GDP: Graph-Based Dynamic Personalization for Multimodal Large Language Models

Supplementary Material

A. Additional Experiments

A.1. Experimental Details

We test our dataset on other algorithms. For visual recognition, we restructure the test data into a format compatible with the RAP and Yo'LLaVA algorithms to generate results. When testing LLaVA, we only input the necessary questions and images; for LLaVA-Retrieve, we use pipeline similar to RAP without fine-tuning; while running LLaVA-LoRA, we design a fine-tuning dataset specifically for the corresponding questions. Since the code for the Personalization Toolkit is not open source, we replicate it based on the method described in the paper. In the Preference QA task, we incorporate historical interaction information (including images and conversations) as background knowledge into the prompt for other models to reference. Since we set the maximum number of historical interactions to 5, the historical interaction information generally does not exceed the model's maximum token limit.

We split the training and testing data. The image data used for fine-tuning LLaVA is mainly selected from movie and TV show clips, while the test data is sourced from the Internet with authorization. We choose LLaVA-13B as the base model and conduct inference and fine-tuning on a single A100 GPU.

In our experiments, we set $\alpha = 0.05$ to indicate the impact of semantic similarity, emphasizing semantically consistent associations, $\beta = 0.5$ for emphasizing the influence of co-occurrence frequency, and $\lambda = 0.05$ to introduce moderate temporal decay, ensuring that recent interactions are prioritized without completely ignoring longer-term dependencies. We set $\gamma_1 = 0.5$ to ensure a high consistency between recommendations and user intent; $\gamma_2 = 0.2$ introduces a constraint on node reliability; and $\gamma_3 = 0.3$ ensures that highly important nodes within the graph structure are effectively considered.

A.2. Experiments on RAP Dataset

We conduct the visual recognition test on the RAP dataset. As shown in Table 4, our method performs slightly better than other methods in terms of recognition effectiveness.

A.3. Identity Recognition Mechanism.

As shown in Table 5, face recognition has recognition capability comparable to other methods(RAP) in simple visual conditions, resulting in similar accuracy(0.909 and 0.919). In complex scenarios, they all tend to fail, where graph augment provides additional gains(0.970 vs 0.919). The above discussion is based on positive accuracy, because the gains on negative samples come from PE Agent. As defined in our paper, negative accuracy can reflects identity confusion. PE Agent avoiding overconfident guesses, resulting in high accuracy on negative samples in Table 5.

Table 4. Quantitative evaluation on visual recognition.

Method	Positive	Negative	Weighted
LLaVA	0.000	**1.000**	0.500
LLaVA+Retriever	**1.000**	0.002	0.501
LLaVA-LoRA	0.878	0.744	0.811
Yo'LLaVA+Prompt	0.933	0.753	0.843
Per.Toolkit+Prompt	0.922	0.778	0.850
RAP+Prompt	0.956	0.758	<u>0.857</u>
Ours (GDP)	<u>0.989</u>	<u>0.951</u>	**0.970**

Table 5. Recognition result comparison under different settings.

Metric	RAP	w/o Graph Augment	GDP
Positive	0.909	0.919	**0.970**
Negative	0.709	0.938	**0.943**
Weighted	0.809	0.928	**0.956**

Table 6. Ablation study on memory retrieval variables.

Metric	w/o Graph	w/o Conf.	RAG	GDP
Consistency	0.781	0.766	0.657	**0.832**
Update	0.748	0.684	0.592	**0.835**
Weighted	0.764	0.725	0.624	**0.834**

A.4. Comparisons against More Baselines.

The corresponding comparative results of ablation studys are reported in Table 6. Table 6 shows that gains appear only when relational modeling and confidence-aware state evolution are combined.

A.5. Evaluation on Recent LLMs.

We conduct experiments on two representative MLLMs, Qwen3-VL and InternVL in Table 7. Our method demonstrates superior personalization capability. We also systematically compare with existing unimodal LLM-based personalization methods in Table 7. The results show that our method achieves consistent gains in preference modeling.

Table 7. Comparison across different personalization systems.

Metric	OPPU	Hydra	P.P.	Qwen3-VL	InternVL	GDP
Consistency	0.696	0.645	0.662	0.681	0.645	**0.832**
Update	0.578	0.559	0.564	0.663	0.627	**0.835**
Weighted	0.636	0.601	0.612	0.672	0.636	**0.834**

B. More Details of Dataset Collection

B.1. Overall Pipeline of Dataset Collection

Personal Data aims to extract aligned multimodal user interaction records from TV dramas or online videos, serving as the initial data source for subsequent user graph construction.During the Data Acquisition Phase, automated scripts like DownSub are used to collectively gather video subtitle files. Keyframes are extracted using ffmpeg according to preset time intervals or scene changes. These keyframes are then associated with their corresponding context text, ensuring high alignment between visual information and subtitles. We specifically retain additional contextual information to facilitate more accurate reasoning by the subsequent Large Language Models (LLMs) in context.Upon completion of data preparation, we proceed to post-processing. The Role Identification and Clustering Phase involves manual annotation to name and confirm each interacting character (e.g., "Anna," "John"), thereby achieving the standardization of character IDs. Subsequently, the data undergoes rigorous Semantic Filtering and Quality Auditing. In the semantic filtering step, the original dialogues and corresponding visual/contextual information are input into a large language model. The LLM determines, based on preset evaluation rules or prompts, whether the dialogue content contains effective information valuable for user profiling, outputting an "available" or "unavailable" label. Only the available data is retained. Quality control adopts a combination of automated and manual auditing: first, automated detection filters out data that is too short, duplicated, or corrupted; then, human reviewers finalize the process by confirming the authenticity, consistency, and absence of potential bias in the samples.Finally, in the Dataset Organization and Partitioning Phase, effective user interaction records are grouped according to user identity, ensuring each group contains the interaction history of only one user, and arranged sequentially by time to simulate a realistic dialogue scenario. To ensure sufficient context for user modeling, we discard records from users with fewer than four interactions.

B.2. Instructions for Dataset Collection

Table 9 presents the prompt template used for generating or evaluating question-answering models. It typically includes clear instructions, context, and questions, designed to guide the model in generating accurate answers.

Table 10 displays the prompt format used for the Supervised Fine-Tuning (SFT) dataset. This dataset consists of high-quality instruction-answer pairs, used to train the model to follow instructions and produce outputs that align with human preferences.

C. Examples of the Experiment Results

In Table 8, we show an example of preference updating, highlighting how the model updates user preferences based on interactions. Our model can update user information in real-time and provide accurate responses based on the understanding of users.

In Fig. 6, we present an example of improved recognition, demonstrating the enhanced recognition capabilities of the model after incorporating user preferences. In the second image, the face of the person is occluded, and existing methods struggle to address the person recognition challenge in such cases. In our proposed algorithm, since the scenes described in both images are identical and the person's attire remains consistent, the individual in the first image can be clearly identified. This information is then propagated through the graph network to assist in face recognition in the second image.

D. Examples of the GDP Dataset

Existing personalized Multimodal Large Language Models (MLLMs) rely on static user information and struggle to update preferences dynamically during interactions. To overcome this, we propose focusing on user preferences as a core aspect of personalization, particularly through the Personal Dataset, which includes images and conversations. This dataset aims to advance personalized tasks in visual recognition and preference understanding such as personalized question answering. Additionally, we introduce GDP datasets for better preference identification.

Our database is divided into two parts: profile and history. As shown in Table 15, the profile contains basic information about individuals, such as their portrait, name, occupation, hobbies, etc. As shown in Table 14, the history includes images and conversations related to the individuals, which are used to construct the graph memory.

- Table 11 shows an example of a conversation. It is sampled from real interaction scenarios and used for the construction of Personal Data.

- Table 12 provides an example of Visual Recognition questions. It is used to evaluate the model's ability to integrate visual information and natural language for identity recognition.
- Table 13 presents an example of Preference QA datasets. It asks about the user's personalized information and is used to assess the model's ability to extract, memorize, and update user preferences.

E. Limitations

While the Graph-based Dynamic Personalization (GDP) framework demonstrates considerable efficacy in establishing a robust model of user identity and preferences, its operational reliance on extensive data introduces significant limitations, primarily concerning user privacy and data security. The model's success hinges on its ability to process and learn from highly sensitive types of user interaction data. This means the GDP model must maintain a comprehensive, centralized repository of deeply personal information about a user's habits, interests, and emotional state. This significantly increases the risk of unauthorized access or misuse, challenging user trust.

Future research and development efforts must therefore pivot towards exploring and integrating state-of-the-art privacy-preserving mechanisms to decouple data utility from individual identity risk. Specifically, we recommend Differential Privacy, which offers a strong, formal guarantee that the inclusion or exclusion of any single user's data does not significantly affect the final model outcome. Additionally, Federated Learning allows the model to be trained directly on user devices without ever requiring the raw, sensitive data to leave the user's secure local environment. By adopting these methods, researchers can mitigate ethical and legal risks, fostering user trust while preserving the framework's effectiveness.

Table 8. User profile and memory update examples.

<Wang>is an office worker. He also has a great sense of fashion...

History: <Wang>is a man. He is an office worker. He has a great sense of fashion. He is sharp-tongued, yet incredibly kind-hearted. He loves beef and vegetables. He is also a vlogger.
User: What do I like?
GDP: You like beef and vegetables.

Memory Updated !
User Input: It is tasty. <Image>
User: What do I like?
GDP: You like beef, vegetables and ice cream.

<JinShu>a doctor with superb medical skills, and he is very gentle and reliable...

History: <JinShu>is a man. He is a doctor with superb medical skills, and he is gentle and reliable. He currently lives in Yunnan. Nowadays, he is serving the local people with his medical expertise while enjoying the beautiful natural scenery and slow-paced life of Yunnan.
User: What do I like?
GDP: You like beautiful natural scenery and slow-paced life.

Memory Updated !
User Input: I highly recommend this rose coffee; it's not greasy, not bitter, and very refreshing. <Image>
User: What do I like?
GDP: You like beautiful natural scenery and slow-paced life in Yunnan Province and rose coffee.

Conversation: She is taking photos under Eiffel Tower.
Persons: **<BanMa>**
Nodes:
• scene: grassy field
• object: a white jacket
• object: Eiffel Tower
• activity: facing the camera

Conversation: She is taking a photo under Eiffel Tower.
Persons: hard to detect ⟶ **<BanMa>**
Nodes:
• scene: grassy field
• object: jacket
• object: Eiffel Tower
• activity: posing for a picture

Figure 6. Example of improved recognition.

Table 9. Prompt template for Preference QA task.

- **Role:**
  You are an intelligent agent capable of providing personalized services.
- **Context:**
  Below is image information of the same individual across multiple scenarios. You are requested to generate five question-answer pairs that require integrating this user's information graph to be answered.
- **Information Provided:**
  A summary of the user is provided below: `[summary]`
- **Requirements:**
  - You must role-play as this user when posing questions (use natural expressions when referring to the user).
  - Questions should be in multiple-choice format, and answers should only include the chosen option(s) without any explanation.
- **Output Format:**
  - Question: xxx
  - Options: xxx
  - Answer: xxx
- **Examples of Outputs:**
  - Question: Which indoor activity makes me happier?
  - Options: A. Socializing B. Shopping
  - Answer: B

Table 10. Prompt template for SFT Dataset.

- **Task:**
  Please analyze the image and extract the following information:
  The conversation is `[conversation]`.
  Output the categories of scene, activity, emotion, object, and preference.
- **Output Format Rules:**
  Each node should contain label and confidence score, in the following format. The confidence score refers to a numerical value that indicates the degree of certainty or reliability of a prediction.
  - node type: [content:confidence score]
- **Examples of Outputs:**
  - scene: [beach:0.89]
  - activity: [playing outdoors:0.92]
  - emotion: [happy:0.98]
  - object: [sunglasses:0.95]
  - preference: [loves the sea:0.94]
- **Requirement:**
  You don't have to explain, just give the answer in the above format.

Table 11. Categorization of conversations.

Dialogue Type	Description and Examples
Sharing Feelings	This is the most common type, directly expressing subjective feelings and emotions about something or the environment. *Examples:* · "So delicious!" · "This museum is so boring." · "I feel so dizzy, this digital exhibition is too dreamlike."
Giving Recommendations	Evaluating a product, food, or place, which may include positive recommendations or negative criticisms. *Examples:* · "This citrus-scented perfume is just okay." · "I highly recommend this rose coffee; it's not greasy, and not bitter." · "This is the best Chinese food I've had here."
Interacting	Engaging with the audience by asking questions or showcasing something to solicit opinions or attract attention. *Examples:* · "Does this look good on me?" · "I finally bought that rose perfume I wanted! Isn't it cute?" · "Guess how much a meal costs at our hospital cafeteria?"
Stating Knowledge	Stating an objective fact, a personal discovery, or a piece of trivia. *Examples:* · "Gaudí was a great artist, but he wasn't recognized before his death and his treatment was delayed. What a pity." · "This sunscreen can be used for tanning." · "That building over there is the hospital where I work."
Sharing Personal Memories	Sharing a personal anecdote, story, or childhood memory. *Examples:* · "Why didn't I have Santa Claus in my childhood? Why?" · "An uncle asked for my name, and I got shy." · "Wrapping up my day as a blogger experiencing a doctor's life."

Table 12. Examples of Visual Recognition questions.

- Is <sks>in this picture?
- Can you see <sks>in this image?
- Is <sks>present in this photo?
- Is <sks>in this shot?
- Can you find <sks>in this crowd?
- Is <sks>visible in this photo?
- Is <sks>in the background of this picture?
- Is <sks>partially visible in this image?
- Can you tell if <sks>is in it?

Table 13. Examples of Preference QA.

PQA Type	Examples
Location-related	• Based on my experiences in France, if I had to choose between exploring the city of Paris again or spending a day at the beach, which location am I more likely to pick? A. Paris B. Beach • In which location am I more likely to be photographed with a camera? A. Beijing B. Finland
Activity-related	• When in Finland, would I prefer to unwind by shopping or chatting outdoors with friends? A. Shopping B. Chatting outdoors with friends • When I reflect on my experiences in Spain, which activity consistently made me happy and content? A. Sightseeing alone B. Outdoor dining
Interest-related	• When recalling my time in France, do my preferences and documented activities suggest I value seeing famous landmarks or capturing memories more? A. Seeing famous landmarks B. Capturing memories • Which of these aligns with my known preferences? A. close-up shot B. museum
Emotional state-related	• To maximize positive emotions, should I return to the beach or the general outdoor setting? A. Beach B. General outdoor • In my outdoor photos, is my dominant emotion more strongly associated with happiness or sadness? A.Sad B. Happy
Object-related	• I need to pick a souvenir that aligns with my tour. Should I get a teddy bear or a miniature Eiffel Tower? A. Teddy bear B. Miniature Eiffel Tower • Which scene feels more authentic to my Spain experiences: holding a skincare bottle while shopping or using fork and knife at an outdoor meal? A. Holding skincare B. Using fork and knife

Table 14. Examples of interaction history.

User	Image	Other Input Information
BanMa		**Place**: Spain **Conversation**: I feel so dizzy, this digital exhibition is like a dream.
Loong		**Place**: Yunnan Province **Conversation**: What a gorgeous wreath!

Table 15. Examples of user profile.

Image	Concept	Profile
	BanMa	She is a kind, principled, and professionally ethical doctor. She is knowledgeable but never shows off. She is emotionally sensitive and often moved to tears by touching stories. She's not particularly skilled in fashion or dressing.
	DaWu	He is a vlogger who loves taking photos. He is good at dancing and works out regularly. He excels at video creation, enjoys communicating with others, and has a wide social circle. He is humorous and lively. He is also a foodie, especially fond of spicy food.
	JinShu	He is a doctor with superb medical skills, and he is gentle and reliable. He currently lives in Yunnan. Nowadays, he is serving the local people with his medical expertise while enjoying the beautiful natural scenery and slow-paced life of Yunnan.
	Loong	He is a content creator in self-media platforms who once worked in the medical field. He is very social and has a wide circle of friends. He loves Primeape in Pokémon so much so that he has adopted it as his net name.
	Niu	He is a well-read and knowledgeable professional at a pharmaceutical company, admired by his peers. He is dedicated to fitness and maintains an excellent physique. At the same time, he possesses a sensitive and introspective nature, given to self-reflection.
	Rou	He is the son of Jinshu and Xiaoshu, and currently attends elementary school. He is a smart, well-mannered yet somewhat shy child who loves sweet foods like ice creams and cakes. He has a lot of friends.
	Wang	He is an office worker. He has a great sense of fashion. He is sharp-tongued, yet incredibly kind-hearted. He loves beef and vegetables. He is also a vlogger. He is one of the best friends of BanMa.
	XiaoShu	She is a doctor from Beijing, now living in Yunnan with her husband and son. She loves drinking tea and grilling tofu. She is an optimistic person who loves to laugh. She has a keen eye for spotting joy in everyday life and has planted two blueberry trees.