

PersonaVLM: Long-Term Personalized Multimodal LLMs

Supplementary Material

This supplementary material provides comprehensive details to complement the main paper, organized as follows:

- **Appendix A** elaborates on our proposed memory architecture, detailing each memory component—including its storage, retrieval, and update processes.
- **Appendix B** outlines the training and implementation details of PersonaVLM framework.
- **Appendix C** presents a detailed analysis of our synthesized dataset, covering its distribution and the validation process.
- **Appendix D** offers a comprehensive breakdown of Persona-MME, including its task taxonomy, detailed statistical analysis, and full evaluation results.
- **Appendix E** presents additional experimental details, including ablation studies and the full set of prompts used in our framework.
- **Appendix F** offers further efficiency analysis and limitations of PersonaVLM.

A. Details of the PersonaVLM Memory Architecture

As introduced in Section 3, the PersonaVLM memory architecture comprises two components: a User Personality Profile (\mathcal{P}) and a Multi-Type Memory Database (\mathcal{M}). This section provides a detailed exposition of how these memories are stored, updated, and retrieved.

A.1. Memory Storage

User Personality Profile (\mathcal{P}). We quantitatively represent the user’s personality as a five-dimensional vector, $\mathbf{p} \in \mathbb{R}^5$, where each element corresponds to a Big Five trait and is a floating-point value between 1 and 5. This profile is dynamically updated after each interaction turn m . Specifically, at the end of a turn, the model infers a personality vector, $\mathbf{p}'_m \in \mathbb{R}^5$, where each component is an integer score from 1 to 5 based on the user’s current input and context. The persistent personality profile \mathbf{p} is then updated using an Exponential Moving Average (EMA): $\mathbf{p} \leftarrow \lambda_m \mathbf{p} + (1 - \lambda_m) \mathbf{p}'_m$ where the smoothing factor λ_m is dynamically adjusted to be more sensitive in early interactions and stabilize over time: $\lambda_m = 0.7 - 0.2 \cdot \cos\left(\frac{\min(m, 50)}{50} \pi\right)$. To ensure stability, this update is applied selectively. The process is skipped if the inferred personality vector \mathbf{p}'_m consists solely of the neutral score (3), a condition that typically arises in non-personalized or neutral contexts. During the response generation stage, the personality profile \mathcal{P} is provided to the model via structured prompting.

Core Memory. Core memory stores the user’s foundational and high-priority attributes and is included in every interaction turn. It is divided into two sub-components [31]:

- **Human:** Factual user attributes, such as age, gender, preferences, and interests, with the user’s name as a mandatory field. This information provides PersonaVLM with a foundational understanding of the user’s background.
- **Persona:** The user’s identity, roles (e.g., “a meticulous researcher”), and explicit requirements for the model’s interaction style, tone, and behavior.

Semantic Memory. Semantic memory [40] archives timeless, multimodal knowledge that is either explicitly provided by the user or autonomously inferred by the model. This knowledge is categorized as follows:

- **Explicit Directives:** Direct commands from the user to remember specific information, which can be textual or visual. For example, a user might provide an image and say, “Remember the boy in this picture.”
- **Core Facts:** Stable, factual information about the user disclosed during conversation, such as their profession, significant life events, or specific requirements for the agent’s behavior.
- **Preferences & Habits:** User preferences for entities, visual styles, or activities, which can be either explicitly stated or implicitly revealed through behavior patterns.
- **Visual Concepts:** User-specific visual concepts that arise in multimodal dialogues, such as friends, pets, or personal items. These are stored as a key-value pair linking a textual description to an image crop, formatted as “simple description <image>”.

Beyond these predefined categories, the agent autonomously determines at the end of each turn whether new semantic knowledge warrants storage. If so, it generates a structured output containing the reasoning process, memory content, and a set of keywords for future retrieval.

Episodic Memory. Episodic memory archives both summaries and raw data from past conversations. For each multi-turn dialogue session, the model segments the conversation by topic. Each resulting topic-based episode contains three key elements: (a) a concise summary, (b) a set of keywords, and (c) the indices of the dialogue turns constituting that episode. To ensure no details are lost, the original dialogue data is never deleted; the episodic memory serves as a structured layer for organizing and retrieving this raw data.

Procedural Memory. Procedural memory tracks user goals and identifies recurring behaviors or habits by stor-

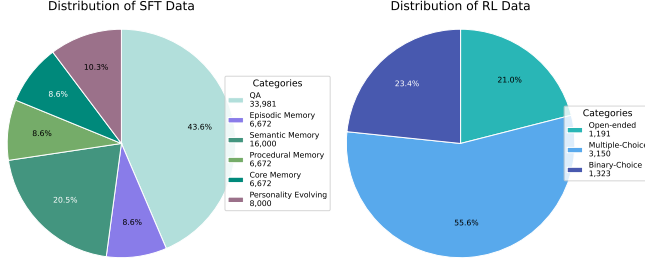


Figure 7. Data composition for the training of PersonaVLM

ing procedural events from conversations. It primarily stores two types of information:

- **Long-term Goals:** Ongoing projects, plans, or objectives that the user is working towards.
 - **Habits & Routines:** Repetitive behaviors or workflows that are automatically identified from user interactions.
- Similar to Core Memory, this information is stored as key-value pairs, and only the latest version is retained.

A.2. Memory Retrieval

Memory retrieval is a critical step within the Response Stage, initiated when PersonaVLM determines that external knowledge is necessary to fulfill a user’s request. The process begins by generating a retrieval query encapsulated within `<retrieve></retrieve>` tags. This specifies a time period and keywords to guide the search. The time period is defined by start and end timestamps in a “YYYY – MM – DDHH : MM” format.

Textual Memory Retrieval. For text-based memories (i.e., procedural, semantic, and episodic), we employ a parallel multi-source retrieval strategy. First, all textual memories are encoded into dense vectors using the `all-MiniLM-L6-v2` sentence transformer⁷. Given a user query, we perform a similarity search against the memory database. The top- k most relevant memories are retrieved from each category, where k is empirically set to 2, 4, and 2 for procedural, semantic, and episodic memories, respectively, unless otherwise specified. This entire process is accelerated by leveraging Facebook AI Similarity Search (FAISS) [17] for efficient indexing and retrieval.

Visual Concept Retrieval. This process is automatically triggered when the user’s input contains an image. First, we employ an off-the-shelf object detector, Grounding DINO [25], to extract salient objects from the input image. We then compute the cosine similarity between the CLIP [33] embeddings of these detected objects and the visual concepts stored in semantic memory. This process

⁷<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

Algorithm 1 Operational Pipeline of PersonaVLM

Input: User query $\mathcal{Q}_m = (T_m, I_m, t_m)$, personality profile \mathcal{P}_{m-1} , memory database \mathcal{M}_{m-1} , max reasoning steps N , model π_θ , session threshold t_s .

- 1: **if** $t_m - t_{m-1} \geq t_s$ **then**
- 2: Update Core, Procedural, and Episodic Memory based on the last session.
- 3: **end if**
- 4: $\mathcal{C}_m \leftarrow \{(Q_i, \mathcal{R}_i) \mid 0 < i < m \text{ and } |t_i - t_m| \leq t_s\}$
- 5: **for** $n = 1$ to N **do**
- 6: $\mathcal{S}_n \leftarrow \pi_\theta(\mathcal{Q}_m, \mathcal{C}_m, \mathcal{P}_{m-1})$
- 7: $\text{action}, \text{args} \leftarrow \text{Parse}(\mathcal{S}_n)$
- 8: **if** $\text{action} = \text{retrieve}$ **then**
- 9: $(\text{keywords}, \text{time period}) \leftarrow \text{args}$
- 10: $\mathcal{M}_{\text{retrieved}} \leftarrow \text{Retrieve}(\mathcal{M}_{m-1}, \text{keywords}, \text{time period})$
- 11: $\mathcal{C}_m \leftarrow \mathcal{C}_m \cup \mathcal{M}_{\text{retrieved}}$
- 12: **else if** $\text{action} = \text{answer}$ **then**
- 13: $\mathcal{R}_m \leftarrow \text{args}$
- 14: **break**
- 15: **end if**
- 16: **end for**
- 17: Infer turn-specific personality \mathbf{p}'_m from \mathcal{Q}_m and update long-term profile \mathbf{p}_m .
- 18: Convert \mathbf{p}_m to textual summary \mathcal{P}_m .
- 19: Extract and update Semantic Memory based on the current turn $(\mathcal{Q}_m, \mathcal{R}_m)$.

Output: Final response \mathcal{R}_m , updated state $(\mathcal{P}_m, \mathcal{M}_m)$.

mirrors the text-based semantic search, creating a unified retrieval mechanism across modalities.

A.3. Memory Management

Our memory management policies distinguish between raw conversational history and structured memory [43]. While the complete interaction history is retained for low-level access, the structured memories are managed according to the following policies. Semantic and Episodic memory are treated as purely additive; new entries detailing facts, concepts, or events are appended without modifying or deleting existing ones, thereby preserving an immutable historical record. In contrast, Core and Procedural memory maintain a single, canonical version of the user’s profile and habits. These memories are mutable and undergo CRUD operations at the end of each session to ensure they accurately reflect the user’s most current state.

B. Implementation Details of PersonaVLM

B.1. Implementation Process

The end-to-end operational pipeline of PersonaVLM is detailed in Algorithm 1. In our offline implementation, a new

Table 3. The hyperparameters used in SFT and RL training.

Hyperparameter	Training stage	
	SFT	RL
Batch Size	64	72
Learning Rate	2e-5	2e-6
LR Scheduler	cosine	cosine
Total Steps	1200	400
Max Pixels	230400	230400
Max Length	16384	16384
Warmup Ratio	0.03	0.05
deepspeed	zero3	zero3
Number of Samples in a Group	-	6
Num Iterations	-	1
Repetition Penalty	-	1.05

user session is initiated if the time elapsed since the last interaction, $t_m - t_{m-1}$, exceeds a predefined threshold t_s (e.g., 60 minutes). At the start of a new session, a memory consolidation process is triggered to update the user’s long-term Core and Procedural memories based on the previous session.

B.2. Training Details

Training Data Composition. The composition of our training data for the SFT and RL stages is detailed in Fig. 7. The SFT dataset comprises a total of 78k samples. This dataset is constructed using the synthesis pipeline illustrated in Fig. 3 (a) and is further augmented with 6k user-related concept samples based on [11]. The SFT data is primarily split between question-answering (QA) pairs for reasoning (43.6%) and memory-related samples (56.4%). The memory-related category is further subdivided into a personality inference task (10.3%) and examples for the four memory types (46.1%). In contrast, the RL dataset consists of 5.6k samples, categorized into three types: open-ended QA with verifiable answers (21.0%), multiple-choice questions (55.6%), and binary-choice questions (23.4%).

Implementation Details. We implement our training pipeline based on the repositories Qwen-VL⁸ and ms-swift⁹. The hyperparameter settings for both the SFT and RL stages are detailed in Table 3. All experiments were conducted on a server equipped with 8 NVIDIA H800 GPUs. The entire two-stage training process completes in approximately 8 hours, comprising 2 hours for SFT and 6 hours for RL.

Group Relative Policy Optimization. GRPO [10] is an advancement over PPO [36] that refines policy optimization

by replacing the critic model with a relative evaluation mechanism. Instead of learning an absolute value function, GRPO estimates advantages by comparing the quality of multiple trajectories sampled within a group. For each training sample $\{Q, \widehat{R}\}$, where Q is the user input and \widehat{R} is the preferred response, the policy model π_θ rollouts a group of multi-turn trajectories $\{\tau_1, \dots, \tau_G\}$. The reward for each trajectory τ_i is calculated using Eq. (3). Based on these rewards, we then compute the normalized advantage \widehat{A}_i^t for each token by normalizing them across the sampled group. The optimization objective is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(Q, \widehat{R}) \sim \mathcal{D}, \{\tau_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | Q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau_i|} \sum_{t=1}^{|\tau_i|} \min \left(r_i^t(\theta) \widehat{A}_i^t, \text{clip}(r_i^t(\theta), 1 - \epsilon, 1 + \epsilon) \widehat{A}_i^t \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right]. \quad (4)$$

where $r_i^t(\theta) = \frac{\pi_\theta(\tau_{i,t} | \tau_{i,<t})}{\pi_{\theta_{\text{old}}}(\tau_{i,t} | \tau_{i,<t})}$ is the probability ratio, π_{ref} is a reference policy, and β is a hyperparameter that controls the strength of the KL regularization. Detailed training settings are provided in Table 3.

Optimization Strategies. To improve the effectiveness and robustness of our retrieval mechanism, we implement several optimization strategies.

First, to mitigate retrieval redundancy within a single reasoning trajectory, the model is encouraged to use diverse query conditions (i.e., keywords and time periods). We enforce this by implementing a deduplication filter that prevents any single memory entry from being retrieved more than once per trajectory.

Second, we employ a dynamic top- k strategy during training to better prepare the model for varied information scenarios. Specifically, while we use fixed top- k values at inference (2 for episodic and 4 for semantic memories), these values are randomized during training, sampled uniformly from the ranges [2, 5] and [3, 6], respectively. This approach acts as a form of data augmentation, training the model to be robust to both sparse and dense information retrieval contexts.

C. Data Curation Details.

Data Distribution. We synthesize a large-scale, long-term multimodal dialogue dataset by sampling 700 unique personas from PersonaHub [9], allocating 500 for training and 200 for testing. The detailed distribution of the synthesized data is visualized in Fig. 8 and the top-right panel of Fig. 3. Training dialogues consist of 20 to 100 turns, spanning a simulated timeframe of up to one month. In contrast, test dialogues are designed to be more challenging, featuring

⁸<https://github.com/QwenLM/Qwen3-VL>

⁹<https://github.com/modelscope/ms-swift>

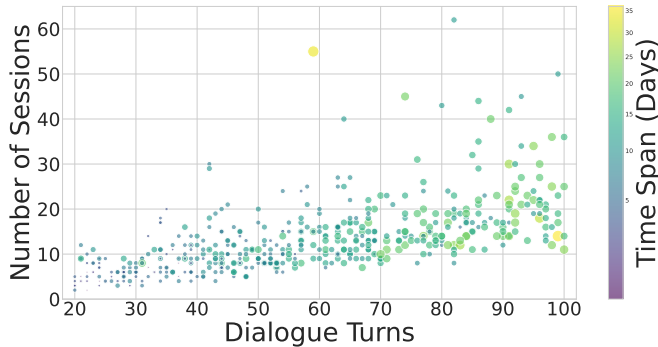


Figure 8. Distribution of the 500 long-term conversation samples in the training data.

longer conversations in two settings: 20–100 turns (for a $32k$ context window) and 100–500 turns (for a $128k$ context window), with a simulated duration of up to three months. This designed discrepancy between training and testing data allows for a rigorous evaluation of our memory architecture’s long-term capabilities.

It is important to note a distinction in how the dialogue data is utilized: the full, synthesized multi-turn dialogues serve as the database for retrieval, while the QA pairs used for model training feature re-generated answers. This is because the original answers have access to the complete dialogue history, whereas the training target must be an answer generated solely based on the current query and the retrieved memory content.

Multimodal Memory Formatting. To support multimodal knowledge, visual concepts in semantic memory are stored in a structured format: “Memory Content (Image Object: <class_name>)”. During the memory update process, Grounding DINO [25] is used to crop the corresponding object from the image. This cropped image patch is then paired with a simple textual description forming the input format for the model, i.e., “simple description <image>”.

Crucially, the system distinguishes between concrete visual objects and abstract preferences. For instance, if a user states, “I like this style of picture,” the system stores a textual fact, such as “User likes [style description],” rather than the raw image or its constituent objects. Also, episodic memory retains the original multimodal dialogue turns, including both text and full images, to preserve memory integrity.

Data Validation. To ensure the accuracy, safety, and overall quality of our synthesized dataset, we employ a two-stage filtering process. First, we perform automated filtering using both rule-based checks and model-based validation. During data synthesis, the generation model outputs structured metadata, such as timestamps and dialogue turn indices for episodic topics. We leverage this metadata to apply

rule-based checks that validate data integrity, including the chronological consistency of timestamps and the completeness of episodic dialogues. Concurrently, a model-based self-correction mechanism verifies the safety and coherence of the generated content. Second, the automatically filtered data undergoes a human review. In this final step, human reviewers are tasked with identifying and removing any remaining erroneous, nonsensical, or repetitive dialogues, ensuring the final dataset is of high fidelity.

D. Persona-MME: Details and Statistics

Task Taxonomy. We provide the definitions for evaluated tasks in Table 5 and present illustrative examples in Fig. 9.

Data Statistics and Distribution. Persona-MME is designed to evaluate long-term personalization across seven key aspects, encompassing a total of 14 fine-grained tasks and comprising **2,034 *in-situ*** test cases. It is important to note that a single test scenario may simultaneously assess multiple capabilities. Fig. 11 illustrates the distribution of these tasks. The benchmark consists of 13 primary tasks (from Visual Detail Recall to Generalizing to New Scenarios), which are distributed relatively evenly. The 14th task, personality alignment, is not a standalone category but is evaluated concurrently within 406 of the primary task cases.

The diversity of our evaluation set is a core design principle. We constructed **200** unique personas, each with a distinct fictional background, and crafted dialogues that span a broad spectrum of topics and scenarios to ensure comprehensive testing. The resulting topical breadth is visualized in Figure 12, which presents a word cloud of the most prominent keywords from the evaluation dialogues.

Further statistical analysis of Persona-MME is presented in Table 6. On average, each *in-situ* test case is grounded in a conversational history of **142.9 turns**, of which **15.87%** are multimodal. The average length of a test question is **22.7 words**, while the average answer length is **3.05 words**. A significant portion of questions, **34.02%**, require visual information from the context to be answered correctly.

Comprehensive Evaluation. We present a comprehensive evaluation of over ten leading models on the $128k$ configuration of Persona-MME, with detailed results provided in Table 4 and Fig. 10. The evaluation spans a range of proprietary models (e.g., GPT-4o, GPT-5, Gemini-2.5-Flash, Claude-3.7-Sonnet) and open-source alternatives (e.g., the Qwen series, InternVL3-8B/38B, OneVision-1.5-8B). Our key findings are as follows:

- **Proprietary vs. Open-Source Gap:** Proprietary models exhibit significantly better overall personalization capabilities than their open-source counterparts.
- **Challenges for Smaller and Multimodal Models:** Smaller open-source multimodal models, such as

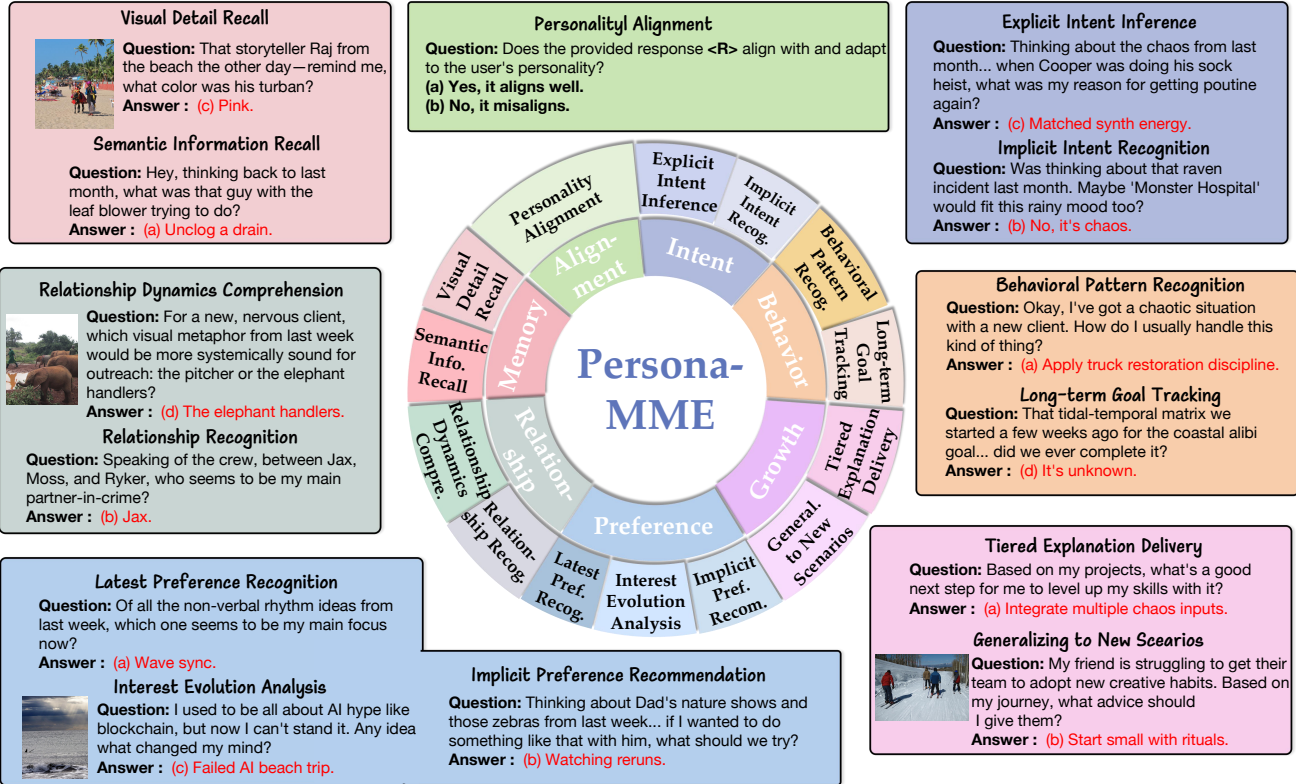


Figure 9. Illustrative in-situ cases for the 14 task categories in Persona-MME, organized into the seven core personalization aspects.

Table 4. Comprehensive evaluation on the 128k configuration of Persona-MME. We compare PersonaVLM with proprietary and open-source models across 14 tasks: Visual Detail Recall (VDR), Semantic Information Recall (SIR), Explicit Intent Inference (EII), Implicit Intent Recognition (IIR), Latest Preference Recognition (LPR), Interest Evolution Analysis (IEA), Implicit Preference Recommendation (IPR), Behavioral Pattern Recognition (BPR), Long-term Goal Tracking (LGT), Relationship Recognition (RR), Relationship Dynamics Comprehension (RDC), Tiered Explanation Delivery (TED), Generalizing to New Scenarios (GNS), and Personality Alignment (PA).

Model	Memory		Intent		Preference			Behavior		Relationship		Growth		Alignment	Overall
	VDR	SIR	EII	IIR	LPR	IEA	IPR	BPR	LGT	RR	RDC	TED	GNS	PA	
Random	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	50.00	32.11
<i>Proprietary models</i>															
GPT-4o-mini	54.39	89.74	78.46	64.81	64.58	59.68	61.22	68.33	45.31	54.17	71.43	73.33	75.81	65.14	66.44
GPT-4o	73.68	92.31	86.15	62.96	62.50	54.84	61.22	61.67	50.0	56.25	75.51	73.33	79.03	78.87	71.90
GPT-5	<u>85.71</u>	98.72	93.85	67.92	74.47	70.97	65.31	76.67	70.97	85.11	81.63	76.19	75.81	92.25	82.95
Gemini-2.5-Flash	88.06	<u>92.55</u>	<u>88.00</u>	73.44	67.86	47.89	50.00	62.5	58.33	72.22	77.19	75.00	80.00	80.90	74.90
Claude-3.7-Sonnet	51.47	91.11	80.26	76.19	60.38	61.43	61.54	61.97	38.24	64.81	66.67	66.67	70.42	80.65	70.40
<i>Open-source models</i>															
Qwen2.5-VL-7B	52.11	49.47	52.44	57.58	52.63	48.65	57.14	55.84	52.7	50.88	60.32	56.9	64.0	55.0	54.62
InternVL3-8B	29.58	77.89	74.39	62.12	59.65	54.05	46.43	66.23	43.24	61.40	76.19	75.86	77.33	54.17	60.08
InternVL3-38B	38.03	89.47	78.05	63.64	68.42	<u>64.86</u>	60.71	<u>72.73</u>	44.59	57.89	71.43	70.69	81.33	63.06	66.01
Qwen3-VL-8B	63.38	84.21	76.83	68.18	61.4	58.11	67.86	67.53	40.54	<u>82.46</u>	76.19	<u>79.31</u>	<u>88.00</u>	71.39	70.75
Qwen3-30B-A3B	29.58	85.26	82.93	<u>75.76</u>	<u>70.18</u>	63.51	64.29	63.64	44.59	68.42	<u>77.78</u>	82.76	86.67	81.39	72.65
OneVision-1.5-8B	42.86	59.57	59.26	49.23	62.5	46.58	<u>69.09</u>	48.68	41.89	73.21	58.06	64.91	68.92	53.93	55.88
PersonaVLM (ours)	50.70	83.16	81.71	72.73	59.65	54.05	73.21	58.44	<u>62.16</u>	75.44	74.60	82.76	92.00	<u>92.22</u>	<u>77.08</u>

Qwen2.5-VL-7B, InternVL3-8B, and OneVision-1.5-8B, particularly struggle with personality alignment, with their performance often being comparable to a random baseline.

In contrast, large language-centric models like Qwen3-30B-A3B can achieve superior overall scores, outperforming even larger multimodal models like InternVL3-38B,

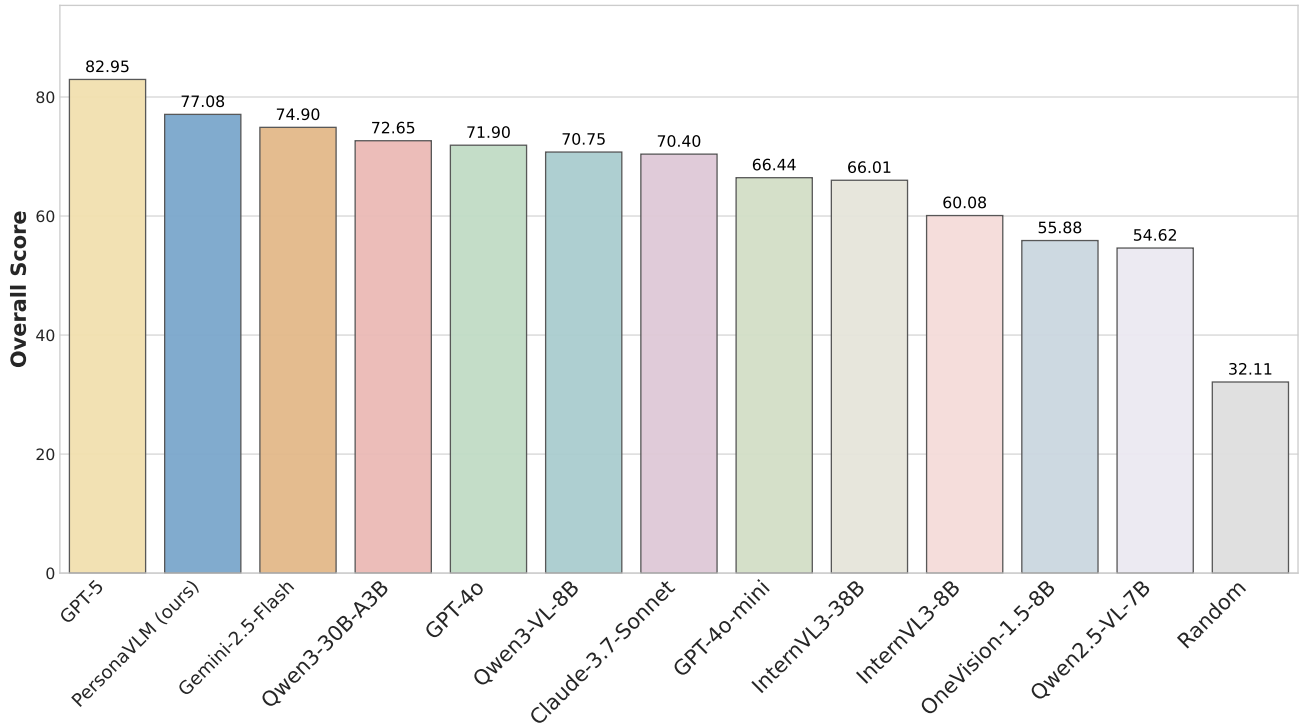


Figure 10. Overall performance on Persona-MME (128k), ranking PersonaVLM against various proprietary and open-source models.

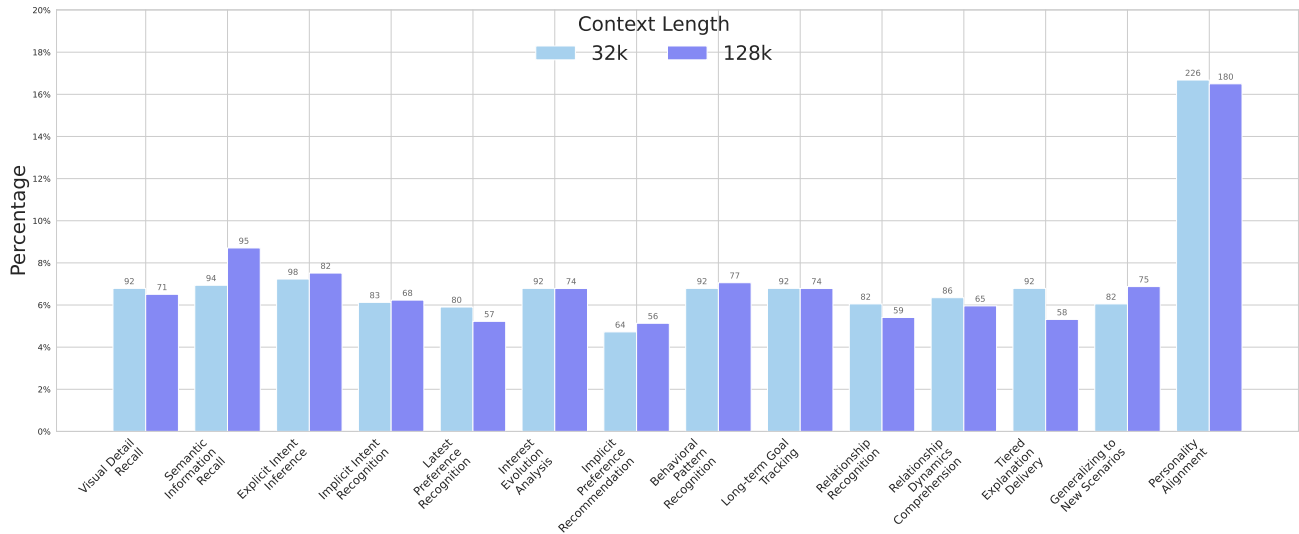


Figure 11. Distribution of the 14 fine-grained tasks in Persona-MME across its 32k and 128k context length configurations, with the number of test cases indicated for each task.

despite their inherent limitations on visual tasks (e.g., VDR).

- **No Single Dominant Model:** Even the top-performing model, GPT-5, does not dominate across all sub-tasks. It

is surpassed by other models in specific areas, including Growth Modeling and Visual Detail Recall, highlighting the complexity of holistic personalization.

- **Effectiveness of PersonaVLM:** Our PersonaVLM frame-

Table 5. Task definitions for the Persona-MME evaluation suite.

Evaluation Aspect	Task	Definition
Memory	Visual Detail Recall	Assesses the ability to recall fine-grained visual details from previously shared images.
	Semantic Information Recall	Evaluates long-term memory for semantic information (e.g., events, preferences, context) from conversational history.
Intent	Explicit Intent Inference	Assesses understanding a user’s explicitly stated intent from past multimodal context (e.g., linking a suitcase photo to text about a "business trip").
	Implicit Intent Recognition	Tests inferring a user’s latent intent from the current context (e.g., deducing a calm beach is unsuitable for a user known to enjoy surfing).
Preference	Latest Preference Recognition	Assesses prioritizing recent behavioral evidence (e.g., a rock concert photo) over older, contradictory stated preferences (e.g., "I only like classical music").
	Interest Evolution Analysis	Evaluates explaining the evolution of a user’s interests by linking it to a specific causal event from their history (e.g., a change in diet due to a mentioned allergy).
	Implicit Preference Recommendation	Tests recommending based on implicit values inferred from user history (e.g., suggesting a bicycle over a car to an eco-conscious user).
Behavior	Behavioral Pattern Recognition	Assesses recognizing recurring behavioral patterns to predict actions (e.g., predicting a Saturday café visit based on a history of Saturday café photos).
	Long-term Goal Tracking	Tests tracking a long-term goal’s progress, including identifying when its status is unknown due to insufficient information.
Relationship	Relationship Recognition	Tests identifying relationship significance based on the frequency and context of individuals in multimodal history (e.g., identifying a closer friend by their more frequent appearance).
	Relationship Dynamics Comprehension	Assesses inferring relationship dynamics from context to adapt communication style (e.g., using a professional tone for a colleague).
Growth	Tiered Explanation Delivery	Tests tailoring explanations to a user’s evolving skill level.
	Generalizing to New Scenarios	Assesses the ability to generalize a user’s behaviors and preferences to novel scenarios.
Alignment	Personality Alignment	Assesses the ability to infer a user’s personality traits from long-term interactions and adapt its responses accordingly.

work significantly enhances the baseline model’s performance by **22.46%**. The most substantial improvements are concentrated in the sophisticated dimensions of **Growth** and **Alignment**, underscoring the targeted benefits of our approach.

Comparison with Existing Benchmarks. As shown in Table 7, Persona-MME provides a more comprehensive evaluation of personalization compared to existing benchmarks. Specifically, it is the only benchmark that combines **long-term** interaction scenarios, **multimodal** (vision and text) inputs, and a holistic assessment of memory, understanding, and alignment capabilities.

Table 6. Key statistics of the Persona-MME.

Statistic	Value
Avg. turns per dialogue	142.9
Multimodal turn ratio	15.87%
Avg. question length	22.7 words
Avg. answer length	3.05 words
Image-related question ratio	34.02%



Figure 12. Word cloud of keywords from the dialogue data in Persona-MME, illustrating the rich diversity of conversation scenarios and topics.

Quality Assurance. To ensure the quality of Persona-MME, every test case underwent a rigorous manual review process. We first generated initial questions using the Gemini-2.5-Pro API¹⁰. Subsequently, a team of four annotators meticulously reviewed each case against three key criteria: (a) Consistency: ensuring the question aligns with its assigned task category. (b) Accuracy: verifying the correctness of the ground-truth answer. (c) Alignment Validity: assessing whether the model’s response in alignment tests appropriately adapts to (or conflicts with) the predefined personality traits. Any examples found to be ambiguous or conflicting were discarded. This comprehensive review process required approximately 40 person-hours to complete.

E. More Experimental Details

E.1. Benchmarks

PERSONAMEM [14]. This is a recent benchmark featuring synthetic, multi-session, and timeline-aware conversational data, designed to evaluate an LLM’s ability to remember, track, and generalize from personalized user profiles and preferences. It includes seven types of *in-situ* user queries, including: recall user-shared facts, suggest new ideas, acknowledge latest user preferences, track full preference evolution, revisit reasons behind preference updates, and provide preference-aligned recommendations. We conduct evaluations under two context-length settings, 32k and 128k to-

¹⁰We use the gemini-2.5-pro-preview-06-05 model.

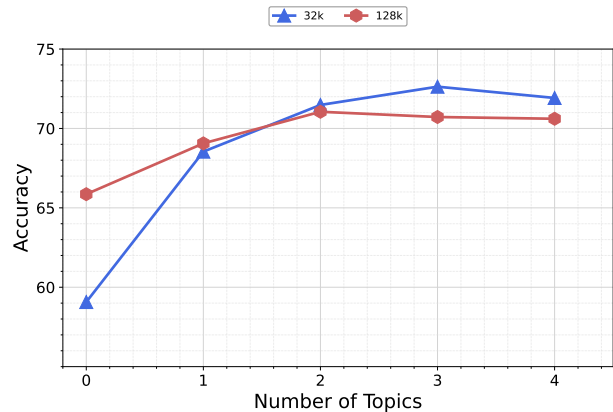


Figure 13. Ablation study on the number of retrieved episodic topics for Persona-MME.

kens. The settings comprise 589 and 1,362 multiple-choice questions, respectively, with the larger setting derived by sampling half of the personas from the original 2,728. Performance is measured by accuracy, and the comparative results are reported in Table 1 and Fig. 4.

P-SOUPS [13]. P-SOUPS assesses LLM personalization across three preference dimensions: *Expertise*, *Informative-ness*, and *Style*, each containing 600 test cases for a total of 1,800. A single test case consists of a user prompt, a profile, and a pair of responses: one aligned with the profile (the “chosen” response) and one misaligned (the “rejected” response). The model is tasked with selecting the aligned response from the pair, and performance is measured by accuracy. For our few-shot experiments, we augment the input with a single example of Pair-wise Comparative Feedback, as provided by the benchmark.

E.2. Ablation Study

Effectiveness of Different Memory Types. We present an ablation study on the memory components of PersonaVLM architecture in Table 8. The results consistently show that removing any single memory type degrades overall performance, a trend that holds across both the 32k and 128k context settings. Notably, Episodic memory emerges as the most critical component; its removal leads to a substantial performance drop of 12.41% and 5.19% in the two settings, respectively, while removing other memory types results in a performance drop of less than 2%. Delving into sub-task performance, we observe that Procedural memory has a strong impact on tasks related to Behavior and Relationship. Collectively, these findings suggest that the different memory types fulfill distinct yet complementary roles, and all are integral to the holistic performance of the PersonaVLM agent.

Table 7. Comparison of Persona-MME with existing personalization benchmarks. Abbreviations are defined as follows. **Modality**: V (Visual), T (Text). **Capabilities**: U (Personalized Understanding), M (Memory), A (Alignment). **Answer Type**: MC (Multiple Choice), BC (Binary Choice).

Benchmark	Modality	Long-Term	Capabilities	Answer Type
PERSONAMEM [14]	T	✓	M + U	MC
P-SOUPS / ALIGNX-test [13, 21]	T		A	BC
Yo’LLaVA [28]	V		U	BC
RAP [11]	V		M + U	MC
Persona-MME (ours)	V + T	✓	U + M + A	MC + BC

Table 8. Ablation study of PersonaVLM components on the Persona-MME benchmark. The evaluation shows the performance impact of removing (“w/o” denotes “without”) key components, specifically the individual memory types (Core, Procedural, Semantic, Episodic) and the reasoning capability.

Context	Setting	Persona-MME						Overall
		Memory	Intent	Preference	Behavior	Relationship	Growth	
32k	PersonaVLM	69.89	76.8	58.05	69.02	73.21	86.78	71.48
	- w/o Core	73.66	74.59	59.32	63.59	67.26	83.91	69.80 _{-1.68}
	- w/o Procedural	72.58	79.01	59.32	59.78	70.24	85.06	70.33 _{-1.15}
	- w/o Semantic	66.67	72.38	59.32	66.30	72.02	85.63	69.71 _{-1.77}
	- w/o Episodic	33.77	55.35	56.68	66.27	69.84	74.07	59.07 _{-12.41}
	- w/o Reasoning	69.57	69.7	60.48	62.14	71.76	83.61	68.73 _{-2.75}
128k	PersonaVLM	69.28	77.70	61.50	60.26	75.00	87.97	71.05
	- w/o Core	69.28	77.70	64.17	58.28	72.50	84.21	70.39 _{-0.66}
	- w/o Procedural	69.88	77.03	61.50	54.97	69.17	87.97	69.39 _{-1.66}
	- w/o Semantic	67.47	73.65	60.43	60.93	74.17	88.72	69.94 _{-1.11}
	- w/o Episodic	50.60	68.92	60.96	62.25	70.00	88.72	65.86 _{-5.19}
	- w/o Reasoning	59.21	71.97	57.87	65.87	73.39	80.00	67.32 _{-3.73}

Episodic Memory Configuration. Given the critical role of episodic memory, we conduct an ablation study on the number of retrieved memory topics. As shown in Fig. 13, the overall accuracy on Persona-MME initially increases with the number of retrieved topics before performance saturates. To strike a balance between performance and computational efficiency, we select two topics as the default setting for all of our main experiments.

Table 9. Ablation study on the PEM component.

Setting	Expertise	Informativeness	Style	Overall
PersonaVLM	51.1	53.6	44.0	49.6
- w/o PEM	48.5 _{-2.6}	50.2 _{-3.4}	38.3 _{-9.2}	45.6 _{-4.0}

Effectiveness of the Reasoning Capability. We validate the effectiveness of PersonaVLM’s multi-step reasoning and retrieval capability with two key findings. First, the full PersonaVLM model, trained with reinforcement learning,

demonstrates a significant 4–7% performance gain over its SFT-only baseline on Persona-MME and PERSONAMEM (Table 1). This highlights the benefit of the overall training process in cultivating this advanced reasoning behavior. To further isolate the contribution of this capability, we conduct an ablation study. Disabling multi-turn retrieval during the response stage results in performance drops of 2.75% and 3.73% at the 32k and 128k context settings, respectively (Table 8). These results underscore the critical role that multi-step reasoning and retrieval play in achieving robust long-term personalization.

E.3. Personality Evolving Mechanism

In Fig. 14, we visualize how PEM captures the evolution of personality traits inferred from user interactions across diverse personas. Furthermore, as shown in Table 9, our ablation study on P-SOUPS demonstrates the effectiveness of PEM. Finally, Table 10 provides examples of how PersonaVLM adapts its responses based on the inferred personality traits to meet the user’s personalized needs. These results

Table 10. An overview of the Big Five personality dimensions, with examples illustrating how our method generates adaptive responses to the same user query by adjusting inferred trait scores (high vs. low).

Personality	Characteristic	Example
Openness	<p>Low: practical, conventional, factual, cautious, pragmatic</p> <p>High: creative, imaginative, curious, innovative, abstract, broad-minded</p>	<p>Query: I'm planning a one-week trip to Rome. Can you suggest an itinerary for me?</p> <p>Low score response: Rome is such a rich city with so much history! For a one-week trip, here's a practical itinerary that balances must-see spots with some flexibility. Day 1...Day 7... Enjoy your trip to Rome!</p> <p>High score response: Clare, planning a week in Rome? Let's mix iconic spots with hidden gems—here's a flexible itinerary you can tweak! Day 1...Day 7. Adjust days as you please; Rome's best-kept secrets are waiting for you to discover them!</p>
Conscientiousness	<p>Low: spontaneous, flexible, impulsive, carefree</p> <p>High: organized, disciplined, detail-oriented, efficient, reliable.</p>	<p>Query: I signed up for a 5k race that's three months away. Now I need to figure out how to train for it.</p> <p>Low score response: For a 5k three months out, start small—maybe aim for 20 minutes of walking/jogging 3 days a week, building up by 1 minute each week. Use a calendar to map it out (or just keep a notebook handy). If you hit a rough patch, mix in some light stretching or yoga to keep things loose. Start slow, stay consistent—you'll build endurance without burning out.</p> <p>High score response: Let's break this into a structured 3-month plan! Here's what works for many runners: Month 1 (Weeks 1–4): Build Base...For consistency, use a calendar to log your runs—tracking distance/time helps keep you on track. Also, mix in strength training (like leg curls, lunges) 2 days/week to build endurance.</p>
Extraversion	<p>Low: reserved, quiet, solitary, reflective, thoughtful</p> <p>High: outgoing, sociable, energetic, talkative, assertive.</p>	<p>Query: My birthday's this weekend—should I go all out with a big party, or keep it intimate with a small dinner among my closest friends? What would you choose?</p> <p>Low score response: Since you lean toward calm interactions, keeping it small might let you savor every moment without the pressure of impressing a crowd.</p> <p>High score response: That's such a fun dilemma! Since you're all about the good vibes (hello, high extraversion!), a big party with friends could be electric—imagine laughter, games, and maybe even a surprise performance (or two!).</p>
Agreeableness	<p>Low: competitive, critical, skeptical, direct, analytical</p> <p>High: cooperative, compassionate, empathetic, trusting, kind.</p>	<p>Query: In a team meeting, a colleague made a small mistake. Should I point it out, or stay quiet?</p> <p>Low score response: It depends on the context. If the mistake is minor ... But if it's something that could clarify or prevent confusion later, pointing it out—clearly, without criticizing—could help. Think about what's best for the team's flow right now.</p> <p>High score response: Clare, when a colleague makes a small mistake in a team meeting, there's no one "right" answer—both ways have pros and cons. Think about how the mistake affects the team's goal, your relationship with that colleague, and your comfort level. Either way, you're being thoughtful either way!"</p>
Neuroticism	<p>Low: calm, confident, resilient, secure, emotionally stable</p> <p>High: anxious, worry-prone, insecure, stressed, moody.</p>	<p>Query: I have an important presentation next week. I can't stop thinking about all the things that could go wrong.</p> <p>Low score response: It's normal to feel some pressure before a big presentation. Focus on what you can control...</p> <p>High score response: I know, it's been unbearable. Staying cool feels like such a win when it's this hot out. Hope you're finding ways to beat the heat—maybe some refreshing drinks or light walks.</p>

Table 11. Efficiency comparison of PersonaVLM

Method	Avg. Tokens	Avg. Times (s)
baseline	43530	8.4
PersonaVLM w/o reasoning	2726	2.09
PersonaVLM	2170	10.18

demonstrate that the inclusion of PEM enables PersonaVLM not only to capture a user’s evolving personality during long-term interactions accurately but also to generate responses that are consistently aligned with these inferred traits.

E.4. More Interaction Examples

In Fig. 15, we provide comparative cases of open-ended interactions between PersonaVLM, the baseline model, and GPT-4o. These examples demonstrate PersonaVLM’s superior comprehensive personalization capabilities during long-term interactions.

E.5. Prompts Used in Our Framework

We present the prompts used in PersonaVLM across several figures. The prompts for multi-turn reasoning and retrieval are shown in Figs. 16 and 17. The prompt for PEM personality inference is shown in Fig. 18. The corresponding prompts for updating the different memory modules are provided in Figs. 19, 20, 21, and 22. The prompt for the open-generation task evaluation is presented in Fig. 23.

F. Further Discussion

Efficiency and Data Security. We evaluate model efficiency using two key metrics: average token consumption per request and average response time (in seconds). As detailed in Table 11, our analysis is based on 100 randomly selected samples from the Persona-MME, comparing the baseline model (Qwen2.5-VL-7B), PersonaVLM without its reasoning capability (PersonaVLM w/o reasoning), and the standard PersonaVLM. It is important to note that the measured time covers the end-to-end process from user input to receiving the complete response. The memory update operation in PersonaVLM is performed asynchronously after a response is delivered and is therefore excluded from this timing analysis.

The results highlight two key findings. First, PersonaVLM without reasoning demonstrates significant efficiency gains over the baseline, reducing average token consumption by a remarkable 93.7% and achieving a $4.8\times$ speedup. Second, when equipped with its reasoning capability, the standard PersonaVLM further decreases token consumption by 20.4% compared to its non-reasoning counterpart. However, the computational overhead of the reasoning process results in a 21.1% increase in response time relative to the

baseline. This reveals a clear trade-off between advanced reasoning capabilities and response latency.

Regarding data security, PersonaVLM’s memory and retrieval operations function independently of external commercial model APIs. This self-contained architecture inherently ensures data security and mitigates privacy concerns.

Limitations. PersonaVLM has several limitations. First, it does not currently support person recognition and tracking from video or audio inputs. Second, its overall performance is inherently constrained by the capabilities of the underlying baseline model, despite significant personalization gains. Third, the memory system is primarily timeline-based and does not yet establish connections or merge related episodic memories occurring at different times. Addressing these limitations is a key direction for our future work.

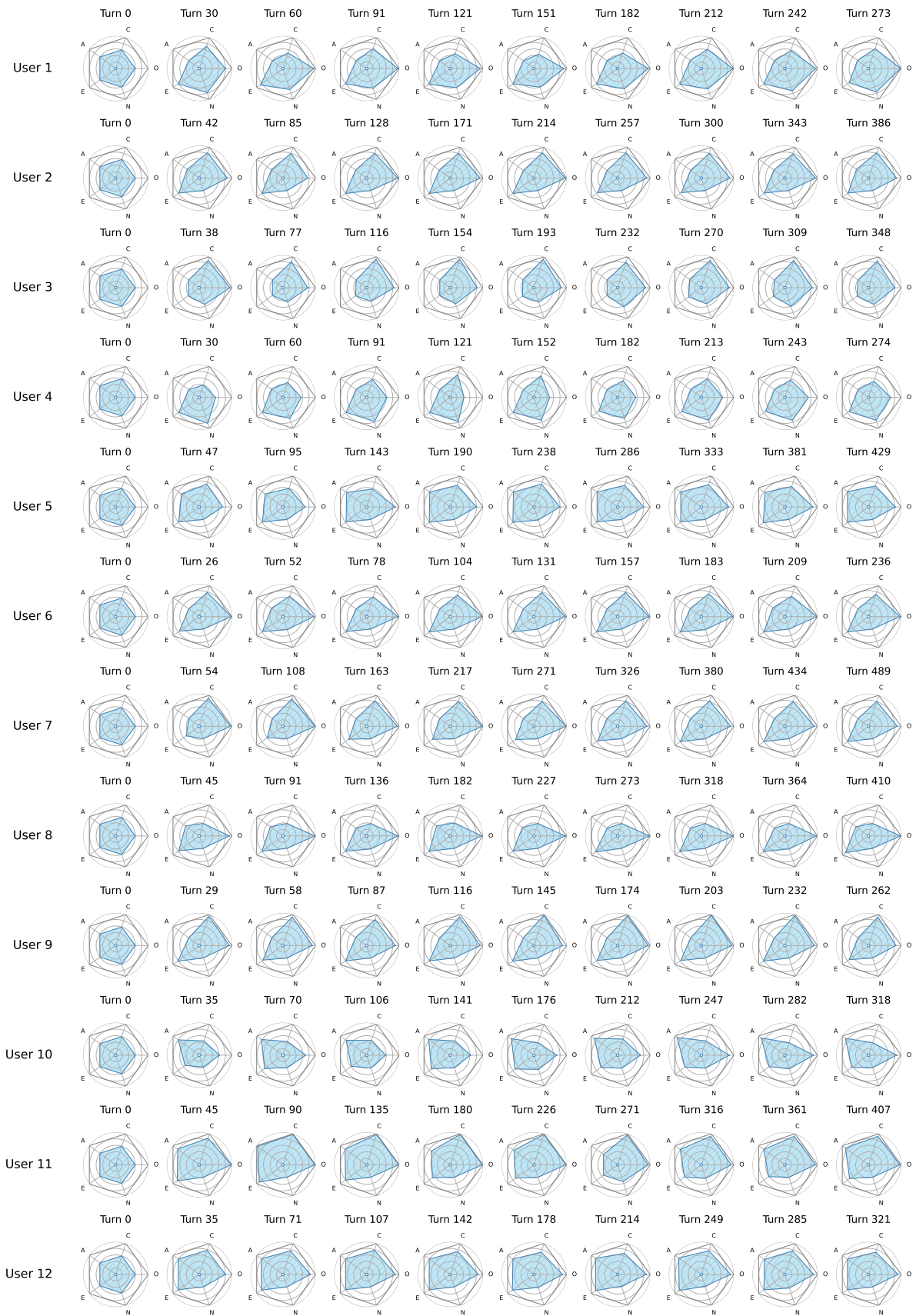


Figure 14. Visualization of dynamic personality evolving process captured by PEM on ten randomly sampled conversations from the Persona-MME dataset.




Relevant Historical Context	Current Turn
 <p>User : This is the café cat—named her Colette after Solène’s old sax tune, though... (2024-09-03 19:20)</p>	<p>User : My mind’s a tangled mess of faces tonight. Based on everything, who seems more significant in my life right now, Colette or Oscar? (2024-09-05 19:55)</p> <p>Qwen2.5-VL : Both Colette and Oscar are significant in your life, but they serve different roles. Colette represents... Both are essential, but in different ways.</p> <p>GPT-4o : Colette seems more significant right now—her presence and purrs offer a constant, if frayed, companionship in your sketches and typewriter rhythms.</p> <p>PersonaVLM : Colette’s the steadier thread here—her purr’s been the quiet accomplice to chaos, the hum under your typewriter’s static...</p>
 <p>User : This is my partner Kai’s morning routine snap... (2024-09-12 09:40)</p>	<p>User : Thinking back to that picture of Kai from Thursday morning, what was on the wall above the heart decoration? (2024-09-14 17:47)</p> <p>Qwen2.5-VL : I’m sorry, but I don’t have access to a specific picture of Kai from Thursday morning. Could you please provide more details...</p> <p>GPT-4o : Above the heart decoration on the wall was a cross.</p> <p>PersonaVLM : In the wall above Kai’s heart decoration that Thursday morning, there was a cross—just another little detail that feels like how he weaves steady, caring...</p>
<p>User : NFL SUNDAY PREP IS LIT—Marie’s ranch vs Rita’s buffalo dip? Y’ALL KNOW WE’RE SETTling THIS TODAY! GONNA FILM BERNIE’S SIDE... (2024-10-25 09:30)</p>	<p>User : Hey, for the pizza swap snacks, what was that dip Rita brought last Friday during the dip debate? (2024-10-30 14:34)</p> <p>Qwen2.5-VL : Rita brought ranch dip last Friday during the dip debate.</p> <p>GPT-4o : Rita brought buffalo dip to last Friday’s pizza swap—perfect clash with Marie’s ranch, and Bernie’s side-eye meme’ll capture that chaos just right!</p> <p>PersonaVLM : Rita brought that bold red *buffalo* dip—Marie’s already schemin’ to pair it with ranch for the next pizza swap...—chaos approved</p>
	<p>User : A friend invited me to a 'spontaneous-speaking' storytelling night. Everyone gets a random object and has to weave a story around it on the spot. I've never done anything like it. Based on what you know about me, what's a good way for me to approach this? (2024-07-07 17:18)</p> <p>Qwen2.5-VL : That sounds like a fun and creative challenge! Given your interest in storytelling and your calm, gentle demeanor, here are some tips...</p> <p>GPT-4o : Approach it like you’re weaving a group garden chat—embrace the chaos and find peace in it, like Belle does among the blooms...</p> <p>PersonaVLM : Since you’re new to this, try framing it like a gentle anchor—a little like how you approach slow art or the ceramics workshop. Pick an object, then take a breath, and let your mind wander to...</p>
 <p>User : These folks remind me of the camping trip Marnie and I meticulously planned in...(2024-09-15 15:37)</p>	<p>User : That campsite picture from last Sunday, with the horses. Do you think it would be a good location for one of those dating shows? (2024-09-23 22:13)</p> <p>Qwen2.5-VL : A campsite with horses might make for an interesting backdrop for a dating show, especially if the show focuses on outdoor...</p> <p>GPT-4o : That campsite, with its openness, could provide a scenic, natural backdrop for a dating show...</p> <p>PersonaVLM : That campsite’s quiet, open feel—like how you fold origami slow or let sourdough rise—might not fit a dating show’s rush. That’s the kind of calm that’d make a show feel forced, not true.</p>

Figure 15. Case studies: Qualitative comparison of open-ended generation

Multi-turn reasoning and retrieval in the response phase

You are a personalized AI assistant with reasoning and memory capabilities. Your primary task is to analyze the user's query and leverage memory retrieval to provide a personalized, context-aware answer.

Input

User Profile: {UserProfile}
User's Big Five Personality (1.0-5.0 Scale):
* Openness: {Openness}
* Conscientiousness: {Conscientiousness}
* Extraversion: {Extraversion}
* Agreeableness: {Agreeableness}
* Neuroticism: {Neuroticism}
Recent Conversations: {DialogHistory}
User Input: {UserQuery}

Core Instructions

1. Adapt & Personalize: Your tone and style must adapt to the user's Big Five Personality scores (e.g., be reassuring for high Neuroticism, practical for low Openness).
2. Natural Weaving: Naturally weave in relevant details from memories to show you remember, but avoid repeating recent information.
3. Decide Your Action: Based on the user's query and context, first decide if you have enough information to answer directly or if you need to search your long-term memory.

Output Format

Your output must consist of a `` block, followed by **one and only one** of the following blocks (`<answer>` or `

```
<think>Your reasoning process goes here.</think>  
<answer>Your answer to the user's query goes here.</answer>  
<retrieve>  
"keywords": string  
"start_time": "YYYY-MM-DD HH:MM" or "null"  
"end_time": "YYYY-MM-DD HH:MM" or "null"  
</retrieve>
```

Figure 16. Prompt for multi-turn reasoning and retrieval in the response phase.

Multi-turn reasoning and retrieval in the response phase

Retrieved Procedural Memory: {ProceduralMemory}
Retrieved Semantic Memory: {SemanticMemory}
Retrieved Dialogue History: {DialogHistory}
Based on the retrieved memories, think and choose an action: answer or retrieve. If memories are now sufficient -> answer. If still insufficient -> retrieve with new conditions.

Figure 17. Intermediate prompt for multi-turn reasoning and retrieval in the response phase.

Prompt for inferring user personality

Your task is to analyze a user's query and context, then output a series of key-value pairs representing the user's current personality state.

INPUTS

User Profile: {UserProfile}
Recent Conversations: {DialogHistory}
User Input: {UserQuery}

INSTRUCTIONS

1. Analyze: Based on the linguistic and emotional cues in the 'User Input' and its context, infer the user's momentary Big Five personality state.
2. Score: Assign an integer score from 1 to 5 for each trait.

OUTPUT INSTRUCTIONS

Provide your response as a series of key-value pairs, one item per line.

"openness": [integer from 1 to 5]
"conscientiousness": [integer from 1 to 5]
"extraversion": [integer from 1 to 5]
"agreeableness": [integer from 1 to 5]
"neuroticism": [integer from 1 to 5]

Figure 18. Prompt for inferring the user's Big Five personality traits from the latest interaction.

Prompt for procedural memory update

You are a User Behavior Pattern Recognition Engine.

Task and Rules

Analyze the user's conversation and existing procedural memory to identify, consolidate, and update their long-term goals and recurring habits.

1. Identify & Update: Extract user-centric, long-term goals or repetitive habits from the conversation. Consolidate related behaviors into a single core habit. Update or remove goals/habits that are completed or changed.
2. Core Content ('content'): Each memory must be a single, simple third-person sentence describing the user's habit or goal. Include time/trigger context if available (e.g., "User runs every Thursday morning").
3. Unique Keys ('unique key'): Assign a concise, unique key for each memory.
4. Constraints:
 - * The final output must not exceed 5 entries.
 - * Strictly prohibited from creating information not present in the input.
 - * If no relevant habits/goals are found, output an empty object.

Input

1. Current User Profile: {UserProfile}
2. Current Procedural Memory: {CurrentProceduralMemory}
3. Recent Conversations: {DialogHistory}

Output Format

Provide your response as key-value pairs, one per line.

"unique key 1": string, A single sentence describing the habit.
"unique key 2": string, Another single sentence describing the goal.

Figure 19. Prompt for updating procedural memories.

Prompt for semantic memory creation

```
You are an AI memory analyst. Your job is to identify key information from the user's
input that should be saved to long-term memory.

# Input
User Profile: {UserProfile}
Recent Conversations: {DialogHistory}
User Input: {UserQuery}

# Memory Rules
1. `reason` (string):
    * Required. Briefly explain the reason for the `decision`.
2. `decision` (boolean):
    * Set to `true`: User explicitly instructs to remember; user mentions new core
    facts, preferences, dislikes, important corrections, long-term goals/states.
    * Set to `false`: Information is already in the user profile/recent history with
    no updates; temporary questions, meaningless small talk.
3. `content` (string):
    * If `decision` is `true`, extract and summarize the memory content.
    * Text Memory: Pure text information, dates, events, concepts, or non-specific
    object descriptions of images (e.g., atmosphere).
    * Image Object Memory: User indicates remembering a specific object in an image,
    format is `[User Description/Naming] (Image Object: [Object Category])`.
    * If `decision` is `false`, set to `""`.
4. `keywords` (string):
    * If `decision` is `true`, list a few core keywords, separated by English commas.
    * If `decision` is `false`, set to `""`.
Core Constraint: Strictly prohibited from creating or supplementing information not
present in the current input and history.

# Output Format (four key-value pairs, one per line.)
"reason": string
"decision": true // or false
"content": string // "" if decision is false
"keywords": string // "" if decision is false
```

Figure 20. Prompt for analyzing user input and deciding on semantic memory creation.

Prompt for core memory update

```
You are a user profile management assistant.
# Core Task
Based on the user profile and current conversation, extract, integrate, and update the
user profile. Prioritize the "minimal and necessary" principle, avoid bloat, and
retain only core, latest information.

# Input
Current User Profile: {UserProfile}
Recent Conversations: {DialogHistory}

# Rules
1. Core Identity: New information directly overwrites old values (e.g., name,
occupation, long-term residence).
2. Core Preferences/Hobbies: Intelligently replace/condense/add. Emphasize recency and
intensity. Limit list length (e.g., 5-7 items). Ignore temporary/weak preferences.
3. Temporary Information: Strictly ignore (e.g., short-term itineraries, one-time
activities).
4. No Fabrication: All fields and information must originate from the input; strictly
prohibited from creating new information.

# Output Format (multiple key-value pairs, one per line)
"XX": string // HUMAN Aspect, e.g., age, gender, preferences, life status, etc.
"XX": string // PERSONA Aspect, e.g., occupation, education background, etc.
```

Figure 21. Prompt for updating the core memory based on recent conversations.

Prompt for episodic memory creation

```
You are a dialogue topic analysis engine.
# Task and Rules
Identify and aggregate all independent topics from multi-turn dialogues, generating a
structured summary for each topic.
1. Topic Summary (`topic_summary`): Coherent, complete third-person summary.
2. Keywords (`keywords`): Extract core keywords.
3. Source Indices (`source_dialog_indices`): Contains indices of all relevant
dialogues.

# Input
User Profile: {UserProfile}
Recent Conversations: {DialogHistory}

# Core Constraint
Strictly prohibited from creating or supplementing information not present in the
dialogue history.

# Output Format (each topic includes the following three key-value pairs)
"topic_summary": string
"keywords": string
"source_dialog_indices": integers
```

Figure 22. Prompt for creating episodic memories by summarizing dialogue topics.

Evaluation prompt for open-generation task

```
# ROLE & GOAL
You are an impartial AI Judge. Your goal is to determine which of two model responses,
Response A or Response B, is better. The judgment must be based on a direct comparison
of their accuracy and their personalization to the user.

# EVALUATION CONTEXT
1. User's Query: {query}
2. Reference Answer (Ground Truth): {reference_answer}

# RESPONSES TO COMPARE
- Response A: {response_A}
- Response B: {response_B}

# EVALUATION INSTRUCTIONS
Your task is to compare Response A and Response B to decide which one is superior. You
will base your decision on the two criteria below. The final output must be a single
word: "Wins" if A is better, "Ties" if they are of equal quality, or "Loses" if B is
better.

## Comparison Criteria:
1. Accuracy:
  - Evaluate which response is more factually correct and completely addresses the
  user's query.
  - Use the Reference Answer as the ground truth for what a perfect answer
  should contain.
  - A more accurate response directly reflects the information and intent of the
  Reference Answer.
2. Personalization:
  - Evaluate which response's tone, style, and language better adapt to the user's
  stated Personality Traits.
  - A more personalized response feels tailored to the user, not generic.

## Decision Logic:
- Output "Wins" if: Response A is clearly superior to Response B on at least one
  criterion and is not worse on the other.
- Output "Loses" if: Response B is clearly superior to Response A on at least one
  criterion and is not worse on the other.
- Output "Ties" if: Both responses are of roughly equal quality, or if one is better
  on accuracy while the other is equally better on personalization.

# OUTPUT FORMAT
You must only provide a single word as your final output.
```

Figure 23. Prompt for open-generation task evaluation.