Benchmarking Multimodal Personalized Reasoning of Vision-Language Models in the Wild

Dohyun Kim [*1] **Junyong Lhim** [*2] **Kwanyong Park** [1]

Abstract

People often make decisions by holistically reasoning over heterogeneous forms of personal information. In this paper, we study the relatively underexplored capability of multimodal personalized reasoning in multimodal large language models (MLLMs). To this end, we introduce `MPRBench`, the first comprehensive benchmark specifically designed to evaluate personalized reasoning across a wide range of tasks and real-world challenges. `MPRBench` consists of 12 sub-tasks and additionally supports interpretable analysis of representative error cases. Through extensive experiments, we find that current personalized MLLMs still struggle with personalized reasoning due to several key challenges, including inaccurate recognition of personalized concepts and sensitivity to irrelevant personal information. We hope that MPRBench will stimulate further research on personalized reasoning in MLLMs.

1. Introduction

In daily life, people frequently interact with personal concepts such as friends, belongings, and pets, relying on their own memories. These memories are inherently multimodal and play a central role in human behavior and decision-making. In other words, humans naturally integrate current context with heterogeneous forms of personal memory, including visual impressions, textual knowledge, past experiences, preferences, and social relationships.

Existing approaches mainly focus on relatively simple tasks, such as recognizing the presence of personal concepts from visual inputs (Nguyen et al., 2024) or reporting their basic status (An et al., 2024). While current personalized models perform well on such benchmarks, these settings remain far from the complexity of real-world user requests. As a result, it is still unclear whether existing models can faithfully handle realistic scenarios and satisfy user needs, especially when deeper reasoning over multimodal personal information is required.

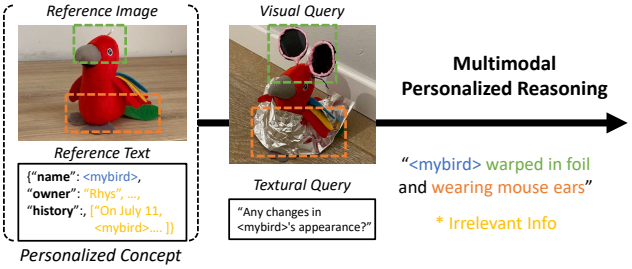

Figure 1. **Conceptual Overview and Problem Setup for Multimodal Personalized Reasoning.**

To bridge this critical gap, we introduce `MPRBench`, a novel and comprehensive benchmark designed to evaluate personalized reasoning capabilities of MLLMs with multimodal personal information. In designing `MPRBench`, we highlight a key challenge of multimodal personalized reasoning: *the ability to reject irrelevant personal information during the reasoning process*. Although MLLMs may have access to a wide range of personal information in both visual and textual forms, not every user request requires all of that information. Accordingly, we divide our benchmark into three groups—visual-specific, text-specific, and multimodal reasoning—encompassing a total of 12 sub-tasks designed to reflect diverse real-world scenarios.

On `MPRBench`, we evaluate a wide range of personalized MLLMs spanning three major categories of prior work: (1) zero-shot, prompting-based personalization (Seifi et al., 2025), (2) methods based on learning specialized personalized tokens (Nguyen et al., 2024; An et al., 2024), and (3) retrieval-based models (Hao et al., 2025; Das et al., 2025). Our findings identify three critical bottlenecks. First, robust recognition of personal concepts remains a fundamental bottleneck for personalized reasoning. Second, current methods struggle to reject irrelevant personal information across diverse scenarios. Third, reasoning that requires sequential and repeated use of personalized information remains highly

[1]Department of Computer Science, University of Seoul, Seoul, South Korea [2]Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, South Korea. Correspondence to: Kwanyong Park <kwanyong.park@uos.ac.kr>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

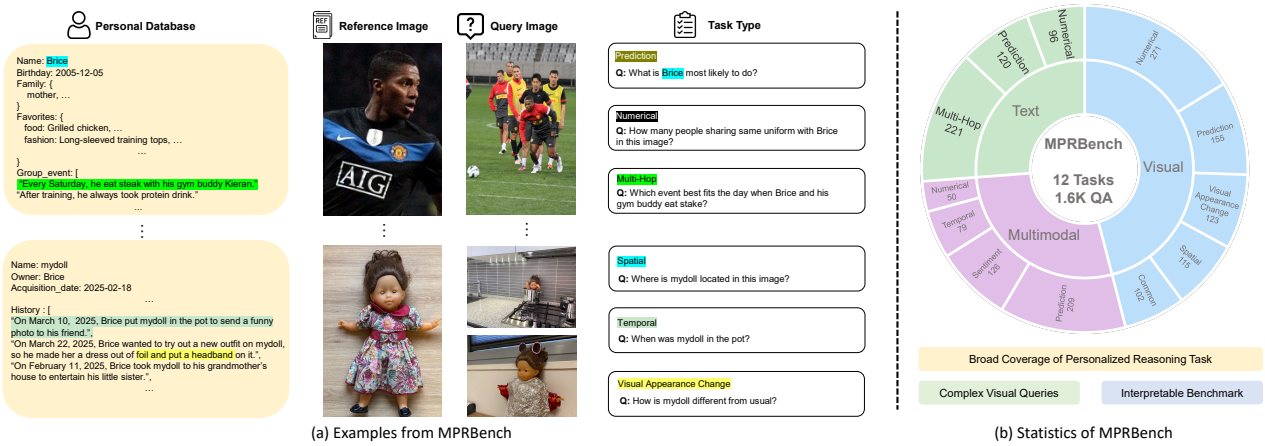

Figure 2. **Illustrative Examples and Statistics of MPRBench.**

challenging for existing models.

2. The MPRBenchmark

2.1. Overview of MPRBenchmark

Our goal is to assess the multimodal reasoning capabilities of personalized MLLMs in real-world scenarios. While existing benchmarks (Nguyen et al., 2024; Alaluf et al., 2024; Das et al., 2025) primarily evaluate elementary tasks, such as identifying personal entities or recognizing their states in visual contexts, MPRBench is designed to challenge higher-level reasoning tasks in which multimodal personal information must be used in diverse ways.

Notions and Problem Setup. We define a user-specific entity as a *personalized concept*, represented by a set of *reference image(s)* and *reference text*. The reference image(s) serve as a visual profile for identification, while the reference text provides structured personal information (e.g., name, preferences, history) in a JSON-like textual format. Given this, the goal of each task is to generate an appropriate answer to a *textual query* that mimics a user request, by jointly reasoning over both the *visual query* and the personalized concept. (See Fig. 1)

2.2. Key Features of MPRBenchmark.

Broad coverage of personalized reasoning tasks. MPRBench enables a holistic evaluation by including tasks that require reasoning over joint or individual modalities. It comprises 12 sub-tasks, allowing for fine-grained analysis of model weaknesses across diverse scenarios.

Complex visual queries. To better reflect realistic reasoning scenarios, we construct image queries with complex scenes in which personalized concepts appear alongside other entities within cluttered backgrounds. This design differs substantially from existing benchmarks, which often rely on personalized-object-centric images (Nguyen et al., 2024) or attribute-centric illustrations that omit the personalized concept itself (Kim et al., 2025).

Interpretable benchmark design. We formulate all tasks as multiple-choice questions, where each option reflects a representative failure case. This design makes it easier to identify major failure modes and provides actionable insights for future improvement.

2.3. Task Definition

MPRBench comprises 12 sub-tasks, each designed to capture a distinct challenge in multimodal personalized reasoning. Representative examples of all sub-tasks are provided in Fig. 2.

Vision-specific reasoning. These tasks require identifying the personalized concept from reference image(s), and reasoning about its current visual state. We include (1) **predictive reasoning** for forecasting future action; (2) **spatial reasoning** for understanding spatial relations; (3) **numerical reasoning** for counting and quantities; and (4) **visual commonsense reasoning** for inferring the rationales. In addition, we introduce (5) **visual appearance change reasoning**, a novel task requires explicit comparison between stored visual memory and the current appearance (e.g., detecting a new hairstyle) in the query image.

Text-specific reasoning. These tasks require reasoning over reference text with user queries, where textual information alone is sufficient for the solution. We include (1) **predictive reasoning** to infer outcomes from preferences or habits; (2) **numerical reasoning** for arithmetic or logical reasoning over numerical personal attributes (e.g., birthdays); and (3) **personalized multi-hop reasoning**, a novel task that requiring models to sequentially retrieve and integrate multiple pieces of personal information across iterative reasoning steps.

Table 1. Evaluation results on the proposed MPRBench.

	Vision-specific Reasoning						Text-specific Reasoning				Multimodal					
Method	Predict	Spatial	Numeric	Commsense	Change	V-Avg	Predict	Numeric	Multi-hop	T-Avg	Predict	Sentiment	Numeric	History	M-Avg	Avg
Random Choice	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Proprietary + Prompting																
GPT-4o (Achiam et al., 2023)+P	91.6	64.3	63.5	71.6	65.0	70.6	97.5	70.1	40.3	62.5	88.1	79.4	50.0	49.4	73.7	69.4
Open-sourced + Prompting																
LLaVA (Liu et al., 2024)+P	55.6	53.9	64.4	55.9	93.1	64.5	68.3	47.9	17.2	37.9	45.2	44.4	50.0	30.4	41.1	51.1
Token-learning based																
Yo'LLaVA (Nguyen et al., 2024)	47.7	28.7	33.9	47.1	42.3	39.0	43.3	30.2	14.0	25.6	30.1	34.9	30.0	25.3	31.9	33.5
MC-LLaVA (An et al., 2024)	56.1	39.9	43.5	39.2	65.0	48.4	57.5	35.4	13.6	30.4	39.7	50.0	34.0	25.3	39.6	41.3
Retrieval based																
Pekit (Seifi et al., 2025)	65.8	48.7	41.3	54.9	95.1	57.8	71.6	47.9	12.7	36.6	24.6	41.3	20.0	32.9	32.1	45.1
Rap (Hao et al., 2025)	46.5	24.3	37.4	35.3	46.3	38.4	50.0	39.6	3.17	24.0	35.5	47.6	20.0	22.8	32.9	33.1
R2P (Das et al., 2025)	60.6	39.1	39.5	50.9	85.4	52.6	30.8	8.34	1.80	11.2	49.9	58.7	40.0	32.9	46.7	40.1

Multimodal reasoning. This group consists of tasks in which both visual and textual personal information are jointly required. We include (1) **predictive reasoning** to forecast actions based on personal preferences or habits under a given visual context; (2) **sentimental reasoning** to infer the cause or rationale of visually expressed emotion in light of personal events or experiences; and (3) **numerical reasoning** to count or aggregate textual attributes associated with concepts present in the visual scene. Finally, we propose (4) **multimodal history reasoning**, a novel task in which the model must infer when a visual query was captured by matching its visual evidence with textual historical records.

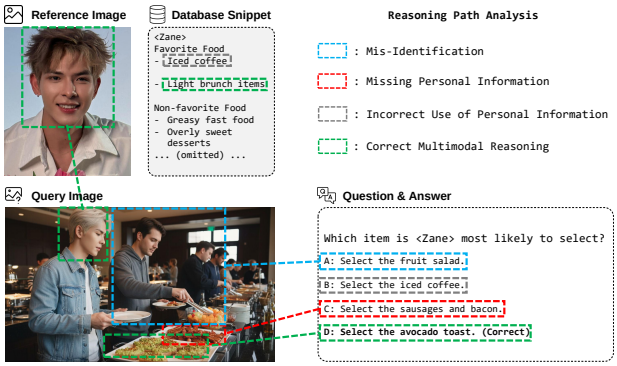

Figure 3. **Diagnostic Analysis of Multimodal Reasoning Paths**

2.4. Dataset Construction

We develop a semi-automated pipeline that minimizes human labor while maintaining rigorous standards. (1) Raw Data Collection: We curate reference images for personalized concepts from open datasets (e.g., Yo'LLaVA (Nguyen et al., 2024), RPC (Wei et al., 2019)) and CC BY 4.0 licensed sources. (2) Textual Reference Construction: We structure textual profiles through a four-step process—visually-grounded, diversity-oriented, context-aware, and group-wise attributes—using Gemini-2.5-Pro to ensure plausibility and consistency. (3) Query Construction: We employ three strategies based on data availability: template-based (using metadata), human-labeled (for complex reasoning), and generative-model-based (using Gemini 2.5 Flash Image for scalable scene generation). All instances undergo Human Verification to filter out ambiguous or incorrect cases.

3. Experiments

3.1. Experimental Setup

Baseline models. We benchmark a wide range of personalization methods: (1) **zero-shot, prompting-based personalization**, which injects personal information as additional context through prompting; (2) **personalized special-token learning methods**, which represent visual identity through learnable token, such as Yo'LLaVA (Nguyen et al., 2024) and MC-LLaVA (An et al., 2024); and (3) **retrieval-based models**, which employ similarity-based retrieval mechanisms for concept recognition, such as Pekit (Seifi et al., 2025), RAP (Hao et al., 2025) and R2P (Das et al., 2025).

Evaluation metrics We report accuracy for each sub-task. To obtain representative metrics, we compute the average accuracy within each reasoning group: V-avg (vision-specific), T-Avg (text-specific), and M-Avg (multimodal). The final overall score, Avg., is defined as the average of these three group-level scores.

3.2. Evaluation Results and Analysis

We conduct extensive evaluations to assess the personalized reasoning capabilities of personalized MLLMs and to analyze the potential causes behind the observed performance trends (Table 1).

1) Can current personalized MLLMs perform personalized reasoning effectively? All baselines remain far from perfect, as reflected in the final Avg. score. This contrasts sharply with the near-perfect performance re-

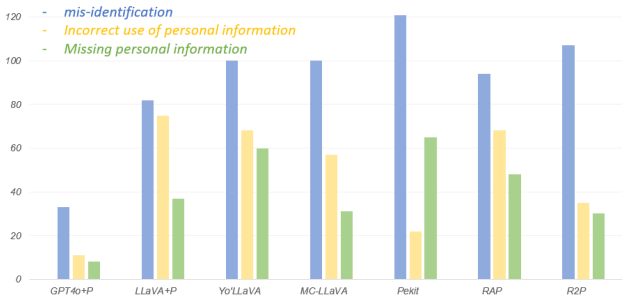

Figure 4. **Analysis on Failure Cases.**

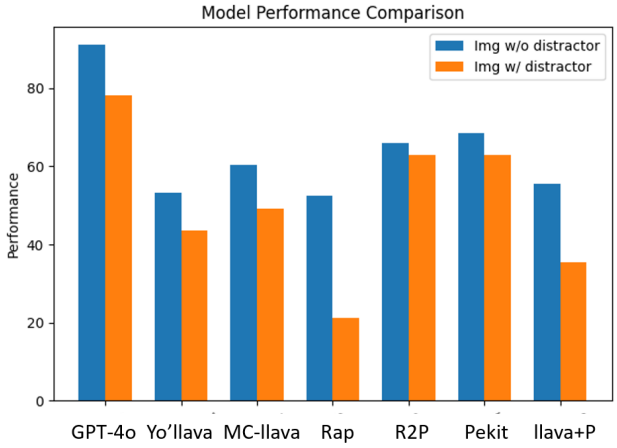

Figure 5. **Analysis on Distractor.**

ported on previous personalized recognition benchmarks such as Yo'LLaVA, suggesting that *current personalized MLLMs still struggle substantially with personalized reasoning tasks*.

More interestingly, specialized methods—both token-learning-based and retrieval-based approaches—often underperform the simple LLaVA+Prompt baseline, contradicting prior performance claims. In addition, RAP, which involves large-scale personalization tuning of the backbone MLLM, shows the weakest performance. We conjecture that tuning on elementary personalized tasks, such as recognition-oriented VQA, do not generalize well to reasoning-intensive settings and may even weaken the intrinsic reasoning capability.

2) What are the common failure modes? Thanks to `MPRBench`'interpretable design, we can analyze model failures by categorizing incorrect choices into three modes as illustrated in Fig 3: (1) **Mis-identification**: confusing the target with other individuals in the scene; (2) **Incorrect Use of Personal Information**: successfully retrieving data but failing to match it with the current visual context; and (3) **Missing Personal Information**: ignoring the database and relying on generic cues. As shown in Fig. 4, the most common failure mode is mis-identification. This indicates that accurate identity anchoring in cluttered, multi-person scenes remains a primary bottleneck-a challenge.

3) Are personalized MLLMs robust to cross-modal distractors? We investigate whether models can ignore irrelevant information from the non-target modality. By introducing hard negative textual distractors—additional events using keywords from incorrect answer choices added to the personal database that do not change the ground truth—we assess their impact on the prediction task. As shown in Fig 5, most models suffer noticeable performance drops, failing to properly filter out plausible but irrelevant information. This result suggests that current personalized MLLMs lack sufficient robustness to cross-modal distractors and are easily misled by irrelvant personal context.

4) Which classes of methods excel at understanding fine-grained visual details of concepts? We evaluate how well models leverage fine-grained visual details (e.g., hairstyle) through the visual appearance change reasoning task ("Change" in Table 1). In this task, token-learning-based methods generally perform worse than retrieval-based methods and simple prompting baselines, both of which directly access the reference image itself. We suspect that compressing rich visual information into a few learnable tokens inevitably causes information loss, leading to weaker performance on fine-grained visual understanding. Note that RAP again appears as an exception, but as discussed earlier, this may be because tuning the MLLM backbone degrades the model's intrinsic reasoning capability.

5) What unique challenges arise in personalized multi-hop reasoning? Among all sub-tasks, personalized multi-hop reasoning is the most challenging. This task introduces the unique difficulty of repeatedly leveraging personal information across multiple reasoning steps. Even propritary models struggle, and most personalized MLLMs perform even worse than random choice. This finding reveals a new weakness of current personalized models and suggests an important direction for future research.

4. Conclusion

In this paper, we introduce MPRBench, a novel benchmark for multimodal personalized reasoning. MPRBench consists of 12 sub-tasks spanning vision-specific, text-specific, and multimodal reasoning. In addition, we adopt an interpretable answer design that enables analysis of representative failure modes. Using MPRBench, we find that current personalized MLLMs still face substantial challenges in personalized reasoning, particularly in accurate personalized recognition and robustness to cross-modal distractors during the reasoning process.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alaluf, Y., Richardson, E., Tulyakov, S., Aberman, K., and Cohen-Or, D. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pp. 73–91. Springer, 2024.

An, R., Yang, S., Lu, M., Zhang, R., Zeng, K., Luo, Y., Cao, J., Liang, H., Chen, Y., She, Q., et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024.

Das, D., Talon, D., Wang, Y., Mancini, M., and Ricci, E. Training-free personalization via retrieval and reasoning on fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9683–9692, 2025.

Hao, H., Han, J., Li, C., Li, Y.-F., and Yue, X. Rap: Retrieval-augmented personalization for multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14538–14548, 2025.

Kim, J., Kim, W., Park, W., and Do, J. Mmpb: It's time for multi-modal personalization. *arXiv preprint arXiv:2509.22820*, 2025.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Nguyen, T., Liu, H., Li, Y., Cai, M., Ojha, U., and Lee, Y. J. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37: 40913–40951, 2024.

Seifi, S., Dorovatas, V., Reino, D. O., and Aljundi, R. Personalization toolkit: Training free personalization of large vision language models. *arXiv preprint arXiv:2502.02452*, 2025.

Wei, X.-S., Cui, Q., Yang, L., Wang, P., and Liu, L. Rpc: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*, 2019.