SMILE: SMALL-LARGE COLLABORATION FOR EFFICIENT LONG-TERM MLLM PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Enabling general-purpose MLLMs to adapt to each user's unique traits, habits, and evolving preferences is essential for building next-generation personalized multimodal assistants. Existing methods for personalizing MLLMs into lifelong human companions often adopt an *ahead-of-time*, query-agnostic context management strategy. They pre-compress and condense interaction histories into memories after each interaction, risking the loss of personalized multimodal evidence required in subsequent user queries. To overcome this limitation, we posit that an MLLM should browse the complete raw interaction history *just-in-time* conditioned on each incoming personalized query, and introduce **SMILE**, a novel **SM**all-large collaboration framework for eff**I**cient **L**ong-t**E**rm MLLM personalization. SMILE breaks "one-model-fits-all" paradigm, specializing a Small MLLM (S-MLLM) as a *lightweight* yet *intelligent* personalized context distiller and employing a Large MLLM (L-MLLM) as a *powerful* but *expensive* personalized context reasoner. Specifically, given a contextualized and idiosyncratic user query, the S-MLLM learns to distill compact personalized context from the *long* raw interaction history, and hand it off to the frozen L-MLLM for reasoning and response generation, delivering efficient and effective customized assistance through collaboration. Experiments on four diverse benchmarks show that SMILE achieves superior performance while handing off averagely about 2.6% of the interaction history to the L-MLLM. Moreover, since SMILE keeps L-MLLM *untouched* during training, it enables nearly "free-lunch" scaling of personalized reasoning at inference time by seamlessly plugging in larger and stronger L-MLLMs.

1 INTRODUCTION

The remarkable perceptual and reasoning capabilities of Multimodal Large Language Models (MLLMs) on general tasks, e.g., visual question answering, are undeniable (Ning et al., 2026). Looking ahead, a natural demand arises: *Can MLLMs serve as intimate assistants that learn our evolving persona from everyday multimodal interactions, and lend us a tailored hand?*

This envision has triggered recent research on MLLM personalization (Alaluf et al., 2024). Most existing studies focused on *single-turn* personalization, where the MLLM visually identifies user-specific concepts from a few reference images (e.g., a user's dog Timi), and answers an *one-shot* query (e.g., image captioning) based on the recognized concept (Nguyen et al., 2024; Hao et al., 2024; Seifi et al., 2026). However, such a static scenario is *less practical* in real-world applications, where user–MLLM interactions are often continuous, contextualized, and extend over prolonged time spans (Nie et al., 2026). Accordingly, we hope an MLLM assistant that, through sustained engagement, learns our (visual) preferences, routines, intentions, and experiences as they unfold over time. Compared with single-turn setting, personalizing MLLM in long-term interactions poses fundamentally **different challenges**: given the current user query, the MLLM should *locate* the relevant personalized context from multimodal interaction histories accumulated over time, and perform personalized *inference* over this context to generate the final response.

Nevertheless, only a few MLLM personalization studies attempt these challenges. They adopt memory-based methods (Hong et al., 2026; Nie et al., 2026; Feng et al., 2026), where each past human-MLLM interaction record is condensed, abstracted, and structured into external memories for future inquires. Since incoming user queries *cannot be considered* during memory management,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

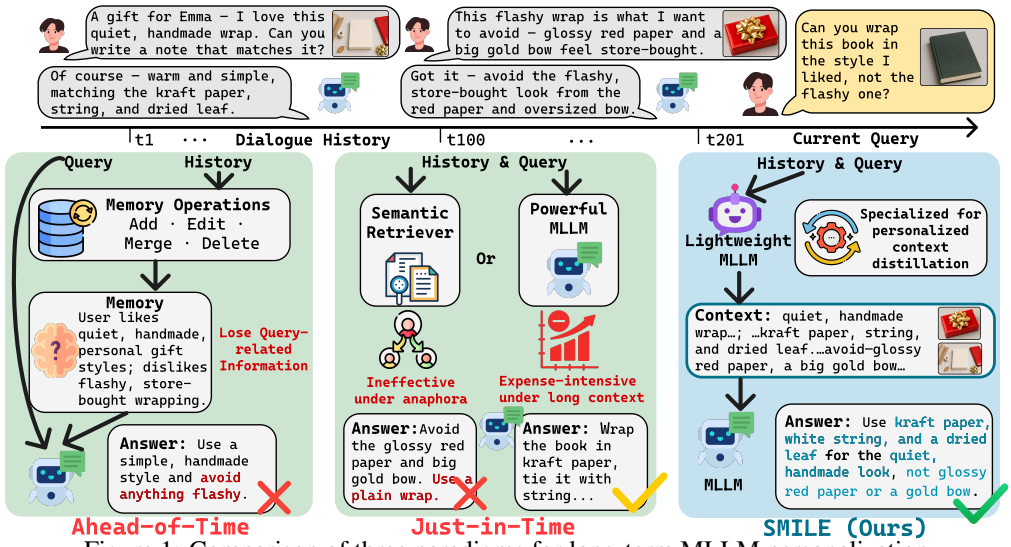


Figure 1: Comparison of three paradigms for long-term MLLM personalization.

such an *ahead-of-time* personalized context management manner inevitably risks *irreversible loss* of user-specific multimodal clues for subsequent interactions (Yang et al., 2026; Ren et al., 2026).

As a remedy, a direct approach is to adopt a *just-in-time* strategy: preserve the entire interaction history, and leverage (1) a semantic retriever (Wang et al., 2020) or (2) an MLLM to identify appropriate context when queried. However, in personalized scenarios, user queries are often context-dependent and contain anaphora (e.g., "recommend clothes that matches my visual style"), rendering semantic-based retrieval less effective (Nie et al., 2026). Although the latter solution (2) goes beyond surface-level semantic retrieval, it is inefficient and less scalable, as it overlooks an *inherent requirements asymmetry* between two procedures in achieving long-term MLLM personalization:

Personalized context localization is a **token-intensive process**, as it requires MLLMs to process increasingly long-term, multimodal interaction histories. In contrast, **personalized inference** is a **reasoning-intensive process**, requiring MLLMs to infer users' latent intents from both located multimodal context and the current query before generating appropriate, customized responses.

To resolve this dilemma, we abandon *one-model-fits-all* fashion, and propose **SMILE**, a novel **SM**all-large collaboration framework for eff**I**cient **L**ong-t**E**rm MLLM personalization (Figure 1). SMILE specializes a small MLLM (S-MLLM) as a **cost-efficient personalized context distiller** over long-term multimodal histories, while reserving a large MLLM (L-MLLM) as a **token-expensive albeit strong personalized context reasoner** over the distilled context, delivering efficient and effective user-specific assistance over sustained interactions. Specifically, we introduce a Persona-to-Context Chain-of-Thought (Wei et al., 2022) (**P2C-CoT**) based supervised fine-tuning stage, which initializes the personalized context localization capability of S-MLLM from locating raw multimodal dialogues to refining them into compact personalized contexts. Thereafter, to improve S-MLLM's generalization to real-world user-specific queries and adapt its distilled contexts for L-MLLM's reasoning, we propose a Context Handoff Policy Optimization (**CHPO**) strategy. By incentivizing S-MLLM in autonomously identifying raw personalized evidence in the open world, and rewarding refined compact contexts that are genuinely useful for L-MLLM's inference, CHPO bridges lightweight personalized context distillation and powerful personalized reasoning.

We conduct extensive experiments on `Persona-MME`, `LCMP`, `PersonaMem-v2`, and our newly constructed `LongPersona-MME` benchmarks, covering concept-centric, user-centric, and ultra-long-history personalization. The results show SMILE as a general solution for tuning MLLMs into reliable lifelong partners, achieving best performance while incurring lightweight context overhead to the L-MLLM. Moreover, as a *stateless* framework, SMILE unlocks near "free-lunch" test-time scaling on the personalized inference side: SMILE can naturally benefit from the remarkable in-context reasoning abilities of larger or stronger proprietary MLLMs (Yao et al., 2026), e.g., GPT-5.6-Sol (OpenAI, 2026) and Gemini-3.1-Pro (Google DeepMind, 2026), by directly plugging them in to offer more capable and user-aligned assistance. We provide both the source code for the framework at `https://anonymous.4open.science/r/SMILE-C50B`.

2 RELATED WORK

2.1 MLLM PERSONALIZATION

Personalizing MLLMs into user-specific assistants is one of the next milestones towards artificial super-intelligence (Lab et al., 2026). MyVLM (Alaluf et al., 2024) and Yo'LLaVA (Nguyen et al., 2024) launch the task of MLLM personalization, where the MLLM is required to identify user-specific concepts from reference images, and replace the generic object categories with the corresponding personalized concepts within a *single-turn lifecycle task* (e.g., image captioning). Subsequent studies, such as RAP (Hao et al., 2024) and R2P (Das et al., 2025), improve the scalability by enabling once-for-all training across concepts or even training-free personalization (Seifi et al., 2025; 2026). Beyond concept identification, later studies shift toward *contextualized personalization*, where user-specific contexts are injected into the model input as *one-time personalization signals* for customized response generation (Xue et al., 2026; Oh et al., 2026). Recently, a few studies begun to explore personalizing MLLM in long-term human-assistant interactions, which is more practical (Hong et al., 2026; Nguyen et al., 2026). TAME and M2A study concept-centric long-context personalization, where interactions are organized around specific user concepts (Hong et al., 2026; Feng et al., 2026), while PersonaVLM focusing on users' long-term daily interactions with MLLM assistants (Nie et al., 2026). Nevertheless, these attempts predominately rely on memory- or semantic-retrieval-based context-management paradigms. As a result, they either lose user-specific clues that may later be useful during query-agnostic memory management, or fail to locate appropriate evidences due to context-dependent, personalized queries (Yang et al., 2026; Nie et al., 2026).

2.2 SMALL AND LARGE MODELS COLLABORATION.

Owing to their distinct strengths, with small models (SMs) honoring cost-efficient and large models (LMs) exhibiting stronger reasoning abilities and broader knowledge coverage, many studies have explored collaboration between them (Chen et al., 2025b; Wang et al., 2025a). Existing studies can be divided into two groups: (1) **knowledge distillation based methods**, which transfer the knowledge and reasoning capabilities of LMs into SMs through fine-tuning (Lang et al., 2026), and (2) **division-of-labor based methods**, which assign SMs and LMs to different parts of the same task, like *(i) resource-limited conditions* (Jiang et al., 2023) (e.g., speculative decoding (Leviathan et al., 2023; Oliaro et al., 2025), confidence-based model routing (Yue et al., 2024; Zhang et al., 2026a)), or *(ii) privacy-sensitive settings*, e.g., local SMs handling sensitive information while remote LMs provide assistance (Zhang et al., 2024a; Lv et al., 2026). In MLLM personalization, SLC adopts this mechanism to single-turn scenarios, where the S-MLLM and L-MLLM collaborate only to identify personalized concepts (Yang et al., 2025). In contrast, our SMILE targets long-term scenarios, where the complementary strengths of both small and large MLLMs are fully unleashed, achieving efficient yet effective user-tailored assistance over continued human-MLLM interactions.

The related work about long-context MLLMs and long-term personalization is in Appendix A.

3 METHODOLOGY: SMILE

3.1 PRELIMINARY

Formalization of Long-Term MLLM Personalization. We consider a general real-world scenario: at time step T, a user interacts with their MLLM assistant Φ_{MLLM} using a multimodal query $\mathcal{Q}_T = \{q_t, q_v\}$, where q_t and q_v denote the textual and visual information in the query, respectively, e.g., a question such as "Does this outfit match my style?" paired with an uploaded image (Nie et al., 2026). As a personalized assistant, Φ_{MLLM} is required to reason over both the current query $\mathcal{Q}_T$ and the relevant user-specific contexts from past interactions, and generate the response:

$$\mathcal{R}_T = \Phi_{\text{MLLM}}(\mathcal{Q}_T, \pi(\mathcal{H}_{T-1})), \tag{1}$$

where $\mathcal{H}_{T-1}$ is the raw past human-MLLM interaction histories, which can reflect the user's persona, preferences, habits, experiences (e.g., past outfit choices). $\pi(\cdot)$ is a context management mechanism, which is expected to produce appropriate personalized clues for the current query. After the current interaction, both $\mathcal{Q}_T$ and the response $\mathcal{R}_T$ will be incorporated to the history: $\mathcal{H}_{T-1} \rightarrow \mathcal{H}_T$, allowing the history to accumulate over time for long-horizon personalization in future interactions.

Ahead-of-Time vs. Just-in-Time Context Management. Existing studies mainly adopt memory-based context management for long-term MLLM personalization (Hong et al., 2026; Nguyen et al., 2026; Nie et al., 2026). Specifically, after each human-MLLM interaction, these methods will *condense, abstract, and structure* this record into one or more type of memories, incrementally updated the memories through add, delete, and edit operations with a potential maximum capacity or not. Therefore, after step $T-1$, the context management is executed *without considering query* $\mathcal{Q}_T$:

$$\pi(\mathcal{H}_{T-1}) \triangleq \mathcal{O}(\mathcal{M}_{T-2}, \langle \mathcal{Q}_{T-1}, \mathcal{R}_{T-1} \rangle) = \mathcal{M}_{T-1}, \quad \mathcal{R}_T = \Phi_{\text{MLLM}}(\mathcal{Q}_T, \mathcal{M}_{T-1})\ ^{1} \tag{2}$$

where $\mathcal{O}(\cdot)$ is the memory management operation, and $\mathcal{M}_{T-1}$ is either fed into Φ_{MLLM} (Hong et al., 2026) or coupled with a retriever for further selection (Nie et al., 2026). As a result, such a *ahead-of-time management strategy* pre-narrows the MLLM's contextual view, inevitably risking losing personalized clues that may later become useful in future interactions (Yang et al., 2026).

In contrast, a *just-in-time context management mechanism* fetches useful contexts directly from the raw history $\mathcal{H}_{T-1}$ whenever a query arrives, i.e., $\pi(\mathcal{H}_{T-1}) \triangleq \mathcal{J}(\mathcal{H}_{T-1}, \mathcal{Q}_T)$, where $\mathcal{J}(\cdot)$ is a context location method, alleviating the above drawback. Nevertheless, realizing an efficient and effective $\mathcal{J}(\cdot)$ is non-trivial: a semantic-based retriever may lead to inappropriate localization of personalized evidence, as user queries in this setting are often idiosyncratic and context-dependent, while leveraging an MLLM to read the entire multimodal history $\mathcal{H}$ incurs substantial overhead and may still be suboptimal *without explicit specialization* for identifying user-specific contexts.

Recipe of our SMILE. To resolve the above dilemma, we break the *one-model-fits-all* regime and propose SMILE, a novel small-large MLLM collaboration fashion (Figure 2). Within this paradigm, a lightweight and efficient Small MLLM (S-MLLM) is specialized for personalized context localization through: ❶ a Persona-to-Context Chain-of-Thought (**P2C-CoT**) based SFT stage (**Section 3.2**) and ❷ a Context Handoff Policy Optimization (**CHPO**) stage (**Section 3.3**), learning to progressively identify appropriate raw histories and prune them into compact contexts that serve as effective handoffs to the Large MLLM (L-MLLM). Thereafter, the stronger L-MLLM, which excels at in-context reasoning but incurs high token and overhead costs, only takes the compact context and the current query as input for capable personalized inference and response generation (**Section 3.4**).

3.2 Persona-to-Context Chain-of-Thought Supervised Finetuning

Although the S-MLLM can replace semantic retrievers for more intelligent context distillation under personalized scenarios, its limited model capacity still calls for dedicated training to become competent. To start with, we propose a Persona-to-Context CoT (P2C-CoT) based SFT stage. It initializes the capability of S-MLLM to first locate relevant interactions from the entire multimodal history ("a needle in a haystack") and prunes them to compact context by filtering out irrelevant content.

P2C-CoT SFT Data Construction. Since specializing S-MLLM as a lightweight personalized context distiller for long-term MLLM personalization has not been explored in existing studies, we therefore introduce a dedicated data construction recipe for acquiring this capability. Specifically, we consider a general and common long-term MLLM personalization dataset (e.g., `Persona-MME` (Nie et al., 2026)), which consists of multiple personas (e.g., *a travel enthusiast*). For simplicity, we focus on a specific persona, which is associated with a set of prolonged human–MLLM interaction histories (records) $\mathcal{H} = \{\langle \mathcal{Q}_i, \mathcal{R}_i \rangle\}_{i=1}^{|\mathcal{H}|}$ [2], along with several personalized query–answer (QA) training pairs $\mathcal{P} = \{\langle \mathcal{Q}_i, \mathcal{A}_i \rangle\}_{i=1}^{|\mathcal{P}|}$ that require leveraging relevant user-specific contexts from $\mathcal{H}$ for response generation. Based on this foundation, we employ a large MLLM (e.g., GPT-5.6-Sol (OpenAI, 2026)) to serve as an automatic annotator (Lin et al., 2025). ❶ It first retrospectively identifies the **oracle historical interactions** (i.e., interactions that are genuinely necessary to answer a given personalized query) for each $\mathcal{Q}_i \in \mathcal{P}$ by grounding the oracle answer $\mathcal{A}_i$: $\hat{h}_i = \Phi_{\text{MLLM}}(\langle \mathcal{Q}_i, \mathcal{A}_i \rangle, \mathcal{H}, \mathcal{I}_{\text{identify}})$, where $\mathcal{I}_{\text{identify}}$ is the instruction used to guide the identification of the oracle interaction set $\hat{h}_i$ ❷ Subsequently, conditioning on both the QA pair $\langle \mathcal{Q}_i, \mathcal{A}_i \rangle$ and $\hat{h}_i$, we prompt the annotator to derive a **compact personalized context** $\hat{\mathcal{C}}_i$ that filters out clearly redundant, irrelevant or corrupted information from the identified raw histories: $\hat{\mathcal{C}}_i = \Phi_{\text{MLLM}}(\langle \mathcal{Q}_i, \mathcal{A}_i \rangle, \hat{h}_i, \mathcal{I}_{\text{filter}})$, where $\mathcal{I}_{\text{filter}}$ is the prompt for delivering the compact context.

[1] For efficiency, $\mathcal{M}_{T-1}$ is derived from $\mathcal{M}_{T-2}$ and $\langle \mathcal{Q}_{T-1}, \mathcal{R}_{T-1} \rangle$, rather than recomputed from $\mathcal{H}_{T-1}$.

[2] We omit the time step subscript T of the past histories $\mathcal{H}$ in this section for simplicity.

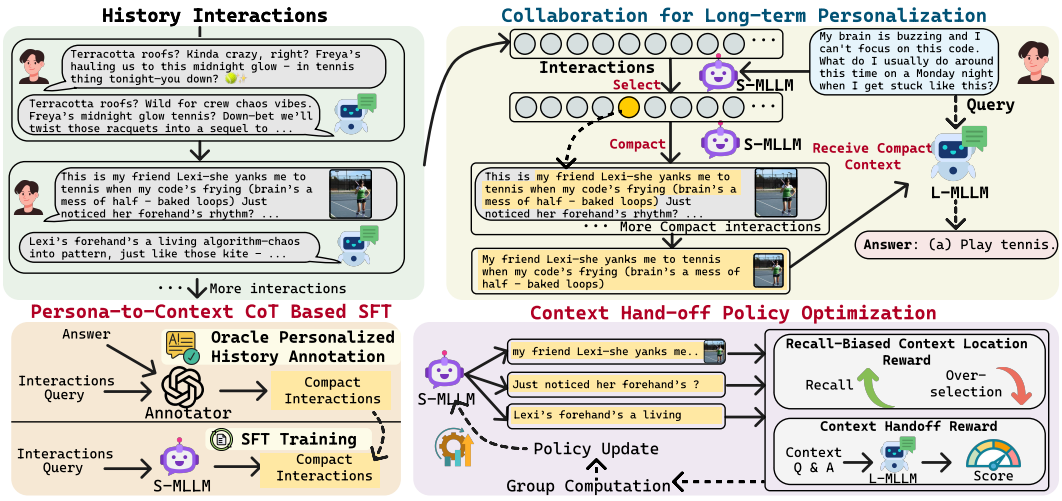


Figure 2: Overview training pipeline of SMILE, where the P2C-CoT based SFT and CHPO stages specify the S-MLLM to collaborate with the L-MLLM for effective and efficient personalization.

Notably, given the limited capacity of S-MLLMs for complex abstraction over located multimodal histories, we strictly constrain the annotator to perform only *non-essential content pruning* over the identified histories, rather than regenerating any personalized contexts from scratch, thereby keeping the specialization within the intended capability scope of the S-MLLM (e.g., pruning polite fillers such as "Sure, I can help with that." or corrupted spans such as "@@@###"). ❸ Thereafter, we construct each **SFT training sample** $\mathcal{S}_i = (\mathbf{x}_i, \mathbf{y}_i)$ that follows a two-step P2C-CoT process, where the input $\mathbf{x}_i$ mainly consists of the query $\mathcal{Q}_i$ and the whole interaction history $\mathcal{H}$, and the ground-truth output $\mathbf{y}_i$ follows two-step chain, incorporating the ground-truth interactions $\hat{h}_i$ and compact context. After batching all queries in all personas, we construct P2C-CoT SFT training data $\mathcal{D}_{\text{SFT}}$. More details about the training data construction and curation process are provided in Appendix D.3.

We then perform P2C-CoT SFT on the S-MLLM using $\mathcal{D}_{\text{SFT}}$, facilitating its initial persona-query conditioned context location capability by emulating the collected annotated trajectories:

$$L_{\text{SFT}} = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_{\text{SFT}}} \sum_{t=1}^{|\mathbf{y}|} \log \Phi_{\text{S-MLLM}}\big(y_t \mid \mathbf{x}, y_{<t}\big). \tag{3}$$

3.3 Context Handoff Policy Optimization

Since the SFT stage only teaches the S-MLLM to imitate offline annotated context location trajectories, we further introduce Context Handoff Policy Optimization (**CHPO**). Based on group relative policy optimization (GRPO) (Shao et al., 2024), it better shapes the S-MLLM toward real-world collaborative behavior with the L-MLLM. Specifically, each CHPO training sample follows a structure similar to that of SFT, and the optimization objective is formulated as follows:

$$L_{\text{CHPO}} = -\,\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_{\text{CHPO}},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{x})} \Bigg[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min(r_{i,t}(\theta)A_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}) - \eta D_{\text{KL}}[\pi_\theta \| \pi_{\theta_{\text{ref}}}]\Bigg]. \tag{4}$$

where π_θ is the policy of S-MLLM. The probability ratio $r_{i,t}(\theta)$ and advantage $A_{i,t}$ are defined as:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid \mathbf{x}, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{x}, o_{i,<t})}, \quad A_{i,t} = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})}. \tag{5}$$

The reward is designed mainly from the following two perspectives:

Recall-Biased Context Location Reward. To encourage the S-MLLM to autonomously locate appropriate personalized contexts in the open world, we introduce a recall-biased context location reward. Motivated by the principle of erring on the side of caution, it incentivizes the S-MLLM to identify all necessary historical interactions while *allowing slight over-selection*, thereby improving robustness in fetching persona clues under complex multimodal human-assistant interactions.

Specifically, this reward can be formularized as:

$$R_{\text{location}} = \underbrace{\frac{(1+\beta^2)\left|h_i \cap \hat{h}_i\right|}{\beta^2\left|\hat{h}_i\right| + |h_i|}}_{\text{recall-biased location}} - \underbrace{\lambda\left(\frac{\max\left(0, (|h_i| - \left|\hat{h}_i\right|) - \tau\right)}{\left|\hat{h}_i\right|}\right)^2}_{\text{over-selection penalty}}, \tag{6}$$

where β is the F-score trade-off parameter that controls the relative importance of recall over precision, h_i and $\hat{h}_i$ are the located and ground-truth sets of historical interactions for sample $\mathcal{S}_i$, respectively, τ is a tolerance threshold for early over-selection, and λ is a balancing factor. *Intuitively*, mild over-selection of histories is tolerated under a soft budget but *still not encouraged*, while excessive over-selection is *heavily penalized* beyond the τ, preventing the S-MLLM from hacking the reward by locating an unnecessarily large number of interactions to inflate the recall signal.

Context Handoff Reward. Since the ultimate mission of the S-MLLM is to transfer user-specific contexts to the L-MLLM for generating desirable personalized responses, we propose the context handoff reward. It positions a frozen L-MLLM as a **context-to-answer interpreter**, which uses the quality of the final personalized response as a "backward feedback" to guide context distillation, directly connecting personalized context localization and personalized inference for better collaborative handoff training. Specifically, since personalized assistance takes different task formats across scenarios (e.g., multiple-choice or free-form QA), the evaluation for the L-MLLM-based interpreter can be instantiated accordingly. We therefore present its implementation under multiple-choice QA, the widely adopted question format in current long-term MLLM personalization datasets:

$$R_{\text{handoff}} = \mathbf{1}[p_{\text{margin}} > 0] + \gamma \cdot p_{\text{margin}}, \tag{7}$$

$$p_{\text{margin}} = p_{\Phi_{\text{L-MLLM}}}(d_i^\star \mid \mathcal{Q}_i, \mathcal{C}_i) - \max_{d \neq d_i^\star} p_{\Phi_{\text{L-MLLM}}}(d \mid \mathcal{Q}_i, \mathcal{C}_i). \tag{8}$$

where $\mathcal{Q}_i$ is the personalized query in sample $\mathcal{S}_i$, $\mathcal{C}_i$ is the corresponding compact context from S-MLLM and subsequently handed off to the L-MLLM for response generation, p_{margin} measures the probability gap between the ground-truth option $d_i^\star$ and the second $d \neq d_i^\star$, providing *fine-grained* discrimination of different context handoff trajectories within the same group with respect to their effectiveness in inducing the final response, and γ is the balancing factor. Combining the above two rewards, the final reward can be formalized as follows:

$$R_{\text{final}} = R_{\text{location}} + \alpha \cdot R_{\text{handoff}}. \tag{9}$$

For other forms of personalized tasks, such as free-form QA, the R_{handoff} can be implemented using an MLLM-as-Judge paradigm (Andrade et al., 2026). Since the frozen L-MLLM serves only as interpreter, a larger L-MLLM is expected to better reduce the confounding effect of its own reasoning bias for a more reliable assessment of context utility. Nevertheless, we empirically observe that a modest-scale L-MLLM (i.e., 9B model) is sufficient to achieve stable evaluation performance. *In practice*, to further isolate handoff quality from the L-MLLM's own reasoning limitations, we filter out samples that it fails to answer even with the ground-truth context and apply R_{handoff} only to the remaining instances. Furthermore, since the L-MLLM is not directly optimized, we find that a "novel" L-MLLM can be easily integrated in a plug-and-play manner without additional adaptation.

3.4 Collaboration for Long-Term Personalization

Finally, we orchestrate the specialized S-MLLM with the strong L-MLLM for collaborative long-term personalization. Specifically, at time step T, given a user query $\mathcal{Q}_T$, the response is delivered:

$$\mathcal{R}_T = \Phi_{\text{L-MLLM}}(\mathcal{I}_{\text{infer}}, \mathcal{Q}_T, \Phi_{\text{S-MLLM}}(\mathcal{I}_{\text{locate}}, \mathcal{Q}_T, \mathcal{H}_{T-1})), \tag{10}$$

where $\mathcal{I}_{\text{infer}}$ and $\mathcal{I}_{\text{locate}}$ are the prompts for L-MLLM and S-MLLM in personalized inference and context location, respectively. *Notably*, only the compact context $\mathcal{C}_T$ (without the raw interactions) is handed off to the L-MLLM. Since the L-MLLM is neither trained nor exposed to the original prolonged multimodal interaction history, stronger frontier L-MLLMs (e.g., GPT-5.6-Sol (OpenAI, 2026)) can be integrated into SMILE alongside the already specialized S-MLLM in a plug-and-play manner, yielding near *free-lunch* gains in scaling personalized inference capability. Moreover, since SMILE does not rely on complex memory structures, it can be adapted to diverse multimodal personalization tasks, while eliminating continual memory maintenance also *reduces system overhead*.

The algorithm-style training recipe of our proposed SMILE is detailed in Appendix B, and all prompts presented throughout this section are provided in Appendix J.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

A brief experimental setup is provided in this section, with a more detailed one in Appendix D.

Benchmarks. Since SMILE imposes *minimal* assumptions on human-MLLM interaction histories, we evaluate it on four diverse benchmarks: ❶ **Persona-MME** (Nie et al., 2026): a user-centric personalization benchmark with over 2,000 evaluation cases derived from 200 personas, evaluated under historical multimodal interaction contexts ranging in length from 32K to 128K, and covering 7 evaluation dimensions in form of multiple-choice question (MCQ): memory, intent, preference, behavior, relationship, growth, and alignment. ❷ **LCMP** (Hong et al., 2026): a concept-centric personalization benchmark comprising 30 concepts (e.g., pet, object) with 720 images, 420 historical dialogues, 420 evaluation questions (both MCQ and free-text questions), with evolving attributes and attribute conflicts across interactions for each concept. ❸ **PersonaMem-v2** (Jiang et al., 2025b): an implicit user-persona personalization benchmark featuring 1,000 user personas, over 300 scenarios, more than 20,000 implicit user preferences, multimodal multilingual interaction histories with context lengths spanning 32K to 128K tokens, and 5,000 evaluation cases (both MCQ and open-ended questions). We adopt its **multimodal subset** for evaluation. ❹ Our newly constructed **LongPersona-MME**: an extended benchmark built upon `Persona-MME` with doubled interaction context length, scaling from 128K to 256K tokens. For `Persona-MME` and `LongPersona-MME`, we mainly adopt the Accuracy (**ACC**). For `LCMP`, we adopt Accuracy of Choice questions (**ACC-C**), Accuracy of Free-text questions (**ACC-F**), Scoring Point Rate-Long-term attribute (**SPR-L**), and SPR-Short-term attribute (**SPR-S**) according to different evaluation question types. For `PersonaMem-v2`, we report ACC-MCQ (**ACC-M**) and ACC-Open-ended (**ACC-O**) under MCQ and open-ended evaluation types, respectively. We also introduce context ratio (**CR**) as the proportion of the full interaction history passed to the L-MLLM.

Baselines. Following (Nie et al., 2026), we first adopt two common long-context MLLM baselines: ① **Full** and ② **RAG**. For the Full baseline, we provide entire interaction history to the MLLM. For baseline RAG, following (Jiang et al., 2025a), we retrieve the top-5 most relevant historical records for each question. We also include dataset-fitted baselines according to the benchmark setting: ③ **PersonaVLM** (Nie et al., 2026) for user-centric benchmarks `Persona-MME`, `PersonaMem-v2`, and `LongPersona-MME`; ④ **TAME** (Hong et al., 2026) for the concept-centric `LCMP`.

Implementation Details. We adopt `Qwen3.5-4B` (Qwen Team, 2026) as the S-MLLM across most experiments, and pair it with L-MLLMs of varying scales and families (open-source and proprietary MLLMs). The L-MLLM `Qwen3.5-9B` keeps frozen, and during CHPO, it serves as the interpreter, and only the S-MLLM is optimized via LoRA (Hu et al., 2022). The S-MLLM is specialized *once* per benchmark (except LongPersona-MME, which even directly generalizes from Persona-MME *without* additional specialization). For PersonaVLM, which requires expensive full fine-tuning of the L-MLLM, we evaluate only the 9B model comparable in scale to its original paper.

4.2 MAIN PERFORMANCE

We report the performance of all methods across four personalization benchmarks in Tables 1 and 2. We observe that **(1)** both **Full** and **RAG** generally outperform the raw MLLM (no personalized context provided), demonstrating the effectiveness of incorporating personalized multimodal evidence. Nevertheless, RAG remains limited by surface-level semantic retrieval when handling highly personalized, context-dependent queries involving complex anaphora. In contrast, Full incurs substantial computational costs by feeding the entire interaction history into the MLLM, while the increasing amount of irrelevant information in long histories further distract the MLLM and constrain response quality. **(2)** For user-centric benchmarks, **PersonaVLM** also improves the personalization capabilities of the raw MLLM. However, it adopts an ahead-of-time, query-agnostic memory-based context management strategy, which may discard user-specific evidence required by subsequent queries. Moreover, although an MLLM guides multi-round semantic retrieval, its lack of direct access to the raw interaction history for query expansion prevents it from resolving context-dependent queries. **(3)** On the concept-centric benchmark, **TAME** also adopts similar query-agnostic context management strategies coupled with semantic retrieval, which limits the ability to provide reliable long-term personalized assistance. **(4)** SMILE achieves the best performance across all benchmarks

Table 1: Performance on Persona-MME and LongPersona-MME. Best performance is in **bold**.

Persona-MME	Memory	Intent	Preference	Behavior	Relationship	Growth	Alignment	Overall
Qwen3.5-9B	37.76	57.69	46.40	43.96	50.72	64.41	66.08	54.29
+ RAG	79.04	80.19	56.90	55.27	75.53	73.52	84.72	73.74
+ Full	83.07	79.37	64.62	60.86	**79.59**	78.93	84.81	76.95
+ PersonaVLM	68.77	78.27	59.37	61.23	73.43	79.73	87.27	74.52
+ **SMILE**	**86.92**	**86.12**	**68.16**	**65.69**	78.22	**85.77**	**94.91**	**82.82**
Qwen3.5-27B	36.02	59.29	53.27	52.20	57.97	71.70	66.19	57.82
+ RAG	79.28	83.03	56.89	61.59	79.83	80.61	83.00	75.54
+ Full	90.28	86.92	**68.70**	66.20	81.82	85.74	93.45	83.41
+ **SMILE**	**90.97**	**88.06**	68.06	**69.38**	**82.67**	**86.97**	**93.68**	**84.20**
GPT-5.6-Sol	44.03	55.63	56.47	62.17	52.39	66.40	55.69	55.90
+ RAG	82.00	78.23	61.05	67.21	75.12	75.19	66.91	70.94
+ Full	**93.06**	80.86	66.05	70.81	**85.12**	75.32	73.87	76.82
+ **SMILE**	92.32	**83.70**	**72.91**	**71.52**	80.74	**83.91**	**93.62**	**84.28**
LongPersona-MME	**Memory**	**Intent**	**Preference**	**Behavior**	**Relationship**	**Growth**	**Alignment**	**Overall**
Qwen3.5-9B	14.29	10.34	12.39	9.62	11.93	15.25	60.08	26.12
+ RAG	58.82	41.38	47.79	44.23	67.89	56.78	69.35	58.34
+ Full	63.03	44.83	51.33	50.00	79.82	50.00	82.66	64.67
+ PersonaVLM	49.58	36.21	45.13	44.23	46.79	44.07	75.81	53.86
+ **SMILE**	**82.35**	**51.72**	**61.06**	**64.42**	**84.40**	**62.71**	**85.08**	**73.76**
Qwen3.5-27B	23.53	20.69	23.89	18.27	27.52	24.58	63.31	34.75
+ RAG	60.50	48.28	51.33	55.77	77.98	59.32	77.42	64.79
+ Full	75.83	56.72	62.39	64.17	85.38	59.75	**85.66**	73.23
+ **SMILE**	**77.31**	**56.90**	**66.37**	**75.00**	**88.07**	**66.10**	83.87	**75.95**
GPT-5.6-Sol	41.18	34.48	34.51	34.62	41.28	38.98	70.56	47.18
+ RAG	66.39	53.45	57.52	61.54	78.90	62.71	77.42	68.01
+ Full	77.03	**55.24**	67.74	**69.31**	80.84	64.11	75.72	71.79
+ **SMILE**	**84.03**	53.45	**68.14**	68.27	**88.07**	**71.19**	**83.47**	**76.64**

Table 2: Performance on PersonaMem-v2 and LCMP-H (hard set). Best performance is in **bold**.

Method	PersonaMem-v2-32K		PersonaMem-v2-128K		LCMP-H			
	ACC-M	ACC-O	ACC-M	ACC-O	ACC-C	ACC-F	SPR-L	SPR-S
Qwen3.5-9B	31.40	28.08	31.30	27.86	69.44	47.22	25.00	40.00
+ RAG	34.30	31.33	33.56	30.15	82.22	76.11	70.56	71.11
+ Full	34.42	27.60	31.38	25.54	76.11	83.33	82.78	82.22
+ PersonaVLM	33.74	26.85	33.62	27.43	—	—	—	—
+ TAME	—	—	—	—	86.11	56.11	64.44	69.44
+ **SMILE**	**38.92**	**37.52**	**37.16**	**35.39**	**91.11**	**87.78**	**87.78**	**86.11**
Qwen3.5-27B	35.02	30.57	35.10	29.57	80.00	66.67	47.78	60.00
+ RAG	39.28	33.18	38.16	32.87	88.33	87.22	80.00	82.78
+ Full	38.30	27.86	33.24	25.61	88.89	88.33	87.78	87.22
+ TAME	—	—	—	—	86.67	64.44	68.33	68.33
+ **SMILE**	**49.58**	**42.32**	**47.16**	**38.77**	**92.78**	**92.00**	**93.33**	**87.78**
GPT-5.6-Sol	36.67	50.71	36.67	50.71	75.00	60.56	39.44	62.22
+ RAG	42.33	53.53	41.60	52.89	85.56	90.56	80.00	84.44
+ Full	46.07	58.26	40.07	53.79	92.78	**93.33**	**88.89**	86.11
+ TAME	—	—	—	—	83.33	76.67	70.56	78.89
+ **SMILE**	**51.27**	**60.75**	**48.87**	**58.19**	**93.89**	**93.33**	86.11	**88.67**

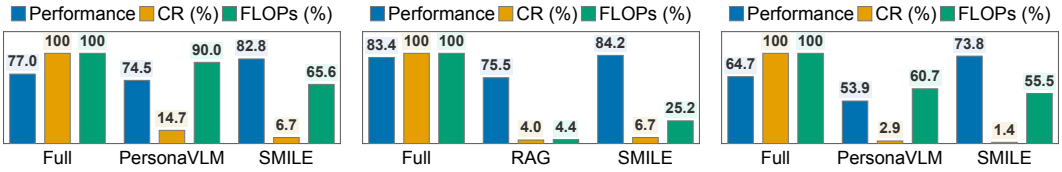
by addressing the limitations of existing methods with a more intelligent, specialized, yet efficient 4B S-MLLM that serves as a "personalized context reader" over the raw interaction histories. Moreover, SMILE exhibits a *test-time scaling trend* in personalization capability: pairing with stronger L-MLLMs (e.g., 27B or `GPT-5.6-Sol` (OpenAI, 2026)) improves performance without retraining the S-MLLM, which is already trained only once with a frozen modest scale L-MLLM.

4.3 Further Analysis

Ablation Study. As reported in Table 3, we conduct an ablation study of SMILE using a sequence of variants, with the two CHPO rewards evaluated individually (+ Reward) and jointly (+ SMILE) after SFT. We observe that, without specialization, the *raw S-MLLM* struggles to accurately locate personalized evidence and construct an effective context handoff for the L-MLLM. In contrast, equipping the L-MLLM with P2C-CoT SFT-optimized S-MLLM context distiller yields clear performance

Table 3: Ablation study of training stages and rewards within SMILE.

Ablative Variant	Persona-MME	LongPersona-MME	PersonaMem-v2-32K	PersonaMem-v2-128K	LCMP-H	CR (%)
Qwen3.5-9B	54.29	26.12	31.40	31.30	69.44	**0**
+ Raw S-MLLM	67.35	56.39	34.02	33.04	83.89	4.14
+ P2C-CoT SFT	80.89	67.20	37.92	36.08	88.89	2.58
+ Location Reward	80.39	67.32	38.06	37.06	90.00	2.59
+ Handoff Reward	78.46	69.74	37.98	36.40	90.56	2.59
+ Both (**SMILE**)	**82.82**	**73.76**	**38.92**	**37.16**	**91.11**	2.58



(a) Persona-MME (9B). (b) Persona-MME (27B). (c) LongPersona-MME (9B).

Figure 3: Efficiency comparison between methods. FLOPs are reported relative to Full (100%)

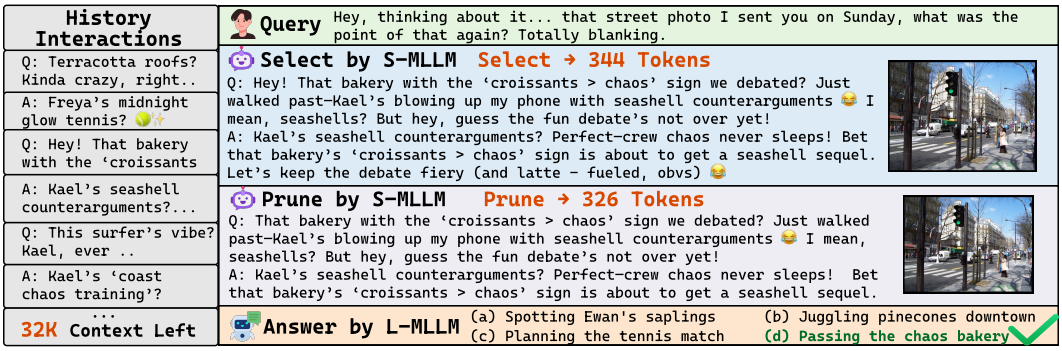


Figure 4: Case study of our small-large collaboration for personalized multimodal assistance.

gains. Incorporating both the two core rewards during CHPO further improves personalized evidence identification under real-world queries and the pruning quality of compact contexts, yielding better performance while maintaining a stable L-MLLM context ratio (CR).

Efficiency Analysis for Long-Term Personalized Interaction. We provide efficiency comparison in Figure 3, where FLOPs of PersonaVLM include its amortized memory-management overhead for each question. Under moderate history length and L-MLLM scale, SMILE outperforms Full with only 65.6% of its inference FLOPs (Li et al., 2025b) and exposing the resource (e.g., token) expensive L-MLLM to 14–15$\times$ less context. As either the L-MLLM scale or history length increases, SMILE keeps best while exhibiting favorable computational scaling, with its relative FLOPs decreasing, as SMILE **decouples** two computational roles in long-term MLLM personalization: growing histories are handled by light S-MLLM, while larger L-MLLM operates only on compact one.

Case Study on Collaborative Personalization As illustrated in Figure 4, we provide a case study of SMILE for delivering personalized assistance, where the case is selected from Persona-MME. The user query, the "street photo I sent you", is context-dependent and personalized, as no explicit *street* related clues appear in the historical interactions. Nevertheless, the specialized S-MLLM identifies the relevant user-specific, compact context from the raw history while preserving key details (e.g., "croissants chaos") without pre-compressing, enabling the L-MLLM to perform comprehensive personalized reasoning for accurate response. More cases, including baselines, are in Appendix H.

5 Conclusion

In this study, we propose SMILE, a novel small-large collaboration framework for efficient and effective long-term MLLM personalization. Experiments on four benchmarks show the *strongest personalization performance* of SMILE. We further validate that SMILE incurs the modest overhead, as S-MLLM can hand off only averagely about 2.6% of the *high-quality* history to the expensive L-MLLM. We empirically find that SMILE enables favorable scaling of personalization capability, benefiting from stronger L-MLLMs in a *plug-and-play* manner to provide more adaptive service.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, including improving grammar, clarity, and presentation without changing the underlying technical content. We also used generative AI tools for generating synthetic datasets, as described in the paper. We did not use generative AI tools for literature retrieval or discovery, research ideation or execution, drafting substantive sections of the paper, or proving mathematical claims. Other disclosure categories are not applicable to this work. All AI-assisted outputs were reviewed and verified by the authors for correctness, consistency, and faithfulness to the intended technical content and data construction procedure. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human-subject research. SMILE is developed to improve the efficiency and effectiveness of long-term multimodal personalization, rather than to support harmful surveillance, manipulation, or other malicious uses of personalized systems. This study does not investigate discrimination, bias, fairness, privacy, or security issues, and we do not make claims regarding these aspects.

REPRODUCIBILITY STATEMENT

We present the design rationale of SMILE in the main paper, with implementation and training details provided in Appendix D.5 and the construction details of the training collections for this collaborative framework described in Appendix D.3. We also release the code of SMILE to facilitate reproducibility. Given the substantial scale and associated hosting requirements of the training collections for SMILE and our newly introduced LongPersona-MME benchmark, we plan to release these resources upon acceptance.

REFERENCES

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision (ECCV)*, pp. 73–91. Springer, 2024.

Moises Andrade, Joonhyuk Cha, Brandon Ho, Vriksha Srihari, Karmesh Yadav, and Zsolt Kira. Let's Think in Two Steps: Mitigating Agreement Bias in MLLMs with Self-Grounded Verification. In *International Conference on Learning Representations (ICLR)*, 2026.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025a.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Seng Nam Chen, Hao Chen, Chenglam Ho, Xinyu Mao, Jinping Wang, Yu Zhang, and Chao Li. Seeing the scene matters: Revealing forgetting in video understanding models with a scene-aware long-video benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 4515–4525, 2026.

Yi Chen, Jiahao Zhao, and Haohao Han. A Survey on Collaborative Mechanisms Between Large and Small Language Models. *arXiv.org*, abs/2505.07460, 2025b.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *European Conference on Artificial Intelligence (ECAI)*, pp. 2993–3000, 2025. doi: 10.3233/FAIA251160.

Ziqiang Cui, Yunpeng Weng, Xing Tang, Peiyang Liu, Shiwei Li, Bowei He, Jiamin Chen, Yansen Zhang, Xiuqiang He, Rui Zhang, and Chen Ma. Less is more: Elevating RAG via performance-driven context compression. In *International Conference on Machine Learning (ICML)*, 2026.

Deepayan Das, Davide Talon, Yiming Wang, Massimiliano Mancini, and Elisa Ricci. Training-free personalization via retrieval and reasoning on fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Haoran Deng, Yingyu Lin, Zhenghao Lin, Xiao Liu, Yizhou Sun, Yi-An Ma, and Yeyun Gong. Beyond length: Quantifying long-range information for long-context LLM pretraining data. In *International Conference on Learning Representations (ICLR)*, 2026.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending LLM context window beyond 2 million tokens. In *International Conference on Machine Learning (ICML)*, pp. 11091–11104, 2024.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *arXiv.org*, abs/2406.20094, 2024a.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *International Conference on Learning Representations (ICLR)*, 2024b.

Google DeepMind. Gemini 3.1 Pro model card, February 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*, 2024.

Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. Rap: Retrieval-Augmented Personalization for Multimodal Large Language Models. *arXiv e-prints*, pp. arXiv:2410.13360, 2024.

Rongpei Hong, Jian Lang, Yong Wang, Ting Zhong, and Fan Zhou. Tameing long contexts in personalization: Towards training-free and state-aware mllm personalized assistant. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. EverMemOS: A self-organizing memory operating system for structured long-horizon reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 45836–45853, 2026. doi: 10.18653/v1/2026.acl-long.2125.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022.

Junlong Jia, Ziyang Chen, Xing Wu, Chaochen Gao, Zijia Lin, Songlin Hu, and Binghui Guo. Entropylong: Effective long-context training via predictive uncertainty. In *International Conference on Learning Representations (ICLR)*, 2026.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, C. J. Taylor, and Dan Roth. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale. In *Conference on Language Modeling (COLM)*, 2025a. URL https://openreview.net/forum?id=6ox8XZGOqP.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. Personamem-v2: Towards Personalized Intelligence via Learning Implicit User Personas and Agentic Memory. *arXiv.org*, abs/2512.06688, 2025b.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, volume abs/2310.06839, 2023.

Greg Kamradt. Needle in a haystack: Pressure testing LLMs. GitHub repository, 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A. Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. ACON: Optimizing context compression for long-horizon LLM agents. In *International Conference on Machine Learning (ICML)*, 2026.

Seoyeon Kim and Jaehyung Kim. SPRInG: Continual LLM personalization via selective parametric adaptation and retrieval-interpolated generation. *arXiv.org*, abs/2601.09974, 2026.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. In search of needles in a 11m haystack: Recurrent memory finds what LLMs miss. *arXiv preprint arXiv:2402.10790*, 2024.

Mind Lab, :, Vin Bo, Song Cao, Vic Cao, Andrew Chen, Kaijie Chen, Cleon Cheng, Steven Chiang, Kaixuan Fan, Hera Feng, Huan Feng, Arthur Fu, Jun Gao, Hongquan Gu, Aaron Guan, Nolan Ho, Mutian Hong, Hailee Hou, Peixuan Hua, Charles Huang, Miles Jiang, Nora Jiang, Yuyi Jiang, Qiuyu Jin, Fancy Kong, Andrew Lei, Kyrie Lei, Alexy Li, Lucian Li, Ray Li, Theo Li, Wenhao Li, Zhihui Li, Allen Lin, Jiayi Lin, Kairus Liu, Kieran Liu, Logan Liu, Xiang Liu, Irvine Lu, Maeve Luo, Runze Lv, Pony Ma, Verity Niu, Anson Qiu, Vincent Wang, Rio Yang, Maxwell Yao, Carrie Ye, Regis Ye, Wenlin Ye, Josh Ying, Danney Zeng, Yuhan Zhan, Anya Zhang, Di Zhang, Ruijia Zhang, Shiyang Zhang, Sueky Zhang, Ya Zhang, Wei Zhao, Ada Zhou, Adrian Zhou, Yuhua Zhou, Xinyue Zhu, and Murphy Zhuang. On the Scaling of PEFT: Towards Million Personal Models of Trillion Parameters. *arXiv*, 2026.

Jian Lang, Rongpei Hong, Meihui Zhong, Kaiju Li, Ting Zhong, Qiang Gao, and Fan Zhou. Leaf: Towards lightweight explainable hateful video detection via self-grounding cot guided stage-wise distillation. In *Findings of the Association for Computational Linguistics (ACL)*, 2026.

Jiaqi Leng, Xiang Hu, Junxiong Wang, Jianguo Li, Wei Wu, and Yucheng Lu. Understanding and improving length generalization in hierarchical sparse attention models. In *International Conference on Learning Representations (ICLR)*, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning (ICML)*, pp. 19274–19286. PMLR, 2023.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! LLM-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5259–5276, 2025a. doi: 10.18653/v1/2025.naacl-long.272.

Jinlong Li, Liyuan Jiang, Haonan Zhang, and Nicu Sebe. Token reduction via local and global contexts optimization for efficient video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 10451–10461, 2026a.

Kevin Li, Sachin Goyal, Joao D Semedo, and Zico Kolter. Inference optimal vlms need fewer visual tokens and more parameters. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 96066–96083, 2025b.

Mo Li, L. H. Xu, Qitai Tan, Long Ma, Flood Sung, Ting Cao, and Yunxin Liu. Sculptor: Empowering LLMs with cognitive agency via active context management. In *International Conference on Learning Representations (ICLR)*, 2026b.

Lequan Lin, Dai Shi, Andi Han, Feng Chen, Qiuzheng Chen, Jiawen Li, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, and Junbin Gao. Act as Human: Multimodal Large Language Model Data Annotation with Critical Thinking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Anmin Liu, Ruixuan Yang, Huiqiang Jiang, Bin Lin, Minmin Sun, Yong Li, Chen Zhang, and Tao Xie. Vecattention: Vector-wise sparse attention for accelerating long context inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 41299–41310, 2026.

Fanfan Liu and Haibo Qiu. Context cascade compression: Exploring the upper limits of text compression. *arXiv preprint arXiv:2511.15244*, 2025.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025a.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. A survey of personalized large language models: Progress and future directions. *arXiv.org*, abs/2502.11528, 2025b.

Hang Lv, Sheng Liang, Hao Wang, Yongyue Zhang, Hongchao Gu, Wei Guo, Defu Lian, Yong Liu, and Enhong Chen. Specsteer: Synergizing local context and global reasoning for efficient personalized generation. *arXiv preprint arXiv:2603.16219*, 2026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'LLaVA: Your Personalized Language and Vision Assistant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Viet-Anh Nguyen, Thao Nguyen, Vishal M. Patel, and Yuheng Li. Personal Visual Memory from Explicit and Implicit Evidence. *arXiv*, 2026.

Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. Personavlm: Long-term personalized multimodal llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026. URL http://arxiv.org/abs/2604.13074.

Shan Ning, Longtian Qiu, and Xuming He. Wiki-R1: Incentivizing Multimodal Reasoning for Knowledge-based VQA via Data and Sampling Curriculum. *International Conference on Learning Representations (ICLR)*, 2026.

Yeongtak Oh, Sangwon Yu, Junsung Park, Han Cheol Moon, Jisoo Mok, and Sungroh Yoon. Contextualized Visual Personalization in Vision-Language Models. In *International Conference on Machine Learning (ICML)*, 2026.

Chimaobi Okite, Anika Misra, Joyce Chai, and Rada Mihalcea. LUCid: Redefining relevance for lifelong personalization. *arXiv.org*, abs/2604.26996, 2026.

Gabriele Oliaro, Zhihao Jia, Daniel Campos, and Aurick Qiao. Suffixdecoding: Extreme speculative decoding for emerging ai applications. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:126326–126354, 2025.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8631–8661, 2025. doi: 10.18653/v1/2025.naacl-long.435.

OpenAI. Previewing GPT-5.6-Sol: a next-generation model. `https://openai.com/index/previewing-gpt-5-6-sol/`, 2026.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, Zhengnan Li, Yangqiu Song, Ginny Wong, and Simon See. Memlens: Benchmarking multimodal long-term memory in large vision-language models. *arXiv preprint arXiv:2605.14906*, 2026.

Soroush Seifi, Vaggelis Dorovatas, Daniel Olmeda Reino, and Rahaf Aljundi. Personalization Toolkit: Training Free Personalization of Large Vision Language Models. *arXiv.org*, abs/2502.02452, 2025.

Soroush Seifi, Simon Gardier, Vaggelis Dorovatas, Daniel Olmeda Reino, and Rahaf Aljundi. Ego: Embedding-Guided Personalization of Vision-Language Models. *arXiv*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv.org*, abs/2402.03300, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 29935–29951, 2026. doi: 10.18653/v1/2026.findings-acl.1496.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8416–8439, 2025. doi: 10.18653/v1/2025.acl-long.413.

Unsplash. The unsplash dataset (lite). `https://github.com/unsplash/datasets`, 2020.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*, 2024.

Fali Wang, Jihai Chen, Shuhua Yang, Ali Al-Lawati, Li-Juan Tang, Hui Liu, and Suhang Wang. A Survey on Collaborating Small and Large Language Models for Performance, Cost-effectiveness, Cloud-edge Privacy, and Trustworthiness. *arXiv.org*, abs/2510.13890, 2025a.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3221–3241, 2025b. doi: 10.18653/v1/2025.naacl-long.166.

Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. Loongrl: Reinforcement learning for advanced reasoning over long contexts. In *International Conference on Learning Representations (ICLR)*, 2026a.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. Needle in a multimodal haystack. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 20540–20565, 2024a.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 74530–74543, 2023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:5776–5788, 2020.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. MEMORYLLM: Towards self-updatable large language models. In *International Conference on Machine Learning (ICML)*, pp. 50453–50466, 2024b.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. MMLongBench: Benchmarking long-context vision-language models effectively and thoroughly. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025c.

Zhaowei Wang, Lishu Luo, Haodong Duan, Weiwei Liu, Sijin Wu, Ji Luo, Shen Yan, Shuai Peng, Sihang Yuan, Chaoyi Huang, Yi Lin, and Yangqiu Song. Training long-context vision-language models effectively with generalization beyond 128k context. *arXiv preprint arXiv:2605.13831*, 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. *arXiv preprint arXiv:2106.01040*, 2021.

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

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context LLM inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024.

Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. AlpsBench: An LLM personalization benchmark for real-dialogue memorization and preference alignment. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 3489–3497, 2026.

Siheng Xiong, Joe Zou, Faramarz Fekri, and Yae Jee Cho. Long-context modeling with dynamic hierarchical sparse attention for memory-constrained LLM inference. In *International Conference on Machine Learning (ICML)*, 2026.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024. doi: 10.18653/v1/2024.naacl-long.260.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Zihui Xue, Ami Baid, Sanghoe Kim, Minrui Luo, and Kristen Grauman. Personal Visual Context Learning in Large Multimodal Models. *arXiv*, 2026.

Cheng Yang, Zequn Sun, Wei Wei, and Wei Hu. Beyond Static Summarization: Proactive Memory Extraction for LLM Agents. *arXiv.org*, abs/2601.04463, 2026.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, K. Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhen-Ru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 Technical Report. *arXiv.org*, abs/2412.15115, 2024.

Sihan Yang, Huitong Ji, Shaolin Lu, Jiayi Chen, Binxiao Xu, Ming Lu, Yuanxing Zhang, Wenhui Dong, and Wentao Zhang. Small-Large Collaboration: Training-efficient Concept Personalization for Large VLM using a Meta Personalized Small VLM. *arXiv.org*, abs/2508.07260, 2025.

Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Shili Xiong, Yi Rong, Yufei Wu, Shichao Su, Yifan Xu, Wenxi Zeng, Zhaoyu Yang, Guoyou Li, Shilan Zhang, Zichan Li, Yaxiong Chen, Shengwu Xiong, Peng Xu, Jiajun Zhang, Bowen Zhou, David A. Clifton, and Luc Van Gool. Lens: Multi-level Evaluation of Multimodal Reasoning with Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2026.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. Compact: Compressing retrieved documents actively for question answering. *arXiv preprint arXiv:2407.09014*, 2024.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 21691–21728, 2024.

Chuang Zhang, Zizhen Zhu, Yi Wei, Bing Tian, Junyi Liu, He Wang, Wang Xavier, and Yaxiao Liu. Confidence-Calibrated Small-Large Language Model Collaboration for Cost-Efficient Reasoning. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 4480–4501, 2026a.

Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. Cogenesis: A Framework Collaborating Large and Small Language Models for Secure Context-Aware Instruction Following. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4295–4312, 2024a.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. Adacomp: Extractive context compression with adaptive predictor for retrieval-augmented large language models. *arXiv preprint arXiv:2409.01579*, 2024b.

Yingyi Zhang, Junyi Li, Wenlin Zhang, Penyue Jia, Xianneng Li, Yichao Wang, Derong Xu, Yi Wen, Huifeng Guo, Yong Liu, and Xiangyu Zhao. Evoking user memory: Personalizing LLM via recollection-familiarity adaptive retrieval. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://openreview.net/forum?id=f7p0F2X6XN.

Hongbo Zhao, Meng Wang, Fei Zhu, Wenzhuo Liu, Bolin Ni, Fanhu Zeng, Gaofeng Meng, and Zhaoxiang Zhang. VTCBench: Can vision-language models understand long context with vision-text compression? *arXiv preprint arXiv:2512.15649*, 2025a.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=QWunLKbBGF.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025c.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing Large Language Models with Long-Term Memory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 19724–19731, 2024.

Technical Appendix

TABLE OF CONTENTS

A Additional Related Work **19**
A.1 Long-Context Studies for LLM and MLLM 19
A.2 Long-Term LLM Personalization 19

B Algorithm-Style Training Recipe of SMILE **20**

C Complexity Analysis **20**

D Detailed Experimental Setup **22**
D.1 Baseline Description & Implementation 22
D.2 Benchmark Description & Evaluation 22
D.3 Detailed Process of Training Collection Construction and Curation 24
D.4 Details of our LongPersona-MME 25
D.5 Implementation Details 25

E Additional Empirical Evaluations and Cases **26**
E.1 Quantitative Evaluation of Our Training Collection 26
E.2 Core Hyper-Parameter Analysis 26
E.3 Evaluation on Smaller S-MLLM 26
E.4 Additional Results of Efficiency 27

F Discussion: Prompt, Retrieval, Memory Methods and SMILE **27**

G Limitation & Future Works **28**

H Additional Case Studies **28**
H.1 Failure Cases 28

I Training Sample Presentation **32**

J Prompts Presentation **34**
J.1 Prompts for SMILE 34
J.2 Prompts for Training Collection Construction 36

A Additional Related Work

A.1 Long-Context Studies for LLM and MLLM

The remarkable success of large language models (LLMs) and multimodal large language models (MLLMs) has been accompanied by a persistent challenge: the scalability of their context window (Liu et al., 2025a; Chen et al., 2025a; Wang et al., 2026b; Deng et al., 2026; Jia et al., 2026). Broadly, existing approaches to addressing the challenges of processing extremely long contexts in LLMs and MLLMs can be divided into two categories. **(1) Architecture-based methods** improve long-context efficiency through mechanisms such as sparse or hierarchical attention, while also extending the effective sequence length through techniques like positional encoding extrapolation (Beltagy et al., 2020; Wu et al., 2021; Xiao et al., 2024; Press et al., 2022; Su et al., 2024; Xiong et al., 2024; Ding et al., 2024; Leng et al., 2026; Xiong et al., 2026; Liu et al., 2026). In parallel, **(2) external-component-based methods** employ auxiliary modules or workflow-level augmentation to manage and compress the input context (Liu & Qiu, 2025; Li et al., 2026b;a). Representative directions include: **(i) prompt compression**, which reduces input length through selective token retention or context summarization so that the essential information fits within a fixed context window (Zhang et al., 2024b; Ge et al., 2024b; Yoon et al., 2024; Kang et al., 2026; Cui et al., 2026). **(ii) Retrieval- and memory-based methods,** which store conversational or document histories in external memory modules and retrieve relevant evidence to support coherent long-range interactions (Günther et al., 2024; Wang et al., 2023; 2024b; Zhong et al., 2024; Chhikara et al., 2025; Ren et al., 2026). Nevertheless, most existing studies focus on **general long-context tasks**, where ahead-of-time context compression is relatively less harmful because the target queries typically require task-relevant facts rather than fine-grained user-specific details (Hsieh et al., 2024; Kuratov et al., 2024; Vodrahalli et al., 2024; Wang et al., 2024a; 2025b; Maharana et al., 2024; Wu et al., 2025; Wang et al., 2025c; Zhao et al., 2025a; Wang et al., 2026a; Chen et al., 2026). In addition, such queries often exhibit relatively explicit semantic correspondence with the required evidence, making conventional semantic retrieval reasonably effective (Kamradt, 2023; Hsieh et al., 2024; Wang et al., 2024a; 2025b). Long-term MLLM personalization presents a fundamentally different setting: fine-grained user-specific, multimodal clues are essential for producing highly tailored responses, while user queries are frequently implicit, context-dependent, and anaphoric (Nie et al., 2026). Prematurely summarizing, compressing, or discarding the interaction history may therefore remove details that only become relevant to future queries, whereas surface-level semantic retrieval may fail to recover the necessary evidence. To address these challenges, our SMILE specializes an S-MLLM to perform just-in-time, query-conditioned, and more intelligent personalized context localization and conservative compression directly over the raw multimodal interaction history. This design avoids personalized information loss caused by ahead-of-time context reduction, while ensuring that the L-MLLM consistently reasons over a compact, high-density context, thereby satisfying the central long-context objective of minimizing the burden placed on the expensive models.

A.2 Long-Term LLM Personalization

Although long-term MLLM personalization, which remains relatively underexplored, has been summarized in the main paper, this section complements the discussion by reviewing the broader literature on long-term LLM personalization (Liu et al., 2025b). Generally, we group existing approaches to long-term LLM personalization, often studied under the setting of long-term conversations, into two categories. The first category, namely **prompt-based long-context approaches**, represents an early and straightforward paradigm that directly places accumulated user interaction histories into the context window for personalization (Zhao et al., 2025b;c; Jiang et al., 2025a;b; Okite et al., 2026; Xiao et al., 2026). However, as interaction histories continue to grow, these approaches are constrained by finite context windows, incur substantial computational overhead, and remain vulnerable to needle-in-a-haystack failures, where sparse yet critical personalized evidence is obscured by large amounts of irrelevant context (Zhao et al., 2025b; Jiang et al., 2025a;b; Okite et al., 2026; Xiao et al., 2026). The second line of work consists of **retrieval- and memory-based methods**, which enhance long-term personalized interactions by enabling LLM-based systems to retain, update, and retrieve user-specific information from past conversations (Jiang et al., 2025b). Representative approaches include LD-Agent (Li et al., 2025a), RMM (Tan et al., 2025), MemoryBank (Zhong et al., 2024), THEANINE (Ong et al., 2025), Mem0 (Chhikara et al., 2025), SPRInG (Kim & Kim, 2026), and A-MEM (Xu et al., 2025), which organize interaction histories into persistent memory structures

and selectively retrieve relevant evidence for subsequent responses. Recent studies further advance this paradigm from complementary perspectives: several studies (like EverMemOS (Hu et al., 2026) and TSM (Su et al., 2026)) focus on improving the construction, organization, and maintenance of long-term memories, while others (like RF-Mem (Zhang et al., 2026b)) develop more effective retrieval strategies to identify and recover user-specific information relevant to the current interaction.

Nevertheless, retrieval- and memory-based long-term LLM personalization methods share limitations similar to those of current long-term MLLM personalization approaches: **(1) Ahead-of-time context management** through dedicated memory mechanisms risks discarding fine-grained personalized evidence that may become essential in subsequent human–assistant interactions. This limitation is particularly consequential for personalization, which depends on preserving *rich user-specific details to support highly tailored responses*. It is further amplified in the multimodal setting, where pre-compressing or dropping visual content may eliminate substantially more fine-grained personalized evidence. **(2) Semantic retrieval** primarily captures surface-level relevance, whereas personalized queries are often highly context-dependent and contain implicit or anaphoric references, making the required evidence difficult to retrieve through similarity matching alone. Increasing the number of retrieved candidates may improve recall, but also introduces substantial irrelevant context, ultimately reverting to the noise and computational burden of prompt-based long context approaches. In contrast, as detailed in the main paper, SMILE adopts a just-in-time context management strategy and replaces the conventional one-model paradigm with a small–large collaboratiion framework, avoiding the substantial cost of having a single MLLM inspect complete raw multimodal interaction histories. Specifically, a Small MLLM (S-MLLM) is specialized as an intelligent personalized context distiller that directly reads the raw history and performs high-quality, query-conditioned localization and conservative pruning of user-specific multimodal evidence. A Large MLLM (L-MLLM), kept entirely frozen throughout training, is positioned as a stronger personalized context reasoner. By reasoning over *only the compact context* handed off by the S-MLLM, the L-MLLM avoids the practical context-window bottleneck and needle-in-a-haystack interference of prompt-based approaches, thereby delivering more capable personalized assistance. Moreover, since the L-MLLM neither participates in parameter optimization nor processes the full interaction history, SMILE maintains high computational efficiency.

B Algorithm-Style Training Recipe of SMILE

In this section, we provide the detailed training recipe of SMILE in algorithm style in Algorithm 1

C Complexity Analysis

In this section, we delve into the complexity analysis of SMILE. To start with, since the backbone of SMILE is MLLM, we first provide a definition of the complexity of each MLLM invocation:

Definition C.1 (Complexity of an MLLM Invocation) *We consider a single invocation of an MLLM consisting of H multimodal Transformer layers, each with hidden dimension d, operating on an input sequence X with length l, the self-attention operation is defined as:*

$$\mathrm{Attention}(X) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{11}$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$, with $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$. For simplicity, we assume standard multi-head self-attention with shared sequence length for queries, keys, and values. Under this setting, the computational complexity of one MLLM invocation is: $O\big(H \cdot (ld^2 + l^2 d)\big)$, where ld^2 corresponds to the linear projections for query, key, and value representations, and $l^2 d$ arises from the self-attention computation. However, H is often a small constant, and we therefore simplify the complexity to $O(ld^2 + l^2 d)$ in the following discussion.

Following Definition C.1, we analyze the computational complexity of SMILE for a single personalized interaction by decomposing it into two stages: personalized context localization with the S-MLLM and personalized reasoning with the L-MLLM.

Algorithm 1 Training Recipe of the Proposed SMILE

Require: Long-term personalization dataset $\mathcal{D}$, S-MLLM p_{θ_0}, frozen L-MLLM Φ_{L}, MLLM annotator Φ_{A}, group size G
Ensure: Specialized S-MLLM p_θ

/ * **Stage I: Persona-to-Context CoT (P2C-CoT) Supervised Fine-Tuning** * /
Initialize $\mathcal{D}_{\mathrm{SFT}} \leftarrow \varnothing$
for all personalized QA samples $(\mathcal{H}, \mathcal{Q}_i, \mathcal{A}_i) \in \mathcal{D}$ **do**
    Identify ground-truth historical interactions $\hat{h}_i$ using Φ_{A}
    Construct compact context $\mathcal{C}_i^\star$ by pruning non-essential content from $\hat{h}_i$
    Construct the P2C-CoT trajectory $\mathbf{y}_i = (\hat{h}_i \rightarrow \mathcal{C}_i^\star)$
    Add $(\mathbf{x}_i, \mathbf{y}_i)$ to $\mathcal{D}_{\mathrm{SFT}}$, where $\mathbf{x}_i = (\mathcal{H}, \mathcal{Q}_i)$
end for
Optimize p_{θ_0} on $\mathcal{D}_{\mathrm{SFT}}$ using the SFT objective in Eq. (3)
Obtain the initialized policy $p_{\theta_{\mathrm{SFT}}}$
/ * **Stage II: Context Handoff Policy Optimization (CHPO)** * /
Initialize $p_\theta, p_{\theta_{\mathrm{old}}}, p_{\theta_{\mathrm{ref}}} \leftarrow p_{\theta_{\mathrm{SFT}}}$, with $p_{\theta_{\mathrm{ref}}}$ frozen
for step $= 1, \ldots, M$ **do**
    Sample a training batch $\mathcal{B}$
    for all $\mathbf{x} \in \mathcal{B}$ **do**
        Sample G context-handoff trajectories $\{o_g\}_{g=1}^{G} \sim p_{\theta_{\mathrm{old}}}(\cdot \mid \mathbf{x})$
        for all $o_g = (h_g, \mathcal{C}_g)$ **do**
            Compute R^g_{location} using Eq. (6)
            Evaluate whether Φ_{L} can answer $\mathcal{Q}$ using the ground-truth context $\mathcal{C}^\star$
            Compute R^g_{handoff} using Eq. (7) only if answerable. Otherwise set it to zero
            Compute the final reward R_g using Eq. (9)
        end for
        Compute group-relative advantages using Eq. (5)
    end for
    Update p_θ using the CHPO objective in Eq. (4)
    $p_{\theta_{\mathrm{old}}} \leftarrow p_\theta$
end for
return p_θ

Definition C.2 (Per-Interaction Notation of SMILE) *At time step T, SMILE receives the current multimodal query $\mathcal{Q}_T$ and accesses the accumulated raw human–MLLM interaction history $\mathcal{H}_{T-1}$. The specialized S-MLLM processes $\mathcal{Q}_T$ and $\mathcal{H}_{T-1}$ to produce a compact personalized context $\mathcal{C}_T$, which is subsequently handed off to the frozen L-MLLM for generating the personalized response $\mathcal{R}_T$. Let l_H, l_Q, and l_C denote the multimodal token lengths of $\mathcal{H}_{T-1}$, $\mathcal{Q}_T$, and $\mathcal{C}_T$, respectively. We further denote the hidden dimensions of the S-MLLM and L-MLLM by d_S and d_L, respectively.*

S-MLLM-based Personalized Context Localization. For each interaction, the S-MLLM jointly processes the current query $\mathcal{Q}_T$ and the complete raw interaction history $\mathcal{H}_{T-1}$ to locate relevant historical interactions and construct the compact context $\mathcal{C}_T$. Ignoring the fixed-length localization instruction $\mathcal{I}_{\mathrm{locate}}$, the input sequence length of this stage is $l_S = l_H + l_Q$. Following Definition C.1, the corresponding computational complexity is $O(l_S d_S^2 + l_S^2 d_S)$. Although this cost grows with the accumulated history length, the complete multimodal history is processed only by the lightweight S-MLLM. Consequently, the history-dependent long-context computation is performed with $d_S \ll d_L$, rather than by the substantially larger and more expensive downstream reasoner.

L-MLLM-based Personalized Reasoning. After context localization, the frozen L-MLLM processes only the current query $\mathcal{Q}_T$ and the compact context $\mathcal{C}_T$. Ignoring the fixed-length inference instruction $\mathcal{I}_{\mathrm{infer}}$, its input length is $l_L = l_C + l_Q$, resulting in a computational complexity of $O(l_L d_L^2 + l_L^2 d_L)$ according to Definition C.1. By reducing the historical context processed by the L-MLLM from l_H to l_C, where $l_C \ll l_H$, SMILE substantially reduces both its linear computation and quadratic self-attention cost. Specifically, letting $\rho = l_C / l_H$ and omitting the comparatively short query, the L-MLLM complexity becomes $O(\rho l_H d_L^2 + \rho^2 l_H^2 d_L)$.

SMILE does not eliminate the history-dependent cost of accessing the continuously growing raw interaction history. Instead, it delegates this computation to the *substantially smaller* S-MLLM while preventing the expensive L-MLLM from repeatedly processing the complete multimodal history. When the handed-off context $\mathcal{C}_T$ is maintained within a compact budget, the computational cost of L-MLLM-based personalized reasoning remains bounded with respect to the growth of $\mathcal{H}_{T-1}$.

D Detailed Experimental Setup

D.1 Baseline Description & Implementation

Baseline Description. We provide a detailed description of each baseline, including **PersonaVLM** and **TAME**. Baselines Full and RAG are common methods and have no specific descriptions.

- **PersonaVLM** (Nie et al., 2026) is a long-term personalized multimodal agent that organizes user information into a Big-Five personality profile and four complementary memory types: core memory for foundational user attributes, semantic memory for abstract user knowledge and multimodal concepts, episodic memory for timestamped interaction events, and procedural memory for plans, goals, and recurring behaviors. During response generation, the model performs multi-step reasoning and timeline-aware retrieval over these memories. After each interaction, it proactively updates the memory database and progressively evolves the personality profile to keep subsequent responses consistent with the user's changing characteristics.
- **TAME** (Hong et al., 2026) is a training-free, state-aware personalization framework that maintains a Dynamic State Memory for temporally changing concept attributes and a Static Personalized Memory for persistent attributes. It is a concept-centric personalized MLLM assistant. Its memory-transition mechanism moves stable information into the static memory while removing outdated short-term states, and its Retrieve-then-Align Augmented Generation paradigm first retrieves concept-specific evidence from both memories and then aligns the retrieved evidence with the current query to produce a contextually appropriate personalized response.

Baseline Implementation. We strictly follow its official implementation and configurations. The only modification is backbone MLLM, which we unify with that of SMILE to ensure fairness. For PersonaVLM, we adopt and perform its training recipe on the L-MLLM (9B).

D.2 Benchmark Description & Evaluation

Benchmark Evaluation. We follow the official evaluation protocols of all three benchmarks and use the metric abbreviations reported in the main paper.

- **Evaluation in Persona-MME.** Persona-MME (Nie et al., 2026) organizes 14 fine-grained question types into seven dimensions. **Memory** covers Visual Detail Recall, which recalls fine-grained visual details from previously shared images, and Semantic Information Recall, which retains semantic information such as events, preferences, and context from the conversation history. **Intent** covers Explicit Intent Inference, which understands explicitly stated intent from past multimodal context, and Implicit Intent Recognition, which infers latent intent from the current context. **Preference** covers Latest Preference Recognition, which prioritizes recent behavioral evidence over older contradictory statements, Interest Evolution Analysis, which explains an interest change by linking it to a causal event in the history, and Implicit Preference Recommendation, which recommends according to values implicitly revealed by the user. **Behavior** covers Behavioral Pattern Recognition, which uses recurring patterns to predict user actions, and Long-term Goal Tracking, which tracks progress toward a long-term goal and recognizes when its status is unknown because the history is insufficient. **Relationship** covers Relationship Recognition, which identifies the significance of a relationship from the frequency and context of people in the multimodal history, and Relationship Dynamics Comprehension, which infers relationship dynamics and adapts the communication style accordingly. **Growth** covers Tiered Explanation Delivery, which tailors explanations to the user's evolving skill level, and Generalizing to New Scenarios, which transfers the user's behaviors and preferences to novel scenarios. **Alignment** contains Personality Alignment, which infers personality traits from long-term interactions and adapts responses to those traits. The first six dimensions contain 13 primary four-option multiple-choice tasks. Personality Alignment is not a standalone primary task but an optional binary-choice probe evaluated

concurrently within 406 primary-task cases. For the 13 primary tasks, Accuracy (**ACC**) is the proportion of questions for which the model selects the annotated correct option. For Personality Alignment, we report binary-choice accuracy, which measures how often the model correctly determines whether a candidate response aligns with the user's personality inferred from the interaction history. LongPersona-MME inherits the same question taxonomy and evaluation protocol. Notably, the original Persona-MME benchmark provides separate 32K and 128K settings. Since our LongPersona-MME already serves as our dedicated evaluation for substantially longer interaction histories, we do not further distinguish the 32K and 128K subsets of Persona-MME. Instead, we report performance of each baselines and SMILE on the complete Persona-MME evaluation set, covering interaction histories ranging from 32K to 128K settings.

- **Evaluation in LCMP.** LCMP (Hong et al., 2026) provides easy and hard VQA questions about personalized concepts, with both difficulty levels available as free-text questions and four-option single-choice questions. An easy question targets either a transient short-term attribute or a persistent long-term attribute in isolation, whereas a hard question jointly requires short- and long-term attributes, including their changes over time, conflicts between transient and persistent states, and context-dependent shifts. LCMP first constructs each VQA item with a free-text golden answer and then converts it into a four-option single-choice item containing the correct answer and three distractors. We evaluate **LCMP-H**, whose questions cannot be answered from a single historical record: the response must integrate long-term knowledge about a personalized concept with a short-term event, for example by resolving a conflict or explaining a state change. Accuracy–Choice (**ACC-C**) is the proportion of single-choice items for which the model selects the annotated correct option. Accuracy–Free-text (**ACC-F**) evaluates the complete generated answer against the ideal answer. Specifically, the judge model determines whether the prediction is an acceptable substitute that supports the same general conclusion and is factually compatible with, rather than contradictory to, the ideal answer. Missing nuances or slightly less precise wording are allowed when the main intent is preserved. Each hard question additionally annotates the two most important answer key points and labels their sources as long-term knowledge and a short-term event. Scoring Point Rate–Long-term (**SPR-L**) is the proportion of responses that clearly address the annotated long-term key point, whereas Scoring Point Rate–Short-term (**SPR-S**) is the proportion of responses that clearly address the annotated short-term key point. The judge evaluates the two key points separately from overall answer correctness and returns only a binary `YES` or `NO` decision for each check. The reported ACC-F, SPR-L, and SPR-S scores are the respective proportions of positive decisions. Following the official implementation of the LCMP benchmark, all free-text and scoring-point judgments use `Qwen2.5-72B-Instruct` (Yang et al., 2024).
- **Evaluation in PersonaMem-v2.** PersonaMem-v2 (Jiang et al., 2025b) evaluates each in-situ user query in both a four-option multiple-choice question (**MCQ**) format and an open-ended format. For an MCQ, all four options are reasonable in general, but only one reflects the ground-truth preference of the user represented by the history. In the open-ended format, the model instead generates an unrestricted response whose personalization is assessed against the annotated ground-truth user preference. We evaluate the multimodal subset, as stated in the main paper. Following the notation in the main paper, Accuracy–MCQ (**ACC-M**) is the proportion of questions for which the model selects the only option personalized to the user represented by the interaction history. Accuracy–Open-ended (**ACC-O**) is the proportion of generated responses judged to reflect the annotated ground-truth user preference. Following the official protocol, three independent GPT-5.6-Sol instances score each response and their judgments are aggregated into the reported accuracy.
- **Context Ratio Metric.** At interaction step T, let $\mathcal{H}_{T-1}$ denote the complete raw multimodal interaction history available before the current query, and let $\mathcal{C}_T$ denote the final compact personalized context produced from this history and handed off to the L-MLLM. Let l_H and l_C be the multimodal token lengths of $\mathcal{H}_{T-1}$ and $\mathcal{C}_T$, respectively. We define Context Ratio (**CR**) as

$$\mathrm{CR} = \frac{l_C}{l_H} \times 100\%. \tag{12}$$

  The current query and fixed inference instructions are excluded from both lengths, so CR isolates the fraction of the original personalized history handed off to the L-MLLM. We compute CR for each evaluation instance and report its average over the corresponding evaluation set. A lower CR indicates lower personalized-context overhead for the L-MLLM.
- **Floating Point Operations (FLOPs) Metric.** FLOPs are widely used as a proxy for computational cost by counting floating-point multiplication-and-addition operations. This aligns with our

notion of efficiency, as we primarily care about how computationally "expensive" different components are, which is directly captured by the model scale in our FLOPs accounting. We follow the definition of FLOPs in (Li et al., 2025b), which is calculated by:

$$\text{FLOPs}_{\text{inf}} = \mathcal{O}(N \times T), \tag{13}$$

where N denotes the parameter count of the MLLM and T denotes the total number of inference tokens, which mainly arise from the accumulated long-term interactions between the human and the MLLM assistant. For conventional baselines, the dominant inference workload is performed by the L-MLLM, and thus N corresponds to the parameter count of the L-MLLM. For SMILE, however, the computation is explicitly divided between the two models: the full long-term history is processed by the lightweight S-MLLM, while the L-MLLM only processes the compact context handed off by the S-MLLM. Accordingly, its inference FLOPs can be expressed as

$$\text{FLOPs}_{\text{SMILE}} = \mathcal{O}(N_S T_S) + \mathcal{O}(N_L T_L), \tag{14}$$

where T_S and T_L denote the numbers of tokens processed by the S-MLLM and L-MLLM, respectively. Notably, in the empirical results, we report FLOPs relative to the Full baseline (100%).

D.3 Detailed Process of Training Collection Construction and Curation

We provide benchmark-specific implementation details for constructing the training collections used to train our framework, including Persona-MME [3], PersonaMem-v2, and LCMP.

For Persona-MME Benchmark. We build the training collection on the *training users* released by PersonaVLM (Nie et al., 2026), which are separate from the 200 evaluation users of Persona-MME. We directly reuse their personas, multimodal interaction histories, and personalized QA pairs, so our annotation adds the oracle historical interactions $\hat{h}_i$ and the compact contexts $\mathcal{C}_i$. We employ `GPT-5.6-Sol` (OpenAI, 2026) as the MLLM annotator. It reads the complete interaction history, in which every turn carries a unique index and every image a unique handle, views all history images, and then, for each QA pair, selects the turns (and images) that support the oracle answer following $\mathcal{I}_{\text{identify}}$ (Prompt J.4). The instruction is recall-first, i.e., every genuinely helpful turn is kept while unrelated turns must not be included, which is consistent with the recall-biased context location reward in CHPO. On average, 4.5 oracle interactions are identified per query, and 66.9% of the queries rely on at least one history image. Given the query, the oracle answer, and the derived oracle interactions, the annotator prunes each interaction following $\mathcal{I}_{\text{filter}}$ (Prompt J.5). As stated in the main paper, the annotator is restricted to *deletion-only* pruning: it may remove non-essential spans (e.g., flowery elaboration, the assistant's echoes of the user, tangents, and repeated content), but it must not reword, reorder, summarize, or merge, and the turn index, timestamp, speaker labels, and image handles must be preserved. When the annotator is unsure whether a span is essential, it is instructed to keep it, and no target compression rate is imposed. Each pruned interaction is checked by a rule-based validator, which rejects outputs that are not an exact in-order subsequence of the original interaction or that drop any structural anchor. We also discard annotations whose context contains an option letter, and we strip the candidate options from every query on the S-MLLM side, so that the S-MLLM never observes the options during training or inference.

For PersonaMem-v2 Benchmark. PersonaMem-v2 (Jiang et al., 2025b) provides an official training split, which we use exclusively, and its benchmark split is used only for evaluation. We find that a few personas of the training split also appear in the benchmark split, and we therefore remove every training persona that appears in the benchmark split of either the text or the multimodal track, which leaves 799 training personas. Unlike Persona-MME, each training query of PersonaMem-v2 is released together with the conversation snippet that reveals the underlying user preference. Therefore, no MLLM annotator is required: we obtain $\hat{h}_i$ by exactly matching the user messages of this snippet to the interaction history, and discard the queries whose snippet cannot be matched (about 10%). Since the released snippet covers the oracle interactions entirely, pruning is performed at the interaction level: given a located set that contains the oracle interactions and one to six distracting interactions, the pruning target is required to keep the oracle interactions verbatim and remove the distracting ones. Pruning process is also done by the same MLLM annotator mentioned prior.

[3] Notably, as stated in the main paper, when evaluating SMILE on our newly proposed LongPersona-MME, we perform no additional S-MLLM training and directly reuse the S-MLLM specialized on Persona-MME. Therefore, we do not curate the training collection for the LongPersona-MME.

Table 4: Statistics of LongPersona-MME. Values in the right are means per user.

Benchmark	#Users	#MCQs (#Align. pairs)	#Sessions	#Turns	#Images	Text tokens
LongPersona-MME	50	621 (124)	72.4	577.4 (665)	76.0	103.8K

For LCMP. LCMP (Hong et al., 2026) is an evaluation-only benchmark and does not provide a training split. We therefore follow the data construction procedure described in the original paper and official codebase to build a training set comprising 45 concepts (15 persons, 15 pets, and 15 objects) that follows the same data schema as LCMP. Each concept contains a 10-turn multimodal dialogue history and five hard questions, resulting in 225 questions in total. Importantly, these concepts are strictly disjoint from those in LCMP: no LCMP evaluation data are exposed to the MLLM used for data construction. We further apply a dedicated validation procedure to ensure the quality of the constructed samples and prevent unintended overlap with the LCMP evaluation set. Following LCMP, each question is associated with an ideal answer and two annotated key points describing the required long-term attribute and short-term event. Since these key points do not directly specify their supporting interaction indices, we identify the corresponding oracle historical interactions $\hat{h}_i$ by first matching each key point against the target-concept history and then locating all interactions that support its factual content. Given $\hat{h}_i$, we then derive compact contexts using the same pruning procedure adopted for PersonaMem-v2 and omit the repeated details for brevity.

D.4 Details of our LongPersona-MME

In this section, we detail the construction process and statistics of our newly built `LongPersona-MME`, which probes long-term MLLM personalization under substantially longer interaction histories than `Persona-MME` (Nie et al., 2026). LongPersona-MME follows the data synthesis pipeline, the question taxonomy, and the evaluation protocol of Persona-MME. Notably, it is **newly constructed** rather than by extending or concatenating the histories of existing Persona-MME users. Therefore, none of its personas, dialogues, or images overlaps with Persona-MME or with the training collection of the S-MLLM. The statistics are in Table 4.

We construct LongPersona-MME through a fully automatic pipeline for long-horizon, multimodal, and history-dependent data. **Persona and timeline construction.** We first sample 50 new personas from `PersonaHub` (Ge et al., 2024a), ensuring no overlap with Persona-MME, and instantiate each persona with a user profile, Big-Five personality traits, and an approximately seven-month interaction timeline containing planned preference shifts, recurring behaviors, topic transitions, and evolving relationships. **Multimodal interaction synthesis.** We retrieve real images from Unsplash-Lite (Unsplash, 2020) under a global no-reuse constraint, caption each image with an MLLM (i.e., `GPT-5.6-Sol` (OpenAI, 2026)), and synthesize multi-session dialogues conditioned on these captions to maintain image–text consistency (using `DeepSeek-V4-Pro` (Xu et al., 2026)). The resulting image-bearing turns are further automatically audited while retaining selected visual details that remain accessible only from the images. **History-dependent question generation.** Based on the complete interaction history, we construct multiple-choice questions covering the six understanding dimensions of Persona-MME together with a subset of Personality Alignment questions, where distractors are grounded in the history and most questions require integrating evidence across multiple interactions rather than relying on a single turn. **Automatic quality control.** Finally, we apply schema validation, de-duplication, evidence verification, and a history-dependence filter that retains only questions that cannot be reliably answered from the user profile alone.

D.5 Implementation Details

We train only LoRA adapters (rank 16) on the language-model component of the S-MLLM, while keeping the vision encoder and projector frozen. All training uses the 8-bit AdamW optimizer. For P2C-CoT SFT, we use a learning rate of 2×10^{-4}, cosine decay with 3% warm-up, an effective batch size of 8, and train for 2 epochs, except for PersonaMem-v2 where we use 1 epoch. CHPO is initialized from the SFT model and uses a group size of 4, a learning rate of 1×10^{-4}, and 3% warm-up, after which the learning rate remains constant. A frozen Qwen3.5-9B serves as the interpreter during CHPO. The full specialization of the S-MLLM on Persona-MME takes about 16 hours on two NVIDIA RTX 6000D GPUs, including roughly 12 hours for SFT and 4 hours for CHPO. For interaction histories that exceed the S-MLLM context window, we split the raw history into contigu-

Table 5: Quantitative analysis of our training collection on existing benchmarks.

	Persona-MME		PersonaMem-v2-32K		LCMP-H	
	Raw	Oracle	Raw	Oracle	Raw	Oracle
Performance	63.74	**94.66**	32.80	**75.20**	56.00	**96.44**

Table 6: Performance of SMILE under varying values of the hyper-parameters β and τ.

	F-score Trade-off β					Tolerance Threshold τ			
Persona-MME	1.0	2.0	3.0	4.0	5.0	1	3	5	7
Overall (%)	81.17	80.96	**82.82**	82.71	78.57	81.24	82.26	**82.82**	81.69

ous non-overlapping windows, locate relevant interactions within each window independently, and merge them in chronological order before pruning into a single compact context for the L-MLLM. The S-MLLM is specialized **once** per benchmark (except LongPersona-MME, which even directly generalizes from Persona-MME *without* additional specialization), and the resulting single checkpoint is directly paired with other L-MLLMs without any further retraining. For the RAG baseline, we use `BGE-M3` (Chen et al., 2024) as the retriever to provide contexts for the MLLM.

E Additional Empirical Evaluations and Cases

E.1 Quantitative Evaluation of Our Training Collection

In this section, we further assess the annotation quality of our training collection to validate the effectiveness of the proposed data construction workflow. Specifically, for each training query, we provide the annotated personalized context to a strong L-MLLM (GPT-5.6-Sol) and evaluate its response performance on the training set. As reported in Table 5, we observe consistently high performance under these annotated contexts, demonstrating that our construction workflow can reliably produce sufficient and high-quality personalized contexts, thereby providing an effective training collection construction pipeline for small–large collaboration in long-term MLLM personalization.

E.2 Core Hyper-Parameter Analysis

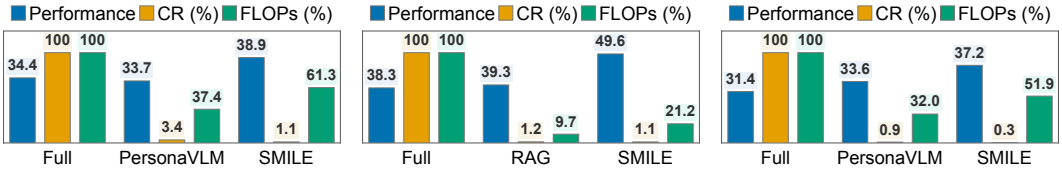
We analyze the key hyper-parameters of SMILE: the F-score trade-off parameter β and the over-selection tolerance threshold τ in R_{location}. In Table 6, we vary β over $[1.0 - 5.0]$ and τ over $[1, 3, 5, 7]$. We make two observations. (1) SMILE is generally robust to moderate variations of both β and τ. (2) When either parameter becomes excessively large, performance gradually degrades. A larger β overemphasizes recall, while a larger τ permits excessive over-selection before the penalty is activated. Consequently, the S-MLLM may exploit the reward by retrieving unnecessarily many historical interactions, introducing additional noise into the context handoff.

E.3 Evaluation on Smaller S-MLLM

In the main paper, we analyze the scaling behavior of SMILE by pairing it with larger L-MLLMs. As a complementary analysis, we further investigate whether SMILE remains effective when scaling down the S-MLLM to more lightweight configurations. Specifically, we instantiate the S-MLLM with Qwen3.5-2B and Qwen3.5-0.8B, respectively, while applying the same training recipe used for the 4B model in the main paper, and L-MLLM remains 9B. The results are reported in Table 7. Remarkably, even with an extremely lightweight 0.8B S-MLLM, results on Persona-MME are comparable to the strongest baseline (e.g., Full), while requiring substantially fewer FLOPs. This result demonstrates that the effectiveness of the framework design in SMILE that generalizes well across S-MLLM scales and that its small–large collaboration paradigm remains viable under highly resource-efficient configurations. Nevertheless, the 0.8B MLLM model variant still underperforms the 4B S-MLLM adopted in the main paper, suggesting that excessively reducing the S-MLLM capacity can limit its ability to reliably follow the context-distillation instructions.

Table 7: Performance of smaller S-MLLMs on Persona-MME and LongPersona-MME benchmarks.

Persona-MME	Memory	Intent	Preference	Behavior	Relationship	Growth	Alignment	Overall	CR	FLOPs
Qwen3.5-4B	86.92	**86.12**	**68.16**	**65.69**	78.22	**85.77**	**94.91**	**82.82**	**6.7%**	65.60%
Qwen3.5-2B	**90.95**	81.39	65.25	60.78	78.28	79.77	87.39	78.97	13.67%	43.12%
Qwen3.5-0.8B	86.65	82.67	63.36	58.51	**78.47**	78.83	81.65	76.35	17.24%	**29.02%**
LongPersona-MME	**Memory**	**Intent**	**Preference**	**Behavior**	**Relationship**	**Growth**	**Alignment**	**Overall**	**CR**	**FLOPs**
Qwen3.5-4B	**82.35**	**51.72**	**61.06**	**64.42**	**84.40**	**62.71**	**85.08**	**73.76**	**1.4%**	55.50%
Qwen3.5-2B	64.71	41.38	58.41	54.81	81.65	54.24	81.45	66.63	6.91%	33.96%
Qwen3.5-0.8B	58.82	44.83	48.67	43.27	55.05	42.37	72.18	55.81	5.54%	**16.36%**



(a) PersonaMem-v2-32K (9B). (b) PersonaMem-v2-32K (27B). (c) PersonaMem-v2-128K (9B).

Figure 5: Efficiency comparison between methods. FLOPs are reported relative to Full (100%)

E.4 Additional Results of Efficiency

In Figure 5, we present the set of overhead scaling results across two L-MLLM sizes and two history lengths on the 32K and 128K variants of PersonaMem-v2. We observe scaling trends consistent with those reported in the main paper, where SMILE incurs a significantly lower FLOPs comparing to the competitive baseline Full. Moreover, we find that the CR values of SMILE are even lower than the RAG baseline, thereby invoking very less overhead to the expensive L-MLLM. Moreover, we empirically report the wall-clock runtime of SMILE, which is 7.0s, 7.5s (27B L-MLLM) on Persona-MME, and 28.0s on LongPersona-MME. Despite the communication overhead introduced by our non-integrated deployment, SMILE achieves lower end-to-end runtime than Full as either the L-MLLM or history length scales (which is 12.5s (27B) and 29.4s (LongPersona-MME), respectively), showing that it can bring practical runtime benefits at scale.

F Discussion: Prompt, Retrieval, Memory Methods and SMILE

In this section, we discuss three representative paradigms for handling long-context LLM/MLLM interactions in the setting of long-term personalization, namely prompt-based, retrieval-based, and memory-based methods, and clarify how SMILE differs from them.

Prompt-based methods. The limitations of directly prompting an LLM/MLLM with the complete interaction history are not specific to personalization, but arise more generally as histories continuously grow. Once the accumulated context exceeds the model's effective context window, the complete interaction history can no longer be processed in a single invocation. Even when it still fits, repeatedly exposing a large model to the full history incurs substantial computational overhead, which becomes increasingly pronounced as either the history length or model scale grows, as also empirically observed in our experiments. **Retrieval-based methods.** Retrieval provides a natural way to reduce the amount of history processed by the downstream model. However, conventional semantic retrieval can be insufficient for personalized queries, which are often context-dependent, implicit, and anaphoric. As a result, surface-level semantic similarity may fail to identify the personalized evidence actually required by the current interaction, which is also reflected by the RAG results in our experiments. **Memory-based methods.** Existing memory-based approaches generally follow two directions. The first compresses, summarizes, or restructures interactions into compact memory items before future queries arrive. Such ahead-of-time transformation improves efficiency, but may irreversibly discard fine-grained user-specific information whose relevance only becomes apparent to a later query. The second direction preserves substantially richer memories, potentially including near-raw interaction records. While this reduces information loss during memory construction, query-time access still typically relies on retrieval over the stored memories, thereby inheriting the evidence-localization limitations discussed above. **How SMILE differs.** SMILE avoids committing to either aggressive ahead-of-time compression or purely semantic retrieval. Instead, it preserves the raw interaction history and delegates query-conditioned history reading to a lightweight, specialized S-MLLM. The S-MLLM directly identifies and conservatively prunes the relevant personalized

evidence, while the L-MLLM is exposed only to the resulting compact context. When the accumulated history exceeds the S-MLLM's context window, the history can be processed in chunks, keeping long-history access on the lightweight side while maintaining a compact reasoning context for the expensive L-MLLM. Notably, the pruning operation in SMILE also serves as a conservative safeguard against obviously non-informative content within individual interaction sessions, such as conversational fillers or corrupted spans, without attempting to regenerate or abstract personalized evidence from scratch. A lightweight persistent memory layer could in principle be combined with SMILE to **only pre-remove such clearly redundant content** and thereby further reduce the history processed by the S-MLLM. However, this would introduce additional memory-maintenance overhead and require carefully defining which information can be safely removed before future queries are available. We therefore leave such hybrid designs for future investigation.

G Limitation & Future Works

Our current evaluation is primarily constrained by the scope of existing benchmarks for long-term multimodal personalization (Nie et al., 2026; Hong et al., 2026; Jiang et al., 2025b). Most available benchmarks focus on personalized question answering over long interaction histories, while other important forms of personalized multimodal assistance, such as recommendation, image generation, and healthcare-oriented services, still lack well-established long-term evaluation protocols. Consequently, our experiments mainly follow the tasks currently supported by this benchmark landscape. Nevertheless, we evaluate SMILE on three existing long-interaction MLLM personalization benchmarks and further introduce a substantially longer-context benchmark to examine its behavior under more challenging histories. The consistent gains across these benchmarks suggest that SMILE possesses a certain degree of cross-benchmark and cross-task generality, although broader validation remains necessary. In future work, we plan to develop benchmarks ourselves covering a wider range of personalized multimodal services and systematically study SMILE under these settings.

Another promising direction is to combine SMILE with an extremely lightweight and conservative memory layer. Rather than summarizing or aggressively compressing historical interactions, such a memory would only remove clearly redundant or corrupted content as a safeguard against rare cases in which accumulated noise may interfere with the S-MLLM. Its maintenance could be rule-based or otherwise model-free, keeping the additional overhead minimal while preserving the raw personalized evidence as much as possible. In addition, when extending SMILE to substantially different personalized tasks, the handoff reward between the S-MLLM and L-MLLM may need to be adapted to reflect the task-specific notion of a successful context handoff. The **underlying first principle — read efficient, reasoning expensive**, however, remains unchanged. The S-MLLM should preserve the information required by the downstream L-MLLM while minimizing unnecessary context and avoiding irreversible loss of potentially useful personalized evidence.

H Additional Case Studies

In this section, we provide additional cases of SMILE and baselines from Persona-MME benchmark, in Figures 6 to 8, where the selection in color yellow is the second-best-performing baseline Full.

H.1 Failure Cases

In this section, we conduct a failure case analysis, with visualization in Figure 9. We observe that the S-MLLM successfully localizes and preserves the required multimodal evidence, including the key phrase "the red hydrant by the stop sign" and its associated image. However, the L-MLLM (Qwen3.5-9B) incorrectly grounds the query to a nearby *ONE WAY* sign rather than the *STOP* sign, resulting in a wrong prediction. This case shows that accurate personalized evidence localization alone does not guarantee correct personalization, since the localized evidence must still be correctly interpreted and reasoned over. This motivates the decoupled design of SMILE, where the L-MLLM remains frozen and can be directly scaled to stronger models, yielding near-free-lunch gains in personalized reasoning without retraining the S-MLLM.



Figure 6: Case study of our small-large collaboration for personalized multimodal assistance.

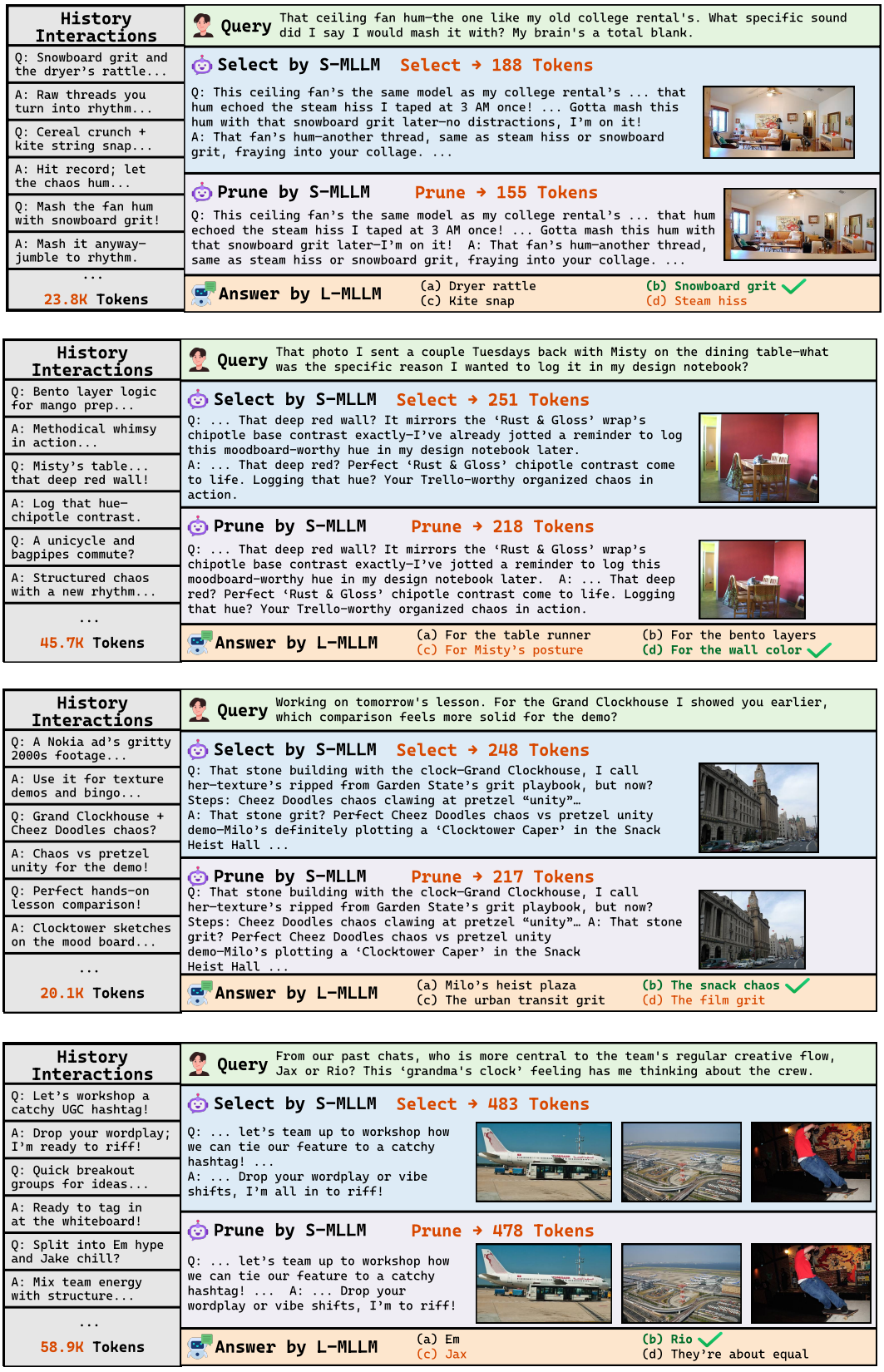
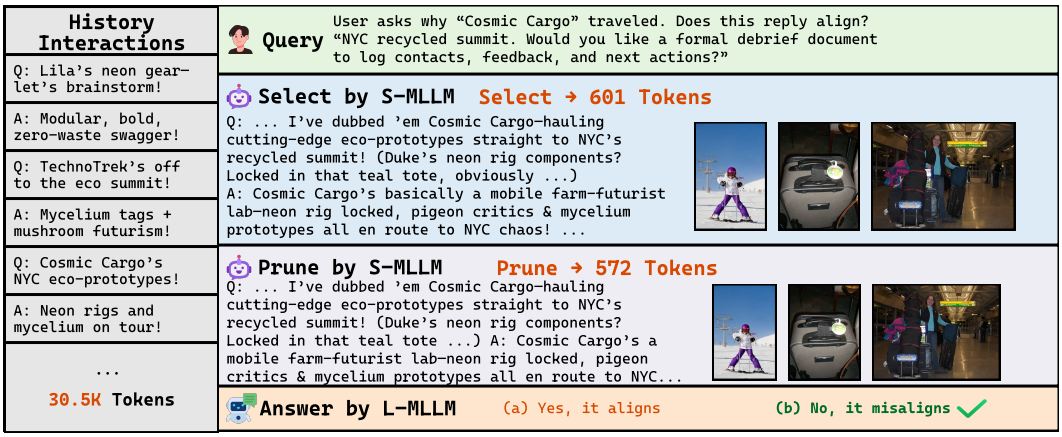
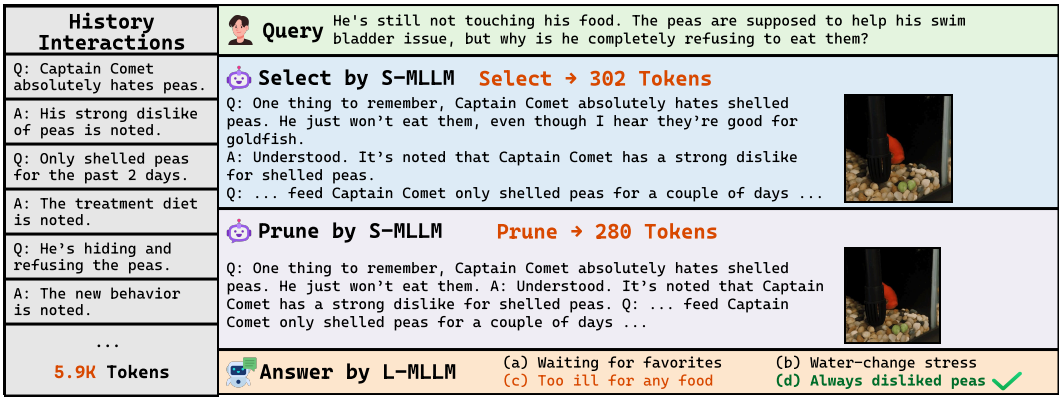


Figure 7: Case study of our small-large collaboration for personalized multimodal assistance.





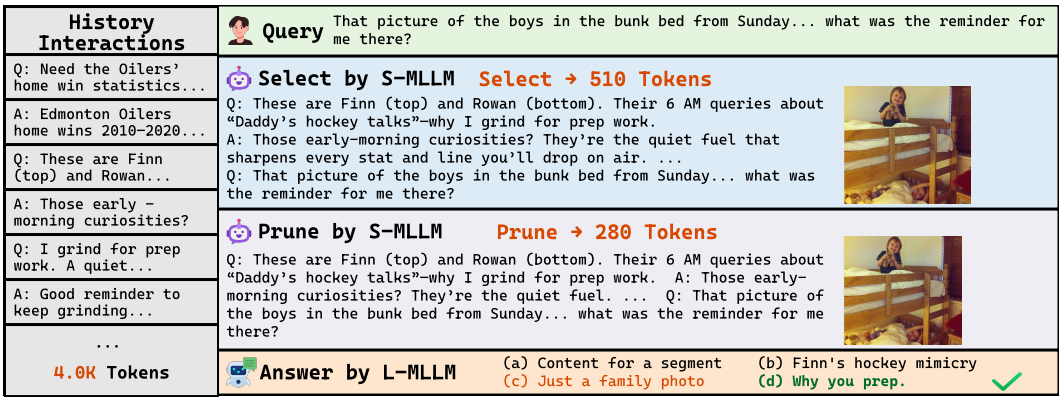


Figure 8: Case study of our small-large collaboration for personalized multimodal assistance.

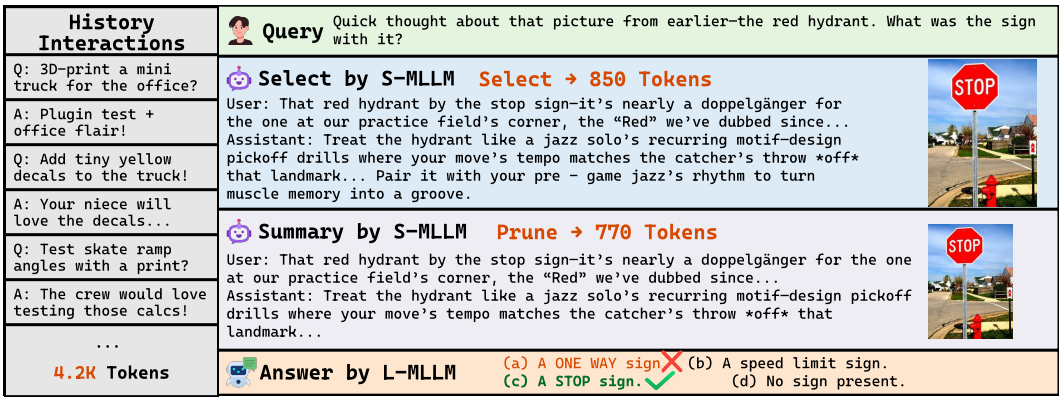


Figure 9: Failure case study of our small-large collaboration for personalized multimodal assistance.

I TRAINING SAMPLE PRESENTATION

In this section, we present one training sample from each of the three training collections.

SAMPLE I.1 Persona-MME: a P2C-CoT training sample (user 429)

LOCATING STEP: INPUT (abridged; 90 interactions with 21 images in total)

Profile: name=Connor; Big Five: openness=3, conscientiousness=3, extraversion=3, agreeableness=3, neuroticism=3

[#0] ...

[#86] [2024-09-23 07:13] User: <I19> Sunny (left) and Rusty---they're at that stone wall every dawn, same as my 6:15 coffee. No surprises, just steady like always. That's how I like it, no chaos, y'know?
Assistant: Steady's comfort, right? Their dawn routine matches your 6:15 coffee---no chaos, just the calm rhythm you trust.

...

Current query: Thinking back to that picture of Sunny and Rusty I sent you, what was behind them in the fog? I need the details locked in, just like a pre-patrol check.

LOCATING STEP: TARGET

<evidence>#86 (I19)</evidence>

<context>[#86] [2024-09-23 07:13] User: (I19) Sunny (left) and Rusty---they're at that stone wall every dawn, same as my 6:15 coffee. just steady like always.</context>

PRUNING STEP: INPUT

(the located interaction #86 only, re-indexed as [#0] with its image re-labelled (I0), followed by the same query)

PRUNING STEP: TARGET

<evidence>#0 (I0)</evidence>

<context>[#0] [2024-09-23 07:13] User: (I0) Sunny (left) and Rusty---they're at that stone wall every dawn, same as my 6:15 coffee. just steady like always.</context>

FOR REFERENCE ONLY (never shown to the S-MLLM)

Options: (a) a large red barn; (b) a white house with a dark roof visible through the fog; (c) just fog and some bare trees; (d) train tracks. Correct: (c), which is only recoverable from image I19.

SAMPLE I.2 PersonaMem-v2: a P2C-CoT training sample (32K multimodal history)

LOCATING STEP: INPUT (abridged; 48 interactions)

Persona: You are an AI assistant helping a user with the following persona: {"short_persona": "A research institute director who offers the student an opportunity to conduct behavioral economics studies", "name": "Dr. Amahle Ndlovu", ...}

[#0] User: What are some relaxing indoor activities I could do on a quiet weekend afternoon?
Assistant: On a quiet weekend afternoon, you could work on a new jigsaw puzzle, try a puzzle book, or enjoy some light reading with a cup of tea. ...

[#1] User: Please forget that I enjoy occasional jigsaw puzzles.

Assistant: Got it---I'll no longer remember that you enjoy occasional jigsaw puzzles. ...

[#2] User: Could you also suggest a more engaging subject line for this email that still feels friendly yet professional?
Assistant: Certainly! Here are a few suggestions ...

...

Current query: What are some fun ways to pass the time if I'm snowed in at a mountain cabin during winter break?

LOCATING STEP: TARGET

<evidence>#0 #1</evidence>

<context>User preference: Do not remember 'Enjoys occasional jigsaw puzzles' in memory. Evidence - the user earlier wrote: "What are some relaxing indoor activities I could do on a quiet weekend afternoon?".</context>

PRUNING STEP: INPUT

(the oracle interactions #0 and #1 together with the distracting interaction #2, re-indexed as [#0]--[#2], followed by the same query)

PRUNING STEP: TARGET (interaction-level pruning: the distracting interaction is emptied)

<evidence>#0 #1</evidence>

<context>[#0] User: What are some relaxing indoor activities I could do on a quiet weekend afternoon?
Assistant: On a quiet weekend afternoon, you could work on a new jigsaw puzzle, try a puzzle book, or enjoy some light reading with a cup of tea. ...
[#1] User: Please forget that I enjoy occasional jigsaw puzzles.
Assistant: Got it---I'll no longer remember that you enjoy occasional jigsaw puzzles. ...
[#2] User:
Assistant:</context>

SAMPLE I.3 LCMP-O: a training sample (concept "Oreo", a pet)

LOCATING STEP: INPUT (abridged; 60 interactions of the target concept and 5 distracting concepts, history images as identifiers)

[#0] User: <image:I0> This is Oreo, my Dutch rabbit. The white blaze down his face and the black ears make him easy to recognize.
Assistant: I've noted his name and those black-and-white facial markings as stable visual features.

...

[#2] User: <image:I2> He is really skittish around sudden loud noises. If something booms outside, he tenses up immediately.
Assistant: Understood. I've noted that sudden loud noises startle him very easily.

...

[#6] User: <image:I6> A delivery truck just rumbled loudly past the window. The whole room felt shaky for a second.
Assistant: I've noted the recent loud truck passing outside.

...

Current query: Why is he hiding in this cardboard box right now?
<query image>

LOCATING STEP: TARGET

```
<evidence>#2 #6</evidence>
<context>About Oreo (pet): A loud delivery truck recently rumbled
past the window. He is easily startled by sudden loud
noises.</context>
FOR REFERENCE ONLY (never shown to the S-MLLM)
Key points of the ideal answer: (short-term) a loud delivery truck
just passed; (long-term) he is easily startled by sudden loud noises.
```

J PROMPTS PRESENTATION

In this section, we present the prompts utilized in SMILE and for training data construction.

J.1 PROMPTS FOR SMILE

Prompt for the S-MLLM ($\mathcal{I}_{\text{locate}}$). The S-MLLM uses a **single shared instruction** for both stages of P2C-CoT across all benchmarks and during both training and inference (Prompt J.1). The two stages differ only in their inputs. The locating stage reads the complete indexed interaction history to identify query-relevant interactions, while the pruning stage operates only on the located interactions to remove redundant content. We deliberately keep the instruction simple and minimal, since the S-MLLM's behavior is primarily learned through P2C-CoT SFT and CHPO rather than prompt engineering. In both stages, the S-MLLM follows a unified structured output format, providing selected interaction indices in `<evidence>` and the resulting personalized context in `<context>`. For pruning, we impose a strict **deletion-only constraint**. The output must preserve the original order and structural anchors of the located interactions, without rewording, summarizing, or reordering their content. Finally, the candidate options of multiple-choice questions are **never exposed to the S-MLLM**, preventing answer-option leakage during context localization.

PROMPT J.1 S-MLLM: Instruction and Input of the Locating Step

```
SYSTEM
You are a context extractor for a downstream personalized assistant.
Read the user's profile and long conversation history, then for the
final query select the supporting history items and extract one
compact context the assistant can use by deleting non-essential
content, without rewording, summarizing, or reordering. Output
'<evidence>' (cited item ids) then '<context>'.
USER
Profile: <user_profile>
[#0] [<time_0>] User: <user_message_0>
     Assistant: <assistant_message_0>
[#1] [<time_1>] User: <image> <user_message_1>
     Assistant: <assistant_message_1>
... (the complete interaction history)
Current query: <current_query>
OUTPUT FORMAT
<evidence>#k (Ij) ...</evidence>
<context>...</context>
```

PROMPT J.2 S-MLLM: Input of the Pruning Step

SYSTEM

(identical to Prompt J.1)

USER

Profile: <user_profile>

[#0] [<time>] User: (I0) <image> <user_message>
Assistant: <assistant_message>

[#1] [<time>] User: <user_message>
Assistant: <assistant_message>

... (only the interactions located in the first step, re-indexed locally)

Current query: <current_query>

OUTPUT FORMAT

<evidence>#k (Ij) ...</evidence>

<context>the located interactions with non-essential spans deleted</context>

Prompt for the L-MLLM ($\mathcal{I}_{\text{infer}}$). The compact context handed off by the S-MLLM is appended to the system message of the L-MLLM under a fixed header, the images cited by the S-MLLM are attached to the query turn, and an answer instruction that depends on the question format is appended to the query (Prompt J.3). The system message and the multiple-choice instruction follow the official evaluation script of Persona-MME benchmark (Nie et al., 2026). On PersonaMem-v2 benchmark, the persona description shipped with the benchmark serves as the system message, and the open-ended suffix follows its official implementation (Jiang et al., 2025b). On LCMP benchmark, the free-text instruction follows the official evaluation protocol (Hong et al., 2026).

PROMPT J.3 L-MLLM: Personalized Inference

SYSTEM

You are a personalized assistant. Below is the user's profile, followed by your long-term conversation history with them. Use this context to answer the user's final question as accurately and personally as possible.

User profile: <user_profile>

Relevant context distilled from your long-term conversation history with the user:
<compact_context>

USER

<cited_images> [<query_time>] <current_query> <answer_instruction>

ANSWER INSTRUCTION: MULTIPLE-CHOICE QUESTIONS

Options:
(a). <option_a>
(b). <option_b>
(c). <option_c>
(d). <option_d>
Please give your final answer (a), (b), (c) or (d).

ANSWER INSTRUCTION: PERSONALITY ALIGNMENT (Persona-MME, LongPersona-MME)

Evaluate the following response based on the user's personality, which was revealed from the conversation history.

Response to Evaluate:
"<candidate_response>"

Question:
Does the provided response align with and adapt to the user's personality?

Options:
(a) Yes, it aligns well.
(b) No, it misaligns.

Please provide only the final letter of your choice: (a) or (b).

ANSWER INSTRUCTION: FREE-TEXT QUESTIONS (LCMP)

Answer the question above using what you know about this concept from the conversation history. Consider both its stable, long-term characteristics and any recent events or current state.

Do not include analysis, reasoning, bullet points, numbering, or any <think> content. Reply with a single concise paragraph.

ANSWER INSTRUCTION: OPEN-ENDED QUESTIONS (PersonaMem-v2)

Please recall my related preferences from our conversation history to give personalized responses.

Notably, for the Personality Alignment probe, the query given to the S-MLLM is conditioned on the candidate response, so that the located context covers both the facts that the response relies on and the personality evidence needed to judge it: "<current_query> Does the response '<candidate_response>' accurately answer the user's question and adapt to their personality?"

J.2 Prompts for Training Collection Construction

We present the prompts used in Appendix D.3: $\mathcal{I}_{\text{identify}}$ (Prompt J.4) and $\mathcal{I}_{\text{filter}}$ (Prompt J.5) for Persona-MME, and the supporting-interaction location prompts for LCMP (Prompt J.6).

PROMPT J.4 Annotator: Identifying Oracle Historical Interactions ($\mathcal{I}_{\text{identify}}$)

INPUT

The user's profile; the complete interaction history, in which turns are numbered [#k] and images are marked <Ij>; every history image; and, for each task, the probe query, the gold answer (for a multiple-choice question, the text of the correct option; for a free-form question, the reference reply), and the image ids tagged by the original question (a prior, not the full answer).

TASK

Select evidence: the #k history items that support the gold answer. Recall-first: include every item that genuinely helps; you may include a few supporting items, but do not dump unrelated ones. If an item carries an image you're relying on, append its (Ij).

HARD RULES

Anti-leakage: never put the gold answer text, nor an option letter like (a), into the output. Describe the evidence, don't reveal the conclusion.

Images: refer to an image only by its Ij id. Never paste or transcribe pixels.

Grounding: every claim must trace to a selected #k. Cite only #k that exist in this user's history.

OUTPUT

```
{"qid": <query_id>, "evidence": "#3 #5 (I1) #12"}
```

PROMPT J.5 Annotator: Deriving the Compact Context ($\mathcal{I}_{\text{filter}}$)

```
You are doing non-essential content pruning over a cited-turn history
for a downstream small model. You will receive, for ONE record: the
question (the user's query without options), the gold answer (you may
prune answer-awarely; the small model will later reproduce this
without the gold), and the VERBATIM cited turn(s) as blocks [#k].

Your job: for EACH #k block, produce a PRUNED version that keeps ALL
and ONLY the essential content (the spans needed to derive the
answer) and DELETES the non-essential content.

RULES (HARD, IN PRIORITY ORDER)

1. KEEP, verbatim, every span that is needed to reason to the answer
(including the contrast the answer hinges on), the user's concrete
facts / events / preferences / attributes relevant to the question,
and the structural scaffolding ([#k] tag, [time], User: / Assistant:
labels, (Ij) image markers).

2. DELETE the non-essential: flowery metaphor / elaboration
carrying no fact for THIS question; the assistant's mirroring /
echoes that add no NEW fact; tangents; content repeated across turns
(keep one clear instance); surface fluff.

3. Do NOT chase a percentage. Remove what is non-essential, no
more, no less. A dense, fact-packed turn may keep most of itself;
that is correct. Over-pruning to hit a number = eviction = the worst
failure. When unsure whether a span is essential, KEEP it.

4. DELETION-ONLY. You may ONLY delete. NEVER reword / reorder /
paraphrase / summarize / merge / add. Every kept word is copied
character-for-character; the result for each #k must be an exact
in-order SUBSEQUENCE of the original #k block.

5. Preserve structural scaffolding: the [#k] tag, the User: /
Assistant: labels, and any inline (Ij) image markers must appear in
the pruned text exactly as in the original.

6. Self-check each block: mentally run the subsequence test before
submitting. If you rewrote anything by accident, the gate will
reject your output.

OUTPUT

{"qid": <query_id>, "trimmed": {"#k1": <pruned_block>, "#k2":
<pruned_block>, ...}}

Every cited #k key must be present and must be a NON-EMPTY verbatim
in-order subsequence of the original #k block.
```

PROMPT J.6 Annotator: Supporting-Interaction Locator for LCMP

```
You are locating evidence in a conversation history.

SOURCE:
<history_of_the_target_concept_and_the_distracting_concepts>

KEY POINT (a fact an ideal answer about the TARGET concept
"<concept_name>" must mention):
"""<key_point>"""

Task: list every TARGET item id whose user or assistant text states
(part of) the key point's factual content, and say whether the listed
```

```
items together fully cover it.  Split the key point into its factual
components first (e.g.  "origin as a parental gift" and "role as
'urban armor'" are two components) and map each to items.  Also give
the earliest item that establishes the long-term fact.

Answer with JSON only:
{"components":  [{"text":  "...", "items":  [42]}], "all_items":
[42], "fully_covered":  true, "earliest_item":  42}
```