g., a cat or a person.
- **Modeled Emotion**: Emotion that the modeler seeks to model onto the concrete objects in the scene. E.g., adding expressive features to facial expressions so that an emotion can be visually seen on the object.
- **Subject.** Concrete, material element(s) of all elements in the scene that the modeler identifies as a dominant, primary component of the entire scene.
- **Character.** Object that, from the description, seems to express or elicit emotions.
- **Observer.** The perspective from a person viewing from outside the image or 3D scene.
- **Participant Observer.** If the modeler identifies, from the raw description, that the scene must allow an observer to merge with the scene so that the observer is also a participant (e.g., "Hand with Reflecting Sphere" by Maurits C. Escher). It corresponds to our dataset attribute *Outward (ad hoc) participant/observer*.
- **Intended Observer's Emotional Response.** The emotions modelers intend to elicit in the observer by looking at the 3D scene. In humor, many times, the intended emotion conflicts with modeled emotions (e.g., a scene of a cat not enjoying a bath is likely to be funny to an observer whose perspective is from the outside of the 3D scene).
- **Faithfulness**: The extent to which the 3D model is 'true' to the modeler's emotional mental models of the scene triggered by the raw description. We use the definition: "true to the facts, to a standard, or to an original" [121], which are the modeler's emotional mental models. Once it is 'true', the 3D model passes the *faithfulness* gateway, as shown in our Gateways diagram. Here, "reliable" relies more on the concrete source (e.g., image), whereas "faithfulness" on the co-creation of someone (modeler) blending together pieces from a concrete source (descriptions) and generated mental models.
- **3D Model Accuracy.** The extent to which the 3D model reflects the explicit and concrete elements triggered by the raw description in the modeler's mental models. Once it reflects those elements, the 3D model passes the *accuracy* gateway, as shown in our Gateways diagram.
- **Missing Elements.** Elements a modeler identifies to be missing from the raw description. Then, modelers use their knowledge and experiences to make assumptions, turning missing elements into assumed elements to pass through both gateways: *accuracy* and *faithfulness*.
- **Assumed Elements.** Elements intentionally added by modelers to the Raw 3D Model to "fill in the gap" left by the missing elements. That enables modelers to shape the 3D model to match their mental models of the scene triggered by the raw description.

Other relevant terms:

- **Abstract.** "Expressing a quality apart from an object" [122]. We use abstraction as an umbrella term that intercepts connotative meanings and emotion and mental models.
- **Connotative vs. Denotative Meaning.** "Connotative meaning refers to the associations, overtones, and feel that a concept has, rather than what it refers to explicitly (or denotes, hence denotative meaning). Two words with the same reference or definition may have different connotations" [4].
- **Explicit, Implicit, and Tacit Knowledge.** "When knowledge has been articulated, then it is explicit knowledge. Otherwise, another question is raised: Can it be articulated? If the answer is yes, then it is implicit knowledge. If the answer is no, then it is tacit knowledge" [38].
- **Emotion-rich message.** Anything that conveys emotional messages that human senses can perceive.
- **Emotions and Feelings.** Both are key concepts for homeostatic regulation: "Feelings are mental experiences of body states. They signify physiological need (for example, hunger), tissue injury (for example, pain), optimal function (for example, well-being), threats to the organism (for example, fear or anger) or specific social interactions (for example, compassion, gratitude or love)". Whereas "Emotions include disgust, fear, anger, sadness, joy, shame, contempt, pride, compassion and admiration, and they are mostly triggered by the perception or recall of exteroceptive stimuli" [123]. Emotions "regulate

social interaction and in extension, the social sphere. In turn, processes in the social sphere regulate emotions of individuals and groups" [22].

- **Emotional Mental Models.** Cover emotions and feelings connected to mental models. Hu and Twidale [124] provide a broad definition of mental models: they "refer to humans' internal representations of the external world that derive from their perception, memories, knowledge, and causal beliefs". As Hu and Twidale [124], we acknowledge that the term "mental models" has a multidimensional nature, and below we provide more context for Emotional Mental Models: "Mental models cause certain expectations/thoughts of how things should look like/work and connect certain emotions with this. Consequently, a mental model is a cognitive and an emotional framework in the brain, influenced by person's personality (genes) and the environment including social variables" [9].
- **Mental Models.** "internal representations of the external world consisting of causal beliefs that help individuals deduce what will happen in a particular situation" [8]. For simplicity, we use 'mental models' as an umbrella term that covers terms such as spatial mental models and mental representations of environments or 'cognitive collages' [125]. Although that simplification is not ideal and considerations on mental simulation and mechanical reasoning [126] are extremely relevant, such an examination falls out of the scope of this manuscript. Likewise, considerations on a distinction between "mental abilities that require a spatial transformation of a perceived object (e.g., mental rotation) and those that involve imagining how a scene looks like from different viewpoints (e.g., perspective taking)" [127].
- **Networked Emotions ("Messy Layers").** Takes into account the social nature of emotions and the messy layers of emotion and emotion regulation. It refers to the view of "emotions as multi-layered processes in which intraindividual processes are tightly coupled and often cannot be separated from interindividual processes" [22]. It "involves the mobilization of affect in online emotional cultures as a transmittable, spreadable, and self-contained resource, bringing out formerly privately shared emotions into online spaces and collective experience" [24].
- **Humor.** "results when the incongruous is resolved (i.e., the punchline is seen to make sense at some level with the earlier information in the joke). Lacking a resolution the individual does not "get" the joke, is puzzled or even frustrated. The resolution phase is a form of problem solving, an attempt to draw information or inferences that make a link between the initial body of the joke or cartoon and its ending" [83].
- **Irony.** Is "determined by the attitudinal element arising from the clash between an epistemic and an observable scenario". We consider verbal and situational irony as different materializations of the same phenomenon: "In both cases, the epistemic scenario is drawn from the speaker‘s certainty about a state of affairs (be it formed through an echo or not), and the observable scenario from the situation that is evident to the speaker" [66].
- **Memes.** "A form of media communicating a thought or idea through some shared understanding" [3].
- **Image Description.** It is an umbrella term for image descriptions in a textual form.
- **Detailed Description.** Our detailed descriptions aim to fully describe images that have complex, abstract messages.
- **Alt-text.** "Alt-text, also known as alternative text, offers textual description of images. These text descriptions are visually hidden but when a blind or visually impaired reader encounters an image while using their screen reader, the alt-text will be read out. Descriptions are generally concise" [32]. They are "text-based descriptions of visual details in an image written primarily for people who are visually impaired (inclusive of blind/low vision)" [34].
- **Caption.** "A caption is a visible text component which accompanies an image and provides additional information. It may describe the image briefly and/or give contextual information about the source. It does not usually describe the image in great detail but instead, works in conjunction with the image" [32].
- **Sense-Making Tasks.** They "consist of information gathering, re-representation of the information in a schema that aids analysis, the development of insight through the manipulation of this representation, and the creation of some knowledge product or direct action based on the insight. In a formula Information → Schema → Insight → Product" [55], and the re-representation may be in the modeler's mind, written or drawn, or even digitally represented.
- **Spatial thinking.** It "involves thinking about the shapes and arrangements of objects in space and about spatial processes, such as the deformation of objects, and the movement of objects and other entities through space. It can also involve thinking with spatial representations of nonspatial entities". And spatial intelligence "can be defined as adaptive spatial thinking" [128].

Appendix B: The Observer-Centered Dataset Attributes

The Concrete Design dimension focuses on concrete characteristics, and it splits into five categories and 14 attributes. The Blend dimension has six categories (two shared with the Emotional design) and 11 attributes (4 from the shared categories). Due to better alignment, we depict the shared categories within the Emotional Design dimension, which focuses on networked emotions and has three categories and 5 attributes (4 shared). In Appendix B, we detail the attributes.

The Concrete Design Dimension categories and attributes are as follows:

1) Logistics: attributes related to handling an image.
   (a) Image ID: image's numbered identification; format: image_#number.
   (b) Source. The memes' source, if available (N/A otherwise).
   (c) Date Processed. The most recent date (month/day/year) an image's attributes were updated/completed.
   (d) Tags: three words/short sentences that help to identify an image.
2) Concrete Elements: image's main subjects.
   (a) Subject(s): those are the concrete, material element(s) of all elements in the scene that have been identified as a dominant(s), primary component(s) of the entire image. There are no fixed attribute options, as they are meant to provide context about what the image is about. E.g., "cat", "person", "cactus that looks like a cat".
   (b) Subjects' number: registers how many subjects are the focus of the image; select a number between $1-10$, "M" if there are more than ten subjects in the image focus, and "N/A" either if there is no clear focus or if subjects are absent.
3) Distortion: if the image shows any distortion.
   (a) Synthetic Component: whether an image has been clearly altered to achieve a certain effect (such as adding a drawing on top of a picture). Select one of the entries: Absent (it does not seem to have been modified in any way), Edited (has clearly been altered), Live (it looks like being modified in real-time while it is being created, similar to M. C. Escher's *Drawing Hands*).
   (b) Reality Divergence or distortion: whether an image deviates from the expected reality in which it is presented. This includes instances where there are synthetic components or staged appearances of objects or creatures performing actions that are not possible in reality. This attribute is binary, with "True" indicating a divergence from reality and "False" indicating that the image adheres to reality. While non-photographic images such as drawings or cartoons may have more flexibility in their realities, the category still considers the context and the physical laws.
4) Image style and location: refer to an image's style and depicted location.
   (a) If an image has multiple styles, select the one that best fits it; if the image does not easily fit into any of the entries, the option "Other" is selected. Select one of the entries: Cartoon, Drawing, Meme, Photograph, or Other.
   (b) Does it show a clear location? Attribute inspired by [33]. Select one of the entries: Indoors: private space, Indoors: public space, Outdoors: private space, Outdoors: public space, or Unclear.
5) Textual Elements: we consider the text's location only, but it could be interesting to add typeface details as well.
   (a) Language: text's original language. Select English, Portuguese, or "N/A" if the image does not contain text.
   (b) Leading Text: text outside of the image that provides context. Select Yes, No, or "N/A" if there is no text.
   (c) Follow-up Text: text that builds off of leading text, providing more context or a punchline. Select Yes, No, or "N/A" if there is no text.
   (d) Integrated Text: any text within the image itself. Select Yes, No, or "N/A" if there is no text.

The Blend Dimension shares two categories with the Emotional Design dimension (both shown with the latter). Categories and attributes are as follows:

1) Resemblance schemes: if there are possible comparisons within an image.
   (a) Resemblance: an umbrella term for visual metaphors, comparison, and personification. Whether a subject in an image appears to imitate something it is not in reality or is compared to something in a way that showcases similarities. For example, an object's shape could naturally resemble that of an animal or human, or it could be artificially manipulated to look like something else, e.g., a cake that looks like a computer. Note: this category refers only to visual comparisons. Select one of the entries: Absent or Present.
   (b) Optical Illusion: if the image contains an element that tricks the viewer's eyes in some way. Select one of the entries: Absent or Present.
   (c) Figure of speech: if the image's textual elements use a "figure of speech", such as a metaphor, personification, or prosopopoeia. Select one of the entries: True, False, Unclear.
2) Outward (ad hoc) Participant/Observer
   (a) Outward Observer: whether the image's observer is assumed to be observing the scene or participating in it in some way. Whether there is an implied observer, who is not explicitly shown in the image but is assumed to exist in order to understand the image's context or meaning (e.g., "POV" memes). Select one of the entries: Absent or Present.
3) Image Context: any relevant contextual information needed to understand the image.
   (a) Context: external factors or circumstances that influence or inform the image's interpretation and meaning. It can include a wide range of concepts, such as cultural

references, historical events, social norms, or even the specific time and place in which the image was created or viewed. No fixed attribute options. E.g., "COVID", a movie's name if knowledge of a certain movie is needed, etc. "N/A" if there are no external contexts necessary for understanding.

(b) Time-situated context: whether the image refers to a specific time frame, such as the pandemic. Select one of the entries: True or False.

4) Call for action: whether it seems to provoke the observer to act.

(a) Call for action: Select one of the entries: True, False, Unclear.

The Emotional Design dimension categories and attributes are as follows:

1) Meaning breakdown: written notes to explain the image and call attention to something particularly unique about the image.

(a) Explanation notes: there are no fixed attribute options, as it should contain short written notes.

2) Emotional-Alignment: this category is shared with the Blend dimension.

(a) Emotional alignment: points to the observer. If the observer is supposed to feel the same/similar emotion as the image's subject (s), then the attribute is considered "aligned". If the intended emotion is different from that of the subject, then the attribute is considered "unaligned". If there is no obvious emotional framing, then is considered "absent". However, if it is ambiguous due to various emotional layers within the image, the attribute is labeled as "Ambiguous". Select one of the entries: Absent, Aligned, Unaligned, Ambiguous.

(b) Irony: whether it conveys irony, either for a humorous effect or not. Select one of the entries: True, False, or Unclear.

3) Humorous Intent and Delivery Method: this category is shared with the Blend dimension.

(a) Intent: whether an image is clearly designed to provide enjoyment or humor. Select one of the entries: True, Neutral, or Opposite (for negative emotions).

(b) Humor Delivery Method: describes how humor is conveyed to the image's observer. Multiple categories can be selected: Absent (the image does not have entertainment/humorous intent), Visual Humor (humor is conveyed using visuals), Textual Humor (is conveyed using text), Pun (humor is conveyed through wordplay), Self-deprecating humor, Other (some form of humor not covered in the previous options).