

# Enabling Chatbots with Eyes and Ears: An Immersive Multimodal Conversation System for Dynamic Interactions

Jihyoung Jang<sup>1\*</sup> Minwook Bae<sup>3\*</sup> Minji Kim<sup>1</sup> Dilek Hakkani-Tür<sup>4</sup> Hyounghun Kim<sup>1,2</sup>

<sup>1</sup>Graduate School of Artificial Intelligence, POSTECH

<sup>2</sup>Department of Computer Science and Engineering, POSTECH

<sup>3</sup>Artificial Intelligence Graduate School, UNIST

<sup>4</sup>University of Illinois Urbana-Champaign

{jihyoung, mzkim, h.kim}@postech.ac.kr minwook09@unist.ac.kr dilek@illinois.edu

## Abstract

As chatbots continue to evolve toward human-like, real-world, interactions, multimodality remains an active area of research and exploration. So far, efforts to integrate multimodality into chatbots have primarily focused on image-centric tasks, such as visual dialogue and image-based instructions, placing emphasis on the “eyes” of human perception while neglecting the “ears”, namely auditory aspects. Moreover, these studies often center around static interactions that focus on discussing the modality rather than naturally incorporating it into the conversation, which limits the richness of simultaneous, dynamic engagement. Furthermore, while multimodality has been explored in multi-party and multi-session conversations, task-specific constraints have hindered its seamless integration into dynamic, natural conversations. To address these challenges, this study aims to equip chatbots with “eyes and ears” capable of more immersive interactions with humans. As part of this effort, we introduce a new multimodal conversation dataset, **Multimodal Multi-Session Multi-Party Conversation (M<sup>3</sup>C)**, and propose a novel multimodal conversation model featuring multimodal memory retrieval. Our model, trained on the M<sup>3</sup>C, demonstrates the ability to seamlessly engage in long-term conversations with multiple speakers in complex, real-world-like settings, effectively processing visual and auditory inputs to understand and respond appropriately. Human evaluations highlight the model’s strong performance in maintaining coherent and dynamic interactions, demonstrating its potential for advanced multimodal conversational agents.<sup>1</sup>

## 1 Introduction

The development of conversation systems has made significant strides in recent years, with conversation

models transitioning from simple rule-based systems to sophisticated models capable of engaging in human-like interactions (Li et al., 2017; Zhang et al., 2018; Rashkin et al., 2019; Adiwardana et al., 2020; Roller et al., 2021; Shuster et al., 2022; Jang et al., 2023). As these systems advance, integrating multimodal capabilities has emerged as a critical avenue for enhancing their realism. Multimodal conversation models, which combine information from multiple sensory modalities such as text, vision, and audio, hold the promise of emulating human communication more effectively by incorporating richer contextual understanding (Zhang et al., 2019; Ahn et al., 2023; Lee et al., 2024b; Park et al., 2024). Despite this potential, existing research has primarily focused on image-based conversations, such as visual dialogue (Shuster et al., 2020; Meng et al., 2020; Wang et al., 2021; Feng et al., 2023) and image-based instructions (Li et al., 2023a; Brooks et al., 2023; Liu et al., 2024; Koh et al., 2024; Yang et al., 2024), emphasizing the visual, or “eyes”, aspect of human perception. Meanwhile, the auditory, or “ears”, aspect remains relatively underexplored, limiting the holistic nature of current multimodal systems (Goel et al., 2024; Kong et al., 2024; Huang et al., 2024; Gong et al., 2024; Tang et al., 2024). However, existing approaches have been rather limited in integrating both “eyes and ears” simultaneously, highlighting a gap in developing truly holistic multimodal conversational agents.

Also, a key limitation of existing approaches lies in their focus on static interactions. In these paradigms, the chatbot typically receives a shared image and responds to questions or prompts about it, similar to exchanging observations or answers about the image rather than experiencing it as part of a shared scene (Das et al., 2017; Kottur et al., 2021). This approach fails to capture the dynamic, real-time nature of human communication. Furthermore, while multimodal integration has been

\* Equal contribution.

<sup>1</sup>Our code, dataset, and model are publicly available at <https://m3c-dataset.github.io/>.

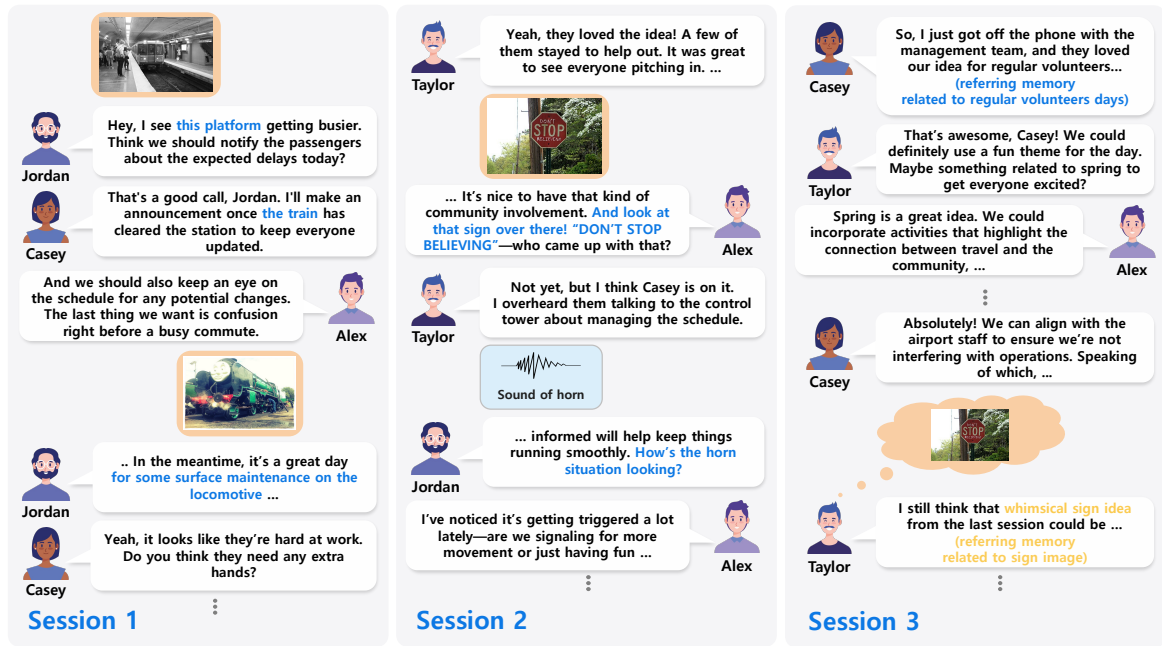


Figure 1: The main speaker (Alex) engages in conversation with two different partners per session, where all speakers simultaneously experience the provided multimodal inputs in the same shared space. In later sessions, the main speaker can meet new partners and continue the conversation.

explored in both multi-party and multi-session conversations (Carletta et al., 2006; Yamamoto et al., 2015; Saha et al., 2018; Lee et al., 2023, 2024b), existing studies are largely task-focused, with little exploration of more complex, dynamic real-world scenarios. These gaps highlight the need for datasets and models that reflect the complexities of natural human conversation, encompassing multiple modalities and interaction dynamics.

To address these challenges, we propose a novel approach for equipping chatbots with “eyes and ears” capable of real-time, multimodal interactions. Central to this effort is the introduction of **Multimodal Multi-Session Multi-Party Conversation (M<sup>3</sup>C)**, a new multimodal conversation dataset designed to capture dynamic and rich conversational settings (Figure 1). Our M<sup>3</sup>C features consecutive sessions involving multiple speakers, with a main speaker interacting with multiple partners in each session. Specifically, extending the mixed-session conversation (Jang et al., 2024), M<sup>3</sup>C allows the main speaker to interact with multiple partners per session while engaging with different partners across sessions, enhancing interaction diversity. Crucially, all speakers in each session observe and interact through simultaneous visual and auditory inputs, fostering a more cohesive and realistic interaction dynamic.

Building on this dataset, we also propose a novel

multimodal conversation model. Our model consists of a dialogue module and a retriever module, enabling it to not only process multimodal inputs from the ongoing session but also store and retrieve multimodal information from previous sessions as part of a multimodal memory. This design seamlessly integrates visual and auditory inputs, enabling coherent and contextually relevant responses in dynamic, real-world-like conversational settings. In human evaluations, our model demonstrates high engagement and immersion with multimodality during conversations.

Here are our contributions:

1. We introduce M<sup>3</sup>C, a new multimodal conversation dataset featuring multiple speakers simultaneously experiencing the same visual and auditory inputs in a shared spatial and temporal context across consecutive sessions.
2. We present a novel multimodal conversation model leveraging a multimodal memory retriever, enabling it to recall and retrieve past visual and auditory elements for contextually rich and coherent responses across sessions.
3. In human evaluations, our model demonstrated deep immersion in multimodal interactions during conversations, resulting in high engagement.

## 2 Related Works

**Multimodal Conversation.** Recent work has explored integrating multiple modalities to enhance conversational interactions (Das et al., 2017; Poria et al., 2019; Saha et al., 2018; Kottur et al., 2021; Lee et al., 2024b), but most studies emphasize the visual aspect, effectively giving chatbots “eyes” while neglecting “ears”. As a result, multimodal conversations often remain image-centric (Shuster et al., 2020; Zang et al., 2021; Lee et al., 2021; Ahn et al., 2023), failing to capture the fluid and dynamic nature of real-world conversations. While some research addresses auditory information, it is typically limited to specific sounds or narrow domains (Goel et al., 2024; Kong et al., 2024; Huang et al., 2024; Gong et al., 2024; Tang et al., 2024), restricting the scope of audio integration. Moreover, existing methods that incorporate dialogue memory and retrieval often focus on single-modality inputs (e.g., text-based summaries), thereby losing critical multimodal context (Li et al., 2023b; Zhang et al., 2024). To address these limitations, we propose a new multimodal conversation dataset and model that jointly processes both visual and auditory stimuli while maintaining a broader contextual memory from images, audio, and conversation history.

**Complex Conversation Scenario.** Previous studies on multi-session conversations mainly concentrate on preserving context in extended one-on-one settings (Bae et al., 2022; Jang et al., 2023; Zhang et al., 2023; Joko et al., 2024), whereas multi-party conversations often focus on coordinating multiple speakers in a single session (Wei et al., 2023; Chen et al., 2023; Gu et al., 2023; Ma et al., 2023; Fan et al., 2024). Jang et al. (2024) bridges these by allowing a main speaker to interact with different partners across sessions. Extending this idea, we introduce multi-party interactions within each session, enabling several speakers to engage simultaneously. This design better reflects the dynamic interplay of real-world conversations. Furthermore, by integrating visual and auditory inputs in every session, we ensure all participants share synchronized stimuli, promoting more cohesive and realistic interaction dynamics.

## 3 M<sup>3</sup>C

We introduce a new machine-generated multimodal conversation dataset, M<sup>3</sup>C. Each episode in M<sup>3</sup>C consists of four speakers engaging in conversations

across three consecutive sessions. Each session is multi-party, with a main speaker interacting with two different partner combinations per session. All speakers share the same spatial and temporal environment, experiencing two instances of multimodal input (visual or auditory) per session as they converse. Since previous machine-generated datasets have demonstrated sophistication and high quality, we follow the same approach by generating our dataset using large language models (LLMs) (Kim et al., 2022; Zheng et al., 2023; Kim et al., 2023; Jang et al., 2023; Xu et al., 2023; Lee et al., 2024a; Jang et al., 2024; Lee et al., 2024b). We employ GPT-4o mini<sup>2</sup> to generate the dataset.

### 3.1 Structuring Multimodality

In our dataset construction, we represent both images and audio through textual captions, since current state-of-the-art LLMs cannot directly process multiple raw image and audio inputs at once. We use COCO (Lin et al., 2014) as the seed image dataset and AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) as the audio datasets. We observe that captions significantly influence the quality of the generated conversations; therefore, we refine the image captions using GPT-4o mini, particularly to include details that are present in the images but missing in the original captions.

### 3.2 Preparing Scenario

Before generating each conversation, we create a corresponding scenario that includes speaker information (such as names and relationships), session-specific partners, two multimodal inputs per session, and specified time intervals between sessions. To ensure thematic consistency, we group candidates’ images and audio clips based on similarity similar modalities and provide them to LLM, which then selects the most fitting modalities while considering the speaker details and the time intervals. To group similar multimodal data, we perform K-means clustering (K=30) based on location tags assigned by GPT-4o-mini based on each caption’s content to facilitate clustering.

### 3.3 Conversation and Memory Generation

We sequentially generate conversation episodes based on the prepared scenarios. According to Jang et al. (2024), when the main speaker meets different partners in each session, it is essential to track

<sup>2</sup><https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

| Datasets                            | Type             | Multiple Sessions | Multiple Speakers | Image (# of Images) | Audio (# of Audios) | # of Sessions | # of Turns |
|-------------------------------------|------------------|-------------------|-------------------|---------------------|---------------------|---------------|------------|
| AMI (Carletta et al., 2006)         | Open-Domain      | ✗                 | ✓                 | ✗                   | ✓                   | 279           | -          |
| VisDial (Das et al., 2017)          | Modality-QA      | ✗                 | ✗                 | ✓ (120K)            | ✗                   | 123K          | 2.4M       |
| MELD (Poria et al., 2019)           | Open-Domain      | ✗                 | ✓                 | ✓                   | ✓                   | 1.4K          | 13K        |
| ImageChat (Shuster et al., 2020)    | Modality-Centric | ✗                 | ✗                 | ✓ (202K)            | ✗                   | 202K          | 401K       |
| MMConv (Liao et al., 2021)          | Modality-Centric | ✗                 | ✗                 | ✓ (114K)            | ✗                   | 5.1K          | 39.8K      |
| PhotoChat (Zang et al., 2021)       | Open-Domain      | ✗                 | ✗                 | ✓ (10.9K)           | ✗                   | 12K           | 156K       |
| MMDD (Lee et al., 2021)             | Modality-Centric | ✗                 | ✗                 | ✓ (13K)             | ✗                   | 17K           | -          |
| MMDialog (Feng et al., 2023)        | Modality-Centric | ✗                 | ✗                 | ✓ (1.53M)           | ✗                   | 1.08M         | 4.92M      |
| MPCHAT (Ahn et al., 2023)           | Modality-Centric | ✗                 | ✗                 | ✓ (153K)            | ✗                   | 15K           | 42.5K      |
| Audio Dialogues (Goel et al., 2024) | Modality-QA      | ✗                 | ✗                 | ✗                   | ✓                   | 163K          | -          |
| MiSC (Jang et al., 2024)            | Open-Domain      | ✓                 | ✓                 | ✗                   | ✗                   | 51K           | -          |
| DialogCC (Lee et al., 2024a)        | Open-Domain      | ✗                 | ✗                 | ✓ (129.8K)          | ✗                   | 83K           | -          |
| LOCOMO (Maharana et al., 2024)      | Open-Domain      | ✓                 | ✗                 | ✓ (2K)              | ✗                   | 1.7K          | -          |
| Stark (Lee et al., 2024b)           | Open-Domain      | ✓                 | ✗                 | ✓ (900K)            | ✗                   | 500K          | -          |
| <b>M<sup>3</sup>C(Ours)</b>         | Open-Domain      | ✓                 | ✓                 | ✓ (24K)             | ✓ (73K)             | 16K           | 2.5M       |

Table 1: Statistics and comparison with other datasets. ‘Type’ defines the scope of the conversation; Modality-QA focuses on question-answering, Modality-Centric revolves around discussions centered on a specific modality (such as images or audio), and Open-Domain allows for interactions on a wide range of topics. The ‘-’ denotes unreported data. Our M<sup>3</sup>C integrates both images and audio, creating dynamic, real-time, and immersive experiences.

interactions separately. In this dataset, a memory summary is created after each session from the main speaker’s perspective, integrating experiences along with images and audio from previous sessions to form a multimodal memory. We also employ memory linking to connect related elements, ensuring that linked memories are retrieved together for an enriched conversational context. After generating the conversation and memory, we perform tagging to indicate where modality elements begin in the conversation and which utterances are associated with specific memory elements.

### 3.4 Filtering for Dataset Integrity

A core requirement of our dataset is that all speakers share the same visual and auditory elements, ensuring conversations occur in a shared environment with consistent modality integration. To this end, a filtering process is employed by posing validation questions to a machine to exclude episodes that fail to maintain temporal or spatial consistency (see Appendix A for more details). Through this process, we collect a total of 54K conversation episodes (34K for train, 8K for validation, and 12K for test). Table 1 presents detailed features and statistics of M<sup>3</sup>C, along with a comparison to other datasets. Also, please see Appendix B for the full prompts used in building the dataset and Appendix C for examples of M<sup>3</sup>C (episode and memory).

## 4 Multimodal Multi-Session Multi-Party Conversation Model

We propose a novel multimodal, multi-session, multi-party conversation model capable of perceiv-

ing both images and audio, akin to having “eyes and ears”. It is designed to facilitate interactions among multiple partners, enabling dynamic and coherent conversations while participants change across sessions.

To achieve this, we introduce a dialogue module and a retriever module (see Figure 2 for the overall architecture). The Dialogue Module performs the generation task, which includes producing conversations based on multimodal inputs, generating session memory, and linking past interactions to ensure coherence in multi-session dialogues. The Retriever Module performs the retrieval task, accessing relevant memories from previous sessions based on the ongoing session. This allows the model to integrate past memory, ensuring consistency across interactions. The following subsections detail each module.

### 4.1 Dialogue Module

The Dialogue Module performs various generation tasks, including conversation generation, memory generation, and memory linking. These processes enable the model to engage in coherent and dynamic multi-session interactions while maintaining contextual continuity. During each session, the module generates responses based on the given dialogue history and multimodal inputs, ensuring a natural and context-aware conversation flow. Once a session concludes, the model constructs a memory that captures the dialogue exchanged with different speakers and integrates perceived modalities.

To maintain long-term coherence, the module employs structured memory linking by explicitly associating new memory units with semantically

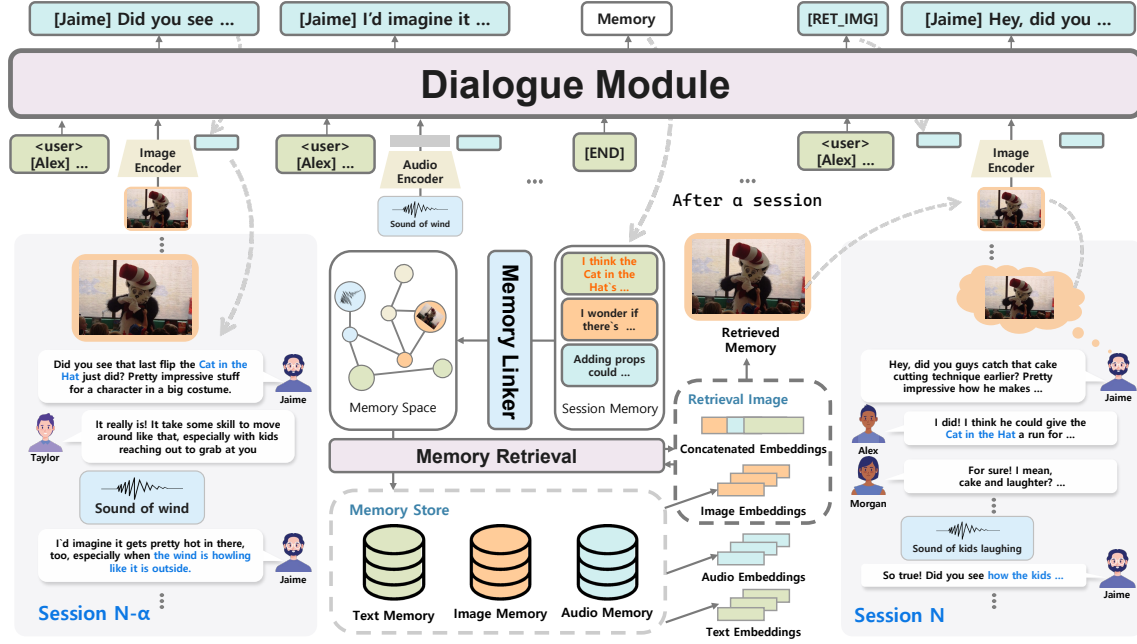


Figure 2: Overall architecture of our model. In Session N- $\alpha$ , Jaime (main speaker) perceives an image of a man in a cat mask and the sound of the wind, engaging in a related conversation with Taylor. In Session N, Jaime recalls this image and linked memory to continue the dialogue. The Retrieval Module selects relevant memories—including images, audio, and chat history—based on the session context.

and perceptually related modalities and contextual cues. This ensures that the memory graph accurately reflects both temporal progression and meaningful relationships across sessions. For example, if the system hears the sound of wind accompanied by dialogue such as, “Be careful, the wind might blow your hat away,” it explicitly links the auditory perception (the sound of wind) to the semantic content of the dialogue. These associations are formed at the time of memory storage—rather than retrieval—enabling contextually relevant activations, where recalling one memory triggers related semantic or perceptual experiences. This process helps the model construct a coherent and evolving memory space. For details on the training process, please refer to Section 5.1.

## 4.2 Retrieval Module

The Retriever Module recalls relevant stored memories—structured through the memory linking process—based on the ongoing session’s multimodal inputs while maintaining each speaker’s perspective and sensory experiences. It retrieves past conversations, images, or audio from their respective memory stores—text from textual stores, images from visual stores, and audio from auditory stores.

To facilitate memory retrieval, the module jointly embeds the entire session—including the current

conversation and the perceived modalities—into a shared representational space. It uses cosine similarity to measure how relevant the linked multimodal memories are to the present context:

$$\text{sim}(c, m_i) = \cos(E_c(c), E_m(m_i)) \quad (1)$$

where  $c$  represents the multimodal conversation context, which may include text, images, and audio, while  $m_i$  represents stored multimodal memory.  $E_c$  refers to the encoder for the conversation context, and  $E_m$  denotes the encoder for memory. During retrieval, we select only one memory with the *top 1* similarity to the given context. This approach helps maintain coherence within ongoing dialogues by consistently recalling contextually relevant memories.

## 5 Experiments

### 5.1 Model Training Details

Our model consists of a Dialogue Module and a Retrieval Module, both built on Qwen2-VL-2B-Instruct (Wang et al., 2024). Since the base model is a Vision-Language Model (VLM), we extend its capabilities using CLAP (Wu et al., 2023), incorporating a linear layer adapter to enable audio comprehension. The training process follows a staged tuning approach: first, we fine-tune the model on

vision-language tasks, treating audio as textual captions. Then, we integrate the linear adapter, allowing the model to process audio inputs directly and enhance its auditory understanding. Also, our model supports model-to-model conversations, enabling multiple instances of the model to engage in dialogue without human intervention. To facilitate this, the model is trained to recognize and manage its own turns in a conversation, ensuring smooth turn-taking and a natural conversation flow. Further implementation details are provided in the Appendix D.

## 5.2 Human Evaluation

We conduct human evaluations to assess both the dataset quality and the model performance. For this purpose, we employ eight professional evaluators through an agency. Evaluators work in groups of three per conversation to measure inter-annotator agreement. We follow the metrics used in previous studies (Lee et al., 2021; Sun et al., 2022; Jang et al., 2023; Lee et al., 2024a; Maharana et al., 2024; Lee et al., 2024b; Jang et al., 2024; Chu et al., 2024a; Sumida et al., 2024; Li et al., 2024; Lee et al., 2024c) for evaluation. All evaluations are based on a 5-points scale. For more details on the evaluation procedures, please refer to Appendix E.

**Dataset Quality.** We randomly select 300 episodes from M<sup>3</sup>C to evaluate overall conversation quality. We provide evaluators with instructions, the full transcripts of the conversations, and the associated modality elements. Rather than supplying captions for these modalities, we present the original image and audio files directly. Evaluators assess the conversations based on coherence and consistency, memorability, modalities alignment, and modalities engagement. Please refer to Appendix E for detailed descriptions of each metric.

**Model Performance.** Our model is trained to enable multi-party conversations among model agents without any human intervention (please refer to Appendix F for an example). To evaluate its performance, we generate 100 conversation episodes in which only model agents participate. Each episode features four agents, and we randomly select a seed episode from our dataset to determine the agents’ names, relationships, the first utterance of the session, and multimodality content for the session.

During the conversation, each agent autonomously decides whether it is their turn to speak. If multiple agents simultaneously determine that

it is their turn, the one with the higher probability takes precedence. Additionally, the main speaker decides at each turn when and which modality appears. If no modality has been introduced by the fifth turn, it is inserted at a random subsequent turn. This setup allows us to assess how effectively the model manages natural, multimodal, multi-party, multi-session conversations without human intervention. We evaluate the model’s performance on the generated episodes using Naturalness, Immersion, and Memorability, and please refer to Appendix E for more detail.

## 5.3 Machine Evaluation

We conduct automatic evaluations for all conversations used in the dataset quality and model performance evaluations by employing a machine evaluator in addition to human evaluators. Specifically, we use o3-mini<sup>3</sup> as the machine evaluator and provide it with the same guidelines and full conversation transcripts given to human evaluators. Unlike human evaluators, who receive direct image and audio files, o3-mini cannot accept audio inputs. Although it can process images, it cannot pinpoint the exact turn at which an image appears during the conversation. Therefore, we substitute modal content by inserting captions in the transcripts at the corresponding points.

Additionally, we also conduct comparative evaluations with other publicly available datasets, selecting those that enable fair and meaningful comparisons, specifically PhotoChat (Zang et al., 2021), DialogCC (Lee et al., 2024a), Stark (Lee et al., 2024b).<sup>4</sup> For these evaluations, we randomly sample 100 conversations from each dataset and use GPT-4o-mini and Claude-3-5-Sonnet<sup>5</sup> as machine evaluators. The models are asked to judge whether the use of multimodality in the dialogues feels natural and immersive. Evaluations are conducted in a multiple-choice format.

## 5.4 Quantitative Evaluation for Model

We also perform automatic metric-based evaluations to measure the performance of the model.

**Retriever Performance.** To quantitatively evaluate the retrieval performance of our retrieval module, we assess its ability to correctly retrieve rele-

<sup>3</sup><https://openai.com/index/openai-o3-mini/>

<sup>4</sup>Only the first sessions of Stark and M<sup>3</sup>C are utilized for a fair comparison.

<sup>5</sup><https://www.anthropic.com/news/claude-3-5-sonnet>

| Metric                    | Human       | Machine     |
|---------------------------|-------------|-------------|
| <b>Dataset Quality</b>    |             |             |
| Coherence and Consistency | 4.81        | 4.99        |
| Memorability              | 4.63        | 4.99        |
| Modalities Alignment      | 4.21        | 4.26        |
| Modalities Engagement     | 4.36        | 4.57        |
| <b>Dataset Overall</b>    | <b>4.50</b> | <b>4.70</b> |
| <b>Model Performance</b>  |             |             |
| Naturalness               | 4.34        | 4.68        |
| Immersive                 | 4.14        | 4.56        |
| Memorability              | 4.35        | 4.46        |
| <b>Model Overall</b>      | <b>4.28</b> | <b>4.57</b> |

Table 2: Evaluation results for both dataset quality and model performance. Human score indicates the average inter-annotator rating among three groups. Machine score indicates the average evaluation score given by o3-mini.

| Dataset          | gpt-4o-mini | claude-3-5-sonnet |
|------------------|-------------|-------------------|
| PhotoChat        | 0%          | 0%                |
| DialogCC         | 1%          | 0%                |
| Stark            | 18%         | 1%                |
| M <sup>3</sup> C | 81%         | 99%               |

Table 3: Comparison results with other datasets using machine evaluators, showing the selection rates of each model for each dataset.

vant memory at specific turns across all conversation episodes in the test split of the M<sup>3</sup>C. However, since most existing baseline multimodal large language models (MLLMs) are limited to bimodal settings (such as text-image or text-audio), and models supporting all three modalities (text, image, and audio) are rare, we conduct modality-specific evaluations to ensure fair comparisons.

**Multi-party Capability.** Our model is designed to autonomously decide the appropriate turn to speak during multi-party conversations, which is essential for agents to interact without human involvement. To evaluate how naturally our model engages in these conversations, we randomly sample 1K episodes from the test split of our dataset.

From each episode’s first session, after the speakers have exchanged six turns of dialogue, we have the model predict which speaker would naturally respond next based on the preceding context. We compare our model’s performance with a baseline (Qwen2-VL-2B-Instruct) using the same input. A prediction is considered correct if it matches the actual speaker at that turn in the dataset.

| Image Only                                  | R@1          | R@5          | MRR          |
|---------------------------------------------|--------------|--------------|--------------|
| Qwen2-VL-2B-Instruct (Wang et al., 2024)    | 66.77        | 92.30        | 77.56        |
| LLaVA v1.5 7B (Liu et al., 2024)            | 62.85        | 90.42        | 74.13        |
| LLaMA-3.2-11B-Vision-Instruct <sup>6</sup>  | 72.41        | 90.62        | 78.90        |
| Ours                                        | <b>92.99</b> | <b>99.09</b> | <b>95.06</b> |
| Audio Only                                  | R@1          | R@5          | MRR          |
| Qwen2-Audio-7B-Instruct (Chu et al., 2024b) | 69.94        | 94.85        | 80.72        |
| Audio Flamingo-2 (Ghosh et al., 2025)       | 63.29        | 85.44        | 70.87        |
| Pengi (Deshmukh et al., 2023)               | 68.33        | 88.75        | 75.12        |
| Ours                                        | <b>92.83</b> | <b>98.19</b> | <b>94.78</b> |

Table 4: Experimental results on the retrieval module’s performance. R@K represents Recall, and MRR indicates Mean Reciprocal Rank.

## 6 Results

### 6.1 Human and Machine Evaluations

Table 2 presents the evaluation results for both M<sup>3</sup>C and our model. The human evaluation score, derived from the average inter-annotator ratings, indicates that both the dataset and the model achieve consistently high scores across all metrics. Machine scores are slightly higher than human scores but show a similar distribution trend. In particular, considering the concept of images or audio appearing simultaneously to speakers, the dataset also shows high scores in modality-related metrics. For model evaluation, where four model agents engaged in conversations without human intervention, the results further demonstrate the model’s ability to conduct human-like interactions. These findings confirm that our model effectively integrates visual and auditory inputs, behaving like it has “eyes and ears” to engage in immersive, contextually coherent conversations. As shown in Table 3, our dataset also demonstrates higher quality compared to the other datasets, indicating that conversations in ours enable an immersive experience in which speakers naturally interact through shared spatial and temporal contexts across multiple modalities. For detailed human evaluation statistics, including average group scores and inter-annotator agreement, please refer to Appendix E.

### 6.2 Model Evaluations

**Retrieval Performance.** Table 4 shows the quantitative performance of the retriever module. We use standard retrieval metrics, including Recall@K (R@1, R@5) and Mean Reciprocal Rank (MRR). As shown in the table, our model effectively identifies the relevant memory elements. Notably, it achieves a much higher R@1 score compared to

<sup>6</sup><https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>

| Model                                    | Rate  |
|------------------------------------------|-------|
| Qwen2-VL-2B-Instruct (Wang et al., 2024) | 10.3% |
| Ours                                     | 85.2% |

Table 5: Performance of next-speaker prediction in a multi-party setting.

| User    | People here look so excited! It looks so fun!                                                                                                                                             |
|---------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|         |                                                                                                          |
| Qwen2   | It does seem like the people in the image are enjoying themselves, as they appear to be rollerblading or inline skating on a street. ...                                                  |
| LLaVA   | Yes! The people in this image look like they're having a blast rollerblading down the street. The open road, the sun, and the sense of freedom definitely add to the fun atmosphere ..... |
| o3-mini | It does look like they're having a blast! From what I can see, there are several people rollerblading (or inline skating) down a street together. ....                                    |
| Ours    | Yes! It's a great day for skating. Do you think they're just having fun, or are they trying to do a trick?                                                                                |

Table 6: A human live chat example showing the different responses from models based on image input.

other models, indicating that it's particularly good at retrieving multimodal memory from ongoing sessions in a multi-session setting.

**Multi-party Capability.** The next-speaker prediction performance shown in Table 5 clearly demonstrates our model's superior ability to identify the appropriate next speaker, highlighting its deeper understanding of multi-party conversational dynamics. These results further underscore the importance of our dataset and model design in enabling more natural and coherent multi-party interactions.

### 6.3 Comparison with Other Models

Unlike existing models, our model enables all speakers to converse as if they are in the same environment, simultaneously observing and experiencing visual elements as they appear during the conversation. To validate this capability, we compare our model against its baseline, Qwen2-VL-2B-Instruct and LLaVA, both without tuning on our dataset, as well as o3-mini. We provide all models with the same context to examine how their

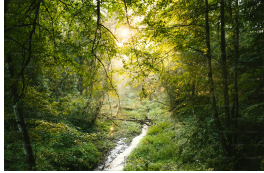
|                  |                                                                                                                                  |
|------------------|----------------------------------------------------------------------------------------------------------------------------------|
|                  |                                               |
| Jamie            | This place is nice. Kinda has that open, peaceful feel to it.                                                                    |
| Taylor           | Yeah, it's got that same relaxed vibe. Also, the trees here are huge!                                                            |
| Alex             | Right? And the way the light comes through the leaves—it's pretty cool.                                                          |
| Jamie            | This kind of reminds me of that time at the park. You know what I mean?                                                          |
| Retrieved Memory | I need to remember that we plan to target the park for our next community clean-up and set a date for this weekend (about Jamie) |
| Alex             | Right! Maybe we can use the park for our clean-up activity. That would be great!                                                 |

Table 7: A human live chat example showing the possibility of a long-term session than the third session in our model.

responses differ. Table 6 presents the comparison examples, showing that our model integrates visual elements directly into the conversation context, responding as if the participants are physically present and viewing the image. In contrast, the other models primarily provide descriptions or fact-based responses about the image, resembling a question-and-answer interaction rather than immersive dialogue. This demonstrates that our model not only benefits from effective training on the dataset but also seamlessly adapts to the given modality, enabling more contextually rich and immersive conversations. Please see Appendix G for additional comparison examples.

### 6.4 More Than Three Sessions

Our dataset consists of three consecutive sessions, and the model is trained on these three sessions. However, since our model manages the contents of previous sessions through a separate memory mechanism, it is able to maintain long-term conversations without being limited by the model's context length. Table 7 presents an example that demonstrates this capability. After the third session is completed and the fourth session begins, the main speaker retrieves information from the first session at an appropriate time and uses it during the conversation. This shows that although our model is trained on three sessions, it can effectively be utilized in longer sessions as well.

### 6.5 Ablation Study

**Audio Caption vs. Raw Audio.** Our model enhances audio comprehension by integrating an au-

| Audio Caption vs. Raw Audio                                                   |                                                                                                                             |
|-------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| .....                                                                         |                                                                                                                             |
| <b>User</b>                                                                   | It's so loud, my head hurts.                                                                                                |
| <b>A helicopter engine runs and wind blows<br/>Clotho dataset - 95383.wav</b> |                                                                                                                             |
| <b>Caption</b>                                                                | I've been thinking about the flight we had last week. It was a blast, but I'm curious about the weather conditions.         |
| <b>Raw Audio</b>                                                              | Yeah, right? I don't think we've seen anything like this before. Did we make plans after our flight earlier to hang around? |

Table 8: A human live chat example during a helicopter ride, illustrating an ablation example of the audio adapter.

dio encoder. To demonstrate the effectiveness of this audio adapter, we conduct ablation experiments comparing the model’s performance on audio inputs with and without the adapter. In the model without the adapter, audio sources are provided as textual captions, while the adapted model processes raw audio clips directly. Table 8 illustrates this comparison, showing that the caption-based model heavily relies on the provided caption, often generating responses that reflect the caption’s content. For example, the caption refers to a previous flight and weather conditions, which may not be contextually linked to the actual audio. In contrast, our adapter-equipped model generates responses that align more closely with the auditory experience, such as acknowledging the loudness of the helicopter engine, demonstrating that the adapter enables the model to perceive and respond to audio more naturally and accurately.

**Multimodal Memory.** Our model leverages a multimodal memory retriever to retrieve multimodal memories based on the given conversational context. To verify the effectiveness of this retriever, we compare model responses with and without the retriever under the same conversational context. Table 9 illustrates this comparison, showing that when the retriever is used, the model references multimodal memories from previous session related to the beach, such as collecting seashells, enriching the conversation with relevant and vivid details. In contrast, without the retriever, the model produces a more generic response focused only on the present view. This demonstrates that our multimodal memory retriever effectively retrieves contextually relevant memories, enhancing the conversational flow and depth.

|                                                                                    |                                                                                                          |
|------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
|  |                                                                                                          |
| <b>Jamie</b>                                                                       | Wow, the view here is incredible! We have to take some photos. Who wants to go first?                    |
| <b>Alex</b>                                                                        | I'm up for it! I've got my camera with me. Just let me know when we're ready.                            |
| <b>Morgan</b>                                                                      | I think we have enough room! I'll just grab my camera, too.                                              |
| <b>Alex</b>                                                                        | Perfect! Let's find the best spot for the shot. Maybe over by that railing?                              |
| <b>Jamie</b>                                                                       | This reminds me of when we were at the beach! The way the sunlight hit the water—it was so perfect.      |
| <b>With Retrieval</b>                                                              |                                                                                                          |
|  |                                                                                                          |
| <b>Alex</b>                                                                        | Yeah! And it'd be fun to add some seashells to the pictures, just like those we collected the last time. |
| <b>Without Retrieval</b>                                                           |                                                                                                          |
| <b>Alex</b>                                                                        | Yes! And we should avoid the crowds today. Just want to capture this view peacefully.                    |

Table 9: An example of a human live chat showing the difference in responses when generating utterances using a multimodal retriever with and without referencing multimodal memory.

## 7 Conclusion

In this work, we address key challenges in multimodal conversation modeling. These challenges include the lack of datasets that capture dynamic and realistic interactions, the limited integration of simultaneous visual and auditory modalities, and the absence of robust models capable of handling complex conversational settings. To overcome these issues, we introduce M<sup>3</sup>C, which enables immersive and authentic interactions by allowing speakers to directly observe and hear synchronized visual and auditory inputs. Building on this, we propose a novel robust multimodal conversation model designed to process and respond to both image and audio modalities concurrently. Trained on M<sup>3</sup>C, our model shows strong performance in maintaining conversational coherence, adapting to diverse multimodal inputs, and managing complex interactions involving multiple partners within each session, as validated through human evaluations.

## Limitations

In building our dataset, we enhance the seed image dataset captions to improve dialogue quality. However, we do not refine the captions for the seed audio dataset. Future work that improves audio captions, similar to image captions, is expected to enhance the dataset’s audio immersion and overall quality. Additionally, our model integrates audio and image recognition by attaching an adapter to the pre-trained VLM model (Qwen), due to the lack of an instruct-tuned model capable of jointly understanding images, audio, and text. The development of an instruct-tuned Visual-Audio-Language Model (VALM) could further improve performance in future research.

## Ethics Statement

We use OpenAI’s text-moderation-latest model (Markov et al., 2023) to review all generated conversations for flags in categories such as sexual content, harassment, hate speech, self-harm, and violence. Any episode flagged in any of these categories is excluded from our dataset. However, we do not conduct additional ethical verification for the COCO, AudioCaps, and Clotho datasets used to provide image and audio modalities, trusting that the original authors have already performed thorough ethical reviews. For human evaluations, we commission the evaluation process to a professional agency to ensure fairness and impartiality, with no direct involvement from the authors. We also confirm that evaluators receive fair compensation and appropriate treatment through the agency. Also, our research results are to be used for research purposes only.

## Acknowledgments

We thank the reviewers for their valuable feedback. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)) and (No.RS-2020-II201336, Artificial Intelligence graduate school support(UNIST)).

## References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. [MPCHAT: Towards multimodal persona-grounded conversation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.
- Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. [Keep me updated! memory management in long-term conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. [PLACES: Prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024a. Cohesive conversations: Enhancing authenticity in multi-agent simulated dialogues. *arXiv preprint arXiv:2407.09897*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024b. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.
- Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.
- Yaxin Fan, Peifeng Li, and Qiaoming Zhu. 2024. [Improving multi-party dialogue generation via topic and rhetorical coherence](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3240–3253, Miami, Florida, USA. Association for Computational Linguistics.
- Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. [MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.
- Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.
- Arushi Goel, Zhifeng Kong, Rafael Valle, and Bryan Catanzaro. 2024. Audio dialogues: Dialogues dataset for audio and music understanding. *arXiv preprint arXiv:2404.07616*.
- Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. [Listen, think, and understand](#). In *The Twelfth International Conference on Learning Representations*.
- Jia-Chen Gu, Chao-Hong Tan, Caiyuan Chu, Zhen-Hua Ling, Chongyang Tao, Quan Liu, and Cong Liu. 2023. [MADNet: Maximizing addressee deduction expectation for multi-party conversation generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7681–7692, Singapore. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Rongjie Huang, Mingze Li, Dongchao Yang, Jia-tong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.
- Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. [Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606, Singapore. Association for Computational Linguistics.
- Jihyoung Jang, Taeyoung Kim, and Hyounghun Kim. 2024. [Mixed-session conversation with egocentric memory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11786–11815, Miami, Florida, USA. Association for Computational Linguistics.
- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2024. [Vlm2vec: Training vision-language models for massive multi-modal embedding tasks](#). *CoRR*, abs/2410.05160.
- Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P De Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 796–806.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. [AudioCaps: Generating captions for audios in the wild](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. [SODA: Million-scale dialogue distillation with social commonsense contextualization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.
- Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. [ProsocialDialog: A prosocial backbone for conversational agents](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

- Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. [Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities](#). In *Forty-first International Conference on Machine Learning*.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. [SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal turn analysis and prediction for multi-party conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 436–444.
- Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyon Myaeng. 2021. [Constructing multi-modal dialogue dataset by replacing text with semantically relevant images](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 897–906, Online. Association for Computational Linguistics.
- Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2024a. [DialogCC: An automated pipeline for creating high-quality multi-modal dialogue dataset](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1938–1963, Mexico City, Mexico. Association for Computational Linguistics.
- Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeong-Jin Oh, Byungsoo Ko, Jonghwan Hyeon, and Ho-Jin Choi. 2024b. [Stark: Social long-term multi-modal conversation with persona commonsense knowledge](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12137–12162, Miami, Florida, USA. Association for Computational Linguistics.
- Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeongjin Oh, and Ho-Jin Choi. 2024c. [Thanos: Enhancing conversational agents with skill-of-mind-infused large language model](#). *arXiv preprint arXiv:2411.04496*.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024. [Hello again! llm-powered personalized agent for long-term dialogue](#). *arXiv preprint arXiv:2406.05925*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International conference on machine learning*, pages 19730–19742. PMLR.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [Dailydialog: A manually labelled multi-turn dialogue dataset](#). *arXiv preprint arXiv:1710.03957*.
- Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. 2023b. [PaCE: Unified multi-modal dialogue pre-training with progressive and compositional experts](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13402–13416, Toronto, Canada. Association for Computational Linguistics.
- Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. [Mmconv: an environment for multimodal conversational search across multiple domains](#). In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 675–684.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: common objects in context](#). In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. [Visual instruction tuning](#). *Advances in neural information processing systems*, 36.
- Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2023. [Enhanced speaker-aware multi-party multi-turn dialogue comprehension](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2410–2423.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of llm agents](#). *arXiv preprint arXiv:2402.17753*.
- Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. [A holistic approach to undesired content detection in the real world](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. [Openvidial: A large-scale, open-domain dialogue dataset with visual contexts](#). *arXiv preprint arXiv:2012.15015*.
- Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. [Let’s go real talk: Spoken dialogue model for face-to-face conversation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16348, Bangkok, Thailand. Association for Computational Linguistics.

- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. [MELD: A multimodal multi-party dataset for emotion recognition in conversations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. [Recipes for building an open-domain chatbot](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. [Image-chat: Engaging grounded conversations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.
- Ryuichi Sumida, Koji Inoue, and Tatsuya Kawahara. 2024. Should rag chatbots forget unimportant conversations? exploring importance and forgetting with psychological insights. *arXiv preprint arXiv:2409.12524*.
- Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. [Multimodal dialogue response generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. [SALMONN: Towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *CoRR*, abs/2409.12191.
- Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.
- Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. [Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. [Baize: An open-source chat model with parameter-efficient tuning on self-chat data](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–

6278, Singapore. Association for Computational Linguistics.

Seiichi Yamamoto, Keiko Taguchi, Koki Ijuin, Ichiro Umata, and Masafumi Nishida. 2015. Multimodal corpus of multiparty conversations in 11 and 12 languages and findings obtained from it. *Language Resources and Evaluation*, 49:857–882.

Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. 2024. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems*, 36.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. [PhotoChat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152, Online. Association for Computational Linguistics.

Heming Zhang, Shalini Ghosh, Larry Heck, Stephen Walsh, Junting Zhang, Jie Zhang, and C-C Jay Kuo. 2019. Generative visual dialogue system via weighted likelihood estimation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. [Mind the gap between conversations for improved long-term dialogue generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10735–10762, Singapore. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. [STICKERCONV: Generating multimodal empathetic responses from scratch](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. [AugESC: Dialogue augmentation with large language models for emotional support conversation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

## A Dataset Filtering

We verify data integrity using GPT-4o-mini with the following questions:

- Is there complete consistency between the environmental, spatial, and temporal features of the settings within the session? For example, it would be contradictory if one setting depicts daytime while the other depicts nighttime, or if spatial features (e.g., location or layout) and time progression are logically inconsistent. (Yes or No)
- Do all sessions within the episode maintain a plausible continuity in time, space, and context? For example, any stated time intervals or implied transitions between settings should be logical and coherent. (Yes or No)
- Are all participants depicted as fully engaging with the setting in real time? References to past or future events should not imply detachment from the present interaction (e.g., avoiding phrases like “for our next scene” or references to reviewing recorded footage). (Yes or No)
- Are all settings within the session entirely realistic? Any elements that seem exaggerated, cartoonish, or overly stylized for natural conversation or interaction should be avoided. (Yes or No)
- Is each setting fully utilized and referenced in the conversation? All settings presented within the session must have a clear role in the dialogue or interaction, without any being neglected. (Yes or No)
- Do all spoken lines reflect the tone and context of natural, real-time interaction? For instance, lines should avoid referring to the setting or events in a way that suggests they are pre-recorded, staged, or viewed from an external perspective. (Yes or No)

Then, we perform basic filtering by checking that all speaker names within each episode are unique, ensuring that the two partners in each session are distinct, confirming that each conversation has at least eight turns, and verifying that all three partners participate in at least one session. Any episodes that fail to meet these criteria are excluded from the dataset.

## B Prompts List

We use gpt-4o-mini to construct our M<sup>3</sup>C. Table 11 shows the prompt for caption refinement, Table 12 shows the prompt for location assignment, Table 13 shows the prompt for scenario generation, Table 14 shows the prompt for caption pair validation, Table 15 shows the prompt for session conversation generation, Table 16 the prompt for shows modality tagging, Table 17 shows the prompt for memory summarization, Table 18 shows the prompt for memory linking, Table 19 shows the prompt for memory tagging, and Table 20 shows the prompts for episode validation.

## C Dataset Examples

Please refer to Table 21, Table 22, and Table 23 for conversation examples of M<sup>3</sup>C episodes, Table 24 for an example of memory summarization, and Table 25 for an example of links between multimodal memories.

## D Implementation Details

We use all pre-trained models through Hugging Face Transformers (Wolf et al., 2020). Our model is built based on the Qwen2-VL-2B model, using exactly 4B parameters (2B for the dialogue module and 2B for the retrieval module). Please see below for the implementation details of each module of the model.

### D.1 Dialogue Module

The Dialogue Module is designed to process multiple modalities for generation tasks such as conversation generation, summarization, and memory linking. By leveraging the multitasking capabilities of the vision-instruct model, the module is fine-tuned with task-specific instructions, system prompts, and prefixes. This allows a single model to efficiently handle diverse tasks like dialogue generation and summarization.

**Conversation Generator.** Our model generates responses by considering several factors, including the speaker’s identity, the ongoing conversation history, and relevant memories. It also supports model-to-model conversations without human intervention, with each model acting as a distinct speaker in a multi-party conversation. When a model wants to speak, it generates a [YES] token. If only one model generates this token, it takes the

### Example of summary generation

|                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                       |
|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| .....                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                       |
|  |                                                                                                                                                                                                                                                                                                                                                                                                       |
| <b>Alex</b>                                                                        | Yeah, it takes a lot of skill to navigate through the track like that. I wonder what kind of techniques they use to keep their balance while pushing off.                                                                                                                                                                                                                                             |
| <b>Morgan</b>                                                                      | It’s all about the rhythm, I guess. Just like in a game of polo, you need to know when to hit and when to pull back.                                                                                                                                                                                                                                                                                  |
| .....                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                       |
| <b>User</b>                                                                        | [SUMMARY]                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Summary</b>                                                                     | <b>I find it fascinating how much skill is required to maintain speed and balance in cross-country skiing, which seems similar to other sports.</b> (from first session, about myself) <sep> Jamie highlighted the importance of rhythm in skiing, comparing it to the game of polo, which makes me think about the coordination needed in both sports. (from first session, about Jamie) <sep> ..... |

Table 10: Example of memory generation across the entire session using the [SUMMARY] token. The memory includes both image-related content and conversation history, with different memory segments separated by the <sep> token.

speaking turn. If multiple models generate the token, the one with the highest probability is selected to speak, ensuring smooth turn-taking and a natural conversation flow. If the model determines that introducing an image or audio is appropriate, it generates [RET\_IMG] or [RET\_AUDIO], respectively. Otherwise, it generates [NO\_RET].

**Memory Summarizer.** Our model summarizes the dialogue history into a personal memory at the end of each session, capturing information about itself and its conversation partners. This process takes the entire session history and a [SUMMARY] token as input, while the system prompt specifies the perspective from which the memory should be generated. When multiple memory entries are created, they are separated using the <sep> delimiter.

Table 10 presents an example of memory generation using the [SUMMARY] token. This memory includes both dialogue and image-related information, with multiple segments separated by the <sep> token. In this example, Alex mentions techniques for maintaining balance in cross-country skiing, and the summary highlights how this skill is simi-

lar to other sports, such as polo.

**Memory Linker.** We adopt a memory-linking approach where newly generated memories are connected to relevant past ones (Jang et al., 2024). While Jang et al. (2024) propose methods for establishing memory connections within a session, extend this approach to link not only dialogues within a session but also multimodality associated with memory. This allows for the formation of a richer memory network that integrates multiple modalities rather than relying solely on text-based memories. The model determines these links by responding [POSITIVE] for relevant memories and [NEGATIVE] otherwise.

**Training Strategy.** The Dialogue Module generates utterances based on audio, images, and text within a session. It also handles tasks such as summarization and memory linking. To train the model for these multi-task operations, the training process is divided into four distinct stages, as outlined below.

#### 1. Session Utterance

The model is trained to process session utterances. Images are directly provided as input without modification. However, audio is converted into captions and represented in the format `<start_audio> caption <end_audio>`. During this stage, the model learns using only the text-based representation of audio.

#### 2. Main Speaker & Memory Retrieval

The model is trained when to output turn tokens and retrieval tokens. It decides when to take the role of the main speaker by producing a [YES] or [NO] token. Similarly, it learns the timing for memory retrieval using tokens like [RET\_IMG] for visual memory, [RET\_AUDIO] for auditory memory, and [NO\_RET] when retrieval isn't needed.

#### 3. Summarization & Memory Linking

The model is trained to generate session summaries to create structured memory. Also, it learns to establish memory links between session memory and observed images/audio, as well as between session memory and chat history.

#### 4. Audio Understanding via CLAP

In the final stage, we enhance the model with a frozen CLAP, using a linear layer adapter

to enable audio comprehension. Instead of using only captions, the model now receives audio embeddings between `<start_audio>` and `<end_audio>`, allowing for direct audio understanding and richer multi-modal interactions.

Each stage progressively enhances the model's capabilities, ensuring it can handle multimodal dialogue generation and memory management. For detailed training settings, please refer to the section below.

**Training Setting.** We apply the LoRA (Hu et al., 2022) to fine-tune Qwen2-VL-2B-Instruct, training with a cross-entropy loss, a maximum input length of 1024, and a learning rate of  $1 \times 10^{-3}$ . We use early stopping with a maximum of 3 epochs. For LoRA configuration, we set  $r = 8$ ,  $\alpha = 16$ , dropout=0.05, and apply it to the `q_proj` and `v_proj` layers. During training stages 1, 2, and 3, we use a batch size of 32, while in stage 4, where the audio adapter is trained, we reduce the batch size to 16. Training is conducted on 8 NVIDIA RTX A6000 GPUs. We employ the following system prompts and sequence inputs for each task.

#### System Prompt for Conversation Generation

"<<SYS>>Please generate the speaker's next utterance. Main-Speaker: [MAIN SPEAKER NAME]-MAIN SPEAKER JOB [SUB SPEAKER NAME]-SUB SPEAKER JOB [MEMORY] MEMORY SENTENCE 1 [LINK] LINK MEMORY SENTENCE [LINK] ... [MEMORY] MEMORY SENTENCE N [LINK] <</SYS>>"

#### System Prompt for Memory Summarization

"<<SYS>>Please review this session's conversation history and summarize the key points [ABOUT WHO] needs for the next session. Separate each memory with <sep>. ...<</SYS>>{Full Sessions}[SUMMARY]"

#### System Prompt for Memory Linking

"<<SYS>>Please determine whether the two memory elements are related or if they reflect an update. [POSITIVE] or [NEGATIVE]<</SYS>>memory sentence 1: {MEMORY 1} memory sentence 2: {MEMORY 2}"

## D.2 Retrieval Module

This module retrieves multimodal memory from prior sessions based on the context of the ongoing

session, preserving speakers’ perspectives and sensory experiences by recalling past conversations, images, and audio. However, effectively retrieving such multimodal memory requires a model capable of jointly encoding diverse modalities. Traditional uni-modal encoding models, such as CLIP (Radford et al., 2021) or CLAP (Wu et al., 2023), are limited to processing specific modality pairs (e.g., image-text, audio-text), making it challenging to represent complex multimodal data, where image, audio, and text are freely concatenated and must be jointly encoded within the same embedding space. To overcome this limitation, inspired by Jiang et al. (2024), we build a unified multimodal retrieval model that directly concatenates image, audio, and text representations within a shared embedding space, enabling efficient cross-modal memory retrieval.

**Training Strategy.** As illustrated in Figure 3, this model is trained using a contrastive learning framework that transforms a vision-audio-language model into a multi-modal embedding model, effectively aligning image, text, and audio representations. To construct a relevant query-target pair, the query consists of the current session’s conversation along with perceived images or audio, represented as  $q^+$ . The target,  $t$ , corresponds to a memory retrieved from previous sessions, containing a single modality. Using a pre-trained VLM, extended with CLAP for audio, we obtain the query and target embeddings,  $(h_{q^+}, h_t)$ , by extracting the final-layer vector representation of the last token and train the embedding model with the standard InfoNCE loss (van den Oord et al., 2018), incorporating in-batch negatives and hard negatives. To enhance its multimodal retrieval capability, the training of the retrieval module is divided into two stages:

#### 1. Text & Image-Based Retrieval

In the first stage, the model is trained without direct audio input. Instead, audio is replaced by its corresponding captions, and only the session’s chat history and images are included for tuning. This allows the model to focus on retrieving memory based on textual and visual context.

#### 2. Audio-Enhanced Retrieval

In the second stage, similar to the Dialogue Module, we integrate the CLAP with a linear adapter to enable direct audio comprehension. By incorporating audio embeddings, the re-

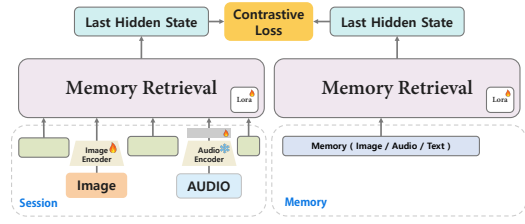


Figure 3: Overview of the retrieval module architecture, which utilizes contrastive loss for memory retrieval.

trieval module learns to process and retrieve memory using full multi-modal context, including raw audio information.

This two-stage training approach ensures that the retrieval module effectively aligns and retrieves multi-modal representations, improving memory recall and interaction coherence across sessions. For detailed training settings, please refer to the section below.

**Training Setting.** We apply the LoRA to fine-tune Qwen2-VL-2B-Instruct, training with a cross-entropy loss, a maximum input length of 1024, and a learning rate of  $1 \times 10^{-3}$ . We use early stopping with a maximum of 3 epochs. For LoRA configuration, we set  $r = 8$ ,  $\alpha = 16$ , dropout=0.05, and apply it to the  $q\_proj$  and  $v\_proj$  layers. During training stage 1, we use a batch size of 32, while in stage 2, where the audio adapter is trained, we reduce the batch size to 16. Training is conducted on 8 NVIDIA RTX A6000 GPUs.

## E Human Evaluation Details

We provide evaluators with guidelines, as shown in Table 26, which include descriptions of the evaluation metrics and scoring criteria. Evaluators study these guidelines and complete a pilot test administered by the evaluation agency to ensure they fully understand the criteria before participating in the evaluation process. Additionally, Table 27 presents detailed human evaluation scores, including the average scores by group and inter agreement scores.

## F Model Examples

Please refer to Table 28 for an example of conversation among models.

## G Example of Model Comparison

Our model enables all speakers to engage in immersive, simultaneous conversations within a shared spatial and temporal context, based on the same

modality input. Table 29 and Table 30 show how our model’s responses differ from those of existing models when processing visual and audio inputs, respectively.

---

|                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Generate a caption describing the image.<br>The caption should begin directly with details about the main objects and the situation.<br>Avoid starting with phrases like “In this image”.<br>Write the caption from the observer’s perspective, focusing solely on the given scene without considering other possible contexts. |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

---

Table 11: Prompt for caption revision.

---

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>For image</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| ###Instruction:<br>1. Given an image caption, identify the location using a single, general word.<br>2. Ensure the location is broadly applicable and captures the essence of the description.<br>3. If the specific location is unclear, make an educated guess based on where the described elements are typically found.<br>4. For example, if objects like a refrigerator, stove, and cooking utensils are visible, the location would be “kitchen”. Similarly, a desk, computer, and books might suggest “office”.<br>5. Avoid adding any additional explanations, introductions, or conclusions.<br><br>###Caption:<br>{CAPTION}<br><br>###Answer: |
| <b>For Audio</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| ###Instruction:<br>1. Given an audio caption, identify the location using a single, general word.<br>2. Ensure the location is broadly applicable and captures the essence of the description.<br>3. If the specific location is unclear, output “none”.<br>4. If the sounds are fictional, unrealistic, or not commonly associated with real-world locations, output “none”.<br>5. For example, sounds like sizzling, a refrigerator hum, and chopping might suggest “kitchen”.<br>6. Avoid adding any additional explanations, introductions, or conclusions.<br><br>###Caption:<br>{CAPTION}<br><br>###Answer:                                        |

---

Table 12: Prompt for specifying location based on caption.

---

###Instruction:

1. Four speakers are involved in a conversation episode, consisting of one main speaker and three partners. Each session will feature the main speaker alongside two different partners.
2. There will be three continuous sessions. In each session, the main speaker and two different partners will witness two settings unfolding in real-time right in front of them. These settings are live observations, experienced not only visually but also audibly, and are not based on past memories, images, or pictures.
3. From a provided list of settings, select two unique settings for each session. Ensure that no setting is repeated across the sessions.
4. There will be a time gap between each session. Choose suitable time intervals from the following options: “a few hours later,” “a few days later”, “a few weeks later”, “a few months later”, and “a couple of years later”.
5. Clearly define the names and relationships of all speakers to provide context for their interactions and to enhance the flow of conversation.
6. During each session, the speakers should engage in discussions that logically connect to the context of the settings they observe, without directly referencing specific details of those settings. For instance, if the setting involves cooking, one might inquire about the dish being prepared.
7. Ensure that the combination of settings in each session does not include contradictory elements, such as differing weather conditions or inconsistent times of day that would not logically coexist.
8. Please do not generate any other opening, closing, and explanations.

###Setting list:

- {MODALITY LIST}

###Response:

- Main speaker name: {insert name}
  - Main speaker relationship: {insert relationship}
  - Partner 1 name: {insert name}
  - Partner 1 relationship: {insert relationship}
  - Partner 2 name: {insert name}
  - Partner 2 relationship: {insert relationship}
  - Partner 3 name: {insert name}
  - Partner 3 relationship: {insert relationship}
  - Scene numbers for session 1: {insert scene number 1}, {insert scene number 2}
  - Two partners' names in Scene 1: {insert partner name 1}, {insert partner name 1}
  - Time interval between session 1 and 2: {insert time interval}
  - Scene numbers for session 2: {insert scene number 1}, {insert scene number 2}
  - Two partners' names in Scene 2: {insert partner name 1}, {insert partner name 1}
  - Time interval between session 2 and 3: {insert time interval}
  - Scene numbers for session 3: {insert scene number 1}, {insert scene number 2}
  - Two partners' names in Scene 3: {insert partner name 1}, {insert partner name 1}
- 

Table 13: Prompt for scenario generation.

---

###Instruction:

1. Two captions are provided, each describing either an image or an audio clip.
2. Determine whether the two captions are aligned and compatible.
3. Assume that the speakers are observing or listening to the described image or audio in real-time within the same context.
4. If the two captions cannot coexist within the same context, output “no” Otherwise, output “yes”.
5. For example, if one caption describes a ski resort and the other describes a meadow, the two captions cannot coexist, and the output should be “no”.
6. Please do not generate any other opening, closing, and explanations.

###Caption:

- {CAPTION A}
- {CAPTION B}

###Answer:

---

Table 14: Prompt for checking modality alignment within a session.

---

###First session conversation:

{FIRST SESSION}

###Second session conversation:

{SECOND SESSION}

###Instruction:

The following is a third-session conversation, {TIME INTERVAL} the second-session conversation, between {MAIN SPEAKER NAME} ({MAIN SPEAKER RELATIONSHIP}), {PARTNER 1 NAME} ({PARTNER 1 RELATIONSHIP}) and {PARTNER 2 NAME} ({PARTNER 2 RELATIONSHIP}).

During the conversation, {MAIN SPEAKER NAME}, {PARTNER 1 NAME} and {PARTNER 2 NAME} participate in the following settings in real-time:

- {CAPTION 1}
- {CAPTION 2}

The following speakers, {MAIN SPEAKER NAME}, {PARTNER 1 NAME} and {PARTNER 2 NAME}, participate in the settings in real-time during the conversation, so we avoid directly describing the settings. Instead, we proceed with the dialogue in a way that shows they are sharing the same settings. We also avoid discussing impressions, feelings, and imagination.

For example, if there is a cutting board and cooking ingredients in the kitchen, we could continue the conversation by asking what kind of dish to make. We maintain the natural flow of the dialogue, not just talking about settings.

Especially, please write the dialogue with the understanding that it is a continuation of the previous sessions. Feel free to reference the previous sessions as needed, and consider the time interval between the sessions. Avoid forcing connections if they don’t naturally flow; the dialogue should reflect the continuity without feeling contrived.

The conversation should begin with {PARTNER 1 NAME} or {PARTNER 1 NAME} initiating it. Also, each speaker’s statements start with their name in brackets. For example, {PARTNER 1 NAME}’s statements start with “[{PARTNER 1 NAME}]”.

The order of the speakers’ utterance is determined randomly. There is no need to follow a fixed order when speaking.

Complete the conversation in exactly that format.

---

Table 15: Prompt for session conversation generation.

---

###Instruction:

1. Two settings and a list of dialogue utterances are provided.
2. Based on the dialogue, identify the utterance number where each of the two settings is first mentioned, seen, or heard by the speakers in real-time, including both visual and auditory elements.
3. Select only one utterance number for each setting.
4. Please do not generate any other opening, closing, and explanations.

###Settings:

- {CAPTION A}
- {CAPTION B}

###Utterances:

{UTTERANCE LIST}

###Response:

---

Table 16: Prompt for modality tagging.

---

###Second session conversation:

- First Setting: {FIRST SETTING}
- Second Setting: {SECOND SETTING}
- {SESSION CONVERSATION}

###Instruction:

1. Please summarize the conversation from {MAIN SPEAKER NAME}'s perspective. The summary should focus on what {MAIN SPEAKER NAME} needs to remember for the next conversation.
2. The summary should include emotions, thoughts, facts, and commitments expressed during the conversation, but only those explicitly revealed during the discussion. Exclude general descriptions or background information about the setting.
3. Do not include content from the setting descriptions in the memory. Only include information shared during the conversation.
4. Ensure the memory entries are concise and focus only on unique, relevant information necessary for the next session. Each memory entry must focus on one person only, without combining multiple speakers into one sentence or perspective.
5. Each memory entry should be a separate sentence or key idea, but avoid summarizing every single statement unless it is crucial for the next session. All memory entries must be separated by a "/". For example, "{sentence 1} / {sentence 2} ...".
6. The memory must be presented from {MAIN SPEAKER NAME}'s perspective, focusing on summarizing the overall key points or themes rather than individual conversational details.
7. Avoid duplicating information already included in the summary. Consolidate similar points into a single entry if possible.
8. If a memory element updates or replaces existing information, ensure the updated version reflects the latest understanding and avoid redundancy.
9. When summarizing {MAIN SPEAKER NAME}'s memory about themselves, start the memory entry with "I".
10. At the end of each memory sentence, specify who the memory is about from {MAIN SPEAKER NAME}'s perspective using the format "(about {name})". For example, "{Memory Sentence} (about {name})". Ensure that the parentheses end with a period to complete the sentence.
11. Do not use group references like "(about all)" or combine multiple people into one memory entry.
12. If there are no memories to summarize, output "no memory".
13. Avoid adding extra explanations, introductions, or conclusions.

###Response:

---

Table 17: Prompt for memory generation.

---

```

###First session setting:
1 - {CAPTION}
2 - {CAPTION}

###Second session setting:
3 - {CAPTION}
4 - {CAPTION}

###{MAIN SPEAKER NAME}'s memory from
first session:
{FIRST SESSION MEMORY}

###{MAIN SPEAKER NAME}'s memory from
second session:
{SECOND SESSION MEMORY}

###Instruction:
1. Summarized conversation memories from the
perspective of {MAIN SPEAKER NAME} and
the settings used in the dialogue are provided.
2. {MAIN SPEAKER NAME} references the set-
tings and memories from previous sessions during
subsequent conversations to ensure seamless con-
tinuity in the dialogue.
3. When referring to memories, the goal is to pro-
vide rich context by connecting related elements.
4. If the provided settings and memories pertain
to the same context or reflect subsequent updates,
they should be connected.
5. The format for expressing connections is
“{{NUMBER}}-{{NUMBER}}”and should be
output on separate lines.
6. Each {{NUMBER}} should contain only a sin-
gle digit. When multiple memories are connected,
write each connection on a new line.
7. Avoid adding extra explanations, introductions,
or conclusions.

###Response:

```

---

Table 18: Prompt for memory linking.

---

```

###Third session conversation:
{SESSION CONVERSATION}

###{MAIN SPEAKER NAME}'s memory:
{MEMORY LIST}

###Instruction:
1. {MAIN SPEAKER NAME} engages in a third
session conversation based on their memory.
2. The utterances by {MAIN SPEAKER NAME}
are informed by their referenced memory.
3. During the conversation, if any part of {MAIN
SPEAKER NAME}'s utterances relies on memory,
that part must be specifically identified.
4. The memory consists of two components: (1)
Summarized information from previous session
conversations, and (2)Real-time observations by
{MAIN SPEAKER NAME} during prior sessions.
5. The output format should follow the structure:
“Utterance Letter-Memory Number” (e.g., “A-3”).
6. Each line should contain a single “Utterance
Letter-Memory Number” entry, separated by a
newline.
7. If an utterance references multiple memory
entries, each reference must be listed on a new line
rather than combining memory numbers on the
same line.
8. If the same information needs to be referenced,
prioritize using the memory from the most recent
session.
9. If no memory is referenced and there is nothing
to output, return “none”.
10. Avoid adding extra explanations, introductions,
or conclusions.

###Response:

```

---

Table 19: Prompt for memory tagging.

---

###Conversation episode:

\* First session:

{FIRST SESSION}

\* Second session:

{SECOND SESSION}

\* Third session:

{THIRD SESSION}

###Instruction:

- Each conversation episode consists of three sessions, with one main speaker and three conversation partners.
- In each session, the main speaker converses with two conversation partners, and the partner combinations may vary between sessions.
- During each session, the speakers either observe or participate in two real-time settings.
- The settings being observed are not images, photos, or past memories but are happening live in front of them, with both visual and auditory elements unfolding in real time.
- As a judge, you will evaluate the conversation episode based on the following criteria.
- Please refrain from adding any other introductions, conclusions, or explanations.

###Metric:

1. Is there complete consistency between the environmental, spatial, and temporal features of the settings within the session? For example, it would be contradictory if one setting depicts daytime while the other depicts nighttime, or if spatial features (e.g., location or layout) and time progression are logically inconsistent. (Yes or No)
2. Do all sessions within the episode maintain a plausible continuity in time, space, and context? For example, any stated time intervals or implied transitions between settings should be logical and coherent. (Yes or No)
3. Are all participants depicted as fully engaging with the setting in real time? References to past or future events should not imply detachment from the present interaction (e.g., avoiding phrases like “for our next scene” or references to reviewing recorded footage). (Yes or No)
4. Are all settings within the session entirely realistic? Any elements that seem exaggerated, cartoonish, or overly stylized for natural conversation or interaction should be avoided. (Yes or No)
5. Is each setting fully utilized and referenced in the conversation? All settings presented within the session must have a clear role in the dialogue or interaction, without any being neglected. (Yes or No)
6. Do all spoken lines reflect the tone and context of natural, real-time interaction? For instance, lines should avoid referring to the setting or events in a way that suggests they are pre-recorded, staged, or viewed from an external perspective. (Yes or No)

###Response:

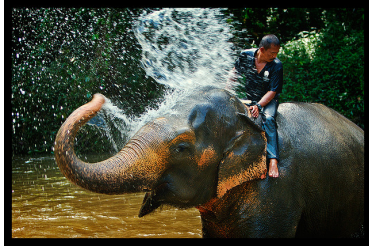
- 1.
  - 2.
  - 3.
  - 4.
  - 5.
  - 6.
- 

Table 20: Prompt for episode validation.

---

**Session 1**

---



**Jamie** Did you see how the elephant just sprayed water? It's like it's having a blast!

**Taylor** Right? And look at the rider; he's handling that balance like a pro. I wouldn't want to fall into that river.

**Alex** I can only imagine how muddy it must be out there! Bet the rider picked the perfect time to hop on.

**Jamie** I wonder if they had to train for this. It looks like a lot of fun, but maintaining that balance can't be easy.



**Taylor** Speaking of fun, check out those guys on the raft! They're about to hit that small waterfall.

**Alex** They look a bit tense, but it seems like they're ready for the rush. I hope they paddle right!

**Jamie** I can't tell if they're nervous or excited. That's such a perfect day for a rafting adventure, though.

**Taylor** Definitely! It feels like a great day for both activities. I'd love to try that raft ride sometime.

**Alex** Me too! But only if I can guarantee not falling out. That water looks pretty wild.

**Jamie** And you know how slippery it can get! But I think with a good crew, we could make it work.

**Taylor** Yeah, teamwork makes the dream work. Just like that elephant and rider out there.

**Alex** Now that's a solid comparison! Might be more at risk of muddy water than a tidal wave, though.

**Jamie** True! Plus, no one would want a trunkful of water splashed on them while they're navigating!

**Taylor** Right! But at the end of the day, it all just adds to the fun. Both experiences seem pretty unforgettable.

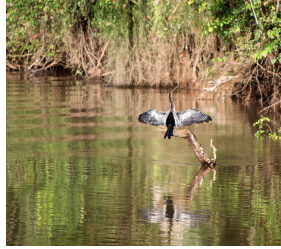
---

Table 21: An example of the first session in the dataset episode.

---

**Session 2**

---



**Jamie** That cormorant over there looks like it's ready for a photoshoot! Its reflection is almost perfect on the water.

---

**Clotho dataset - 39935.wav**

---

**Taylor** I can't believe how still the water is right now. It's like a painting. But just wait until that motorboat speeds by!

**Alex** Right? The contrast between the calmness of the river and the chaos of the engine is pretty interesting.

**Jamie** Totally! Kind of like how we were all raving about the rafting and the elephant the other day. Different vibes, but equally fun!

**Taylor** Speaking of fun, I wonder if those folks on the boat are going to go full throttle as soon as they hit open water.

**Alex** They probably will! I mean, who wouldn't want to crank up the excitement with some splashes like that?

**Jamie** It's like they're showcasing their own version of adventure. You know, I'd still prefer the elephant ride for a more chilled experience.

**Taylor** Fair point! But the adrenaline rush from that motorboat sounds tempting too. I'd want to try both!

**Alex** Agreed. We should plan a day and do both! Maybe we can start with the raft ride, ease into the motorboat energy, and then watch those cormorants dry off.

**Jamie** That sounds like a solid plan! Just hope the motorboat doesn't stir up too much mud and ruin our view of the cormorant.

**Taylor** True! But if it does, it'll give us some splashes to laugh about. Plus, we'll still have that serene moment before it arrives.

**Alex** That balance of calm and excitement is what makes these adventures memorable, right?

**Jamie** Exactly! Just like our last outing. It's all about enjoying whatever comes our way!

**Taylor** And capturing those moments! A bit of mud won't matter when we're having a blast.

**Alex** Alright then, let's keep an eye on the boat and the cormorant. It's the best of both worlds!

---

Table 22: An example of the second session in the dataset episode.

---

**Session 3**

---



**Taylor** Look at this truck! I can't believe they've got a whole advertisement for Moscato right here. It's like they're throwing a party in the middle of the street!



**Alex** Right? It really stands out with that bright pink color. Not to mention that candle on the parking meter—it's like an indoor celebration outdoors!

**Morgan** Totally! It's a fun touch. Speaking of parties, does this give anyone else an idea for a get-together? Maybe we could have a little wine tasting!

**Alex** A wine tasting sounds perfect! We could grab a few different types. I remember our last adventure—we really enjoyed trying new things together.

**Taylor** Yes! And we can recreate the vibe with some snacks. Maybe some cheese and crackers to go along with the wine.

**Morgan** Great idea! I'd also suggest we include some fruit pairs—grapes and strawberries could complement the Moscato well.

**Alex** Definitely! By the way, it's been a while since our last outing. Remember that rafting trip? I think we should make this tasting our next adventure!

**Taylor** Awesome plan! First, the wine tasting, then we can decide on our next outdoor adventure. Who knows, maybe we'll find some activity that even tops the elephant ride!

**Morgan** Sounds exciting! Just think of the great stories we would have. Plus, with the candle on the meter, it feels like we already have a fun theme going.

**Alex** Right? A whimsical birthday party vibe! We could even get some decorations to add to it.

**Taylor** We'll have to check if we can get some balloons too. It would be perfect! You know, make it festive.

**Morgan** Yes! And who knows, maybe we'll attract a few cheerful people passing by. After all, who doesn't love a celebration?

**Alex** Exactly! I say we take a few pictures to document the moment, just like we did when we spotted that cormorant the other day.

**Taylor** Great call! Those memories are always fun to look back on. Plus, it'll give us a reason to gather again soon!

**Morgan** It's settled then! Wine tasting with a little playful theme and photo moments sounds like a fantastic idea!

**Alex** I'm in! Let's plan it out. I can already envision some memorable stories coming out of it, just like our previous outings.

**Taylor** Can't wait! Let's do it!

---

Table 23: An example of the third session in the dataset episode.

---

**Example Session**

---



- Jamie** Did you see that snowboarder? They really nailed that jump!
- Sam** Yeah, it looked awesome! The way they soared above the slope was incredible. I wish I could do that.
- Alex** It takes a lot of practice to get those air tricks down. I just love watching the way they spin mid-air.
- Jamie** For sure! And it seems like the snow conditions are perfect today, too. Really nice and packed.
- Sam** I wonder how the skiers are faring. That jump off the slope was impressive as well.
- Alex** Definitely! I noticed that skier launched off with so much speed. The sun must be ideal for those fast runs.
- Jamie** I hope they have good control on that steep descent. It's always thrilling to watch them tackle those angles.



- Sam** It's interesting how different the techniques are between skiing and snowboarding. One is all about edges, while the other is about balance.
- Alex** Right! Each sport has its own style. I think that blending both could create a unique trick.
- Jamie** Speaking of tricks, we should try to record some of these stunts. It would be great to have a video to look back on later!
- Sam** Great idea! Even if we can't perform them ourselves, at least we can capture the thrill of it all.
- Alex** Let's pick a spot that has a good view of both the skiers and snowboarders.
- Jamie** Perfect! I'll grab my phone. It'll be fun to have some footage to remember this day.
- Sam** I'm in! Maybe we can even get a few shots of those stylish wipeouts—those are always entertaining.
- Alex** Absolutely, let's not forget to cheer them on, too. It makes it more exciting for everyone while we film!

---

**Summarized Memory**

---

1. I enjoy watching snowboarders and skiers perform tricks in the snow. (from first session, about me)
  2. Jamie suggested recording some stunts, and I think it would be fun to have footage to remember this day. (from first session, about Jamie)
  3. Sam is interested in capturing wipeouts, which adds an entertaining element to our video. (from first session, about Sam)
  4. I believe that each sport has its own unique style and that blending both could create something interesting. (from first session, about me)
  5. It's important to cheer on the performers while filming, as it creates a more exciting atmosphere. (from first session, about me)
- 

Table 24: An example of a memory summarized from the perspective of the main speaker (Alex) in a conversation. The summarized memory is combined with modality elements from the session to form a multimodal memory.

---

**Emergency vehicle siren goes off for the entire time**

**Clotho dataset - 44772.wav**

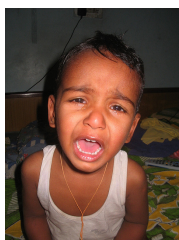
I noted that there was a lot going on that night, including the rally and the lights along the pathway.  
(from first session, about me)

---



Observing the zebras revealed to me that different species have their own unique ways of relaxing in their environments (about Jamie)

---



---

Table 25: Examples of multimodal memory links. The first row shows links between audio clips and text memories, the second row shows links between images and text memories, and the third row shows links between two different modalities.

---

## Multimodal Conversation Evaluation

You are asked to evaluate a multimodal conversation, which integrates multiple modalities such as images or audio within the dialogue. Each conversation episode consists of three consecutive sessions, with two modalities featured in each session. Each episode involves four speakers in total, with three randomly selected to participate in any given session. Your evaluation should focus on both the overall quality of the conversation and the effectiveness with which the modalities are reflected in the dialogue.

### \* For Dataset:

#### **(1) coherence and consistency**

Is the conversation coherent throughout, with no contradictions or logical inconsistencies?

#### **(2) modalities alignment**

Do the modalities in a single session align within the same context of place and time? For example, if one modality features an image of a winter ski resort and another modality includes the sound of a beach ocean, that would be considered inconsistent and incorrect.

**(3) modalities engagement** Are the presented modalities (image, audio) appropriately reflected in the dialogue, and do the speakers actively engage with the modalities in real time?

#### **(4) memorability**

Do conversations in subsequent sessions effectively recall and build upon the content of previous sessions, considering the time intervals between them? Additionally, does the memory summary of the first and second sessions accurately reflect the content of those conversations?

### \* For Model:

#### **(1) naturalness**

Do all speakers demonstrate fluency and natural conversational flow, with smooth transitions between dialogue (turns, sessions) and seamless incorporation of modality cues (images and audio) without awkward pauses or disruptions?

#### **(2) Immersive**

Do the speakers naturally engage with the presented modalities (images and audio) as if they are experiencing them simultaneously in the same spatial and temporal environment, integrating them seamlessly into the conversation to create a realistic and contextually rich dialogue?

#### **(3) memorability**

Do the speakers accurately recall and incorporate relevant information from previous sessions, including both textual and multimodal (visual and auditory) memory, using this information appropriately to maintain context and enhance the flow of the conversation?

### \* Possible Options:

Bad / Slightly Bad / Neutral / Slightly Good / Good

A top score of 5 is reserved for conversations that are entirely flawless, while a score of 1 is assigned when severe issues make a proper evaluation impossible. A rating of 4 indicates an overall excellent performance, whereas a score of 3 reflects the presence of only minor imperfections. A score of 2 is given when notable issues are detected, even though they do not completely undermine the conversation's quality.

---

Table 26: Human evaluation guidelines containing descriptions of metrics and criteria.

| Metric                    | Group 1 | Group 2 | Group 3 | Average | Agreement | Machine |
|---------------------------|---------|---------|---------|---------|-----------|---------|
| <b>Dataset Quality</b>    |         |         |         |         |           |         |
| Coherence and Consistency | 4.84    | 4.85    | 4.73    | 4.81    | 0.54      | 4.99    |
| Memorability              | 4.56    | 4.79    | 4.54    | 4.63    | 0.30      | 4.99    |
| Modalities Alignment      | 4.13    | 4.33    | 4.18    | 4.21    | 0.27      | 4.26    |
| Modalities Engagement     | 4.37    | 4.48    | 4.23    | 4.36    | 0.32      | 4.57    |
| <b>Dataset Overall</b>    | 4.48    | 4.61    | 4.42    | 4.50    | -         | 4.70    |
| <b>Model Performance</b>  |         |         |         |         |           |         |
| Naturalness               | 4.28    | 4.58    | 4.16    | 4.34    | 0.20      | 4.68    |
| Immersive                 | 4.12    | 4.25    | 4.05    | 4.14    | 0.40      | 4.56    |
| Memorability              | 4.20    | 4.68    | 4.17    | 4.35    | 0.28      | 4.46    |
| <b>Model Overall</b>      | 4.20    | 4.50    | 4.13    | 4.28    | -         | 4.57    |

Table 27: Full evaluation results for both dataset quality and model performance, including group-specific scores and agreement. Agreement refers to the rate at which responses match across all conversations from all three groups.

|                                                                                     |                                                                                                                                                                 |
|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Alex</b>                                                                         | Did you see that turn I just made? It felt perfectly sharp!                                                                                                     |
| <b>Taylor</b>                                                                       | I did! That's impressive. That's the key to leaning more, isn't it?                                                                                             |
| <b>Morgan</b>                                                                       | And it's also about control, like you mentioned earlier. Those small adjustments can make a huge difference.                                                    |
| <b>Alex</b>                                                                         | Absolutely! Just like with the way we prepare for a match, small adjustments can lead to big wins.                                                              |
| <b>Morgan</b>                                                                       | It's the same with skiing, isn't it? A slight tweak here and there can make the difference between a perfect turn and a misstep.                                |
| <b>Alex</b>                                                                         | Definitely! And what's interesting is how we can learn from other skiers, just like we can from games or even sports.                                           |
|  |                                                                                                                                                                 |
| <b>Morgan</b>                                                                       | It seems they find the right line, don't they? You really have to know how to read the snow.                                                                    |
| <b>Taylor</b>                                                                       | Right! And just as we need to adapt to the game, we might want to think of how to adapt to the skiing conditions.                                               |
| <b>Alex</b>                                                                         | Speaking of adaptation, I noticed that the skiers seem to change their speed and direction frequently. It's fascinating how they navigate through those shifts. |
| <b>Morgan</b>                                                                       | I wonder how they gauge their speed compared to the rest of the field, just like we must be attentive to our competition.                                       |

Table 28: Example of a conversation among three models.

|                |                                                                                                                                                                                                 |
|----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>User</b>    | Wow, I'm so excited! Where should we go first?                                                                                                                                                  |
|                |                                                                                                                |
| <b>Qwen2</b>   | It looks like you're at an amusement park with lots of fun rides and attractions to choose from! Here's how you might decide where to start: .....                                              |
| <b>LLaVA</b>   | The image shows a lively carnival with a large Ferris wheel and other rides, including a roller coaster. There are also several people walking around and enjoying the festive atmosphere ..... |
| <b>o3-mini</b> | Check out that giant Ferris wheel—it's basically the park's lighthouse! I'd kick things off there .....                                                                                         |
| <b>Ours</b>    | Is that ride near here? That could be a great idea! Let's check the map and see where it is.                                                                                                    |

Table 29: An example of a live human chat comparing how different models respond to image input.

|                 |                                                                         |
|-----------------|-------------------------------------------------------------------------|
| <b>User</b>     | Whoa! What just happened?<br><b>The sharp crack of shattering glass</b> |
| <b>Flamingo</b> | Someone threw a glass bottle at the ground and it broke.                |
| <b>Qwen2</b>    | There was a sudden loud noise, followed by the sound of glass breaking. |
| <b>Ours</b>     | I heard that too! Something smashed nearby.                             |

Table 30: A human live chat example showing the different responses from models based on audio input. For comparison, we use Audio Flamingo-2 (Ghosh et al., 2025) and Qwen2-Audio-7B-Instruct (Chu et al., 2024b).