# LongMP-Bench: A Benchmark for Multimodal Persona Understanding in Long-Term Dialogues

**Zhuoqun Li**[1*], **Zhaopei Huang**[1*], **Wenxuan Wang**[1], **Qin Jin**[1†]
[1]Renmin University of China
{zhuoqunli, huangzhaopei, wangwenxuan, qjin}@ruc.edu.cn

## Abstract

Understanding multimodal user personas over long-term dialogues is essential for personalized and human-like dialogue systems. In realistic interactions, user personas evolve over time and are expressed through both language and visual cues. However, existing benchmarks provide limited support for evaluating such dynamic and multimodal persona understanding, due to shallow persona coverage, weak visual consistency, and static settings. We introduce LongMP-Bench, a benchmark for evaluating models' ability to understand, track, and utilize evolving multimodal user personas in long-term dialogues. We propose a scalable, multi-step data construction pipeline to synthesize extended multimodal interactions, followed by human refinement. The resulting dataset contains long-term conversations from 150 distinct users, each maintaining visual identity consistency while exhibiting progressive persona evolution. Based on LongMP-Bench, we define evaluation tasks for persona tracking, multimodal reasoning, and personalized response generation. Extensive experiments show that current multimodal large language models struggle with long-term persona consistency, persona shifts, and effective multimodal integration. Our data and code are available at https://github.com/skspass/LongMP-Bench.

## 1 Introduction

Personalized dialogue agents aim to engage users in sustained, user-centric interactions grounded in the user's persona—typically a structured representation of the user's background, preferences, and experiences (Chen et al., 2024; Xu et al., 2022b). In real-world online communication, interactions are often multimodal: exchange not only text but also images and other visual content. These multimodal signals play a critical role in expressing identity, preferences, and life events, making persona modeling substantially more complex than text-only settings. Crucially, user personas are not static. Over long-term interactions, users reveal new information, update preferences, and undergo changes in interests, habits, and life circumstances. An effective dialogue agent must therefore track persona evolution over time while maintaining identity consistency, and integrate multimodal cues to support personalized and contextually appropriate responses. Evaluating such capabilities requires benchmarks that jointly consider long-term interactions, multimodal inputs, and dynamic personas.

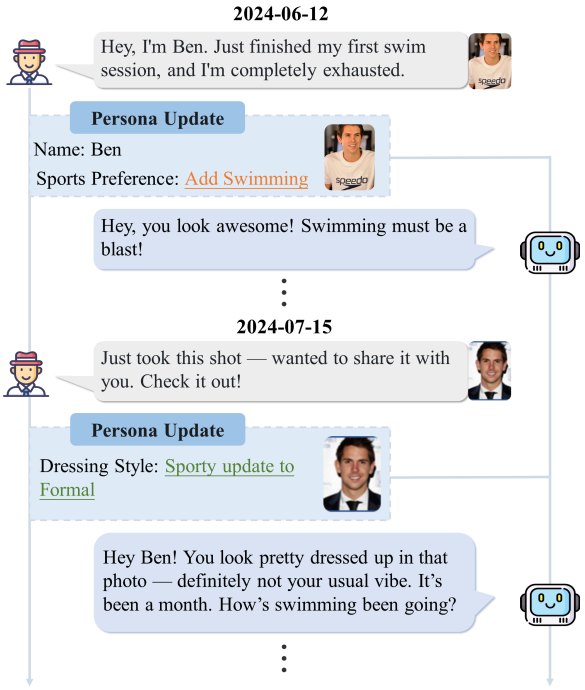



Figure 1: An example from our LongMP-Bench. The user's multimodal persona evolves over long-term interactions while maintaining identity consistency. The dialogue agent is expected to track and update this evolving persona to generate personalized responses.

Recent work has begun to explore multimodal

*Equal contribution.
†Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 23132–23160

personalized dialogue and released corresponding datasets (Maharana et al., 2024; Ahn et al., 2023; Lee et al., 2024). However, existing benchmarks remain limited in their ability to evaluate multimodal persona understanding in realistic settings. First, they often capture only a narrow slice of persona information (e.g., isolated experiences), neglecting preferences and other defining attributes. Second, visual content associated with a single user may lack identity consistency, with images depicting different individuals across turns, undermining coherent persona modeling. Third, personas are typically treated as static, failing to reflect the gradual and evolving nature of user identities in long-term interactions. These limitations collectively hinder systematic evaluation of how well models understand and adapt to evolving multimodal personas.

To address these gaps, we introduce **LongMP-Bench**, a benchmark specifically designed to evaluate multimodal persona understanding in long-term dialogues. LongMP-Bench features multi-faceted user personas that are visually consistent and evolve naturally over extended interactions. We develop a scalable, multi-step data construction pipeline that leverages multimodal large language models and automated tools to generate long-term multimodal conversations, followed by human refinement to ensure coherence and quality. The resulting dataset comprises long dialogues from 150 distinct users interacting with dialogue agents across diverse topics, exhibiting realistic persona development and transitions (Figure 1).

Building on this dataset, we define two complementary evaluation tasks: (i) a question answering (QA) task that probes six key aspects of multimodal persona understanding, and (ii) a response generation task that evaluates models' ability to produce personalized replies grounded in evolving multimodal personas. We conduct extensive experiments benchmarking state-of-the-art long-context multimodal large language models and retrieval-augmented generation (RAG) systems with different memory granularities. Our results reveal substantial performance gaps, particularly in tracking persona evolution and integrating multimodal context over long horizons.

In summary, our contributions are threefold: 1) We propose LongMP-Bench, a benchmark that captures realistic, evolving multimodal personas and provides systematic evaluation tasks for long-term personalized dialogue. 2) We develop a scalable pipeline for constructing visually grounded, dynamic persona dialogues. 3) We provide extensive empirical analysis exposing limitations of current models. We will release all data and code to foster future research on personalized multimodal dialogue systems.

## 2 Dataset Construction

To build LongMP-Bench, we design a multi-stage, scalable data construction pipeline (Figure 2). It begins with generating multimodal personas (comprising both textual and visual information), followed by dynamic topic generation and multi-turn dialogue synthesis. Automated validation and human refinement steps are incorporated throughout to ensure high-quality outputs. Below we summarize each pipeline component; implementation details, prompting strategies, and comparisons with related data generation pipelines (e.g., Stark) are provided in Appendix A.

### 2.1 Multimodal Persona Creation

Each user in LongMP-Bench is assigned a multimodal persona composed of consistent identity images, persona texts, and contextually relevant visual artifacts representing their belongings, preferences, or experiences.

**Persona Images Collection.** We first collect multiple images for each user from four public datasets: Yo'LLaVA (Nguyen et al., 2024), MyVLM (Alaluf et al., 2024), PODS (Sundaram et al., 2025), and CelebA (Liu et al., 2015). All images associated with a single user are drawn exclusively from the same dataset and correspond to the same individual, ensuring consistent appearance, environment, and personal belongings across long-term interactions.

**Persona Texts Generation.** We generate two types of textual personas for each user based on their corresponding images. 1) Demographic Attributes, which include basic characteristics such as age, gender, and occupation. These are generated using LLaMA-4-Scout[1], a model recognized for its effectiveness in personalization tasks. 2) Extended Attributes, which are derived from the PEACOK knowledge graph (Gao et al., 2023). Each relation in PEACOK is mapped to one of five key persona dimensions: *characteristics*, *habits*, *plans*, *experiences*, and *relationships*. To enhance granularity, we further define fine-grained sub-dimensions within each main dimension. For each user, sev-

[1] https://github.com/marketplace/models/azureml-meta/Llama-4-Scout-17B-16E-Instruct

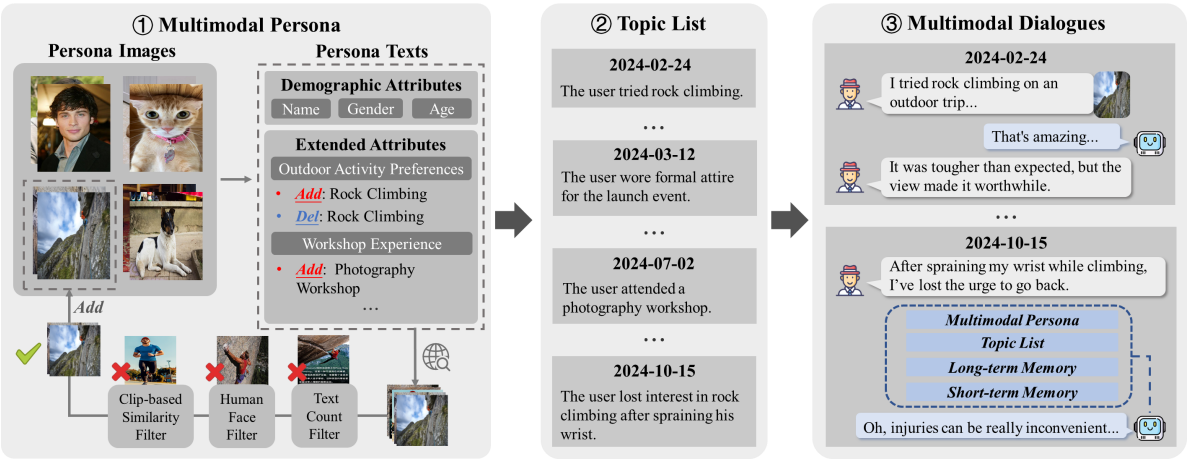



Figure 2: Overview of our data generation pipeline for creating multimodal personas that are visually consistent and dynamically evolving, along with topic lists and multimodal dialogues.

eral dimensions are randomly selected to the focus of the generated textual persona. To support dynamic and evolving user personas, each aspect of Extended Attributes can undergo a sequence of persona shifts: *add*, *update*, or *delete*.

**Persona Images Augmentation.** To visually ground non-identity aspects of personas, we retrieve relevant images based on persona texts that describe attributes beyond user identity. We first use Qwen-Max[2] to extract salient keywords from sampled persona texts, leveraging its strong textual understanding capability. These keywords are then used to perform image retrieval through Bing Search via the icrawler toolkit[3]. The retrieved images undergo a three-stage filtering process: 1) a Haar cascade classifier[4] is applied to remove images containing unintended human faces; 2) EasyOCR[5] is used to filter out text-heavy images; and 3) CLIP ViT-L/14 (Radford et al., 2021) ranks and retains the most semantically relevant images.

## 2.2 Topic List Construction

While multimodal personas define high-level events, we require concrete and expressive topics to guide dialogue synthesis. To this end, we use Qwen-Max to expand each persona change operation into a natural-language topic. For example, an operation like "add Spanish" may be transformed into a topic such as "I recently started learning Spanish." These generated topics are then arranged in chronological order to form a coherent and narratively consistent topic list.

[2] https://qwenlm.github.io/blog/qwen1.5/
[3] https://pypi.org/project/icrawler/
[4] https://github.com/opencv/opencv
[5] https://github.com/JaidedAI/EasyOCR

## 2.3 Multimodal Dialogue Generation

To simulate realistic user-agent interactions that reflect both persona expression and natural topic shifts, we adopt a two-stage dialogue construction process: first generating persona-grounded content, then introducing contextual distractions to preserve persona integrity and reduce hallucinations.

**Persona-Grounded Dialogue Generation.** For each session, we generate multi-turn dialogues grounded in a topic and associated persona via structured prompts including: (1) the session topic and corresponding multimodal persona, (2) the immediate dialogue context, (3) a short-term memory summarizing recent sessions, and (4) a long-term memory storing relevant multimodal persona details from earlier interactions. Generation is conditioned on image captions rather than raw visual inputs to maintain grounding while preventing visual leakage.

**Insertion of Distractions.** To better emulate natural conversations, we introduce contextually related but persona-irrelevant distractions (e.g., casual remarks or anecdotes) inspired by self-disclosure theory (Derlaga and Berg, 1987). Qwen-Max is used to select suitable distractions and determine optimal insertion points, enhancing dialogue realism without contaminating core persona information.

## 2.4 Dataset Refinement and Analysis

To ensure data quality, we first automatically filter redundant, excessively short, or overly long dialogues. Human annotators then review the remaining samples to refine the alignment between textual content and visual elements. The final dataset comprises 150 users, each with an average of 153

| Dataset | Persona Modality | Visually Consistent | Evolving Persona | Avg. sessions per conv. | Avg. turns per conv. |
|---|---|---|---|---|---|
| MMDialog (Feng et al., 2023) | - | ✗ | ✗ | 1 | 4.6 |
| LoCoMo (Maharana et al., 2024) | T | ✗ | ✗ | 19.3 | 304.9 |
| MPChat (Ahn et al., 2023) | T, V | ✗ | ✗ | 1 | 2.8 |
| Stark (Lee et al., 2024) | T, V | ✗ | ✗ | 5.56 | 58.16 |
| LongMP-Bench | T, V | ✓ | ✓ | 14.03 | 153.32 |

Table 1: Comparison of LongMP-Bench with existing multimodal dialogue datasets. 'conv.' and 'Avg.' denote conversation and average, respectively. 'V' and 'T' indicate the presence of visual and textual modalities, respectively. In LongMP-Bench, each conversation comprises multiple sessions, supporting long-term interaction modeling.

dialogue turns and 9 images. Table 1 compares LongMP-Bench with existing multimodal dialogue datasets. Ours is the only dataset that ensures visual consistency, evolving personas, and long-term multimodal interaction, making it a better proxy for real-world scenarios.

To validate data quality, we conduct a human evaluation along four dimensions: persona completeness, internal consistency, transition rationality, and persona-dialogue alignment. Following prior work (Elangovan et al., 2025), we adopt a 1–3 Likert scale for scoring. LongMP-Bench achieves an average rating of 2.73/3.0, confirming its high quality and fidelity. Detailed evaluation procedures are provided in Appendix A.4.

## 3 LongMP-Bench

To systematically evaluate a model's ability to understand and generate persona-consistent responses in long-term multimodal interactions, we construct **LongMP-Bench**, a benchmark built upon the dataset described in Section 2. LongMP-Bench consists of two types of tasks: (1) *Persona Understanding*, which measures how well a model comprehends user personas and their evolution across multimodal dialogues, and (2) *Response Generation*, which assesses the model's ability to produce personalized, context-grounded responses reflecting ongoing persona dynamics.

### 3.1 Persona Understanding Task

The Persona Understanding task is formulated as a question-answering (QA) problem to evaluate multimodal persona comprehension from six perspectives, covering both intra-session and inter-session reasoning. Figure 3 illustrates representative QA types.

**Single-Session Grounded QA.** This setting assesses the model's ability to extract and reason about multimodal persona within a single dialogue session, where relevant clues are interwoven with distractive content. It focuses on three key capabilities: 1) *Persona Detail Extraction*: identifying fine-grained persona attributes from visual and textual cues in the dialogue; 2) *Abstention*: avoiding unsupported inferences when persona information is incomplete or missing; and 3) *Distraction Resistance*: filtering out misleading or irrelevant content to maintain accurate persona understanding.

**Multi-Session Grounded QA.** This setting evaluates the model's ability to track and reason about persona evolution across multiple dialogue sessions. Each QA pair is timestamped and accompanied by historical sessions. It tests three key capabilities: 1) *Persona Temporal Grounding*: identifying the timing and duration of persona states change to maintain temporal coherence; 2) *Time-Aware Persona Tracking*: determining a user's persona state at specific time points or intervals; and 3) *Personalized Concept Recognition*: discerning whether entities (e.g., user, pet, or personal object) referenced across different sessions correspond to the same individual. To increase difficulty, we include hard-negative images that visually resemble but are unrelated to the user's persona, requiring fine-grained visual discrimination.

### 3.2 Response Generation Task

This task evaluates the model's ability to generate responses grounded in both current multimodal input and accumulated persona history. We sample dialogue turns that satisfy two conditions: 1) the user shares an image associated with a personalized concept in the current turn; 2) the image reflects a notable change in persona state relative to previous sessions, which should be captured in the model's response. This setup enables rigorous testing of a model's ability to integrate multimodal cues and track fine-grained persona evolution for personalized, context-aware generation.

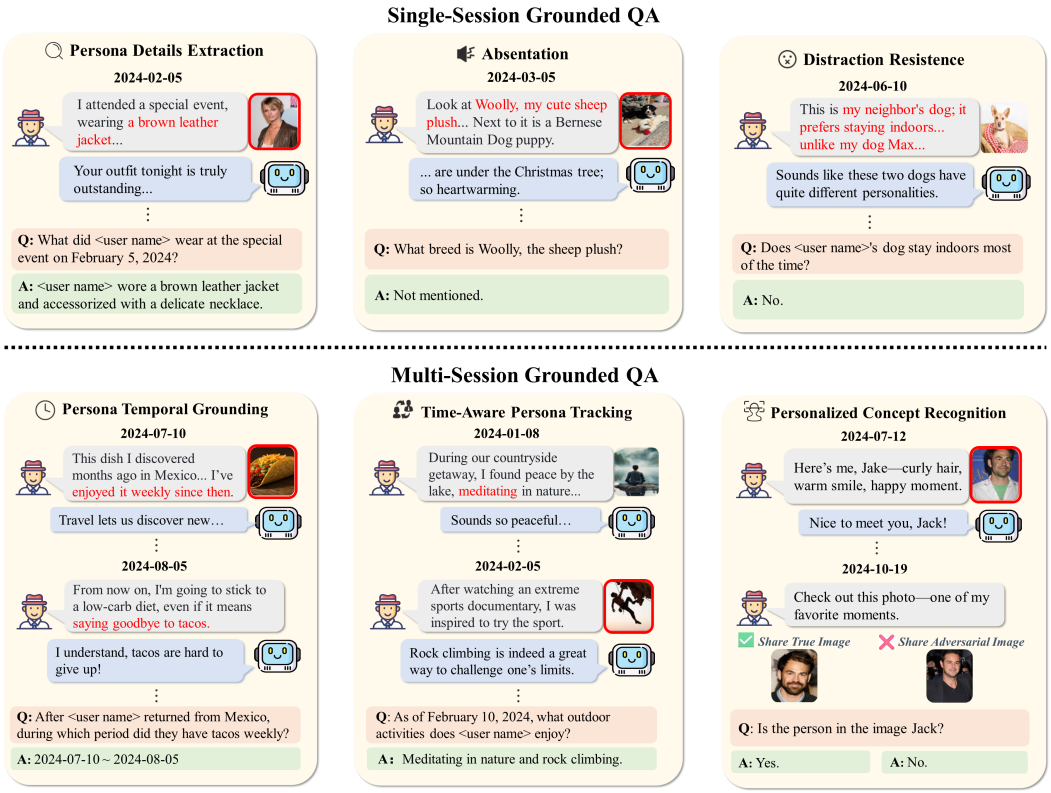



Figure 3: Examples of six diverse question types in LongMP-Bench. For each type, we show the supporting evidence from the conversation history, with key text highlighted in red and key images outlined with red borders. Questions and answers are shown below each example.

## 4 Experimental Setup

### 4.1 Baselines

We evaluate several advanced multimodal large language models (MLLMs): 1) open-source models, including Qwen2-VL (Wang et al., 2024), LLaVA-OneVision (Li et al., 2025), Llama-4-Maverick (Meta, 2025), and Mantis (Jiang et al., 2024); and 2) closed-source models, including GPT-4o[6] and Qwen-VL-Max[7].

Additionally, we examine three retrieval-augmented generation (RAG) strategies using Qwen-VL-Max: *Turn-level RAG* retrieves individual dialogue turns relevant to the current question. *Session-level RAG* expands the retrieval scope to full dialogue sessions for broader context. *Fact-level RAG* leverages Qwen-VL-Max to automatically extract user-specific atomic facts from raw dialogue history, organizing them into structured records indexed by category (e.g., food preferences). Each fact is tagged with an action (e.g., add/delete) and timestamp, forming a timeline of evolving persona states. During retrieval, the model fetches relevant entries and recalls the full update chain within the same category (not just the latest fact) to enhance contextual consistency. Implementation details are provided in Appendix C.1.

### 4.2 Evaluation Metrics

**Question Answering.** For visually grounded QA tasks, traditional lexical or semantic matching metrics often fail to capture nuanced correctness. Therefore, we adopt an LLM-as-a-Judge approach for all QA sub-tasks except for Personalized Concept Recognition (PCR). Specifically, we use Gemini-1.5-Pro (Team et al., 2024) to evaluate the alignment between a model's answer and the reference answer on a 6-point Likert scale (0: completely irrelevant; 5: highly relevant), which we refer to as the *LLM-Score*.

To validate this automatic evaluation, we conduct a human assessment on a sampled subset of model responses. The LLM's judgments exhibit strong alignment with human ratings (Spearman's $\rho = 0.9003$, Kendall's $\tau = 0.8264$), surpassing all other objective metrics. For the PCR task, where the answers are binary (yes/no), we report accuracy directly. Additional evaluation details and meta-

[6]We use the `gpt-4o-2024-11-20` version.

[7]We use the `Qwen-VL-Max-2025-04-08` version.

| Model Type | Model | Method | Single-Sess. Grounded | | | Multi-Sess. Grounded | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PDE | ABS | DR | PTG | TPT | PCR-P | PCR-N |
| Open-source MLLMs | LLaVA-OneVision | Full-Context | 1.356 | 2.102 | 1.858 | 0.032 | 0.465 | 39.36 | 74.50 |
| | Mantis | | 1.671 | 3.553 | 1.830 | 0.524 | 0.776 | **56.88** | 55.36 |
| | Llama-4-Maverick | | **3.357** | 3.225 | 2.978 | **2.353** | **1.219** | 42.33 | 66.38 |
| | Qwen2-VL | | 2.520 | **4.607** | **3.091** | 0.070 | 1.079 | 51.78 | **80.86** |
| Closed-source MLLMs | GPT-4o | Full-Context | 3.722 | 4.124 | 3.587 | **2.823** | **2.050** | 58.86 | 82.46 |
| | Qwen-VL-Max | | **3.885** | **4.185** | **3.848** | 2.255 | 1.833 | **63.26** | **84.94** |
| | Qwen-VL-Max | Turn-level RAG | 3.698 | **4.418** | 3.893 | 2.374 | 1.943 | 65.30 | **93.10** |
| | | Session-level RAG | 3.105 | 4.411 | 3.796 | 0.601 | 1.064 | 64.34 | 92.76 |
| | | Fact-level RAG | **4.070** | 4.211 | **4.088** | **3.284** | **2.280** | **65.82** | 92.86 |

Table 2: **Results of the persona understanding task.** We report recognition accuracy for Personalized Concept Recognition (PCR), while the other subtasks are reported as LLM-Score (0–5, higher is better). Sess.: Session, PDE: Persona Details Extraction, ABS: Abstention, DR: Distraction Resistance, PTG: Persona Temporal Grounding, TPT: Time-Aware Persona Tracking, PCR-P: Personalized Concept Recognition–Positive (user's actual photo); PCR-N: Personalized Concept Recognition–Negative (a visually similar but incorrect photo).

evaluation analyses are provided in Appendix C.2.

**Response Generation.** For response generation, we evaluate the overall semantic similarity between generated and reference responses using BLEU (Papineni et al., 2002) and BERTScore (Zhang* et al., 2020), two widely used metrics in text generation. To further assess personalization, we introduce a *Personalization Score*, which evaluates whether the generated response appropriately reflects the user's persona without contradicting their current persona state. We additionally define an *Alignment Score* to assess the semantic coherence between the generated and reference responses. We use the same LLM-based Likert-scale evaluator as QA, but with distinct task-specific criteria focusing on user-persona consistency and reference-response semantic alignment.

### 4.3 Experimental Details

Non-RAG baselines use the full dialogue history as input. In contrast, RAG-based models operate with a condensed context: the maximum input token budget is set to 3k tokens, typically allowing up to 10 fact-level retrieved items. Retrieval is performed using Qwen's embedding APIs[8]: `multimodal-embedding-v1` for images and `text-embedding-v3` for text.

For all generation methods, we use greedy decoding with a temperature of 0 and top-p set to 1. The maximum output length is capped at 512 tokens. All experiments are conducted on a local machine with 8 NVIDIA A6000 GPUs.

[8]Embeddings obtained via https://help.aliyun.com/zh/model-studio/embedding

## 5 Experimental Results

### 5.1 Persona Understanding Task

Table 2 presents the main results for the persona understanding task. Additional results evaluated using the exact match F1 score are provided in Appendix D.1. Our key findings are as follows: 1) **Closed-source MLLMs generally outperform open-source models in multimodal persona understanding.** Across nearly all question types, closed-source models achieve higher accuracy, particularly on multi-session grounded questions, highlighting their superior capabilities in integrating multimodal information. Interestingly, in the abstention subtask, GPT-4o and Qwen-VL-Max underperform compared to the open-source Qwen2-VL, likely due to stronger models being more prone to overconfidence and hallucinations. 2) **MLLMs show asymmetric performance in personalized concept recognition.** Models consistently achieve high accuracy (>80%) on negative cases (distinguishing different identities) but lower accuracy (<65%) on positive cases (recognizing the same individual), revealing limitations in maintaining consistent visual identity representations over time—a critical component for coherent multimodal persona understanding. 3) **Fact-level RAG enables more effective multimodal persona understanding.** Among RAG strategies, fact-level RAG achieves the best overall performance, especially on multi-session grounded questions. In comparison, turn-level RAG suffers from fragmented memory and often misses semantically important entries, while session-level RAG introduces noise due to coarse granularity and incomplete cover-

| Model Type | Model | Method | BLEU-3 | BLEU-4 | BERTScore | P-Score | A-Score |
|---|---|---|---|---|---|---|---|
| Open-source MLLMs | LLaVA-OneVision | Full-Context | 1.053 | 0.603 | 82.12 | 3.745 | 1.651 |
| | Mantis | | 1.158 | 0.669 | 82.16 | 3.717 | 1.727 |
| | Llama-4-Maverick | | **2.147** | **1.237** | **83.29** | 3.732 | **2.444** |
| | Qwen2-VL | | 2.038 | 1.172 | 83.13 | **3.832** | 2.314 |
| Closed-source MLLMs | GPT-4o | Full-Context | **2.263** | 1.217 | 83.77 | **4.036** | 2.669 |
| | Qwen-VL-Max | | 2.253 | **1.255** | **83.85** | 4.008 | **2.686** |
| | Qwen-VL-Max | Turn-level RAG | 2.195 | 1.104 | 83.67 | 3.891 | 2.468 |
| | | Session-level RAG | 2.756 | 1.414 | 83.59 | 3.933 | 2.715 |
| | | Fact-level RAG | **3.808** | **2.264** | **84.18** | **4.159** | **2.876** |

Table 3: **Results of the response generation task.** P-Score: Personalization Score, A-Score: Alignment Score.

age of long sessions under fixed token limits. Notably, fact-level RAG also surpasses full-context baselines, indicating that even the most advanced MLLMs struggle to directly process long-term multimodal context effectively.

### 5.2 Response Generation Task

As shown in Table 3, closed-source models again outperform open-source models, and RAG-based methods consistently surpass full-context baselines in response generation. These results mirror the trends observed in the persona understanding task, underscoring that accurate multimodal persona understanding is crucial for generating personalized and coherent responses.

## 6 Analysis

To better understand the capabilities and limitations of MLLMs in multimodal persona understanding over long-term dialogue, we conduct a systematic analysis guided by two research questions.

**Q1: To what extent can MLLMs effectively integrate vision and language for multimodal persona understanding?** We first perform a modality ablation study on our QA tasks. Detailed results are provided in Appendix D.2 and Table 10. The findings confirm that textual and visual inputs are complementary within our benchmark. We further compare performance when feeding raw images versus their textual captions (Table 4). Our key observations are as follows:

**(i) Current models struggle to integrate visual and textual information effectively.** Replacing images with captions consistently improves performance across most subtasks. This suggests that MLLMs often rely on unimodal textual reasoning rather than performing true cross-modal integration. When high-quality text is available, models tend to deprioritize visual cues.

**(ii) Visual inputs better preserve identity consistency in visual contexts.** In the personalized concept recognition task, using captions instead of raw images results in an average 7.99% gain in accuracy on positive cases but a 19.31% drop on negative cases. This indicates that captions can capture general concepts but often fail to encode fine-grained visual identity cues (e.g., facial features, clothing details) necessary for accurate differentiation. Consequently, substituting images with captions undermines the model's ability to maintain consistent identity representations over time.

**Q2: To what extent can MLLMs effectively track evolving personas in long-term dialogues?** To assess how well MLLMs track evolving persona traits, we analyze model performance across conversations that vary in the number and types of persona changes. We find:

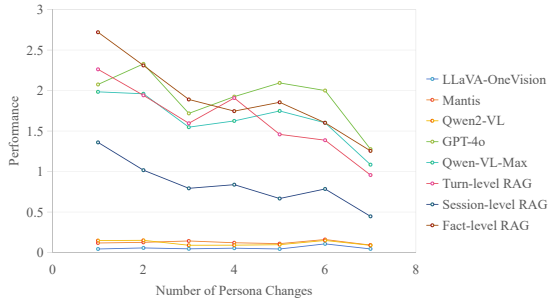
**(i) Tracking persona evolution across multiple changes remains a key challenge.** As shown in Figure 4a, model performance degrades significantly as the number of persona changes increases. Open-source models, in particular, struggle across all settings. We observe that models often conflate the current persona state with information from neighboring temporal contexts, suggesting limitations in modeling temporal continuity and change.

**(ii) MLLMs are particularly insensitive to the termination of persona states.** This issue is especially pronounced in full-context settings, as illustrated in Figure 4b. Many models fail to recognize when a persona state has ended, frequently retaining outdated information and erroneously assuming that prior states persist.

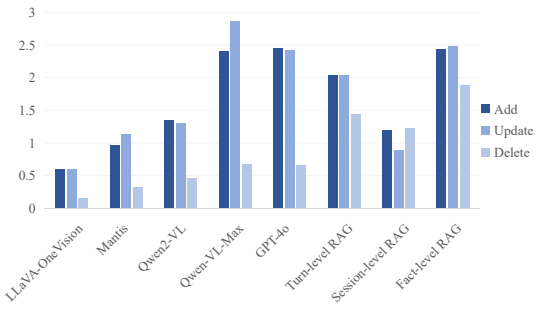
**(iii) Retrieval-Augmented Generation can mitigate this issue.** Fact-level RAG helps reduce the persistence of obsolete states, likely because it filters out irrelevant historical information and focuses attention on temporally appropriate facts.

| Model | Method | Input Type | Single-Sess. Grounded QA | | | Multi-Sess. Grounded QA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PDE | ABS | DR | PTG | TPT | PCR-P | PCR-N |
| Qwen2-VL | Full-Context | Image | 2.520 | 4.607 | 3.091 | 0.070 | 1.079 | 51.78 | **80.86** |
| | | Image Caption | **2.697** | **4.815** | **3.275** | **0.165** | **1.151** | **58.92** | 64.80 |
| Qwen-VL-Max | Full-Context | Image | 3.885 | 4.185 | 3.848 | 2.255 | 1.833 | 63.26 | **84.94** |
| | | Image Caption | **3.923** | **4.389** | **3.922** | **2.455** | **1.881** | **76.00** | 50.74 |
| Qwen-VL-Max | Fact-level RAG | Image | 4.070 | 4.211 | 4.088 | 3.284 | 2.280 | 65.82 | **92.86** |
| | | Image Caption | **4.368** | **4.255** | **4.276** | **3.293** | **2.283** | **69.90** | 85.20 |

Table 4: **Results of the persona understanding task with image vs. image caption inputs.** "Image" denotes the use of original images from the dialogue history, while "Image Caption" uses corresponding textual descriptions. Higher scores indicate better performance. Bold values indicate the best result per column.




(a) Performance across different numbers of persona changes.




(b) Performance across different persona change types.

Figure 4: Model performance under varying numbers and types of persona changes.

A more detailed and deeper analysis of model limitations, distilled into four main error patterns, is provided in Appendix E.

## 7 Related Works

**Long-Term Dialogue.** Among recent advances in long-term dialogue modeling, several studies have aimed to construct datasets that preserve persona information across conversation sessions (Xu et al., 2022a; Wu et al., 2025; Zhang et al., 2023; Jang et al., 2023). However, most approaches fail to adequately account for evolving user characteristics such as shifting interests. They also typically define persona in purely textual terms, neglecting the value of visual information. These limitations hinder the development of dialogue agents capable of understanding real-world evolving multimodal personas in long-term dialogue, which our work addresses by modeling persona evolution over time.

**Multimodal Personalized Dialogue.** Recent work integrates visual and textual inputs to model user personas in multimodal dialogue. MPCHAT (Ahn et al., 2023) is a dataset that captures episodic memories through aligned text and image modalities. STARK (Lee et al., 2024) features dialogue data with simulated long-term image-sharing behaviors. LoCoMo (Maharana et al., 2024) extends modeling to open-domain dialogues with life events. However two key limitations remain (1) personas are often narrowly defined either as episodic memories (Ahn et al., 2023) or generalized information (Maharana et al., 2024) lacking a holistic persona definition and (2) visual consistency is limited due to reliance on web-sourced or synthetic images. These issues constrain the realism of multimodal personas. We address these gaps through a novel data generation pipeline.

## 8 Conclusion

In this work, we introduce LongMP-Bench, the first benchmark designed to rigorously evaluate multimodal persona understanding in long-term dialogues. We develop a scalable data synthesis pipeline to construct persona-grounded multimodal dialogues and define multiple targeted evaluation tasks based on this dataset. Through extensive experiments, we uncover that current MLLMs struggle both to fuse visual and textual cues and to track evolving personas over time, revealing key challenges for building truly personalized and human-like dialogue agents.

## Limitations

While our benchmark covers a diverse range of persona dimensions, certain aspects—such as the Big Five personality traits and evolving emotional states—remain underrepresented. In addition, single-reference evaluation metrics have inherent limitations for open-ended dialogue, as multiple reasonable responses typically exist in such conversational settings. Future work may extend the scope of persona attributes and explore more reliable evaluation metrics to enable more comprehensive modeling of human-like characteristics in dialogue agents.

## Ethical Statement

In this paper, we utilize MLLMs to generate user personas and dialogues. We have taken careful measures to ensure that all generated content is free from harmful, biased, or offensive material. Furthermore, all personas presented in this work are entirely fictional, and there is no disclosure of real users' identities, addresses, or contact information.

Additionally, we acknowledge that the source datasets employed (e.g., Yo'LLaVA (Nguyen et al., 2024), CelebA (Liu et al., 2015)) may introduce biases related to age, gender, or cultural representation. We also note limitations in demographic coverage, such as the under-representation of certain regions or age groups, and recognize their potential impact on the diversity of synthetic personas. We frame this as an important direction for future work, including the integration of more diverse source data to mitigate such biases.

## Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62576347)

## References

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.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. 2024. Myvlm: Personalizing vlms for user-specific queries. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XIII*, page 73–91, Berlin, Heidelberg. Springer-Verlag.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665, Torino, Italia. ELRA and ICCL.

Valerian J Derlaga and John H Berg. 1987. *Self-disclosure: Theory, research, and therapy*. Springer Science & Business Media.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

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.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6569–6591, Toronto, Canada. Association for Computational Linguistics.

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.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeong-Jin Oh, Byungsoo Ko, Jonghwan Hyeon, and Ho-Jin Choi. 2024. 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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed: 2025-06-27.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. 2024. Yo'llava: Your personalized language and vision assistant. In *Advances in Neural Information Processing Systems*, volume 37, pages 40913–40951. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shobhita Sundaram, Julia Chae, Yonglong Tian, Sara Beery, and Phillip Isola. 2025. Personalized representation from personalized generation. In *The Thirteenth International Conference on Learning Representations*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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. *arXiv preprint arXiv:2409.12191*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

# A Data Generation Pipeline

In this section, we provide a detailed description of the data generation pipeline.

## A.1 Multimodal Persona

### A.1.1 Source Image Datasets

To maintain long-term visual consistency in appearance and context, we gather multiple image instances for each user. The resulting collections, are compiled from four publicly available datasets with consistent user identities: Yo'LLaVA (Nguyen et al., 2024), MyVLM (Alaluf et al., 2024), PODS (Sundaram et al., 2025), and CelebA (Liu

| |
|---|
| **Demographic Attributes** |
| Name, Gender, Age, Nationality, Language, Marital Status, Birthplace, Residence |
| **Characteristic** |
| Food Preferences, Drink Preferences, Shopping Preferences, Movie Preferences, Music Preferences, Reading Preferences, Travel Preferences, Work Environment Preferences, Technology Product Preferences, Outdoor Activity Preferences, Time Management Style, Communication Style, Problem-Solving Style, Fashion Style, Personal Values |
| **Habit** |
| Leisure Activities, Commute Options, Work Routine, Health & Fitness Habits, Financial Habits |
| **Plan** |
| Career Advancement Goals, Financial Goals, Hobby Development Goals, Family Life Goals, Community Involvement Goals, Environmental Sustainability Goals, Travel Plans, Health & Fitness Plans, Learning Plans |
| **Experience** |
| Social Media Experience, Offline Interaction Experience, Travel Experience, Academic Experience, Work Experience, Internship Experience, Volunteer Experience, Conference/Workshop Experience, Competition Experience, Achievement Experience |
| **Relationship** |
| Family Members, Friends, Colleagues, Romantic Partners, Neighbors, Professional Network, Mentors, Fans or Followers, Pets, Personal Belongings |

Table 5: Fine-grained persona sub-dimension.

et al., 2015). Below, we provide an overview of each dataset and its role in our data pipeline.

**Yo'LLaVA** This dataset is designed to facilitate the personalization of Large Multimodal Models (LMMs), enabling them to recognize and reason about specific visual subjects in conversational settings. It was constructed by collecting personalized images from volunteers, each of whom provided 4–5 images of a specific subject (e.g., a pet, a toy, or a family member), along with textual descriptions highlighting unique visual attributes (e.g., "My dog has a yellow collar and a scar above the left eye"). These descriptions were used to generate a set of personalized questions and answers that test the model's ability to identify and describe the subject accurately. Each subject is associated with a special token (e.g., <sks>) and a set of latent tokens (e.g., <token1><token2>...<tokenk>), which are used to embed the subject into the model. The dataset supports visual recognition and conversational reasoning tasks, such as "Is <sks> in this photo?" and "What should I get <sks> for their birthday?".

**MyVLM** This dataset facilitates the personalization of vision-language models (VLMs), enabling them to recognize and describe user-specific visual concepts such as personal objects or individuals (e.g., a user's pet or themselves). It was constructed by collecting a diverse set of personalized image-caption pairs from volunteers, where each subject is represented by 3–5 images and a corresponding caption that incorporates the subject in a contextualized manner (e.g., "I am wearing a red jacket at the beach" or "My dog is playing in the garden"). These captions are designed not only to describe the visual content but also to reflect personal experiences and relationships. Each subject is associated with a unique identifier, which can be used to train models to recognize and refer to the subject consistently across different images and tasks. The dataset supports both personalized image captioning and visual question answering, allowing models to answer questions such as "What is <user> doing in this photo?" or "Is <sks> in this image?".

**PODS** This dataset evaluates personalized representation learning under distribution shifts, focusing on the fine-grained recognition of user-specific objects. It was explicitly constructed to assess how well models can learn from limited real examples (as few as three) and generalize to unseen instances of the same object. The dataset contains 100 distinct personal objects, each represented by a small set of real images. The dataset supports a variety of downstream tasks, including classification, retrieval, detection, and segmentation, making it a systematic benchmark for evaluating personalized vision systems.

**CelebA** The CelebA dataset is a large-scale facial attribute dataset containing over 200,000 celebrity images, each annotated with 40 binary facial attributes, such as "smiling", "bearded", "arched eyebrows", and "receding hairline". The images are collected from real-world scenarios and exhibit variations in pose, expression, lighting, and occlusion. It has become a widely used dataset for tasks such as facial attribute prediction, face recognition, and representation learning.

### A.1.2 Implementation Details

We begin by prompting LLaMA-4-Scout-17B-16E-Instruct, conditioned on the consistent image clusters, to generate demographic attributes using the

prompt illustrated in Figure 5. To ensure a multi-faceted and diverse representation of personas, we define a fine-grained set of persona sub-dimension, which are presented in detail in Table 5. Using the generated demographic attributes together with the consistent image clusters, we then sample specific entities from predefined sub-dimension to construct a structured persona profile. For the acquisition of complementary persona images via text-to-image retrieval, we employ Qwen-Max to convert the extended attributes into keyword-based queries. The corresponding prompts used in this process are provided in Figure 6.

### A.2 Topic List

To construct concrete topics for interaction simulation and model the temporal evolution of the persona, we employ Qwen-Max to expand each type of change — addition, update, and deletion — into fully developed topics, using the prompting strategy illustrated in Figure 7. All update and delete operations must reference previously introduced persona traits to maintain logical coherence. Timestamps are assigned to persona updates using an LLM that ensures realistic temporal spacing based on the type of event. Each topic is associated with detailed event descriptions, ensuring a coherent and realistic progression of the persona over time.

### A.3 Multimodal Dialogue

We use a personalized dialogue agent to generate conversations, with each conversation consisting of multiple sessions. Dialogue is generated sequentially, with each session conditioned on a structured prompt. This prompt includes the topic and multimodal persona information to be expressed in the current session, as well as the immediate dialogue context. In addition, it incorporates short-term memory, which summarizes the previous session to provide context for the current interaction, and long-term memory, which contains relevant multimodal persona details retrieved from the previous sessions. The short-term memory helps ensure continuity in the dialogue, while the long-term memory maintains consistency in persona details across sessions. The prompt given to the LLM for generating short-term memory is depicted in Figure 8, while the prompt for generating long-term memory is shown in Figure 9. In order to ensure that the text does not excessively disclose image information during the conversation, We generate a caption based on the image and use this caption to facilitate dialogue generation, following the prompt shown in Figure 10. The prompts used for conversation generation are shown in two separate figures: the user's prompt is depicted in Figure 11, and the dialogue agent's prompt is shown in Figure 12. When inserting distraction utterances into an original dialogue, the optimal insertion point is determined by identifying natural discourse boundaries—such as topic shifts or expressions of curiosity—where a distraction can be introduced with minimal disruption to conversational flow. This strategy ensures that the inserted distractions mimic realistic conversational perturbations while preserving the structural coherence of the dialogue. The complete prompt template used for generating and inserting such distractions is provided in Figure 17.

### A.4 Dataset Refinement

To ensure high-quality and coherent multimodal personas and dialogues, we implement a two-stage dataset refinement process, consisting of rule-based filtering and human-based filtering.

#### A.4.1 Rule-based Filtering Implementation

The following rule-based filters are applied during preprocessing to ensure dialogue quality:

1. Remove dialogues where the same speaker utters more than one consecutive turn.
2. Remove dialogues with fewer than 5 or more than 25 turns.
3. Remove utterances that are too short (< 2 words) or overly long (> 150 words).

#### A.4.2 Human Filtering Implementation

As part of the dataset refinement process, we conducted a human filtering stage to enhance the quality and consistency of the multimodal personas and their associated dialogues. This stage primarily involved two types of interventions:

1. **Ensuring Visual Consistency Across Persona Images**

   We identified cases in which the synthesized personas contained inconsistent images—for instance, variations in age, gender, or appearance across images attributed to the same individual. To maintain long-term visual coherence, annotators either removed the inconsistent images or replaced them with more suitable alternatives that aligned with the overall visual identity of the persona.

2. **Improving Naturalness of Image-Related Dialogues**

   We also refined the dialogues that referenced visual content to enhance conversational fluency and contextual appropriateness. In particular, annotators edited utterances that appeared unnatural, overly artificial, or misaligned with the visual content being described. This step ensured that the multimodal interactions remained coherent and grounded in the actual image information, thereby improving the overall realism and quality of the dialogues.

This human filtering process was conducted iteratively, with annotators receiving detailed guidelines and illustrative examples to ensure consistency in decision-making.

#### A.4.3 Human Evaluation of Data Quality

To further assess the quality of the synthesized multimodal personas and dialogues, we conducted a human evaluation focusing on four key aspects: **multimodal persona multifacetedness**, **consistency**, **transition reasonableness**, and **dialogue alignment**. We randomly sampled 60 persona-dialogue pairs from the dataset and recruited three annotators with backgrounds in computer science. Prior to annotation, the annotators were provided with detailed instructions explaining the evaluation criteria, as well as the characteristics of the personas and dialogues.

- **Multimodal Persona Multifacetedness**: Whether the persona covers multiple aspects of personal information, including appearance, personality, interests, and life experiences.
- **Multimodal Persona Consistency**: Whether the information across modalities (e.g., text and image) and textual components is internally coherent and mutually supportive.
- **Transition Reasonableness**: Whether personal changes are logically and contextually supported.
- **Dialogue Alignment**: Whether the dialogue content aligns with the persona's described attributes, behavior, and linguistic style.

A total of 3 participants completed the task. The participants were primarily based in China, with ages ranging from 18 to 24 years old, and a balanced gender distribution. All participants held a bachelor's degree and demonstrated fluent English proficiency. Each participant was compensated with USD $5 upon completion of the 40-minute task, corresponding to an hourly rate of USD $7.50. Given that participants were primarily based in China, this compensation aligns with fair pay standards in the region.

Each annotation task included a persona description — comprising both image and textual information, potentially involving multiple self-disclosure transitions — and a dialogue between the persona and a chatbot. Before starting the task, participants were required to read and agree to an informed consent form. The form explained the purpose of the study, how the data would be used, and assured participants of anonymity and voluntary participation. Users who did not consent could not proceed with the task.

Annotators were asked to rate each sample on a 3-point scale (1 = poor, 2 = moderate, 3 = high), with detailed rating guidelines provided for each dimension. All participants received detailed instructions before starting the evaluation task. The instructions included:

- The purpose of the task: to evaluate the quality of dialogue responses generated by AI models.
- An estimate of the task duration: approximately 40 minutes.
- Information on data usage: all responses will be used solely for research purposes and anonymized.
- A statement on voluntary participation: participants could withdraw at any time without penalty.

The average score across all dimensions was 2.73 out of 3.0, indicating that the proposed data synthesis pipeline effectively generates high-quality and reliable multimodal personas and dialogues. These results highlight the dataset's reliability and suitability for modeling complex, dynamic persona-grounded conversations in a multimodal setting.

### A.5 Comparison with Stark's Data Generation Pipeline

While our work draws inspiration from synthetic dialogue generation frameworks such as Stark (Lee et al., 2024), the LongMP-Bench pipeline differs fundamentally in both design goals and technical implementation. We highlight four key distinctions below.

First, **support for evolving multimodal personas**. Stark assumes static user profiles with fixed attributes. In contrast, LongMP-Bench explicitly models dynamic persona evolution across multiple dimensions—including appearance, habits, demographic characteristics, social relationships, and future aspirations—with updates jointly reflected in both textual descriptions and visual representations.

Second, **visual identity consistency**. LongMP-Bench employs an identity-stable image pool for each persona, ensuring consistent facial features and biometric attributes across sessions. In Stark, images are generated or sampled independently per turn, which can result in visual inconsistencies for the same user identity (e.g., variations in facial structure or gender presentation).

Third, **topic-guided dialogue generation**. Rather than anchoring interactions to discrete life events, LongMP-Bench structures persona updates around topical dimensions (e.g., "career", "health", "travel"), enabling more diverse and realistic modeling of both transient experiences and enduring personal traits.

Fourth, **dialogue length and temporal scope**. Dialogues in LongMP-Bench are significantly longer—averaging three times the number of turns compared to Stark—thereby simulating extended, multi-session interactions. This design necessitates models to maintain coherence over both short-term context and long-term memory, a requirement not imposed by Stark's single-session setting.

In summary, LongMP-Bench is not a minor adaptation of existing pipelines but a purpose-built framework for studying long-term, multimodal persona evolution—addressing limitations of prior benchmarks in scope, realism, and temporal depth.

## B Dataset

### B.1 Dataset Statistics

The full dataset statistics are summarized in Table 6. Notably, the dialogue response generation task and the personalized concept recognition task are based on the same set of dialogue turns, resulting in an equal number of instances.

| Metric | Value |
|---|---|
| **Conversation Statistics** | |
| Total # Conversations | 150 |
| Total # Sessions | 2105 |
| Total # Turns | 22998 |
| Total # Images | 1366 |
| Avg. # Images per Conversation | 9.11 |
| Avg. # Sessions per Conversation | 14.03 |
| Avg. # Turns per Session | 10.93 |
| **QA Task Statistics** | |
| # Questions. Persona Details Extraction | 1472 |
| # Questions. Abstraction | 1375 |
| # Questions. Distraction Resistance | 1832 |
| # Questions. Persona Temporal Grounding | 2214 |
| # Questions. Time-Aware Persona Tracking | 1681 |
| # Questions. Personalized Concept Recognition | 404 |
| **Response Generation Task** | |
| # Turns for Response Generation | 404 |

Table 6: Dataset and benchmark statistics.

### B.2 Dataset License

Our dataset, including multimodal personas, topics, and multimodal dialogues, will be released under the CC BY-NC 4.0 license[9] and is intended for research purposes only.

## C Experimental Setup

### C.1 Fact-level Retrieval-Augmented Generation (RAG)

We implement a fact-level RAG framework to support dynamic and contextually grounded interactions with the multimodal persona. This framework enables the system to retrieve fine-grained, semantically relevant user information during conversation and incorporate it into the response generation process.

- **Storage Granularity:** Each memory entry is represented as a structured unit containing the following fields: a timestamp, an action (e.g., add/delete), a category (e.g., leisure activities), content, a set of associated persona facts, and optionally, an image reference. These entries are generated per session, with new facts being summarized in the context of previously stored information—referred to as *Associative*

[9] https://creativecommons.org/licenses/by-nc/4.0/

*Memory*—to ensure long-term coherence in the persona representation.

- **Memory Organization:** The memory is structured around the multimodal persona and categorized into semantic types (e.g., food preferences). Within each category, entries are stored in chronological order. Basic demographic attributes is re-summarized at the beginning of each session based on newly added or removed facts, while other sub-dimension retain their historical entries to preserve contextual continuity.

- **Retrieval Mechanism:** Given a user query, the system performs dual-path retrieval—one for textual content and one for image content—based on the semantic meaning of the query. Retrieved results are then merged and further enriched by recalling related facts within the same semantic category, optionally constrained by a time window.

  For example, when a query relates to "reading books", the system retrieves the most recent update (e.g., deletion of the activity) and also recalls prior entries in the same category (e.g., initial adoption of reading). This ensures a comprehensive understanding of the user's evolving preferences over time.

This fact-level RAG design enables the system to generate responses that are not only grounded in the user's multimodal persona but also temporally and contextually coherent. An example of the memory structure is illustrated in Figure 13.

### C.2 Evaluation Metrics

To ensure a reliable and consistent evaluation of the QA task, we designed a carefully crafted prompting strategy to employ Gemini-1.5-Pro as an automated scoring model. The full evaluation prompt is detailed in Figure 14. This prompt explicitly instructs the model to assess answer quality across multiple dimensions, such as factual accuracy, relevance, and coherence.

To validate its alignment with human judgment, we randomly sampled 180 questions along with their predicted answers and corresponding ground-truth answers. We then computed the correlation between various automatic evaluation metrics and human ratings, as shown in Table 7.

Our analysis reveals that Gemini-1.5-Pro demonstrates the highest correlation with human scores, outperforming conventional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). This result highlights its effectiveness as an evaluation model for the QA task, particularly in capturing nuanced aspects of answer quality that align with human perception.

| Metric | Spearman $\rho$ | Kendall $\tau$ |
|---|---|---|
| LLM-Score | **0.9003** | **0.8264** |
| BLEU-1 | 0.5177 | 0.4005 |
| BLEU-2 | 0.6148 | 0.4749 |
| BLEU-3 | 0.4452 | 0.3421 |
| BLEU-4 | 0.3647 | 0.2847 |
| ROUGE-1 | 0.6140 | 0.4792 |
| ROUGE-2 | 0.6403 | 0.5244 |
| ROUGE-L | 0.5871 | 0.4568 |
| METEOR | 0.6094 | 0.4654 |
| BERTScore-P | 0.3672 | 0.3080 |
| BERTScore-R | 0.4072 | 0.3206 |
| BERTScore-F1 | 0.4225 | 0.3468 |
| Vocab-Precision | 0.4749 | 0.3605 |
| Vocab-Recall | 0.5323 | 0.4189 |
| Vocab-F1 | 0.6088 | 0.4802 |

Table 7: Comparison of different metrics using Spearman's rank correlation coefficient ($\rho$) and Kendall's tau ($\tau$).

For the response generation task, we employ widely used automatic evaluation metrics, including BLEU and BERTScore (Zhang* et al., 2020), using their default implementations in standard Python libraries. These metrics provide a general indication of lexical overlap and semantic similarity, respectively. While these metrics capture lexical overlap and semantic similarity, they are insufficient for evaluating user-centric qualities. To address this, we leverage LLM-as-a-Judge, which offers fine-grained, scalable, and cost-effective assessments that align closely with human judgments. Specifically, we introduce two complementary evaluation metrics: Personalization Score, which assesses whether the generated response appropriately reflects the user's persona without contradicting their current persona state, and Alignment Score, which measures the consistency between the generated response and the reference response. Both scores are obtained via the same LLM-as-a-Judge protocol and 6-point Likert scale described in the QA evaluation. To further validate this approach, we conducted a small-scale human study (n=50, covering 33% of the test set), showing strong correlation with LLM-based scores (Spearman's $\rho > 0.83$), thereby confirming their reliability as proxies. For completeness, we present evaluation prompts in Figure 15 and Figure 16.

| Model Type | Model | Method | Single-Sess. Grounded | | | Multi-Sess. Grounded | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PDE | ABS | DR | PTG | TPT | PCR-P | PCR-N |
| Open-source MLLMs | LLaVA-OneVision | Full-Context | 0.175 | 0.421 | 0.279 | 0.011 | 0.048 | 39.36 | 74.50 |
| | Mantis | | 0.235 | 0.712 | 0.296 | **0.060** | 0.122 | **56.88** | 55.36 |
| | Qwen2-VL | | **0.369** | **0.922** | **0.499** | 0.023 | **0.142** | 51.78 | **80.86** |
| Closed-source MLLMs | GPT-4o | Full-Context | 0.604 | 0.825 | 0.611 | **0.627** | **0.378** | 58.86 | 82.46 |
| | Qwen-VL-Max | | **0.632** | **0.838** | **0.662** | 0.517 | 0.312 | **63.26** | **84.94** |
| | Qwen-VL-Max | Turn-level RAG | 0.515 | **0.884** | 0.588 | 0.564 | 0.327 | 65.30 | **93.10** |
| | | Session-level RAG | 0.437 | 0.868 | 0.597 | 0.265 | 0.183 | 64.34 | 92.76 |
| | | Fact-level RAG | **0.635** | 0.842 | **0.681** | **0.752** | **0.380** | **65.82** | 92.86 |

Table 8: **Results of the persona understanding task.** Bold values highlight the best performance in each column. Higher values indicate better performance, with recognition rates for personalized concept recognition and remaining QA scores evaluated based on F1 scores for exact match.

| Model | Method | Input Type | Single-Sess. Grounded QA | | | Multi-Sess. Grounded QA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | PDE | ABS | DR | PTG | TPT | PCR-P | PCR-N |
| LLaVA-OneVision | Full-Context | Image | 1.356 | **2.102** | 1.858 | 0.032 | 0.465 | 39.36 | **74.50** |
| | | Image Caption | **2.465** | 1.745 | **3.030** | **0.140** | **0.852** | **41.32** | 70.92 |
| Mantis | Full-Context | Image | 1.671 | 3.553 | 1.830 | 0.524 | **0.776** | 56.88 | **55.36** |
| | | Image Caption | **2.011** | **4.436** | **2.270** | **0.587** | 0.724 | **68.62** | 38.02 |
| Qwen2-VL | Full-Context | Image | 2.520 | 4.607 | 3.091 | 0.070 | 1.079 | 51.78 | **80.86** |
| | | Image Caption | **2.697** | **4.815** | **3.275** | **0.165** | **1.151** | **58.92** | 64.80 |
| GPT-4o | Full-Context | Image | 3.722 | **4.124** | 3.587 | 2.823 | 2.050 | 58.86 | **82.46** |
| | | Image Caption | **3.912** | 4.058 | **3.703** | **2.879** | **2.056** | **62.62** | 67.08 |
| Qwen-VL-Max | Full-Context | Image | 3.885 | 4.185 | 3.848 | 2.255 | 1.833 | 63.26 | **84.94** |
| | | Image Caption | **3.923** | **4.389** | **3.922** | **2.455** | **1.881** | **76.00** | 50.74 |
| Qwen-VL-Max | Turn-Leve RAG | Image | 3.698 | 4.418 | 3.893 | 2.374 | 1.943 | 65.30 | **93.10** |
| | | Image Caption | **3.841** | **4.425** | **3.960** | **2.416** | **1.995** | **68.36** | 84.70 |
| | Session-Level RAG | Image | 3.105 | **4.411** | 3.796 | 0.601 | 1.064 | 64.34 | **92.76** |
| | | Image Caption | **3.778** | 4.375 | **4.106** | **1.102** | **1.271** | **68.10** | 82.40 |
| | Fact-level RAG | Image | 4.070 | 4.211 | 4.088 | 3.284 | 2.280 | 65.82 | **92.86** |
| | | Image Caption | **4.368** | **4.255** | **4.276** | **3.293** | **2.283** | **69.90** | 85.20 |

Table 9: **Results of the persona understanding task with image and image caption inputs.** "Image" refers to using the original image in the conversation history, while "Image Caption" refers to using the caption of the image instead. Higher scores indicate better performance. Bold values highlight the best performance in each column.

# D Results

## D.1 Supplementary Evaluation of Multimodal Understanding

While semantic similarity metrics (e.g., Gemini-1.5-Pro scoring) provide a more nuanced evaluation of visually grounded responses, we also report exact match F1 scores for completeness. Table 8 presents the exact match F1 results across different question types. These results complement the semantic evaluation reported in the main text, offering a more systematic view of model performance across both literal and interpretive QA tasks, and further substantiating the findings presented in Section 5.1.

We report the full evaluation results comparing the use of original images versus image captions as input. Table 9 summarizes the performance across multiple dimensions and is consistent with the findings reported in Section 6.

## D.2 Modality Contributions to Multimodal Persona Understanding

To understand the relative importance of different modalities in multimodal persona understanding within long-term dialogues, we conduct ablation studies by removing either textual or visual inputs. The results are summarized in Table 10. Our findings are as follows:

**Textual information plays a central role in multimodal persona understanding.** Removing textual input leads to a substantial performance drop across all subtasks, while removing visual

| Model | Input Type | Single-Sess. Grounded QA | | | Multi-Sess. Grounded QA | | | |
|---|---|---|---|---|---|---|---|---|
| | | PDE | ABS | DR | PTG | TPT | PCR-P | PCR-N |
| Qwen-VL-Max | Full-Context | **3.885** | 4.185 | **3.848** | **2.255** | **1.833** | 63.26 | 84.94 |
| | w/o Image | 3.170 | 4.324 | 3.227 | 1.678 | 1.474 | **66.34** | 33.60 |
| | w/o Text | 1.266 | **4.727** | 2.373 | 1.425 | 0.997 | 12.60 | **91.20** |

Table 10: **Ablation results on persona understanding tasks with different input modalities for Qwen-VL-Max.** "w/o Image" and "w/o Text" denote the removal of image or text inputs from the conversation history. Bold values highlight the best performance in each column.

input results in only a minor decline. This suggests that textual content contains the primary cues for modeling personas.

**Text and vision are complementary modalities in our benchmark.** The best performance is consistently achieved when both modalities are used together, indicating that the benchmark effectively captures a model's ability to integrate and reason over multimodal persona information.

## E Deeper Diagnostic Analysis of Model Failure Patterns

To provide deeper insights into model behavior, we conduct a structured diagnostic study examining the systematic failure modes of current MLLMs in tracking evolving personas over time.

Specifically, we identify and elaborate on four key limitations:

- **Temporal misalignment**: Models conflate user states across different time points.
- **Insensitivity to state transition order**: They fail to reconstruct the correct sequence of attribute changes.
- **Failure to interpret vague temporal references**: They struggle with expressions like "recently" or "after that phase".
- **Inaccurate boundary detection for state duration**: They misidentify when a persona trait begins or ends.

These patterns suggest that MLLMs often treat dialogue history as a flat memory, retrieving facts without reasoning about when they changed. As a result, they struggle with timing, ordering, and duration of personal traits—revealing weakness in temporal reasoning. This diagnostic analysis moves beyond raw performance metrics and provides deeper insights into the limitations of current models.

**Prompt Template for Generating Demographic Attributes:**
Generate consistent and realistic personal data including **Demographic Attributes**, **Educational Background**, and **Employment Information** based on the provided **Physical Appearance** shown in the image.
Ensure that all generated information is logically and contextually consistent with each other and with the visual characteristics in the image.

**Input parameters include:**

- `[image]`: An image showing the person's **Physical Appearance**.
- `[Reference Information]`: Optional reference information for contextual consistency.
- `[Available Names]`: A list of available names from which you must **randomly select one**, ensuring that the selected name is not previously used (i.e., no duplication).

The output must be in **strict JSON format**, matching the following structure:

```
{
  "Name": "Alice",
 "Demographic Attributes": "Alice Johnson is a 28-year-old single American woman,
  fluent in English.",
  "Educational Background": "She holds a Master's degree in Data Science from
  Stanford University, graduating in 2020.",
  "Employment Information": "She has 5 years of experience as a Data Analyst
  in the technology industry.",
  "Physical Appearance": "Alice is 5'7\" tall, weighs 130 lbs, with brown eyes
  , black hair, and light skin."
}
```

Ensure that:

- All fields are present and correctly formatted.
- The selected name is grammatically and culturally consistent with the generated personal data.
- No additional text or explanation is included—only the raw JSON object should be returned.

Figure 5: Prompt template for generating speaker demographic attributes from an image.

**Prompt Template for converting extended attributes into keyword-based queries:**
Generate a single phrase from the given sentence that is highly relevant to the topic and suitable for visual representation. Ensure that the images retrieved using this phrase do not contain any faces, hence focus on extracting a phrase centered around an object or scene rather than a person. Ensure that the output does not mention any person's name or pronouns such as I, he, his, her, she.

**Input parameters include:**

- `[Sentence]`: A sentence related to the topic.
- `[Topic]`: The topic associated with the sentence.

The output should follow this structure:

```
Sentence: [Sentence]
Topic: [Topic]
Extracted Phrase: [Generated phrase focusing on objects or scenes]
```

Ensure that:

- The generated phrase does not contain names or pronouns.
- The focus of the phrase is on objects or scenes rather than people.
- All fields are present and correctly formatted.

Figure 6: Prompt template for converting extended attributes into keyword-based queries.

**Prompt for Generating Persona Topic Descriptions**
You are required to process a set of change events related to a person's profile and generate detailed descriptions for each event, along with valid dates. Each change can be one of the following types: **addition**, **deletion**, or **update**.
Each event must be described in **one or two realistic and coherent sentences**, explaining the plausible event that led to the change. In addition, you must assign a **specific date** in the format `YYYY-MM-DD` for each event, ensuring that:

- Dates are arranged in **chronological order**.
- Events are consistent with the person's **Demographic Attributes** and are **realistic and believable**.

The input includes:

- `[Demographic Attributes]`: The **Demographic Attributes** of the person.
- `[change_events]`: The input JSON containing attributes and a list of change events under the `"change_events"` key.

Your output must be a **valid JSON object**, with each attribute containing both `"change_events"` and `"change_dates"` arrays. Ensure that:

- No additional text or explanation is included.
- Only the **raw JSON object** is returned.
- All events are consistent with the individual's background and lifestyle.

Figure 7: Prompt for generating persona topic descriptions.

**Prompt for Generating Short-Term Memory:**
You are given the following information:
- Previous conversation summary between [name] and chatbot: [summary].
- Current date: [current_date].
- Most recent dialogue exchange between [name] and chatbot: [dialogue].
**Task:** Generate a concise summary (150 words or fewer) that integrates both the historical and current conversations between [name] and the chatbot.
**Requirements:**

- Include key facts about [name] (e.g., personal details, preferences, lifestyle, profession).
- Mention relevant topics and time references.
- Highlight any notable updates or changes in interests, habits, or life events.
- Avoid redundancy and focus on meaningful, actionable insights.

**Output:** A clear, well-structured summary in natural language.
**Example Output:**
Leena, a Senior Environmental Analyst specializing in sustainability, has engaged in various conversations with the chatbot over several months, discussing topics such as stress management, sous-vide cooking, local artisan markets, and sustainable practices like buying in bulk and using reusable bags. In previous discussions, she explored eco-friendly options such as an induction cooktop, sustainable resource management, and renewable energy. By October 2024, Leena focused on managing her busy schedule and personal well-being. On November 20, 2024, Leena shared her enthusiasm for updating her grocery shopping habits by taking advantage of seasonal sales. She plans to stock up on organic produce, bulk grains, and is considering using reusable silicone bags to reduce plastic waste.

Figure 8: Prompt for generating short-term memory by integrating historical and current conversations.

**Prompt for Generating Long-Term Memory:**
**Instructions:**

- Please extract all possible personal information about [name] from the provided conversation and initial personal information.
- For each piece of personal information, reference the exact source using the format: "session_id:dialogue_id".
- The extracted information should be concise, factual, and suitable for database storage.
- Exclude subjective interpretations or abstract observations (e.g., "the speaker is supportive").
- Ensure no detail is omitted from either the conversation or the initial personal information.
- Do not repeat or duplicate existing persona entries.
- Output the result in JSON format.
- Escape all double-quote characters within string values using backslash (\).

The output should follow this structure:

```
{
  "personal_belongings": {
    "info": [
      "Leena has new running shoes named Swift Shadows, which she styled with
      a vintage camera on a wooden coffee table."
    ],
    "source": [
      ["2:4", "2:8"]
    ]
  },
  "shopping_preferences": {
    "info": [
    "Leena likes to buy fresh vegetables and ingredients to try out new recipes.",
    "Leena has found a new passion for supporting local communities and artisans.",
    ],
    "source": [
      ["1:4", "1:8"],
      ["8:10", "8:12"]
    ]
  }
}
```

Ensure that:

- The JSON output includes only factual, extractable personal details.
- Each piece of information is tagged with its source using the format: "session_id:dialogue_id".
- Double quotes within strings are escaped using backslash (\).
- The structure remains consistent and suitable for long-term storage in a persona database.

Figure 9: Prompt for generating long-term memory by extracting personal information from conversations.

**Prompt for Generating Image Caption:**
**Instructions:**

- Analyze all visual aspects of the image, including facial features, expressions, poses, clothing, colors, objects, and environmental context.
- Highlight distinctive and meaningful characteristics that may resonate with the user on a personal or emotional level.
- Use natural, vivid, and concise language to convey a clear mental image.
- Where relevant, identify cultural, contextual, or emotional implications inferred from the image.
- Focus on key elements; avoid unnecessary or overly detailed descriptions.

The output should follow this structure:

```
Image Caption: [Description of the image focusing on key elements]
```

Ensure that:

- The generated caption captures the essence of the image while adhering to the provided instructions.
- The language used is accessible and evocative, aiming to engage the viewer emotionally.
- All aspects mentioned in the instructions are considered during the caption generation process.

Figure 10: Prompt for generating image caption focusing on visual analysis and emotional resonance.

**Prompt for Generating User Dialogue:**
You are roleplaying as the user, [name], in a personalized conversation with your chatbot.
Last conversation date: [last_date]. Today's date: [current_date]. The multimodal persona information is: [multimodal persona information]. Today's topic is: [topic].
Summary of previous conversation:
SUMMARY: [conversation_summary]
RELEVANT_PERSONA: [relevant_persona]
Your task:

- Reveal information gradually, across multiple turns.
- Do not repeat or share all information at once.
- Each message must be under 20 words.
- Keep responses natural, relevant, and aligned with the topic.
- Mention personal or general topics as appropriate.
- End the conversation with "Bye!" when appropriate.
- You need to generate [turns_left] more turns before ending.

Only output the dialogue line for the user. Do not include any notes, explanations, or formatting.
UserDialogueStart

Figure 11: Prompt for user role dialogue generation.

**Prompt for generating Dialogue Agent Dialogue:**
You are roleplaying as the chatbot in a personalized conversation with the user, [name].
Last conversation date: [last_date]. Today's date: [current_date]. The multimodal persona information is: [multimodal persona information]. Today's topic is: [topic].
Summary of previous conversation:
SUMMARY: [conversation_summary]
RELEVANT_PERSONA: [relevant_persona]
Your task:

- Stay focused on the current topic.
- Do not introduce unrelated topics.
- Gradually encourage the user to reveal personal information.
- Keep each message under 20 words.
- Do not repeat previously mentioned details.
- End the conversation naturally with "Bye!" when the user signals closure.
- You need to generate [turns_left] more turns before ending.

Only output the dialogue line for the chatbot. Do not include any notes, explanations, or formatting.
ChatbotDialogueStart

Figure 12: Prompt for personalized dialogue agent generation.

**An example fact-level memory structure**


```
{

  "memory": {
    "Leisure Activities": [
      {
        "timestamp": "2024-02-04",
        "action": "add",
        "content": "reading books",
        "persona_facts": {
          "interests": ["reading"],
          "activities": ["Sophia enjoys reading after work,
          finding it both relaxing and enriching."]
        },
        "image_ref": "image_2.png"
      },
      {
        "timestamp": "2024-05-06",
        "action": "del",
        "content": "reading books",
        "persona_facts": {
          "interests": ["was previously fond of reading"],
          "activities": ["Sophia just decided to step back from reading
       due to increasing work commitments and a desire to explore other hobbies."]
        },
        "image_ref": "image_3.png"
      }
    ],
    "Food Preferences": [
      {
        "timestamp": "2024-01-07",
        "action": "add",
        "content": "preferring vegetarian dishes",
        "persona_facts": {
          "dietary_preferences": ["vegetarian"],
          "reasons": ["Sophia enjoys the health benefits and aligns with
          personal ethics towards sustainability."]
        }
      }
    ]
  }
}
```


Figure 13: An example fact-level memory structure.

**Prompt for Gemini-1.5-Pro Style Evaluation**
You are a professional evaluator tasked with assessing whether the generated answer aligns with the standard answer. Evaluate based on semantic accuracy, logical consistency, and information completeness.
**Scoring Criteria (0–5):**

- **5**: Fully consistent with the standard answer. Perfect match in semantics, logic, and completeness.
- **4**: Mostly consistent. Minor differences in phrasing or minor redundancy, but meaning is intact.
- **3**: Partially consistent. Contains key information but has minor omissions or unclear logic.
- **2**: Significant inconsistencies. Contains some correct information but with notable errors.
- **1**: Minimal relevance. Core information is incorrect or incomplete.
- **0**: Completely unrelated or entirely incorrect.

**Evaluation Output Format (JSON):**

```
{
  "evaluation": {
    "score": 4,
   "reason": "The generated answer is mostly consistent with the standard answer.
   However, there is a slight difference in wording between 'nuclear reaction' and
    'nuclear fusion'."
  }
}
```

**Input:**
- Question: [Question]
- Standard Answer: [Standard Answer]
- Generated Answer: [Generated Answer]

Please evaluate the generated answer strictly according to the criteria and return the result in JSON format.

Figure 14: Prompt for Gemini-1.5-Pro evaluation.

**Prompt for Personalization Score Evaluation**
You are a professional evaluator tasked with assessing whether the generated response appropriately reflects the user's persona without contradicting their current persona state. Consider both persona consistency and relevance.
**Scoring Criteria (0–5):**

- **5**: Perfectly reflects the persona and fully consistent with current persona state.
- **4**: Mostly consistent with persona, with only minor irrelevance.
- **3**: Partially reflects persona but misses some key aspects or has minor contradictions.
- **2**: Weak reflection of persona; notable contradictions or irrelevance.
- **1**: Barely related to persona; largely inconsistent.
- **0**: Completely unrelated to persona or fully contradictory.

**Evaluation Output Format (JSON):**

```
{
  "evaluation": {
    "score": 4,
    "reason": "The response generally aligns with the persona of being a
    professional researcher, but it misses details about long-term dialogue
    interest."
  }
}
```

**Input:**
- Persona: [User Persona]
- Generated Response: [Response]

Figure 15: Prompt for LLM-as-a-Judge evaluation of **Personalization Score**.

**Prompt for Alignment Score Evaluation**
You are a professional evaluator tasked with assessing whether the generated response is consistent with the ground-truth response. Evaluate based on semantic similarity, logical consistency, and informativeness.
**Scoring Criteria (0–5):**

- **5**: Fully consistent with the ground truth. Perfect semantic and logical match.
- **4**: Mostly consistent. Minor differences in wording or redundancy, but meaning intact.
- **3**: Partially consistent. Contains key information but some omissions or unclear logic.
- **2**: Significant inconsistencies. Some correct information but major errors.
- **1**: Minimal overlap. Core information incorrect or incomplete.
- **0**: Completely unrelated or entirely incorrect.

**Evaluation Output Format (JSON):**

```
{
  "evaluation": {
    "score": 5,
    "reason": "The generated response exactly matches the ground truth in
    content and structure."
  }
}
```

**Input:**
- Ground Truth Response: [Reference Response]
- Generated Response: [Response]

Figure 16: Prompt for LLM-as-a-Judge evaluation of **Alignment Score**.

**Prompt for Generating Fused Responses with Natural Distraction Insertion**
You are an expert conversational agent. Your task is to generate a single, fluent utterance by seamlessly integrating a *distraction response* into an *original response*. The insertion must occur at a natural discourse boundary—such as a topic shift, an expression of curiosity, or a natural pause—so that the resulting response remains coherent and minimally disruptive to conversational flow.
Do **not** simply concatenate the two inputs. Instead, use appropriate linguistic devices (e.g., "By the way…", "That reminds me…", "Speaking of which…", or a rhetorical question) to bridge the original content and the distraction smoothly. The core meaning of the original response must be preserved.
**Input:**

- **Original Response:** [Original Response]
- **Distraction Response:** [Distraction Response]

**Output Instructions:** Return only the final fused response as a single natural-language sentence or utterance. Do not include any labels, explanations, or formatting.
**Example:**
Input:
- Original Response: We should book the tickets early to get a discount.
- Distraction Response: Did you hear that they're opening a new planetarium downtown?
Output:
- We should book the tickets early to get a discount. Speaking of cool stuff, did you hear they're opening a new planetarium downtown?

Figure 17: Prompt for generating a fluent, fused dialogue response by inserting a distraction into the original response at a natural discourse boundary.