TAMEing the Open-World Personalization: Towards Open-Set Personalized MLLM Assistant

Rongpei Hong[1] **Jian Lang**[1] **Ting Zhong**[1*] **Fan Zhou**[1,2]
[1]University of Electronic Science and Technology of China
[2]Intelligent Digital Media Technology Key Laboratory of Sichuan Province

Abstract

Personalizing Multimodal Large Language Models (MLLMs) as intelligent assistants has recently emerged as an active research topic, where an MLLM is expected to go beyond generic, *one-size-fits-all* replies, and instead generate responses grounded in user-specific objects, entities, and concepts encountered during interactions. Nevertheless, existing studies mainly focus on personalizing MLLMs on a *static, predefined set* of personalized concepts and require users to manually and frequently extend this set to accommodate newly emerging concepts, which is infeasible in dynamic-world scenarios. To address this limitation, we recast conventional MLLM personalization as an **Open-Set problem**, and propose **TAME-O**, the first long-context open-set personalized MLLM assistant. Specifically, TAME-O introduces a training-free, plug-and-play Concept Reconciliation Skill (**CR Skill**). It enables the assistant to automatically unlock novel concepts during prolonged human–MLLM interactions while preventing confusion with previously unlocked ones, allowing the MLLM to continually grasp more and more concepts and progressively refine the modeling of known ones. Extensive experiments under the open-set personalization setting underscore the efficacy of our method.

1 Introduction

Imagine a future intelligent assistant that, through prolonged interactions, can not only remember you, but also gradually uncover and internalize your personal entities that are never specified in advance: your blue cup, and your yellow puppy Timi. When the assistant "sees" these entities in later conversations, it can visually recognize them and generate "personalized responses" that are grounded in the appropriate contexts, rather than *one-size-fits-all* answers.

All these imaginations are closely tied to the research field of multimodal large language model (MLLM) personalization [1, 2, 3, 4], which aims to enable general-purpose MLLMs (e.g., GPT-5 [5], Qwen3.5 [6]) to understand user-specific concepts, also referred to as "personalized concepts"[1] (e.g., Bob's dog), instead of treating them as generic objects. Most existing studies, such as MyVLM [1] and R2P [7], focus on the capability of MLLMs in visual identification of personalized concepts, followed by a simple substitution of the corresponding concept in the output generation. Nevertheless, such static applications are *overly simplified* and fail to cater to real-world demands, where user-MLLM interactions are prolonged and dynamic, and dialogues become more complex due to extended contexts. As a remedy, TAME [4] explored long-context personalized dialogues with MLLMs by modeling long- and short-term attribute shifts of each concept over time, enabling the personalization experience to progressively improve over extended human-MLLM interaction.

However, all existing methods, including TAME, can only operate under the *closed-set personalization setting*: the MLLM is personalized on a fixed and static set of user-specific concepts that are

*Corresponding author: zhongting@uestc.edu.cn.

[1]In this work, we use the terms "concept", "personalized concept", and "user-specific concept" interchangeably.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

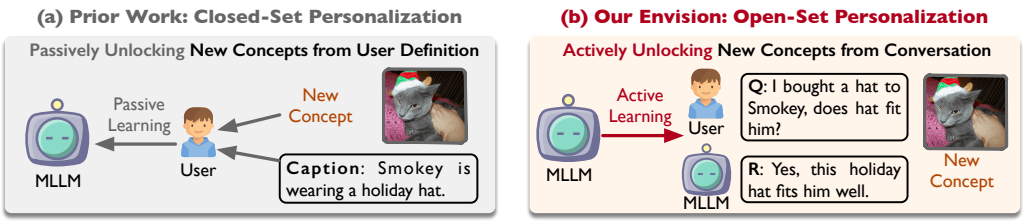


Figure 1: Paradigm comparison between (a) Closed-Set Personalization, where the MLLM passively acquires novel concepts from user provision, and (b) Open-Set Personalization, where the MLLM actively grasps new concepts from everyday interactions.

manually predefined by the user. In real-world scenarios, users frequently encounter new personalized concepts [8, 7], and this rigidity severely undermines the flexibility of existing methods, which require users to frequently register the "novel personalized concepts" for the MLLM by hand (cf. Figure 1(a)). In contrast, an ideal MLLM assistant should *gradually unlock more and more novel personalized concepts automatically* during continued human–MLLM interaction, while progressively refining the modeling of previously known concepts. We term this capability **Open-Set MLLM Personalization**, an appealing property for MLLM assistants in the dynamic world (cf. Figure 1(b)).

To achieve open-set personalization, the core challenge lies in enabling the MLLM assistant to reliably *resolve* newly emerging personalized concepts with existing ones over extended, long-context interactions. On one hand, when novel personalized concepts are misidentified as existing ones, the assistant may *accumulate erroneous contexts* for the existing concepts. On the other hand, misclassifying an existing user-specific concept as novel not only prevents the assistant from leveraging accumulated contexts during generation, but also leads to *concept fission*, where an existing concept is erroneously forked into two or more concepts.

To address these two symmetric problems, we draw upon the insight that user-specific concepts (e.g., visual appearance) evolve over time. As a result, the past dialogues that are *temporally proximal to the current interaction* often provide more reliable cues for identification than distant history. Accordingly, when comparing a newly appearing concept with those grasped from past dialogues, more recent memories should be emphasized for robust concept maintenance.

To this end, we propose **TAME-O**, the first **O**pen-set, long-context personalized MLLM assistant that can automatically **TAME** novel user-specific concepts as they appear during ongoing interactions, beyond a predefined closed-set of concepts. TAME-O endows the MLLM assistant with a novel training-free Concept Reconciliation Skill (**CR Skill**) that mediates between incoming concepts and stored knowledge. Specifically, the CR Skill equips the MLLM with a new Time-Aware Dynamic Memory (**TAD Memory**), which stores personalized concepts along with their associated attributes extracted from past interactions, where each attribute is associated with a *temporal marker* to reflect when it was observed. Subsequently, when a concept is involved in the current interaction, the CR Skill retrieves candidate relevant concepts from the TAD Memory by considering concept attributes in a *temporally weighted manner*, prioritizing memories closer to the current interaction. These retrieved candidates, together with the current user query and recent dialogue contexts, are jointly reasoned by the MLLM to determine whether the concept aligns with an existing one or represents a new concept, enabling the assistant to effectively grasp novel user-specific concepts while continuously refining unlocked ones. The CR Skill is designed to be flexible and can be readily integrated into existing personalized MLLM assistants (e.g., TAME) in a *plug-and-play* way, granting them open-set personalization capability in the dynamic world. We list our contributions as follows:

- We formalize the problem of Open-Set MLLM Personalization, moving beyond static concept dictionaries to enable continuous, autonomous concept acquisition in dynamic environments.
- We propose TAME-O, the first open-set, long-context personalized MLLM assistant. TAME-O introduces a novel plug-and-play CR Skill that can be integrated into existing MLLM assistants to activate their open-set capability, automatically managing the lifecycle of personalized concepts by detecting the new and refining the old.
- Experiments on both existing LCMP-H and LCMP-O, a dedicated benchmark for open-set MLLM personalization newly constructed in this study, showcase the efficacy of TAME-O. Code is provided at `https://github.com/ronpay/TAME-O`.

2 Related Work

MLLM Personalization. MLLMs have been widely adopted as assistants, while most mainstream models still operate under a one-size-fits-all paradigm, treating all users as interchangeable and failing to adapt to individual identities, preferences, or interaction histories [3, 5, 9, 10, 11, 12]. MyVLM [1] and Yo'LLaVA [2] pioneered the concept of MLLM personalization, which enables an MLLM with generic knowledge to respond to user visual queries in a personalized manner. They achieved personalization by training a set of parameters to visually memorize each concept [13, 14]. To alleviate the frequent retraining for each newly emerging concept, studies such as RAP [8], PVIT [15], R2P [7], and PLVM [16] explored memory-based or tuning-free mechanisms, where a set of user-predefined personalized concepts is first stored in memory, and an incoming concept is then treated as a query to visually match against these stored concepts for personalized response generation. However, all these methods primarily focus on the visual identification of personalized concepts, that is, recognizing an object in an image as user-specific and replacing a generic object reference with the corresponding personalized concept name for captioning or single-turn question-answering. Such overly simplified settings fail to meet real-world requirements, where human–MLLM interactions are persistent and highly contextualized, and personalization must account for complex dependencies from extended interaction history rather than isolated, static inputs. To remedy such limitation, TAME [4] proposed the first long-context MLLM personalized assistant, which fosters MLLM to capture the variations in both long- and short-term attributes of user-specific concepts. Nevertheless, existing methods (including TAME) are limited to only responding to a *manually* predefined set of user-specific concepts and fail to adapt to dynamic real-world settings. In contrast, we propose TAME-O, which activates Open-Set Personalization capability through a training-free, plug-and-play Concept Reconciliation Skill (CR Skill), which effectively and autonomously unlocks novel personalized concepts while avoiding potential conflicts with previously "memorized" ones. The additional literature reviews for memory-augmented LLM and open-set recognition are in Appendix G.

3 Methodology: TAME-O

Revisiting TAME. We begin by re-formalizing the pipeline of TAME [4], the first long-context personalized MLLM assistant. TAME considers a continuous stream of human–MLLM interactions: $\mathcal{S} = \langle \mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_i, \ldots \rangle$, where each $\mathcal{Q}_i$ denotes an atomic single-turn vision–language dialogue occurring at time step i, consisting of a user query q_i and the corresponding assistant response a_i. When responding to q_i, the MLLM assistant is required to leverage the potential contexts from the past dialogues tied to the personalized concept appearing in $\mathcal{Q}_i$: $a_i = \Phi_{\text{MLLM}}(q_i, \pi(\langle \mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_{i-1} \rangle))$, where $\pi(\cdot)$ denotes a context extraction strategy that filters out relevant information for current concept from the interaction history. To achieve this, TAME endows the assistant with double memories, a static memory $\mathcal{M}_S$ and a dynamic memory $\mathcal{M}_D$, which are designed to maintain the long-term and short-term attributes of personalized concepts from past dialogues, respectively. Each memory stores items as a tuple $(\texttt{Cid}, \texttt{Attr})$, where `Cid` is the concept identifier that is used to uniquely identify a personalized concept, while `Attr` denotes an attribute extracted from a specific dialogue turn. Subsequently, when facing an incoming personalized concept, TAME first identifies its corresponding concept identifier `Cid`, and then aggregates the memory items (attributes) that are marked with the same `Cid` from $\mathcal{M}_S$ and $\mathcal{M}_D$ to support personalized response generation.

However, TAME implicitly assumes a **closed-set** personalization setting, where a set of N user-specific concepts is predefined and fixed throughout the interaction: $\mathcal{C} = \{c_1, c_2, \cdots, c_N\}$, where $\mathcal{C}$ is the static concept set. When encountering a novel concept c_{N+1}, TAME requires the user to manually initialize this concept and provide the corresponding concept identifier for proper memory management. Consequently, such a setting is ill-suited to dynamic real-world scenarios, where new personalized concepts frequently arise, and requiring users to explicitly specify concept identifiers introduces significant overhead and undermines usability.

Open-Set MLLM Personalization Formalization. We consider a more real-world setting, in which personalized concepts are not predefined by users but are instead **automatically grasped** by the MLLM assistant through extended interactions. Specifically, the concept set is defined as: $\mathcal{C}_{i-1} = \{c_1, c_2, \cdots, c_{K_{i-1}}\}$, where $\mathcal{C}_{i-1}$ is the dynamic concept set at time step $i-1$ that contains K_{i-1} concepts unlocked by the assistant up to the current step (empty at start). At time step i, the concept referred to in the dialogue $\mathcal{Q}_i$ is judged to be either **novel** or **existing**. If

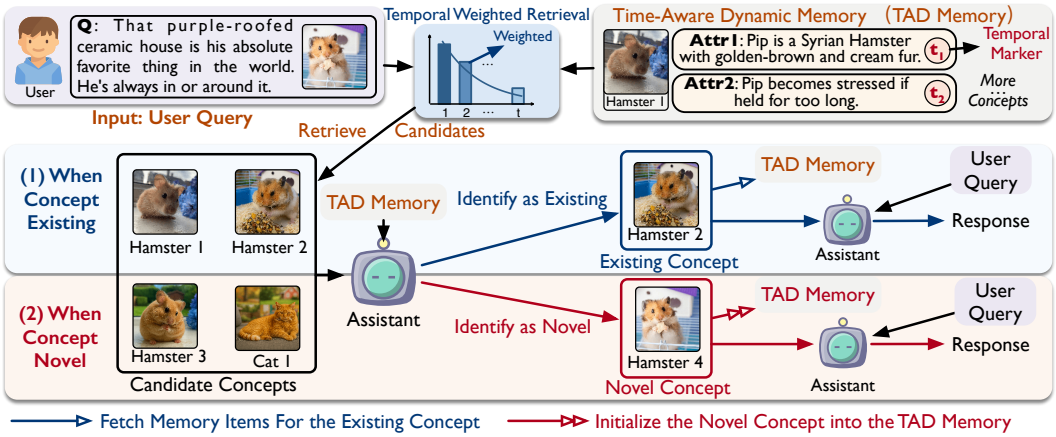


Figure 2: Paradigm of TAME-O. When receiving a new dialogue, the CR skill first retrieves candidate personalized concepts from the TAD Memory in a temporally weighted manner. The retrieved candidates are then provided to the MLLM, together with the user query and recent dialogue context, to determine whether the referenced concept is existing or novel, achieving open-set personalization.

it is identified as a novel concept, the personalized concept set is updated by the assistant as: $\mathcal{C}_i = \{c_1, c_2, \cdots, c_{K_{i-1}}, c_{K_i}\}$, where c_{K_i} is the new concept appearing in $\mathcal{Q}_i$.

Paradigm of TAME-O. To address this challenging setting, TAME-O provides existing personalized assistants with a novel training-free and plug-and-play Concept Reconciliation Skill (CR Skill), which handles novel concepts while mitigating potential conflicts with existing ones. Specifically, TAME-O first equips the MLLM assistant with a Time-Aware Dynamic Memory (TAD Memory), which captures variations of personalized concepts in a temporally annotated manner (cf. Section 3.1). Subsequently, when a new interaction occurs, the CR Skill evaluates the incoming concept against memory items using a temporally weighted matching strategy, and provides a set of candidate concepts to the MLLM for final resolution of this concept (cf. Section 3.2). We further provide an example of integrating the CR Skill into the existing personalized assistant TAME (cf. Section 3.3). Finally, we discuss potential concept fission issues and present corresponding mitigation solutions (cf. Section 3.4). The framework of TAME-O is in Figure 2.

3.1 Time-Aware Dynamic Memory

We consider that after time step $i-1$, the personalized concepts that have been identified by the MLLM assistant are represented as a dynamic concept set $\mathcal{C}_{i-1} = \{c_1, c_2, \ldots, c_{K_{i-1}}\}$, where K_{i-1} denotes the number of personalized concepts discovered up to time step $i-1$. Given a new dialogue $\mathcal{Q}_i$, a simple approach to determining whether a novel personalized concept appears is to compare the information of the concept extracted from a_i against the grasped concepts in memory. However, in the long-context interactions, the information (also referred to as attributes) of each concept often varies over time [4]. Consequently, early attributes of a concept may become outdated and unfaithful for depicting such concept, whereas more recent attributes can better reflect its current state.

Inspired by this observation, we introduce a novel training-free Concept Reconciliation Skill (**CR Skill**), which first endows the MLLM assistant with a new Time-Aware Dynamic Memory (**TAD Memory**) to store each attribute in a temporally-annotated manner. Specifically, taking completing a dialogue $\mathcal{Q}_k$ and storing the newly acquired information tied to the corresponding personalized concept as an example, the TAD Memory $\mathcal{M}_T$ operates as follows:

$$(\texttt{Cid}_k, \texttt{Attr}_k, \delta_k) = \Phi(\texttt{Cid}_k, \mathcal{Q}_k, \mathcal{M}_T, \mathcal{P}_u)\,, \tag{1}$$

$$\mathcal{M}_T \leftarrow \Psi(\mathcal{M}_T, \texttt{Cid}_k, \texttt{Attr}_k, \delta_k, t_k)\,, \tag{2}$$

where $\texttt{Cid}_k$ is the identifier of the concept referred to in $\mathcal{Q}_k$ (its acquisition is discussed in the **next section**), $\texttt{Attr}_k$ is the attribute information extracted from $\mathcal{Q}_k$, δ_k is the memory execution indicator determined by the MLLM $\Phi(\cdot)$ for this attribute $\texttt{Attr}_k$ that follows the update actions adopted in TAME (i.e., add or modify an attribute), $\mathcal{P}_u$ is the prompt for memory update, and $\Psi(\cdot)$ is a memory update operation that updates the TAD Memory according to the indicator δ_k. t_k is the temporal

marker for $\texttt{Attr}_k$, which is separately maintained for each concept and set to 0 for the first stored memory item (attribute). It is incremented by 1 and assigned to each new attribute that is added or updated. Consequently, TAD Memory stores each attribute in a triple tuple: $(\texttt{Cid}_k, \texttt{Attr}_k, t_k)$. If $\texttt{Cid}_k$ is determined to be novel, then the concept set is updated to $\mathcal{C}_i = \{c_1, c_2, \cdots, c_{K_{i-1}}, c_{K_i}\}$.

3.2 Open-Set Concept Reconciliation

When starting a new dialogue $\mathcal{Q}_i$, the assistant is required to identify the concept appearing in user query q_i, resolving whether it is novel or existing and giving its appropriate identifier. Specifically, the CR Skill first retrieves a set of candidate personalized concepts by comparing the visual [2] and textual information extracted from q_i with the information of each concept c_k stored in the TAD Memory in a **temporally weighted manner**, emphasizing the memory items that are closer to the q_i:

$$S(c_k \mid q_i) = \frac{1}{Z_k} \sum_{j=1}^{N_k} \exp\left(-\lambda \cdot \frac{t_k^{\max} - t_k^{(j)}}{N_k}\right) \cdot s_k^{(j)}, \tag{3}$$

$$s_k^{(j)} = \cos(\phi(\mathcal{T}_i), \phi(\mathcal{T}_k^{(j)})) + \cos(\phi(\mathcal{V}_i), \phi(\mathcal{V}_k^{(j)})). \tag{4}$$

Here, $S(c_k \mid q_i)$ denotes the temporally weighted similarity score between the information extracted from q_i and the N_k historical memory items associated with the personalized concept c_k in $\mathcal{M}_T$: $\{(\mathcal{T}_k^{(j)}, \mathcal{V}_k^{(j)}, t_k^{(j)})\}_{j=1}^{N_k}$. $\mathcal{T}_i$ and $\mathcal{V}_i$ represent the textual and visual information extracted from the incoming user query q_i, while $(\mathcal{T}_k^{(j)}, \mathcal{V}_k^{(j)})$ correspond to the textual and visual attributes stored in the j-th memory item of concept c_k. $t_k^{(j)}$ is the temporal marker for that item, and $t_k^{\max} = \max_j t_k^{(j)}$ denotes the most recent temporal marker of attributes for concept c_k. $s_k^{(j)}$ is the similarity score between the current query and the j-th memory item, computed using cosine similarity in the embedding space defined by $\phi(\cdot)$. λ is the temporal decay coefficient, and Z_k is the normalization term ensuring a scaled concept-level similarity score, defined as:

$$Z_k = \sum_{j=1}^{N_k} \exp\left(-\lambda \cdot \frac{t_k^{\max} - t_k^{(j)}}{N_k}\right). \tag{5}$$

Subsequently, the CR skill selects the Top-K candidate personalized concepts $\mathcal{C}_K$ with the highest similarity scores from the grasped concept set $\mathcal{C}_{i-1}$ after time step $i-1$. For all candidates, the assistant feeds their most recent memory items, together with the current query q_i and the most recent W dialogue turns, into the MLLM to determine whether the referred concept corresponds to an existing personalized concept or a novel one:

$$\texttt{Cid}^* = \Phi\left(\{\mathcal{M}_k^{\mathrm{r}}\}_{c_k \in \mathcal{C}_K}, q_i, \langle \mathcal{Q}_{i-W}, \ldots, \mathcal{Q}_{i-1} \rangle, \mathcal{P}_r\right). \tag{6}$$

Here, $\texttt{Cid}^*$ denotes the concept identifier predicted by the MLLM for the personalized concept referred to in the incoming dialogue $\mathcal{Q}_i$. The output $\texttt{Cid}^*$ can either correspond to an existing concept identifier in the candidate concept set $\mathcal{C}_K$ or indicate a novel concept, which will be initialized with a new identifier. $\mathcal{M}_k^{\mathrm{r}}$ represents the recent R memory items associated with concept c_k, which reflects its latest attribute state. q_i denotes the current user query in the dialogue turn, and $\langle \mathcal{Q}_{i-W}, \ldots, \mathcal{Q}_{i-1} \rangle$ denotes the most recent W dialogue turns serving as recent conversational context. $\mathcal{P}_r$ is the prompt that instructs the MLLM to perform concept reconciliation and to output either an **existing concept identifier** or a **novel one**. Consequently, the CR skill completes the concept reconciliation.

3.3 Plug-and-Play Operation

We take the integration of TAME-O (i.e., the training-free CR Skill) into TAME as an example. Specifically, we retain TAME's static memory $\mathcal{M}_S$ while replacing its dynamic memory $\mathcal{M}_D$ with TAD Memory $\mathcal{M}_T$. After obtaining the determined concept identifier $\texttt{Cid}^*$ in $\mathcal{Q}_i$ in Eq. (6), if the concept is **an existing one**, the assistant leverages this identifier to locate relevant memory items from both memories, and aggregates the corresponding information for personalized response generation (following the original TAME pipeline). If the concept is determined to be **novel**, a new identifier will be assigned to this concept and stored in the TAD Memory, therefore **unlocking new concept**

[2] We first perform a visual grounding to identify the concept in the query (cf. Implementation Details)

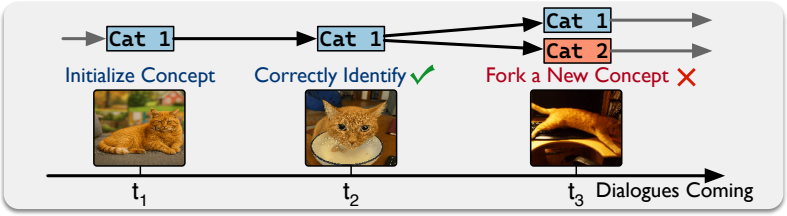


Figure 3: A concept fission example for concept `Cat 1`.

automatically. Notably, when integrated with TAME, we also incorporate information from the static memory for candidate concept retrieval. The final similarity score for each candidate concept is obtained by averaging the similarity scores from the static memory and the TAD Memory.

3.4 Discussion of Concept Fission and Its Mitigation

Although the CR Skill effectively resolves novel and existing concepts during ongoing interactions, it may still misidentify an existing concept as a novel one. In such cases, the original concept unit is unintentionally forked into multiple concept units with different identifiers, a phenomenon we refer to as **concept fission**. This leads to memory chaos, where no single concept unit retains complete contextual information, ultimately resulting in degraded personalization performance. To stably mitigate this issue, we introduce a Periodic Concept Fusion Strategy (**PCF Strategy**), which is performed every fixed interval of T interaction steps. Specifically, after each interval T, we compute pairwise similarities between all personalized concepts based on their stored memory items, and select the Top-L most similar concept pairs as candidates for fusion:

$$S(c_m, c_n) = \frac{1}{|\mathcal{M}_m|\,|\mathcal{M}_n|} \sum_{i=1}^{|\mathcal{M}_m|} \sum_{j=1}^{|\mathcal{M}_n|} s_{m,n}^{(i,j)}, \tag{7}$$

$$s_{m,n}^{(i,j)} = \cos(\phi(\mathcal{T}_m^{(i)}), \phi(\mathcal{T}_n^{(j)})) + \cos(\phi(\mathcal{V}_m^{(i)}), \phi(\mathcal{V}_n^{(j)})). \tag{8}$$

Here, $S(c_m, c_n)$ is the pairwise similarity between two concepts c_m and c_n. $\mathcal{M}_m$ and $\mathcal{M}_n$ represent the memory sets associated with concepts c_m and c_n, respectively. $s_{m,n}^{(i,j)}$ is the similarity between the i-th memory item of concept c_m and the j-th memory item of concept c_n, computed using cosine similarity in the embedding space defined by $\phi(\cdot)$. The overall concept-level similarity $S(c_m, c_n)$ is obtained by averaging the similarities over all memory item pairs, yielding a robust measure of semantic closeness between the two concepts. Subsequently, the Top-L concept pairs are derived:

$$\mathcal{F} = \text{Top-}L\Big\{(c_m, c_n) \;\Big|\; S(c_m, c_n)\Big\}_{m \neq n}. \tag{9}$$

Afterward, we further introduce a similarity threshold τ to identify the most potentially forked concept pairs in $\mathcal{F}$, and leverage the MLLM assistant with prompt $\mathcal{P}_f$ to evaluate each concept pair to perform the fusion and then update the memories. The prompts $\mathcal{P}_r$ and $\mathcal{P}_f$ are presented in Appendix J, while the pseudocode is provided in Appendix A. The complexity analysis of TAME-O is in Appendix B.

4 Experiments

We provide a brief version of the experimental settings, with a detailed version in Appendix C.

Benchmarks. Since open-set personalization is inherently intertwined with long-context dialogues, we first adopt an existing long-context personalized MLLM benchmark **LCMP-H** [4], which consists of 30 personalized concepts (Pet (10), Object (10), and Person (10)), 420 historical dialogue turns, 720 personalized images, and 180 hard questions. In addition, we introduce a new benchmark tailored for long-context open-set personalization, termed **LCMP-O**. Compared with LCMP-H, LCMP-O poses a more challenging open-set scenario by incorporating a *larger number* of personalized concepts (45 concepts) with more *visually similar appearances*. Specifically, its concepts are evenly split into Pet (15), Object (15), and Person (15), where pets and objects are organized into groups of three concepts sharing the same breed or object class, and persons are selected with similar ages and an approximately 1:1 gender ratio. Following LCMP [4], we generate 10 historical dialogue turns for

each concept to reveal its long-term attributes and short-term events, as well as 5 hard questions requiring reasoning over both types of attributes. Overall, LCMP-O contains 450 historical dialogue turns, 675 personalized images, and 225 hard questions, with more details in Section C.2.

Open-Set Personalization Setup. To reflect the open-set setting, we adopt a **concept-incremental evaluation protocol**. Specifically, personalized concepts are introduced to MLLMs sequentially rather than all at once. For each concept, the MLLM is first exposed to its associated historical dialogues in a *temporally sequential manner*, one dialogue turn at a time step [3]. After all dialogue turns for the concept have been processed, the MLLM is instantly evaluated on the corresponding QA set tied to that concept. Moreover, in contrast to [4], the concept identifiers of historical dialogues under the open-set setting **are not** provided to the assistant. Instead, the assistant is required to autonomously and incrementally unlock new concepts.

Baselines. We first compare TAME-O against several state-unaware MLLM personalization baselines, including Backbone (the native MLLM without personalization adaptation), MyVLM [1], Yo'LLaVA [2], RAP [8], PeKit [17], and R2P [7]. Notably, we do not apply the open-set setting to these baselines, as their state-unaware design prevents them from directly supporting unlocking novel concepts automatically across interactions. In addition, we construct a new open-set personalization baseline, TAME†, by extending the long-context baseline TAME [4] with a simple retrieval-based concept unlocking and resolution mechanism. Specifically, TAME† retrieves candidate concepts from the memory and treats the Top-1 retrieved concept as the matched existing concept, if its similarity exceeds a predefined threshold. Otherwise, the concept is considered novel. We further experiment with TAME† w/ MLLM decision (i.e., the w/o TAD Memory variant in Table 2).

Evaluation Metrics. Following LCMP, we employ evaluation metrics: **(1)** Accuracy of Choice questions (**ACC-C**), **(2)** Accuracy of Free-text questions (**ACC-F**), **(3)** Scoring Point Rate Long-term attributes (**SPR-L**), and **(4)** Scoring Point Rate Short-term attributes (**SPR-S**), which measure the proportion of responses that explicitly reference the relevant attribute variation of the personalized concept. Moreover, we propose **(5)** Concept Reconciliation Accuracy (**CRA**) as a new metric for evaluating the open-set personalization capability. Formally, CRA can be defined as:

$$\mathrm{CRA} = \frac{1}{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{H}|} \mathbb{I}\big(\hat{\texttt{Cid}}_i = \texttt{Cid}_i\big)\,, \tag{10}$$

where $\mathcal{H}$ is the set of historical dialogues, $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the prediction is correct and 0 otherwise, $\hat{\texttt{Cid}}_i$ is the concept identifier determined by the MLLM assistant in i-th historical dialogue. $\texttt{Cid}_i$ denotes the ground-truth concept identifier for the i-th dialogue. If the referred concept is an existing one (already grasped), $\texttt{Cid}_i$ corresponds to the identifier of that concept. If it is novel, $\hat{\texttt{Cid}}_i$ is expected to be a previously unused identifier. If an existing concept has been forked into multiple identifiers (concept fission), any predicted identifier $\hat{\texttt{Cid}}_i$ that corresponds to one of the forked identifiers is treated as a correct match. Moreover, we propose the metric **(6)** Concept Fission Rate (**CFR**) to measure the fission rate of the concepts at different time steps t:

$$\mathrm{CFR}(t) = \frac{\sum_{c_i \in \mathcal{C}_t} \max\Big(0,\ |\hat{\mathcal{I}}_{c_i}^{(t)}| - 1\Big)}{\max(t - |\mathcal{C}_t|, \epsilon)}, \tag{11}$$

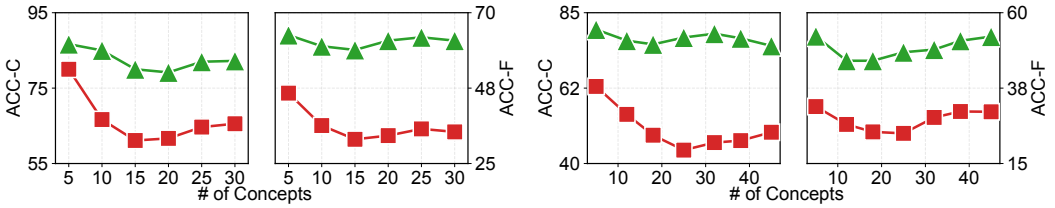
where $\mathcal{C}_t = \{c_1, \ldots, c_{N_t}\}$ denotes the set of ground-truth personalized concepts that exist at time step t, $\hat{\mathcal{I}}_{c_i}^{(t)}$ is the set of concept identifiers derived by the MLLM assistant at time step t associated with the ground-truth concept c_i, which contains the correct identifier and may unintentionally include forked identifiers introduced due to concept fission. As a result, $\mathrm{CFR}(t)$ is designed to characterize the temporal evolution of concept fission degree during long-term interactions.

Implementation Details. We adopt the MLLM `Qwen3.5-9B` [6] as our main backbone, while also evaluating TAME-O on another MLLM `InternVL3.5-8B` [18] for generalizability. Since TAME-O is plug-and-play, we adopt TAME as the base personalized assistant for evaluation. In addition, we also evaluate its scalability on other MLLM personalization methods to unlock their open-set capability (cf. Appendix F). When visually grounding the concept in user's query, we first prompt the MLLM to extract open-set category descriptions from the query, and then use the extracted categories to perform grounding on the query image with `Grounding-DINO` [19]. If no concept is

[3] We define time steps globally across concepts, where each dialogue turn constitutes a time step.

Table 1: Performance comparison on the LCMP-H and LCMP-O datasets. The best results are in **black bold**. Higher values of all metrics indicate better performance. TAME† is the modified TAME.

Methods	MLLM Backbone	Training Free	State Aware	Open Set	LCMP-H					LCMP-O				
					ACC-C	ACC-F	SPR-L	SPR-S	CRA	ACC-C	ACC-F	SPR-L	SPR-S	CRA
Backbone	InternVL3.5-8B	✓	✗	✗	40.00	20.00	12.78	26.11	—	32.00	18.67	16.44	7.11	—
	Qwen3.5-9B				36.11	20.56	17.22	32.78	—	38.67	22.22	29.78	14.67	—
MyVLM	LLaVA-v1.6-7B	✗	✗	✗	40.00	13.33	8.89	12.22	—	46.67	21.33	18.22	6.67	—
Yo'LLaVA	LLaVA-v1.5-13B	✗	✗	✗	38.89	9.44	8.89	17.22	—	56.44	6.67	20.89	9.78	—
RAP	LLaVA-v1.5-13B	✗	✗	✗	41.11	4.44	33.33	23.89	—	34.22	20.89	19.11	17.33	—
	Phi3-V-3.8B				46.67	9.44	15.00	20.00	—	51.56	22.22	28.89	15.11	—
PeKit	InternVL3.5-8B	✓	✗	✗	61.11	49.44	21.11	35.00	—	52.89	40.89	29.78	20.00	—
	Qwen3.5-9B				68.33	56.67	35.56	48.33	—	56.00	33.78	19.56	10.22	—
R2P	InternVL3.5-8B	✓	✗	✗	52.78	35.56	8.89	25.00	—	65.78	40.89	10.22	26.22	—
	Qwen3.5-9B				66.67	55.56	45.00	48.33	—	60.89	41.78	20.00	36.44	—
TAME†	InternVL3.5-8B	✓	✓	✓	53.89	39.44	22.78	42.78	33.08	58.22	19.11	34.67	15.11	28.73
	Qwen3.5-9B				65.56	34.44	22.22	42.22	34.59	49.33	32.44	50.67	27.11	31.63
TAME-O	InternVL3.5-8B	✓	✓	✓	75.00	**61.67**	51.11	57.78	69.17	70.67	38.22	56.44	51.11	66.37
	Qwen3.5-9B				**82.22**	**61.67**	**64.44**	**70.00**	**72.18**	**75.11**	**52.89**	**72.00**	**59.11**	**71.49**



(a) LCMP-H. (b) LCMP-O.

Figure 4: Performance comparison between TAME† and TAME-O with respect to a continually increasing number of concepts. ■ denotes TAME†, while ▲ denotes TAME-O.

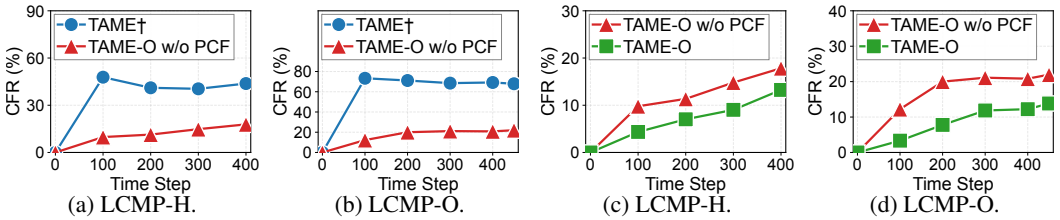
Table 2: Ablation study on core components in TAME-O with Qwen3.5-9B.

Module	Variant	LCMP-H					LCMP-O				
		ACC-C	ACC-F	SPR-L	SPR-S	CRA	ACC-C	ACC-F	SPR-L	SPR-S	CRA
CR Skill	w/o TAD Memory	77.78	57.22	60.56	68.89	69.67	70.67	48.44	68.44	53.33	66.59
	w/o MLLM Decision	72.78	48.89	40.00	52.22	51.13	46.22	37.78	52.44	29.33	46.33
	w/o CR Skill	66.11	38.33	24.44	45.56	36.84	52.44	34.67	50.67	29.33	33.41
PCF Strategy	w/o PCF Strategy	78.89	59.44	61.11	67.22	69.92	73.78	47.56	70.22	54.67	68.15
TAME-O	**ALL**	**82.22**	**61.67**	**64.44**	**70.00**	**72.18**	**75.11**	**52.89**	**72.00**	**59.11**	**71.49**

grounded, we provide the whole image to the MLLMs for concept reconciliation. In addition, we adopt `Qwen3-VL-Embedding-2B` [20] to generate embeddings for both visual and textual inputs.

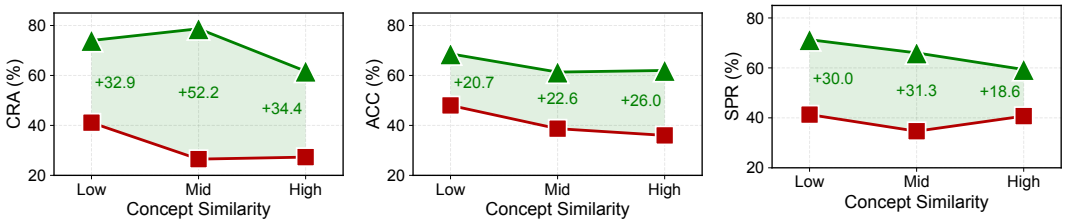
4.1 Open-Set Personalization Performance

From the results in Table 1, we observe that: **(O1)** Training-based methods, such as MyVLM and Yo'LLaVA, perform poorly on both benchmarks, as encoding long-context personalized information with a limited number of lightweight parameters is insufficient to capture extended interaction histories. In contrast, RAG-based, training-free methods such as R2P and PeKit achieve stronger performance by storing personalized concepts in external memory while keeping model parameters fixed. However, all these baselines operate under a closed-set assumption and lack the ability to automatically incorporate novel personalized concepts. **(O2)** TAME is the SOTA baseline for long-context personalization. However, under the open-set setting, TAME† exhibits suboptimal performance, since simply treating all historical attributes equally fails to identify appropriate matched concepts. Moreover, the concept fission problem further undermines the effectiveness of TAME†. In contrast, TAME-O achieves the best performance for two reasons: the CR Skill accurately reconciles incoming personalized concepts by treating the stored concept attributes in a temporally weighted manner, emphasizing the most recently observed concept information. Moreover, the PCF Strategy further stably fuses fissioned concepts, mitigating the concept fission problem and enabling the better localization of sufficient contextual information for each concept. Furthermore, Figure 4 shows that TAME-O maintains stable and effective open-set personalization performance as the total number of concepts continually increases.



(a) LCMP-H. (b) LCMP-O. (c) LCMP-H. (d) LCMP-O.

Figure 5: Comparison of CFR among TAME†, TAME-O w/o PCF Strategy, and TAME-O.



(a) Reconciliation accuracy. (b) QA accuracy. (c) Scoring point rate.

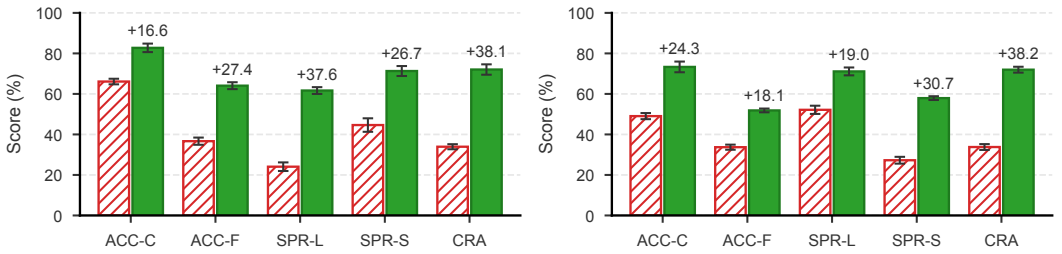
Figure 6: Performance on LCMP-O with concepts stratified by confusability. ACC and SPR average ACC-C/ACC-F and SPR-L/SPR-S, respectively. ■ denotes TAME†, while ▲ denotes TAME-O.

4.2 Further Analysis

Ablative Analysis. We conduct comprehensive ablations on the core components within TAME-O, with results in Table 2. For **CR Skill**, we evaluate the variants: **(1) w/o TAD Memory**, where the Time-Aware Dynamic (TAD) Memory is reduced to a static memory and all temporal markers associated with concept attributes are removed; **(2) w/o MLLM Decision**, where the Top-1 retrieved concept is directly treated as the matched concept if its similarity score exceeds a predefined threshold, and otherwise regarded as novel; **(3) w/o CR Skill**, where both the TAD Memory and the MLLM-based decision mechanism are removed, and a simple time-unaware retrieval strategy is applied, with the Top-1 retrieved concept always selected as the matched existing concept if it exceeds a predefined threshold. We observe that variants (1)–(3) exhibit clear performance drops, indicating that the accurate reconciliation of novel and existing user-specific concepts is vital for open-set personalization. For **PCF Strategy**, we entirely remove it to design a variant: **w/o PCF Strategy**. We observe that this variant incurs a clear performance drop, as the degree of concept fission becomes increasingly severe over time without any mitigation, leading to memory chaos.

Concept Fission and Fusion Analysis. We quantitatively evaluate the severity of the concept fission and how well TAME-O mitigates the fission problem. Specifically, we report the Concept Fission Rate (CFR), as defined in Eq. (11), over time steps on the historical dialogues of LCMP-H and LCMP-O. From Figure 5(a)–(b), we observe that the CFR of TAME† rapidly reaches 47.8% and 73.3% on LCMP-H and LCMP-O and then remains high, while that of TAME-O w/o PCF Strategy gradually increases. Nevertheless, TAME-O w/o PCF Strategy has a much lower CFR than TAME†, since the CR skill mitigates concept fission at its source by more accurately reconciling incoming open-set personalized concepts with existing ones. From Figure 5(c)–(d), we observe that with the PCF strategy equipped, the CFR never exceeds that without it and finally drops from 17.8% to 13.2% (LCMP-H) and from 22.0% to 13.8% (LCMP-O), indicating that the PCF strategy effectively fuses potentially forked concepts. Moreover, we empirically observe that the PCF Strategy performs fusion with relatively high accuracy, achieving a strong fusion success rate (cf. Appendix E for details).

Robustness to Concept Confusability. Since visually similar personalized concepts are where open-set reconciliation is most likely to fail, we further break down the performance on LCMP-O by concept confusability. Specifically, we measure the confusability of each concept by its nearest-neighbour similarity to the other concepts, computed with the same encoder and text-plus-visual similarity score as the PCF Strategy (Eq. (7)), and split the 45 concepts into three equal-sized tertiles (Low, Mid, and High). As shown in Figure 6, TAME-O outperforms TAME† on every metric in every similarity bin, with CRA gains of 32.9–52.2 points. Moreover, TAME-O maintains relatively stable performance as concept confusability increases, keeping its CRA above 60 even in the most confusable bin. In contrast, the CRA of TAME† already drops sharply at medium similarity, as its threshold-based Top-1 matching is easily misled by similar concepts.



(a) LCMP-H. (b) LCMP-O.

Figure 7: Performance (mean ± std) across five concept appearance orders. Numbers above the bars are the gains over TAME†. ▨ denotes TAME†, while ■ denotes TAME-O.

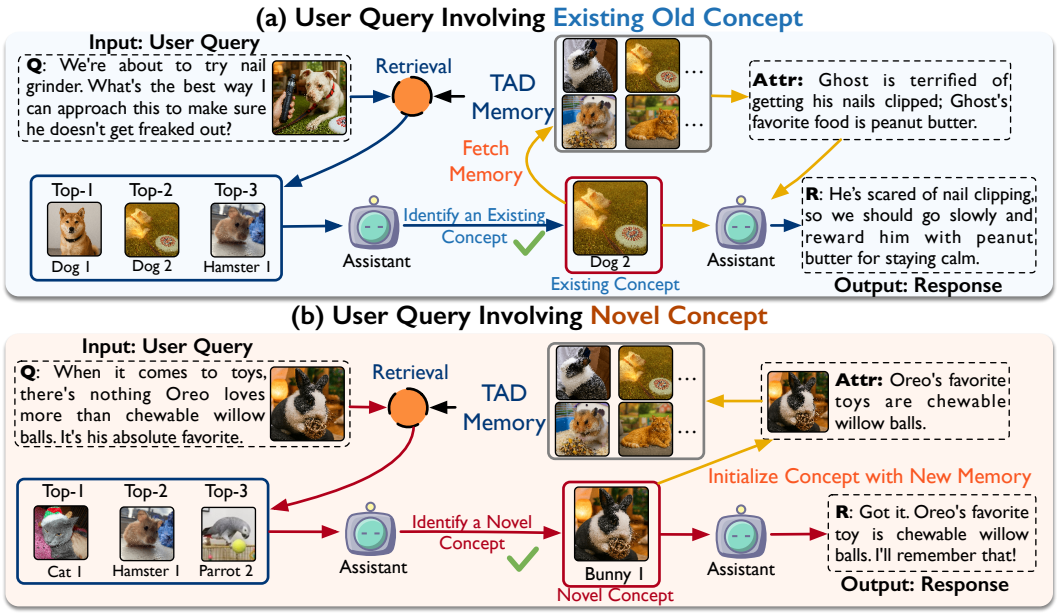


Figure 8: Case study on resolving an existing concept (left) and a novel concept (right).

Stability across Concept Appearance Orders. Under the open-set setup, personalized concepts emerge sequentially. As a result, different appearance orders lead to different sets of previously acquired concepts in memory for the same incoming concept. This, in turn, affects the retrieved candidate concepts and the difficulty of determining whether the incoming concept is existing or novel. Here, we report the mean performance and variability of TAME-O and TAME† across five concept appearance orders in Figure 7. We observe that TAME-O maintains stable performance across different concept appearance orders and consistently outperforms TAME†.

Case Study: Old and Novel Concept Reconciliation. In Figure 8, we visualize two cases to showcase the procedure of TAME-O in tackling the open-set personalization problem. On the top, the incoming concept is correctly identified by the CR skill as an existing one (`Dog 2`), after which the relevant memory items are retrieved for the personalized response. On the bottom, the CR skill correctly identifies the incoming concept as novel, responds without retrieving existing memories, and unlocks this new concept with an initialized identifier `Bunny 1` in the TAD Memory.

5 Conclusion

In this study, we pioneered the research in long-context MLLM personalization into the open-set setting. To this end, we proposed TAME-O, which automatically tames novel personalized concepts from everyday interactions while resolving potential conflicts with previously memorized ones. Extensive experiments in the open-set setting on LCMP-H and LCMP-O benchmarks demonstrate that TAME-O effectively integrates incoming personalized concepts even under visually similar concepts, mitigates concept fission, and remains robust to varying concept arrival orders.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62572097). This work was also supported by the Center for HPC, University of Electronic Science and Technology of China. The authors declare no competing interests.

References

[1] 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)*, pages 73–91. Springer, 2024.

[2] 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.

[3] Jaeik Kim, Woojin Kim, Woohyeon Park, and Jaeyoung Do. Mmpb: It's time for multi-modal personalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[4] 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.

[5] OpenAI. Gpt-5 system card. https://openai.com/index/gpt-5-system-card, August 2025.

[6] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

[7] 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.

[8] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. RAP: Retrieval-Augmented Personalization for Multimodal Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[9] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[10] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. UniCTokens: Boosting personalized understanding and generation via unified concept tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[11] Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'chameleon: Personalized vision and language generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 14438–14448, 2025.

[12] 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.

[13] Ruichuan An, Sihan Yang, Renrui Zhang, Ming Lu, Tianyi Jiang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, Shanghang Zhang, and Wentao Zhang. MC-LLaVA: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

[14] Ruichuan An, Kai Zeng, Ming Lu, Sihan Yang, Renrui Zhang, Huitong Ji, Hao Liang, and Wentao Zhang. Concept-as-Tree: A controllable synthetic data framework makes stronger personalized VLMs. In *European Conference on Computer Vision (ECCV)*, 2026.

[15] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized Visual Instruction Tuning. In *International Conference on Learning Representations (ICLR)*, 2025.

[16] Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. PLVM: A tuning-free approach for personalized large vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3671–3680, 2025.

[17] Soroush Seifi, Vaggelis Dorovatas, Matteo Cassinelli, Fabien Despinoy, Daniel Olmeda Reino, and Rahaf Aljundi. Personalization Toolkit: Training Free Personalization of Large Vision Language Models. *Transactions on Machine Learning Research (TMLR)*, 2026.

[18] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision (ECCV)*, pages 38–55, 2024.

[20] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

[21] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3498–3505, 2012.

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023.

[24] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm Agents Are Experiential Learners. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 19632–19642, 2024.

[25] Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A Survey on the Memory Mechanism of Large Language Model-based Agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025.

[26] Zeyuan Liu, Jeonghye Kim, Xufang Luo, Dongsheng Li, and Yuqing Yang. Exploratory Memory-Augmented LLM Agent via Hybrid On- and Off-Policy Optimization. In *International Conference on Learning Representations (ICLR)*, 2026.

[27] A. Maharana, Dong-Ho Lee, S. 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)*, pages 13851–13870, 2024.

[28] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-Memory: Recalling and Post-thinking Enable LLMs with Long-Term Memory. *arXiv preprint arXiv:2311.08719*, 2023.

[29] 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)*, pages 19724–19731, 2024.

[30] Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco De Nadai, and Hugues Bouchard. Bridging Search and Recommendation in Generative Retrieval: Does One Task Help the Other? In *ACM Conference on Recommender Systems (RecSys)*, pages 340–349. ACM, 2024.

[31] Mufan Xu, Gewen Liang, Kehai Chen, Wei Wang, Xun Zhou, Muyun Yang, Tiejun Zhao, and Min Zhang. Memory-augmented Query Reconstruction for LLM-based Knowledge Graph Reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, pages 24068–24084, 2025.

[32] Yibo Wang, Yong Rui, and Min-Ling Zhang. Classifier-induced Reciprocal Points for Multi-label Open-set Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 26643–26651, 2026.

[33] Yimin Fu, Zhunga Liu, and Jialin Lyu. Reason and Discovery: A New Paradigm for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(7):5586–5599, 2025.

[34] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(7):1757–1772, 2012.

[35] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 4401–4410, 2021.

[36] Baile Xu, Furao Shen, and Jian Zhao. Contrastive open set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 10546–10556, 2023.

[37] Jun-Yi Hang and Min-Ling Zhang. Binary decomposition: A problem transformation perspective for open-set semi-supervised learning. In *International Conference on Machine Learning (ICML)*, 2024.

[38] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 2307–2316, 2019.

[39] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13480–13489, 2020.

[40] Yunrui Guo, Guglielmo Camporese, Wenjing Yang, Alessandro Sperduti, and Lamberto Ballan. Conditional variational capsule network for open set recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 103–111, 2021.

[41] Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. Taskbench: Benchmarking large language models for task automation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 4540–4574, 2024.

[42] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 15(3):1–45, 2024.

[43] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from Downstream and Be Yourself in Multimodal Large Language Models Fine-Tuning. In *International Conference on Machine Learning (ICML)*, 2025.

[44] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *International Conference on Machine Learning (ICML)*, 2025.

[45] Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual Attention Never Fades: Selective Progressive Attention ReCalibration for Detailed Image Captioning in Multimodal Large Language Models. In *International Conference on Machine Learning (ICML)*, 2025.

[46] Tzu-Tao Chang and Shivaram Venkataraman. LV-XAttn: Distributed Cross-Attention for Long Visual Inputs in Multimodal Large Language Models. In *International Conference on Machine Learning (ICML)*, 2025.

[47] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

[48] Chia-Hao Kao, Cheng Chien, Yu-Jen Tseng, Yi-Hsin Chen, Alessandro Gnutti, Shao-Yuan Lo, Wen-Hsiao Peng, and Riccardo Leonardi. Bridging Compressed Image Latents and Multimodal Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

[49] Tianle Gu, Zeyang Zhou, Kexin Huang, Dandan Liang, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Xingge Qiao, Keqing Wang, Yujiu Yang, Yan Teng, Yu Qiao, and Yingchun Wang. Mllmguard: A Multi-dimensional Safety Evaluation Suite for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[50] Ziqiang Liu, Feiteng Fang, Xi Feng, Xeron Du, Chenhao Zhang, Noah Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. Ii-Bench: An Image Implication Understanding Benchmark for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[51] Jian Lang, Rongpei Hong, Ting Zhong, Leiting Chen, Qiang Gao, and Fan Zhou. From shallow humor to metaphor: Towards label-free harmful meme detection via lmm agent self-improvement. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 588–599, 2026.

[52] FENGYUN WANG, Sicheng Yu, Jiawei Wu, Jinhui Tang, Hanwang Zhang, and Qianru Sun. 3d question answering via only 2d vision-language models. In *International Conference on Machine Learning (ICML)*, 2025.

[53] Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *International Conference on Machine Learning (ICML)*, 2025.

[54] Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. Reasoning multi-agent behavioral topology for interactive autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 92605–92637, 2024.

[55] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7370–7392, 2024.

[56] Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. On the Way to LLM Personalization: Learning to Remember User Conversations. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 61–77. Association for Computational Linguistics, 2025.

[57] Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. PersoBench: Benchmarking Personalized Response Generation in Large Language Models. *arXiv preprint arXiv:2410.03198*, 2024.

[58] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213, 2018.

[59] Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10803–10812, 2022.

[60] 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)*, 2025.

[61] Dongshuo Liu, Zhijing Wu, Dandan Song, and Heyan Huang. A Persona-Aware LLM-Enhanced Framework for Multi-Session Personalized Dialogue Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 103–123, 2025.

Appendix

Content

A Pseudocode of TAME-O **17**

B Complexity Analysis of TAME-O **17**

B.1 Complexity of CR Skill 17

B.2 Complexity of PCF Strategy 17

C Detailed Experimental Settings **18**

C.1 Benchmark: LCMP-H 18

C.2 Benchmark: LCMP-O 19

C.3 Baselines 19

C.4 Implementation Details 20

D Hyper-Parameter Analysis **20**

E Quantitative Analysis of the PCF Strategy **20**

F Scalability across Other MLLM Personalized Assistants **21**

G Additional Literature Review **21**

G.1 Memory-Augmented LLMs 21

G.2 Open-Set Recognition 22

G.3 Multimodal Large Language Models 22

G.4 Large Language Model Personalization 23

H Additional Case Study **23**

H.1 Additional Cases of Concept Reconciliation 23

H.2 Cases of Concept Fusion 23

I Failure Case Analysis: Concept Fission **23**

J Prompt Presentation for TAME-O **24**

K Limitations **24**

L Broader Impacts **24**

M License of Assets **25**

Algorithm 1 Python-like Pseudocode for TAME-O

```
# Q_i: dialogue at time step i (q_i, a_i)
# q_i: current user query (before generation)
# M_T: Time-Aware Dynamic Memory (Cid -> [(Attr, t)])
# C: personalized concept set
# K: number of candidate concepts
# W: context window size
# T: fusion interval

# initialize memory and concept set
M_T = {}
C = set()

for i, Q_i in enumerate(dialogue_stream):

    # -------- Encode user query --------
    T_i, V_i = Encode(q_i)

    # -------- Concept Reconcile (CR Skill) --------
    scores = {}
    for c_k in C:
        # temporally weighted retrieval using timestamps t_k^(j), each item
            is (Attr, t)
        scores[c_k] = TemporalWeightedSimilarity(T_i, V_i, M_T[c_k])

    # select Top-K candidate concepts
    C_K = TopK(scores, K)

    # MLLM-based concept reconciliation (Eq. 5)
    Cid* = MLLM_Reconcile(recent_memory(C_K, M_T), q_i, recent_dialogues(W))

    # decide whether concept is novel or existing
    if Cid* == "novel":
        Cid = NewConceptID()
        C.add(Cid)
        M_T[Cid] = [] # initialize memory list
    else:
        Cid = Cid*

    # -------- Personalized Response Generation --------
    # retrieve relevant memory items using resolved Cid
    context = RetrieveMemory(M_T, Cid)
    a_i = MLLM_Generate(q_i, context)

    # -------- Update Time-Aware Dynamic Memory --------
    # delta indicates add / modify of an attribute
    Cid, Attr, delta = MLLM_Update(Cid, Q_i, M_T)

    # update temporal marker for this concept
    t_new = GetNextTimestamp(M_T[Cid])
    M_T = MemoryUpdate(M_T, Cid, Attr, delta, t_new)

    # -------- Periodic Concept Fusion (PCF Strategy) --------
    if i % T == 0:
        # compute pairwise concept similarity
        concept_pairs = PairwiseSimilarity(M_T)
        fusion_pairs = TopL(concept_pairs)

        for (c_m, c_n) in fusion_pairs:
            M_T = MLLM_Fuse(c_m, c_n, M_T, \tau)
```

A Pseudocode of TAME-O

A concise version of Python-style pseudocode for TAME-O is offered in Algorithm 1.

B Complexity Analysis of TAME-O

We delve into the complexity analysis of TAME-O. We first provide a definition of the complexity of each MLLM invocation:

Definition B.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:*

$$\text{Attention}(X) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{12}$$

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.

We then analyze the computational complexity of TAME-O by decomposing it into two major components: (1) the Concept Reconciliation Skill (CR Skill), and (2) the Periodic Concept Fusion Strategy (PCF Strategy).

B.1 Complexity of CR Skill

The CR Skill consists of two stages: temporally weighted candidate retrieval and MLLM-based concept resolution.

Temporally Weighted Retrieval. At time step i, the assistant maintains a dynamic concept set $\mathcal{C}_{i-1}$. Each concept c_k is associated with N_k memory items stored in the TAD Memory. For an incoming query q_i, the CR Skill computes the temporally weighted similarity score $S(c_k \mid q_i)$ by comparing the query representation with all memory items of c_k. The cost of computing a single cosine similarity in the CLIP embedding space is $O(d)$. Therefore, the total cost of temporally weighted retrieval over all concepts is: $O\Big(\sum_{k=1}^{|\mathcal{C}_{i-1}|} N_k \cdot d\Big) = O(|\mathcal{M}_T| \cdot d)$, where $|\mathcal{M}_T|$ denotes the total number of memory items in the TAD Memory. In practice, N_k is bounded to a constant, making this step linear in memory size.

MLLM-based Concept Reconciliation. After retrieval, only the Top-K candidate concepts are retained, where $K \ll |\mathcal{C}_{i-1}|$. The CR Skill then jointly feeds the most recent memory items of these candidate concepts, together with the current user query and a fixed window of recent dialogue turns, into the MLLM for final concept resolution. Importantly, this reconciliation step is performed via a single MLLM invocation, rather than multiple independent calls per candidate. Since the number of candidate concepts K, the number of memory items per candidate, and the dialogue context window size are all strictly bounded, the total input sequence length to the MLLM is also bounded. Consequently, the MLLM-based concept resolution involves a bounded-length attention computation. According to the attention complexity, this step incurs a cost of $O(ld^2 + l^2 d)$, where l is the total token length of the fixed prompt, candidate memory items, and recent dialogue context. As l is independent of the total number of discovered concepts $|\mathcal{C}_{i-1}|$, this step introduces a bounded computational overhead.

B.2 Complexity of PCF Strategy

The PCF Strategy is executed periodically every T interaction steps, rather than at every step. At each execution, the strategy computes pairwise similarities between personalized concepts to identify potentially forked ones.



Figure 9: Cases of personalized concepts from LCMP benchmark.

For each concept pair (c_m, c_n), the similarity $S(c_m, c_n)$ is computed by averaging cosine similarities over all pairs of memory items from $\mathcal{M}_m$ and $\mathcal{M}_n$. The cost of computing the similarity for one concept pair is: $O(|\mathcal{M}_m| \cdot |\mathcal{M}_n| \cdot d)$. Subsequently, computing similarities for all concept pairs yields: $O\Big(\sum_{(c_m,c_n)\in\mathcal{C}\times\mathcal{C},\, m\neq n} |\mathcal{M}_m| \cdot |\mathcal{M}_n| \cdot d\Big)$. For each selected pair, the actual fusion decision is delegated to the MLLM. Similar to the CR Skill, this involves a bounded-length attention computation and incurs a constant overhead. Moreover, since we restrict the number of candidate pairs for concept fusion to the Top-L most similar ones that also exceed a predefined similarity threshold, the overall computational complexity of this step is strictly bounded. **In practice**, we observe that TAME-O incurs about 6.5 seconds runtime cost per user-MLLM interaction, which is acceptable. And baseline TAME† requires about 5 seconds per interaction.

C Detailed Experimental Settings

C.1 Benchmark: LCMP-H

Since open-set personalization is studied in long-context, real-world human–MLLM interactions, rather than in static and overly simplified settings (e.g., personalized image captioning), we first adopt the hard version of the LCMP benchmark (Long-Context MLLM Personalization) [4], a recent benchmark and the only available benchmark specifically designed for long-context MLLM personalization. Below we propose a detailed description of this benchmark.

LCMP is the first benchmark for evaluating personalized multimodal large language models in long-context interactions. It consists of 30 personalized concepts, including Pet (10), Object (10), and Person (10), along with approximately 420 historical dialogue turns, 240 easy questions, 180 hard questions, and 720 personalized images, and includes two evaluation sets: (1) **LCMP-E**, which contains easy VQA questions involving a single type of attribute (either short-term or long-term), and (2) **LCMP-H**, which contains hard questions requiring reasoning over both types of attributes. It measures an assistant's ability to perceive and model variations in both short-term and long-term attributes of user-specific concepts and to generate contextually appropriate responses based on extended dialogue histories. LCMP is constructed through three stages: (1) personalized concept creation, where raw images are sampled and endowed with fine-grained temporal attributes; (2) multi-turn vision–language dialogue generation, which simulates historical interactions that reveal concept attribute changes over time; and (3) evaluation question generation, which produces single-turn VQA tasks requiring reasoning over the preceding dialogue context. The benchmark thus explicitly tests long-context reasoning and personalization beyond static single-turn tasks. We adopt the hard version of the LCMP benchmark, namely LCMP-H, in this study for evaluation. As illustrated in Figure 9, we provide cases of personalized concepts from LCMP-H.

C.2 Benchmark: LCMP-O

To comprehensively evaluate the open-set MLLM personalization capability of the MLLM assistant, we construct a new benchmark: **O**pen-Set **L**ong-**C**ontext **M**LLM **P**ersonalization (**LCMP-O**). The original LCMP benchmark exhibits two limitations in evaluating open-set personalization performance: **(1)** LCMP contains a small number of personalized concepts, which constrains the diversity of the open-set evaluation. **(2)** The personalized concepts in LCMP are characterized by highly distinctive attributes, making it relatively easy for MLLMs to differentiate new concepts from previously grasped ones, softening the challenges of open-set personalization.

Taking the above limitations into consideration, we construct LCMP-O with 45 personalized concepts, deliberately designed to present high intra-class visual similarity and thus pose stronger challenges for open-set personalization evaluation. Following the construction pipeline of LCMP [4], we generate 10 historical dialogue turns and 5 hard evaluation questions per concept, yielding a total of 450 dialogues and 225 evaluation questions.

The 45 concepts span three types, each chosen to maximize within-type visual ambiguity: (1) 15 Pets (cats, dogs, and rabbits), where every three concepts belong to the same breed, making fine-grained visual disambiguation particularly challenging. These cats and dogs concepts are sourced from the Oxford-IIIT-Pet dataset [21], which covers 37 breeds with substantial intra-breed variation in pose, lighting, and deformation. And the rabbits concepts are collected by authors. (2) 15 Persons, sampled from subjects of similar age with an approximately 1:1 gender ratio. Person images are drawn from MS-COCO [22], following the same protocol as TAME [4]. (3) 15 Objects, where every three concepts belong to the same object class. These concepts are sourced from PerVA [17], which contains object concepts that exhibit high visual and semantic similarity. In Figure 13, we provide cases of concepts from the LCMP-O benchmark.

C.3 Baselines

Since open-set MLLM personalization is novel as a setting, we compare TAME-O with traditional closed-set MLLM personalization baselines and a specifically designed baseline for the open-set setting. For traditional baselines, we select MyVLM [1], Yo'LLaVA [2], RAP [8], PeKit [17], and R2P [7], with detailed descriptions as follows:

- **MyVLM** pioneers the personalization of general VLMs so they can understand and talk about concepts unique to a given user. It augments LLaVA [23] with external concept heads that first detect a target entity and then inject a learned concept embedding into the model's mid-level features, guiding the decoder to mention the concept naturally.
- **Yo'LLaVA** frames subject personalization as learning a compact representation from only a few example images. It encodes the subject into a small set of latent prefix tokens that sit between the vision encoder and LLaVA's language decoder, binding visual attributes to text without touching backbone weights.
- **RAP** turns a generic MLLM into a user-aware assistant by adding an external memory layer. A key–value store "remembers" user facts and portraits, a multimodal retriever selects the most relevant keys at inference, and the generator conditions on both the query and retrieved context to tailor its answer.
- **PeKit** argues that strong personalization can be achieved entirely at inference time. It pairs off-the-shelf visual descriptors with retrieval-augmented generation and lightweight visual prompts, injecting the correct identity or object instance into any MLLM decoder without gradient updates.
- **R2P** shows that an MLLM can personalize itself using nothing but its pre-trained weights and a clever retrieval stack. It distills each personalized concept into a sparse fingerprint of discriminative attributes, retrieves candidate fingerprints at run time, and employs chain-of-thought reasoning to verify them before generation.

For open-set personalization method, we additionally create a competitive baseline **TAME**†, a modified variant of the long-context MLLM personalized assistant TAME [4]. TAME is the first long-context MLLM personalization method that explicitly models both the long-term and short-term variations of personalized concepts and retrieves reliable contextual information for each concept to support personalized response generation. We unlock the open-set capability of TAME by retrieving the most relevant concept stored in memory based on the visual and textual information in the query.

Table 3: Performance of TAME-O under varying values of hyper-parameters.

	Parameter λ					Parameter K				
LCMP-H	0	0.1	1	2	5	1	2	3	4	5
ACC-C (%)	78.33	80.56	82.22	82.44	78.67	74.44	82.78	82.22	81.11	80.00
	Parameter λ					Parameter K				
LCMP-O	0	0.1	1	2	5	1	2	3	4	5
ACC-C (%)	72.89	75.89	75.11	75.44	76.11	71.44	74.17	75.11	75.67	75.18

Table 4: CFA of TAME-O over successive time steps on different datasets.

	LCMP-H					**LCMP-O**				
Time Step (t)	50	100	200	300	400	50	100	200	300	400
CFA (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.74	95.65	96.67

A predefined similarity threshold is then applied: if the retrieved similarity exceeds the threshold, the concept is classified as existing; otherwise, it is treated as novel. We conduct a search over the threshold in the range $[0, 1]$ with an interval of 0.1, and select the best-performing value of 0.6 for all experiments with TAME†.

C.4 Implementation Details

We adopt the latest released MLLMs `Qwen3.5-9B` [6], and `InternVL3.5-8B` [18], as the MLLM backbone. Since TAME-O is a plug-and-play framework, we adopt TAME as the base personalized assistant for evaluation. When identifying the concept in user's query, we first prompt the MLLM to extract open-set category descriptions from the query, and then use the extracted categories to perform grounding on the query image with `Grounding-DINO` [19]. In addition, we adopt the `Qwen3-VL-Embedding-2B` [20] to generate embeddings for both visual and textual inputs. For the CR Skill, we set the temporal decay coefficient λ to 1, the number of candidates K to 3, and the number of latest memory items for each concept R to 3. For PCF Strategy, we set the time interval T to 10, the number of fusion candidates L to 3, and the threshold τ to 0.6. For the CFR metric in Eq. (11), we set ϵ to 1. Since each forked identifier is created at one of the $t - |\mathcal{C}_t|$ turns that revisit an already-seen concept, $\mathrm{CFR}(t) \in [0, 1]$. All the experiments are conducted on an NVIDIA RTX 5090 GPU.

Concept Identifiers. In our experiments, each novel concept is assigned an identifier named after its ground-truth concept in the benchmark (e.g., `dog_2`), and each forked concept receives an additional numeric suffix (e.g., `dog_2_1`). These identifiers serve only as evaluation bookkeeping, allowing CRA and CFR to map every predicted identifier back to its ground-truth concept. They are never shown to the MLLM, since an identifier is assigned only after the CR Skill has decided that a concept is novel, and both $\mathcal{P}_r$ and $\mathcal{P}_f$ refer to candidate concepts through anonymous labels, describing each one only by its memory items and portrait (cf. Appendix J). Therefore, the identifiers leak no ground-truth information, and because all compared methods, including TAME†, share the same scheme, none of them benefits from it.

D Hyper-Parameter Analysis

In this section, we evaluate the effect of important hyper-parameters within TAME-O: the temporal decay coefficient λ, and the number of candidates K. As shown in Table 3, our TAME-O exhibits relatively low sensitivity to hyperparameter variations, when $\lambda > 0$ or $K > 1$. We set $K = 3$ and $\lambda = 1$ across all datasets.

E Quantitative Analysis of the PCF Strategy

In the main paper, we offer the Concept Fission Rate (CFR) metric for TAME-O when equipped with the PCF Strategy. Here, we take a further step to empirically evaluate the fusion accuracy of the PCF

Table 5: Performance comparison on the LCMP-H and LCMP-O datasets. Higher values indicate better performance.

Methods	Open-Set Method	LCMP-H					LCMP-O				
		ACC-C	ACC-F	SPR-L	SPR-S	CRA	ACC-C	ACC-F	SPR-L	SPR-S	CRA
PeKit	TAME†	41.67	26.67	18.33	23.33	24.56	38.22	22.22	16.00	10.22	19.82
	TAME-O	**48.33**	**45.56**	**28.89**	**37.22**	**49.62**	**48.00**	**26.22**	**20.00**	**12.44**	**48.33**
R2P	TAME†	36.11	26.11	26.11	23.33	25.56	32.44	21.33	16.00	28.44	23.16
	TAME-O	**42.22**	**31.67**	**36.67**	**35.00**	**54.14**	**50.22**	**34.22**	**17.33**	**30.22**	**52.34**

Strategy. Formally, each fusion operation is defined as an automatic merging process that combines two personalized concepts c_m and c_n stored in the TAD Memory into a unified one. When c_m and c_n correspond to the same ground-truth concept, the fusion operation is considered correct; otherwise, it is regarded as incorrect. Based on this criterion, we define the Concept Fusion Accuracy (**CFA**) as follows.

$$\mathrm{CFA}(t) = \frac{1}{|\mathcal{U}_{\leq t}|} \sum_{(c_m, c_n) \in \mathcal{U}_{\leq t}} \mathbb{I}(\mathtt{GT}(c_m) = \mathtt{GT}(c_n)), \tag{13}$$

where $\mathcal{U}_{\leq t}$ denotes the set of concept pairs (c_m, c_n) that have been accepted by the PCF Strategy and fused into unified concepts up to time step t, $\mathtt{GT}(c_m)$ denotes the ground-truth concept identifier of concept c_m, which corresponds to the identifier assigned when the concept first appears, and $\mathbb{I}(\cdot)$ is a binary indicator function that outputs 1 if the condition holds and 0 otherwise. In Table 4, we present the cumulative CFA at successive time steps while processing the historical dialogues in LCMP-H and LCMP-O. We observe that TAME-O consistently achieves high CFA values, suggesting that the fusion decisions made by the PCF Strategy are reliable.

Note on the CFR Computation. All CFR values in Figure 5 are computed strictly following Eq. (11) with $\epsilon = 1$, recomputed from the logs of the same runs, and plotted without any per-method rescaling. This corrects two inconsistencies in the submitted version. First, the submitted Figure 6 normalized the total number of identifiers $\sum_{c_i \in \mathcal{C}_t} |\hat{\mathcal{I}}_{c_i}^{(t)}|$, rather than the number of forked identifiers, by $t - |\mathcal{C}_t|$, and rescaled the TAME† curve by a factor of 0.55. Second, the logger in the released code normalized the number of forked identifiers by the total number of identifiers instead of by $t - |\mathcal{C}_t|$. Both the plotting script and the released code are now aligned with Eq. (11). Under all three normalizations, the ordering TAME† $>$ TAME-O w/o PCF Strategy $>$ TAME-O holds at every sampled time step.

F Scalability across Other MLLM Personalized Assistants

Since TAME-O is designed as a training-free and plug-and-play framework, we further evaluate its scalability by integrating it with other personalized MLLM assistants beyond TAME [4] (reported in the main paper). Specifically, we integrate TAME-O with two representative training-free, RAG-based personalization methods, PeKit[17] and R2P[7], which differ in their memory organization and retrieval strategies. To enable open-set personalization, we augment both methods with our Concept Reconciliation (CR) Skill and a Time-Aware Dynamic (TAD) Memory that stores dialogue contexts with temporal markers. The CR Skill is used to determine whether an incoming personalized concept is novel or previously observed, while the original memory mechanisms of each method are retained for concept-specific storage and retrieval during response generation. As illustrated in Table 5, we compare the performance of these two baseline methods under the open-set setting by integrating TAME† and TAME-O, respectively. We observe that integrating TAME-O yields superior open-set personalization performance, demonstrating the strong scalability and generality of our method.

G Additional Literature Review

G.1 Memory-Augmented LLMs

Recent research on LLMs explores augmenting models with external memory to store and retrieve past interactions, enabling improved adaptability without parameter updates [24, 25, 26]. Existing studies typically organize memory as structured summaries, temporal events, or reasoning traces to highlight key information while reducing redundancy [27, 28]. To scale memory usage, retrieval mechanisms are introduced to efficiently locate relevant past experiences, often embedding memory

entries as dense vectors and retrieving the most relevant ones via similarity search [29, 30, 31], with cosine similarity and vector indexing tools such as FAISS commonly used in practice. These studies demonstrate that external memory can effectively extend LLM capabilities without retraining by enabling models to accumulate and reuse past knowledge.

Building upon these advances, memory-based methods have also been studied in MLLM personalization, eliminating the need for frequent retraining for novel concepts [8, 4]. Unlike general memory-augmented LLM approaches that treat each memory entry independently, these methods explicitly maintain concept-level memory units to organize information around individual personalized concepts. Approaches such as RAP [8] and PeKit [17] store user-specific concepts in external memory and perform retrieval-based matching, while R2P [7] introduces structured "fingerprints" with reasoning for identification. More recently, TAME [4] extends the memory for concept storage to the modeling of long-context information for each user-specific concept. However, these methods generally assume a closed-set of predefined concepts and fail to extend the memory autonomously by the MLLMs for accommodating novel concepts. In contrast, our TAME-O introduces a time-aware dynamic memory with explicit concept reconciliation, enabling continuous concept acquisition and refinement under open-set, long-context interactions. Moreover, the PCF strategy further refines the memory to mitigate excessive concept fission that could otherwise lead to severe memory chaos.

G.2 Open-Set Recognition

Open-set recognition (OSR) addresses the limitation of the closed-set assumption in traditional deep learning systems, enabling models to both classify known classes and detect inputs belonging to previously unseen classes [32, 33, 34]. The problem was first formally introduced by [34], who proposed a binary SVM-based framework for training and evaluation. Since then, extensive research has been devoted to this area. Existing approaches can be broadly divided into discriminative and generative methods. Discriminative approaches aim to learn more robust representations or decision boundaries to reduce overconfidence on unknown classes, such as placeholder-based representative points, supervised contrastive learning, and diverse representation regularization [35, 36, 37]. In contrast, generative approaches synthesize instances near the boundary between known and unknown classes using techniques such as GANs, variational autoencoders, or prototype-based modeling to better characterize open space risk [38, 39, 40].

Despite the name "open-set," these methods are primarily designed to identify whether an input belongs to an unknown category, typically labeling it as unknown. In contrast, the open-set personalization setting considered in this work requires not only identifying novel user-specific concepts, but also assigning them new personalized concept identities, enabling the assistant to continually expand and refine the set of personalized concepts during interaction.

G.3 Multimodal Large Language Models

Large Language Models (LLMs), built upon extensive world knowledge, have rapidly advanced across a wide range of domains [41, 42]. Building on this foundation, Multimodal Large Language Models (MLLMs) have recently emerged by integrating language decoders with multimodal encoders, enabling LLMs to perceive and reason over diverse data modalities [43, 44, 45, 46, 47, 48, 49, 50, 51]. Among MLLMs, vision–language models have been extensively studied [5, 18], and the fusion of visual and linguistic modalities has driven significant progress in applications such as Visual Question Answering (VQA) [52], multimodal medical analysis [53], and autonomous driving [54]. With these advances, MLLMs are increasingly regarded as reliable human assistants, supporting users in various aspects of daily life. Despite their success, mainstream MLLMs are primarily trained for general-purpose scenarios and fail to adapt to individual users' preferences and personalized contexts [1]. To bridge this gap, recent studies have begun to explore personalized MLLM assistants [2, 15, 4]. Among them, MC-LLaVA [13] extends personalization to multiple concepts via multi-concept instruction tuning, and CaT [14] synthesizes training data to strengthen personalized VLMs, while both still operate on a predefined set of concepts. In this work, we take a step further by addressing the open-set personalization setting and propose the first open-set personalized MLLM assistant TAME-O, which better aligns with users in open-world, long-term interactions.

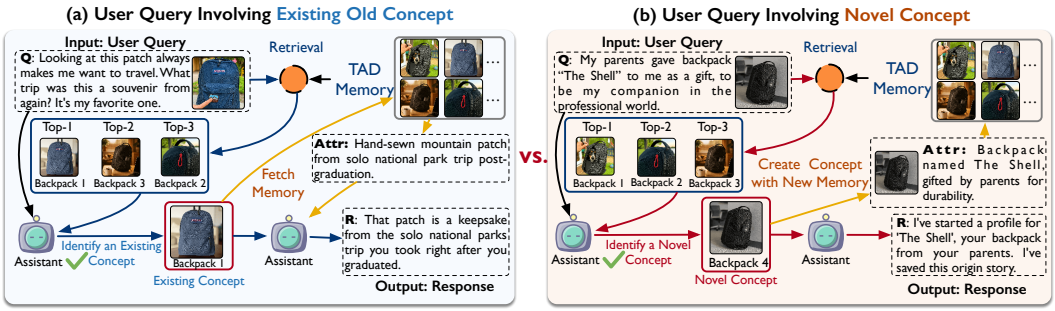


Figure 10: Case study: the procedure of TAME-O in resolving an existing old concept (left) and a novel concept (right).

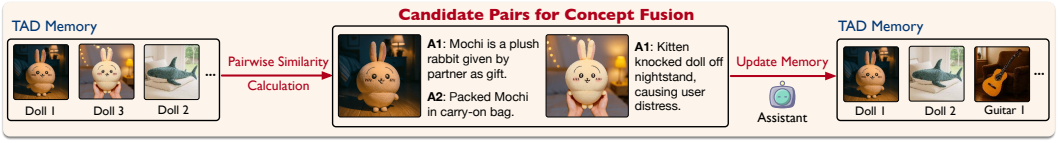


Figure 11: Case study: two concepts that have similar attributes are selected to be fused into one concept.

G.4 Large Language Model Personalization

Large Language Model (LLM) personalization [55, 56, 57, 58, 59] focuses on adapting an LLM's responses to a single user's preferences, interaction history, and persona characteristics (e.g., a preference for concise responses). Early efforts such as PersonaMem [60] introduce benchmarks with numerous simulated profiles to test whether models can leverage dialogue history to infer user preferences and deliver personalized responses. PersoBench [57] further explores persona-aware dialogue generation in a zero-shot setting. Building upon these evaluations, methods like PLUM [56] employ dialogue-based data augmentation to fine-tune LLMs for more consistent user-aware behavior. PALACE [61] advances this line with a topic-aware retrieval mechanism that extracts relevant cues from extended histories and a persona-driven prompt learning module for multi-session adaptation. Nevertheless, LLM personalization focuses on tailoring a model's responses to a single user's preferences, dialogue history, and persona traits (e.g., preferring concise answers). In contrast, MLLM personalization emphasizes enabling personalized conversations grounded in user-related multimodal concepts, such as the user's pets, friends, and frequently mentioned objects.

H Additional Case Study

H.1 Additional Cases of Concept Reconciliation

In Figure 10, we provide additional concept reconciliation cases of TAME-O in resolving novel and existing concepts.

H.2 Cases of Concept Fusion

As illustrated in Figure 11, we provide a case illustrating the concept fusion procedure of our proposed PCF Strategy.

I Failure Case Analysis: Concept Fission

In this section, we present a failure case analysis of TAME-O, focusing on unintentional concept fission, as illustrated in Figure 12. Specifically, a concept originally associated with `Dog 2` — a pet dog undergoing a temporary physical appearance change (e.g., its fur becoming dirty after playing in the mud) — is incorrectly identified as a novel concept, `Dog 4`. This misclassification triggers an unintended concept fission.

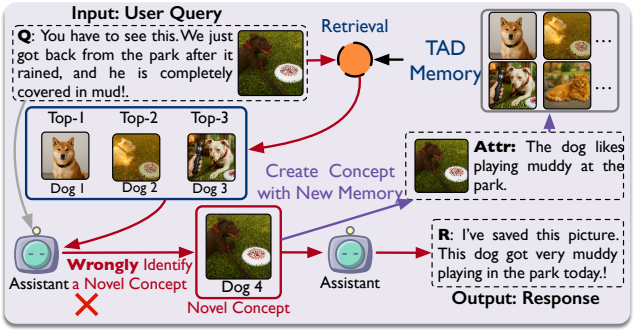


Figure 12: Case study: incorrect concept identification leads to the fission of an already grasped concept.

J Prompt Presentation for TAME-O

In this section, we present the prompts $\mathcal{P}_r$ and $\mathcal{P}_f$ designed for performing concept reconciliation and fusion, respectively.

K Limitations

Reliance on External Modules. TAME-O relies on off-the-shelf grounding and embedding models to localize and retrieve personalized concepts, and their errors may propagate to concept reconciliation. Nevertheless, these modules can be readily replaced by stronger ones as they become available.

Robustness under Noisy Interactions. Our benchmarks are constructed in a controlled manner, whereas real-world interactions can be noisier, e.g., with ambiguous references or low-quality images and descriptions. We leave a systematic evaluation under such noisy interactions to future work.

Indirect Recovery from Incorrect Fusion. Concept fusion may occasionally merge two distinct concepts. Although such a merge cannot be directly reverted, TAME-O can recover from it indirectly: benefiting from the attribute-level and time-aware memory organization, subsequent interactions progressively correct the affected concept memory with recent observations. A direct mechanism to revert incorrect fusions is left to future work.

Scalability and Privacy. The pairwise similarity computation in concept fusion grows quadratically with the number of concepts, although the number of MLLM invocations per fusion round is bounded by L. In large-scale deployment, this cost can be reduced by caching pairwise similarities and recomputing only the pairs involving recently updated concepts, or by approximate nearest-neighbor search. Moreover, long-term personalized memory contains private user information, which calls for local storage and user-controlled deletion in practical deployment.

L Broader Impacts

Since this study focuses on designing open-set personalized MLLM assistant, which is closely related to everyone's daily life, it has broader impacts in both positive and negative aspects. **On the positive side**, open-set MLLM personalization can make future assistants more adaptive, user-aligned, and convenient in daily life. By automatically discovering and maintaining user-specific concepts through long-term interactions, such assistants can provide responses that are more context-aware and personalized, reducing the burden on users to manually register or describe every personal entity. This capability may benefit various real-world scenarios, such as daily assistance, education, accessibility support, elderly care, and long-term human AI collaboration, where understanding user-specific objects, people, pets, or environments is important for providing useful and grounded assistance. **On the negative side**, such systems may introduce risks from both misuse and the inherent limitations of MLLMs. For misuse, the ability to continuously acquire personalized concepts could be abused for intrusive profiling, unauthorized surveillance, or manipulative personalized content generation. For model limitations, our method may inherit hallucination and bias issues from the underlying MLLMs. It may generate responses that are not fully grounded in the input image, dialogue context, or stored personalized memories, and such hallucinations can be harmful in high-stakes scenarios. Moreover,

Prompt $\mathcal{P}_r$ for Concept Reconciliation.

Prompt: You are a concept identification assistant. Your goal is to determine if the user's question refers to one of the EXISTING candidates or to a completely NEW concept.
{**Recent Conversations**}
 # Question
{**User Query**}
Image: {**present / absent**}
 # Candidates (Existing Memories):
[For each candidate concept $c_k \in \mathcal{C}_K$, $k = 1, \ldots, K$:]
Candidate: CANDIDATE_{k}
- Portrait: [IMAGE]
- Static Memory: {**Static Memory Items**}
- Recent Dynamic Memory: {**Recent Dynamic Memory Items**}

 # Instructions
Compare the entity in the Question against each Candidate.

- Check for mismatches in **Name**, **Species**, **Physical Appearance**, or **Core Attributes**.
- If the Question describes an entity that is fundamentally different from a Candidate, they are NOT the same.
- If the Question explicitly mentions a candidate's proper name, select that candidate unless species or stable identity details contradict it.
- Transient details (e.g., accessories, location, mood, pose, or activity) may differ from memory and are not enough by themselves to make a new concept; compare stable identity details (e.g., name, species, breed, color, or markings).
- If the tone of the Question is to introduce a new concept (e.g., a person, pet, or belonging), select "NEW_CONCEPT".
- Context might be similar, but specific details must match.
- Use only the neutral candidate labels shown above. Do not infer identity from labels.
- If NO candidate is a strong match, select "NEW_CONCEPT".

 # Output Format
Output ONLY one of the following:
1. "EXISTING:CANDIDATE_N" (if it matches a candidate)
2. "NEW_CONCEPT"

biases from the vision encoder and language model may lead to unfair or stereotypical interpretations of user-specific entities. Therefore, open-set personalized MLLM assistants should be deployed with appropriate safeguards, usage restrictions, and careful evaluation, especially in sensitive applications.

M License of Assets

Datasets: LCMP (research-use), Oxford-IIIT Pet (CC BY-SA 4.0), MS-COCO (Flickr Terms of Use and image-specific licenses), and PerVA (research-use).

Prompt $\mathcal{P}_f$ for Concept Fusion.

Prompt: You are a concept similarity expert. Your task is to determine if two concepts represent the SAME entity that was incorrectly split (fissioned).
 # CONCEPT A
Portrait: {**Portrait Image or "(No portrait available)"**}
Static Memory (Permanent Facts): {**Static Memory Items**}
Dynamic Memory (Recent Events): {**Dynamic Memory Items**}
 # CONCEPT B
Portrait: {**Portrait Image or "(No portrait available)"**}
Static Memory (Permanent Facts): {**Static Memory Items**}
Dynamic Memory (Recent Events): {**Dynamic Memory Items**}
 # Instructions
Compare the two concepts carefully to determine if they are the SAME entity.
CRITICAL RULES FOR "SAME" VERDICT:

1. Conclude "SAME" only with clear identity evidence: the portraits show the same individual/object, or the memories share a distinctive name, model, or owner-specific identifier.
2. Generic category overlap is not enough: two cats or two similar-looking people can still be different entities.
3. If both concepts have specific names, brands, models, colors, breeds, or personal details that differ, conclude "DIFFERENT".
4. A generic reference (e.g., "the cat") matches a specific concept only when the visual evidence or distinctive memories strongly connect them.
5. Merge an unnamed fragment with few memories into the richer concept when its later details are compatible and no identity evidence contradicts it.

GUIDELINES:

- Assume memories accumulate over time.
- Do not merge based on broad class, scene, pose, or activity alone.
- Later memory fragments may omit the name.
- If the evidence is weak or ambiguous, favor "DIFFERENT".

 # Output Format
1. Brief reasoning (1–2 sentences).
2. Final verdict: "VERDICT: SAME" or "VERDICT: DIFFERENT".



Figure 13: Cases of personalized concepts from LCMP-O benchmark.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We formalize the problem of Open-Set MLLM Personalization, and propose open-set, long-context personalized MLLM assistant TAME-O to address it.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix K.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the code and the LCMP-O benchmark at `https://github.com/ronpay/TAME-O`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide implementation details in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation across five concept appearance orders in Figure 7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide an analysis of computational efficiency in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in Appendix L.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discuss it in Appendix L.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers when using their code or datasets, and we provide the licenses for all datasets used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the code and LCMP-O at `https://github.com/ronpay/TAME-O`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use LLMs to assist in creating the benchmark and focus on open-set personalization MLLM assistant in this paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.