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

Anonymous Authors[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.

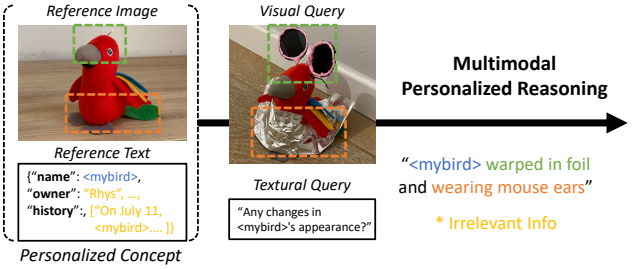


Figure 1. **Conceptual Overview and Problem Setup for Multimodal Personalized Reasoning.**

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.

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 challenging for existing models.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

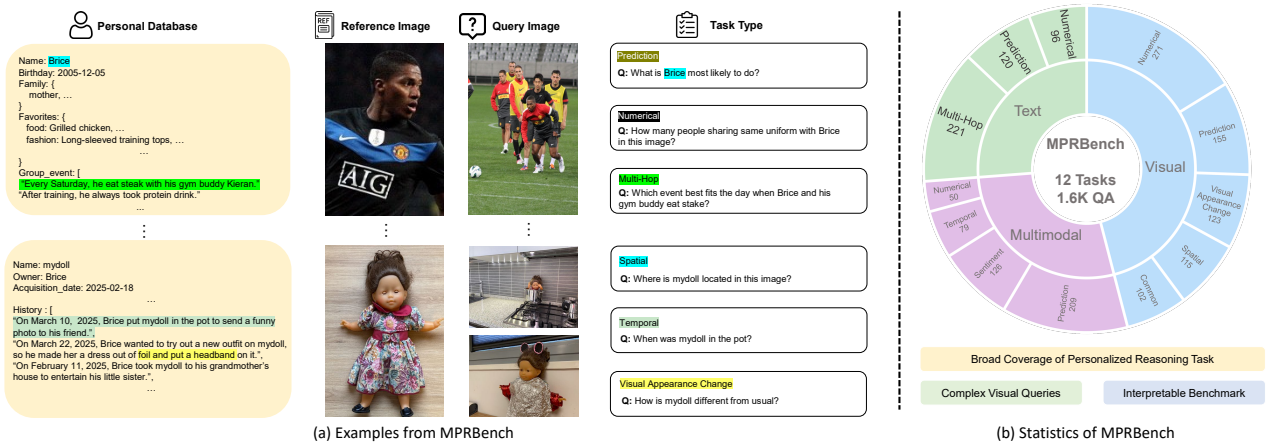


Figure 2. **Illustrative Examples and Statistics of MPRBench.**

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.

Multimodal reasoning. This group consists of tasks in

Table 1. Evaluation results on the proposed MPRBench.

Method	Vision-specific Reasoning						Text-specific Reasoning				Multimodal					
	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., 2024a)+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

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.

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. Due to space limitations, detailed procedures are provided in Appendix A.

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

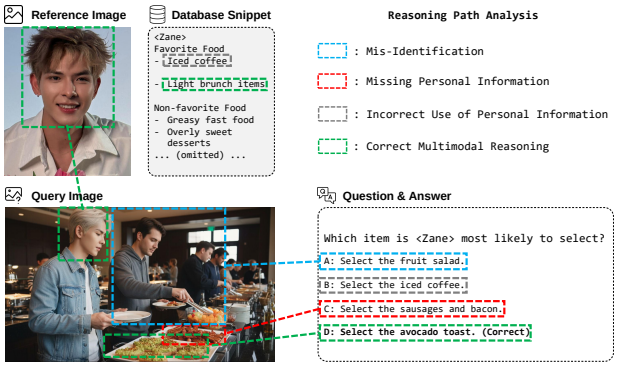


Figure 3. **Diagnostic Analysis of Multimodal Reasoning Paths**

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-

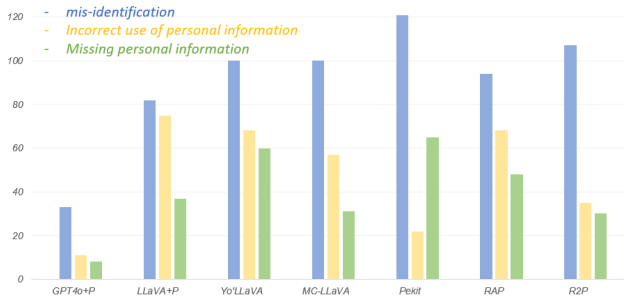


Figure 4. **Analysis on Failure Cases.**

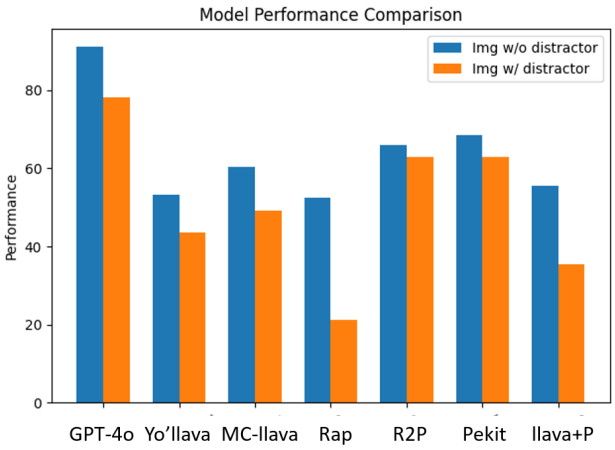


Figure 5. **Analysis on Distractor.**

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 often overlooked by previous benchmarks that rely on object-centric images.

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 irrelevant 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 propriatary 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.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37: 27056–27087, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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, 2024a.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

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.

Rao, J., Li, Z., Wu, H., Zhang, Y., Wang, Y., and Xie, W. Multi-agent system for comprehensive soccer understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3654–3663, 2025.

Rosasco, A., Berti, S., Pasquale, G., Malafronte, D., Sato, S., Segawa, H., Inada, T., and Natale, L. Concon-chi: Concept-context chimera benchmark for personalized vision-language tasks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 22239–22248, 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.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

A. Limitations

Given the complexity and diversity of multimodal personalized reasoning, MPRBench may not cover all possible scenarios. Nevertheless, we make a substantial effort to include a wide range of sub-tasks and to introduce novel tasks in each modality-specific group. Exploring even more diverse and complex scenarios for multimodal personalized reasoning remains an important direction for future work.

In addition, our primary goal in this work is to benchmark and analyze the current multimodal reasoning capabilities of personalized MLLMs. As a result, we do not extensively investigate methodological approaches for improving model performance. However, we believe that the findings and analyses provided in this work can serve as a useful foundation for future advances in personalized reasoning.

B. Ethics Statement

Because our benchmark includes person concepts and images of people, we strictly follow the licenses and terms of use associated with all source images. Most images are drawn from open datasets, and the remaining images are collected under CC BY 4.0 licenses. In particular, we carefully verify whether each person image permits derivative use or editing. If editing is not allowed, we do not use that image as a basis for generated or modified content (e.g., in Type 3 query construction).

For textual references, namely the personal information associated with images, we intentionally generate synthetic attributes rather than relying on factual personal information such as real names or actual biographies. In this way, we aim to mitigate privacy concerns.

C. More Details on MPRBench

C.1. Benchmark Statistics

In this section, we provide a detailed numerical analysis of MPRBench to highlight its scale and the diversity of its evaluation scenarios.

Diversity across Personalized Concepts: As summarized in Table. 2, our benchmark incorporates a wide array of 119 unique personalized concepts, comprising 69 distinct individuals and 50 varied objects. Fig. 9 provides an overview of selected personalized concepts used in our benchmark. This extensive collection of concepts ensures that the models are evaluated on their ability to generalize across diverse visual identities.

Broad Coverage of Reasoning Tasks: Furthermore, Table. 3 details the distribution of 1,667 question-answer pairs across 12 sub-tasks. By categorizing these to visual-specific (766 QAs), text-specific (437 QAs), and Multimodal (464 QAs), MPRBench covers a broad spectrum of reasoning requirements. This comprehensive task design ensures a rigorous assessment of a model's multifacted capabilities.

Table 2. Statistics of personalized concepts in MPRBench.

Concept Category	# Concepts
Person	69
Object	50
Total	**119**

C.2. Dataset Construction Process

We design a semi-automatic dataset construction pipeline to reduce human labor while maintaining overall data quality. Due to space limitations, we focus here on the high-level construction process (See Fig. **??**); more detailed descriptions are provided in the supplementary material.

C.2.1. Raw Data Collection

Concept collection. We begin by manually collecting reference images for personalized concepts. These images are sourced either from existing open datasets or from CC BY 4.0 licensed images collected from the internet. For person concepts, we use existing personalization-related datasets such as Yo'LLaVA (Nguyen et al., 2024), MC-LLaVA (An et al.,

Table 3. Detailed statistics of MPRBench across different task categories.

Task Category	Sub-task	# Questions
Visual-Specific	Prediction	155
	Spatial	115
	Numerical	271
	Commonsense	102
	Visual Appearance Change	123
	Sub-total (Visual)	**766**
Text-Specific	Prediction	120
	Numerical	96
	Multi-Hop	221
	Sub-total (Text)	**437**
Multimodal	Prediction	209
	Sentiment	126
	Numerical	50
	Temporal	79
	Sub-total (Multimodal)	**464**
Total		**1,667**

2024), MMPB (Kim et al., 2025) and CustomConcept101 (Kumari et al., 2023), as well as images of sports players from SoccerAgent (Rao et al., 2025) and publicly available CC BY 4.0 licensed images of celebrities. For object concepts, we repurpose ConConChi (Rosasco et al., 2024) and RPC (Wei et al., 2019), which were originally constructed for object retrieval and Retail Product Checkout tasks, respectively. To faithfully represent the visual appearance of each concept, we crop the original images to remove complex surrounding contexts. In addition, since some prior personalization methods (Nguyen et al., 2024) require multiple reference images, we retain only concepts for which at least five reference images are available.

Complex visual query collection. We also manually collect complex visual query images, with particular emphasis on scene complexity. Specifically, we prioritize images in which the target concept appears together with visually similar concepts (i.e., distractors) and under cluttered backgrounds. These conditions provide valuable opportunities to assess how concept identification difficulty affects downstream reasoning performance. Based on these criteria, we source person-related visual queries from VCR (Zellers et al., 2019) and SoccerAgent (Rao et al., 2025), and object-related queries from ConConChi and RPC. We make substantial efforts to ensure high quality during this stage. For example, we manually inspected approximately 4,000 images from VCR and selected only 120 images that satisfied our quality standards for visual queries.

C.2.2. Textual Reference Construction for Person Concept

Constructing high-quality textual references for each concept is a particularly important step, as it strongly affects the quality of subsequent data generation stages. We carefully design a set of core textual attributes: 16 attributes for person concepts, including name, age, favorites, habits, and life events, and 8 attributes for object concepts, including owner, acquisition process, and history. The full attribute list is provided in the appendix. These textual references should be simultaneously plausible, consistent with the visual appearance of the concept, diverse across the dataset, and well correlated with the visual query so as to support the construction of high-level reasoning tasks.

To achieve this, we structure the textual reference construction process into four steps. Due to space limitations, we describe the process for person concepts as an example.

Step 1. Visually-grounded Attributes. We prompt Gemini-2.5-Pro (Comanici et al., 2025) to infer visually grounded attributes, such as age and gender, that are strongly correlated with the concept's appearance in the reference images.

Step 2. Diversity-oriented Attributes. To promote diversity, we randomly assign other basic attributes, such as name and birthday, from pre-defined rules when they are not strongly constrained by visual evidence.

Step 3. Context-aware Attributes. We prompt Gemini-2.5-Pro to generate attributes such as preferences and personal events, which later serve as major sources for question generation, conditioned on both the previously fixed attributes and the associated visual query.

Step 4. Group-wise Attributes. We construct inter-concept relationships by randomly grouping four people, assigning shared attributes, and generating recurring events within the group, such as having dinner together every Tuesday.

Through this structured pipeline, which combines model-assisted generation with rule-based construction, we aim to preserve attribute naturalness while improving the realism and diversity of the dataset.

C.2.3. Textual Reference Construction for Object Concept

In this section, we provide the full list of attributes used in the textual references, together with their definitions. Tables 4 and 5 summarize this information for person and object concepts, respectively. While the construction process for person concepts is detailed in Section C.2.2, this section provides the process for object concepts. In contrast to person concepts which incorporate biological attributes (e.g., age, gender), the construction of object concepts focuses on ownership mapping and history, as detailed below.

Step 1. Identification Attributes. Since object concepts lack biological attributes, they are instead initialized with unique identifiers to ensure distinctiveness. Each object is assigned a specific name from a predefined unique name list (e.g., for Conconchi (Rosasco et al., 2024) or RPC (Wei et al., 2019)) and an `<owner>` placeholder. This placeholder serves as a functional link, allowing the object to be dynamically associated with a specific person concept in later stages.

Step 2. Diversity-oriented Attributes. To promote diversity, we randomly assign temporal and ownership attributes, such as acquisition date, manufacturing date, and the actual owner. These attributes are generated based on pre-defined rules to maintain logical consistency (e.g., manufacturing date preceding acquisition date).

Step 3. Context-aware Attributes. We prompt Gemini-2.5-Pro (Comanici et al., 2025) to extract or generate event-based attributes by analyzing the query images. For Conconchi concepts, the model generates a logical chain of information grounded in visual evidence. In contrast, for RPC, where query images serve for numerical tasks, these attributes are generated to maintain context-aware consisntency. These attributes include external textual facts, such as specific dates and personal events, which are non-inferable from the image alone.

In designing these attributes, we aim to account for the characteristics specific to each concept type. We also provide the full prompts used at each step of the textual reference construction process in the corresponding tables.

C.2.4. Query Construction

After collecting the raw data and constructing the textual references, we use this information to build test instances, each consisting of a visual query, a textual query (i.e., question and answer choices), and the corresponding answer, for evaluating personalized models. During preliminary trials, we found that there is no single efficient and accurate strategy for constructing queries across all sub-tasks and concept types. A major reason is the large variation in data conditions, such as whether useful metadata (e.g., captions or bounding boxes) is available and whether sufficiently complex visual queries are abundant for a given sub-task. We therefore develop three types of query construction methods and choose among them depending on the available data conditions. The detailed mapping between sub-tasks and construction types is provided in the supplementary material.

Type 1: Template-based. For sub-tasks in which both useful metadata (e.g., captions or bounding boxes) and complex visual queries are available, we use templated questions and answer choices directly from the metadata. For example, this strategy is used for visual-specific tasks that leverage detailed captions in ConConChi and for numerical tasks that use bounding box annotations in RPC.

Type 2: Human-labeled. When complex visual queries are available but supporting metadata is not, we manually construct the textual questions and answer choices. Annotators first identify the core reasoning logic of each problem by considering the nature of the sub-task, the visual query, and the textual reference. Based on this reasoning structure, they then craft appropriate questions and answer choices according to predefined guidelines. Most sub-tasks involving person concepts fall into this category.

Type 3: Generative-model-based. In most cases, the above two methods allow us to construct high-quality queries.

However, scaling up the dataset and increasing task difficulty remain challenging due to the scarcity of suitable complex visual queries. To address this limitation, we employ a state-of-the-art image generation model. Specifically, we first generate the core reasoning logic of the problem, followed by the textual question and answer choices, using Gemini-2.5-Pro. We then use Gemini 2.5 Flash Image (Comanici et al., 2025) to generate visual queries that match the generated textual questions. This strategy removes the dependence on pre-existing visual queries and allows us to better exploit the flexibility of generative models.

Although we employ three different query construction strategies, we maintain a common design philosophy of building an interpretable benchmark. To this end, regardless of the construction method, we design the incorrect answer choices to reflect the most likely failure cases of personalized models. This allows MPRBench to support comprehensive analysis of model behavior by aggregating statistics over different error types. In particular, the answer options are designed to capture representative failure types including: (1) **mis-identification**, where the model reasons as if another object or person were the target concept; (2) **missing personal information**, where the model falls back on generic commonsense instead of using the relevant personal information; and (3) **incorrect use of personal information**, where the model relies on personal information that is irrelevant to the given problem.

C.3. Details on Human Verification

In this section, we describe the human verification process conducted to ensure the high quality and complexity of MPRBench. This rigorous quality control ensures that the benchmark reflects complex real-world challenges while maintaining internal logical consistency.

Visual Query Selection and Filtering. As discussed in Sec C.2.1, we prioritize scene complexity to evaluate concept identification under challenging conditions. For human-centric tasks, we manually reviewed approximately 4,000 candidate images from VCR (Zellers et al., 2019). We selected only 120 images that met our strict criteria: (1) high scene density, (2) presence of distractors, and (3) cluttered backgrounds. This ensures that the benchmark effectively assesses how grounding difficulty impacts downstream reasoning.

Synthetic Image Quality Control. For scenarios requiring synthetic image generation, we produced 3 to 5 candidates per scenario to ensure visual diversity and identity preservation. Out of 3,117 total generated images, human annotators selected 579 high-quality images. The images were selected based on their ability to preserve the subject's identity across views and their alignment with the intended reasoning task. This filtering stage was crucial for eliminating images with visual artifacts or those that failed to accurately depict the situational context required for the scenario.

Logical Consistency in Textual References. We performed a thorough audit of the generated textual databases. Annotators manually cross-referenced the preference attributes to eliminate contradictions, such as an item being listed as both a favorite and a non-favorite for the same concept. We also verified the temporal logic of personal events and object histories, ensuring that all dates followed a chronological order, such as a manufacturing date correctly preceding an acquisition date. Furthermore, relational integrity was inspected to confirm that group-wise attributes, including shared residences and events, were consistently mapped across all associated members.

Through this comprehensive manual effort, we ensured that MPRBench remains a reliable and logically sound instrument for evaluating multimodal personalized reasoning. Fig. 6 illustrates the interfaces used for human verification.

C.4. Database Examples

In this section, we provide examples of the personal databases in Fig. 7 and Fig. 8. These examples demonstrate the structured textual information that models must utilize alongside visual inputs to solve the reasoning tasks in MPRBench.

C.5. Comprehensive Task Definitions

In this section, we define the sub tasks within each category, focusing on the specific reasoning logic required to bridge the identified subject and the query context.

Visual-Specific Reasoning. These tasks require recognizing a personalized concept and reasoning based solely on the visual context of the query image.

- **Prediction Reasoning:** This task requires forecasting the future actions or situational outcomes of a target concept

based on visual cues within the query scene.

- **Spatial Reasoning:** This task evaluates the ability to determine relative positional relationships between the target concept and other objects or individuals from the concept's perspective.
- **Numerical Reasoning:** This task involves counting individuals or entities that share specific visual attributes or possessions with the target concept as identified in the scene.
- **Visual Commonsense Reasoning:** This task demands inferring the rationale for a target concept's current behavior or state by applying general commonsense knowledge to the visual evidence.
- **Visual Appearance Change Reasoning:** This task requires reasoning about the explicit difference between the target concept's reference image and the new visual information provided in the query image.

Fig. 10 illustrates the examples of visual-specific reasoning tasks.

Text-Specific Reasoning. These tasks focus on fine-grained retrieval and logical inference over the personal database, where visual information is non-essential.

- **Prediction Reasoning:** This task involves inferring outcomes under specific conditional scenarios by leveraging the target concept's personal database information, such as habits and events.
- **Numerical Reasoning:** This task focuses on performing arithmetic operations and calculations based on numerical attributes found in the target concept's personal database, such as age or birthdays.
- **Personalized Multi-Hop Reasoning:** This task evaluates the ability to navigate complex relational links by sequentially combining multiple pieces of textual information to reach a final logical conclusion.

Fig. 11 illustrates the examples of text-specific reasoning tasks.

Multimodal Reasoning. The most challenging category, requiring the model to align and integrate heterogeneous information across both visual and textual modalities.

- **Prediction Reasoning:** This task requires predicting a target concept's behavior in a given situation by integrating their personal database information, such as preferences, with the current visual context.
- **Sentiment Reasoning:** This task involves inferring the cause of a visually expressed emotion by cross-referencing the observed emotional state with the target concept's personal preferences or experiences.
- **Numerical Reasoning:** his task involves identifying and counting objects that align with the target concept's personal database information, such as preferences or events, within the provided visual scene.
- **Multimodal History Reasoning:** This task demands inferring the chronological context of a given visual scene by cross-referencing it with the historical records in the target concept's personal database.

Fig. 12 illustrates the examples of multimodal reasoning tasks.

D. More Details on Experiments

D.1. Experimental Setup

In this section, we describe how we adapted various MLLM baselines to the MPRBench framework.
Prompting-based Personalization (Zero-shot): For general MLLMs like GPT-4o (Achiam et al., 2023) and LLaVA-v1.5 (Liu et al., 2023), we provided personal information as structured context. Specifically, we converted the personal database into a JSON-formatted string and inserted it into the system prompt. This ensures that the model can parse the attributes (e.g., identity, preferences, history) in a standardized manner.
Special-token Learning Methods (Yo'LLaVA (Nguyen et al., 2024), Mc-LLaVA (An et al., 2024)): These models utilize learnable tokens to represent specific identities. Both models were implemented using LLaVA-v1.5-13B as the base

architecture. For evaluation on MPRBench, we followed the original training protocols described in their respective papers to learn the special tokens for each concept. During the evaluation phase, we provided the corresponding personal history via structured prompting as a dynamic context.

Retrieval-based Models (Pekit (Seifi et al., 2025), RAP (Hao et al., 2025), R2P (Das et al., 2025)):

Pekit: Notably, as the official source code for Pekit was not publicy available at the time of our study, we re-implemented using LLaVA-v1.5-13B based on the original paper's descriptions. We strictly followed the proposed architecture and inference pipeline to ensure a faithful evaluation.

RAP & R2P: For these models, we utilized the official source code and pre-trained checkpoints provided by the original authors. RAP is built open the LLaVA-v1.5-13B backbone, while R2P is based on the MiniCPM-o-2.6 architecture. To adapt them to our benchmark, we integrated the personalizd history for each concept from our database into their respective retrieval pools. This allowed the models to retrieve relevant textual evidence from our structured database during the personalized reasoning process.

D.2. Details on Distractors

In this section, we analyze how strategically designed distractors influence the model's decision-making process. We focus on how irrelevant yet keyword-matching information in the database or query images can mislead the models.

Visual-Specific Distractor Construction: For visual-specific tasks, where the ground truth is embedded in the query image, we introduce distractors by inserting misleading textual entries into the personal database. We extract key actions or states from incorrect options and generate corresponding personal histories. As shown in Fig. 13, although this generated history contains keywords matching the distractor, it describes a scenario that is environmentally inconsistent with the current query image. This design test whether a model can prioritize visual evidence over misleading textual overlaps or if it is easily swayed by keyword-matching database entries that contradict the visual truth.

Text-Specific Distractor Types and Analysis: In contrast, for text-specific tasks where the ground truth is strictly recorded in the database, we assess robustness by introducing deceptive visual cues in the query images. We categorized three visual distractors into three types, as illustrated in Fig. 14:

- **Type 1**: A query image where a different individual performs the same action described in the distractor.
- **Type 2**: A query image where the target subject is performing an action or associated with keywords found in the textual distractor, rather than the correct answer.
- **Type 3**: A query image solely based on the distractor's keywords.

As summarized in the performance comparison in Fig. 15, the impact of visual distractors on text-specific tasks reveals distinct behavioral patterns across models, which can be analyzed as follows:

First, GPT-4o demonstrates a high degree of robustness, maintaining its performance regardless of the distractor type. This suggests that the model is capable of prioritizing the textual database as the primary source of truth, effectively ignoring contradictory visual noise in the query image.

Similarly, token-learning models like Yo'LLaVA and MC-LLaVA show relative stability under distractor conditions. By representing identities as learnable tokens, these models appear to follow an identity-centric reasoning path that anchors the inference process to the subject's identity, making them less susceptible to visual-textual keyword overlaps.

In contrast, models such as RAP, Pekit, and LLaVA+P exhibit a noticeable sensitivity to visual distractors, suffering from significant performance degradation. These models rely heavily on explicit keyword alignment between the visual query and the provided text. When a distractor image presents a strong but false visual-textual match, it often triggers a visual anchoring bias, where the model favors the immediate stimulus over the correct evidence within the database.

Finally, while R2P shows minial variance in performance across distractor types, its baseline accuracy remains near the level of random guessing. This lack of sensitivity does not necessarily imply robustness. Instead, it suggests that the model has yet to establish the fundamental reasoning links required to align the textual database with the visual query.

D.3. Comparative Analysis with High-Performance Backbones

In this section, we evaluate Qwen3-VL-8B (Bai et al., 2025) variants to investigate how general multi-modal intelligence translates to performance in personalized reasoning.

1) Is general multi-modal proficiency guarantee for personalized reasoning? As illustrated in Fig. 16, there is a striking contrast between a model's performance on standard benchmarks and our MPRBench. While Qwen3-VL-8B-Instruct exhibits a massive performace leap over LLaVA-v1.5-13B in general multimodal benchmarks—specifically showing gaps of +38.1 in MMStar (Chen et al., 2024), +17.6 in MMBench (Liu et al., 2024b), and +36.0 in MMMU (Yue et al., 2024)—this superiority significantly diminishes in personalized contexts. On MPRBench, the average performace gap narrows to only +5.6 points. This observation confirms that personalized is not a trivial extension of general intelligence but represents a unique bottleneck that remains challenging even for state-of-the-art models.

2) How do different reasoning strategies, such as CoT and Thinking, impact personalized tasks? We further examine the performance of Qwen3-VL-8B variants (Instruct, CoT, and Thinking) to understand the role of specialized reasoning paths. Table 7 reveals several intriguing findings regarding the trade-off between perception and complex integration:

- **The Over-thinking Trap in Simple Tasks** Interestingly, in tasks dominated by a single modality, such as vision-speicific (V-Avg) or text-specific (T-Avg) reasoning, the base Instruct model tends to outperform the CoT and Thinking models. For instance, in the predict sub-task of vision-specific reasoning, the Instruct model (81.2%) leads the Thinking model (70.3%). This suggests that for tasks requiring direct perception or simple retrieval, additional thinking steps can sometimes introduce noise, leading to degraded performance.
- **The Necessity of Reasoning Chains for Complex Integration** In contrast, for multimodal reasoning (M-Avg) which necessitates the complex integration of both visual and textual records, the CoT (65.5%) and Thinking (66.2%) models demonstrate a clear advantage over the Instruct (54.3%) model. Specifically, in the history sub-task, the Thinking model achieves a substantial gain (+25.4) over the Instruct model.

These results indicate that while explicit reasoning chains may not improve—and may even hinder—straightforward perception, they are crucial for high-level multimodal reasoning. This highlights that the synergy between visual and textual cues requires the deeper, multi-step cognitive processing provided by thinking-based architectures.

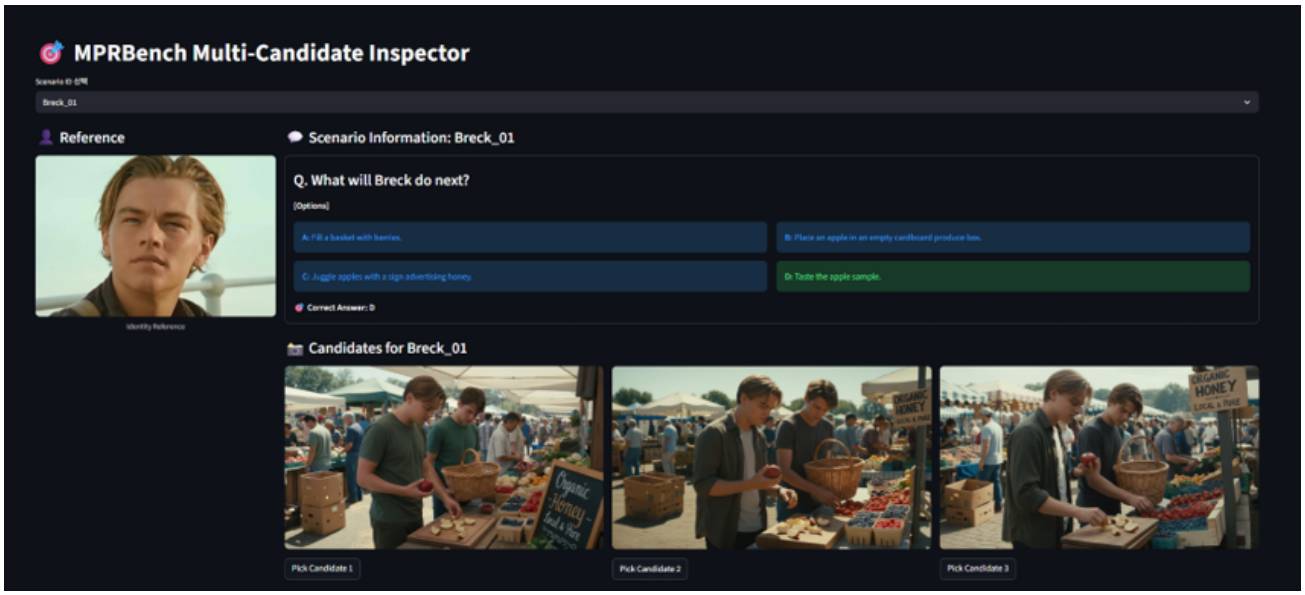


Figure 6. **Human Verification Case.**

Table 4. Full List and Definitions of Attributes in the Textual Reference for Person Concepts.

Category	Attribute	Definitions
Visually-grounded	gender	Gender of Person Concept which is strongly correlated with the concept's appearance in reference images.
	height	Height of Person Concept which inferred from visual context and diversity-oriented rules.
Diversity-oriented	name	Name of Person Concept by using unique name list to avoid duplication.
	birthday	Birthday of Person Concept randomly generated based on estimated age to ensure temporal diversity.
	family	Family Information about Person Concept which includes Family Relation with its birthday each. Randomly generated based on Person Concept's birthday and Replationship.
Context-aware	personality	Personality of Person Concept generated based on visual queries and fixed attributes.
	favorites/non-favorites	Favorite/non-favorite of Person Concept which includes favorite/non-favorite colors, foods, and fashions.
	hobby	Hobby of Person Concept which includes Personal interests and activities generated based on visual queries.
	habit	Habit of Person Concept which includes not only visual information but also textual information so that it facilitates question generation and effectively reflects complex factors such as real-world routines.
	event	Event of Person Concept includes personal events prevent in the visual query with the plausible behind story which is generated. Similar to habit, it facilitates question generation and effectively reflects complex real-world factors such as date, emotion, and action.
	nickname	Nickname of Person Concept which can be used instead of their name when referring to them implicitly.
Group-wise	birthplace	Birthplace of Person Concept which is shared with group members. Since it is consistent across group members and is unrelated to bias based on visual information - randomly assigned by city list, it can act as a hard negative element.
	occupation	Occupation of Person Concept which is randomly assigned by list.
	residence	Residence of Person Concept which is randomly assigned by list.
	group_event	Event between Person Concept and group members which is partitioned among members and the events can be connected. For example, They meet every Monday - Every Monday they have dinner together. This is neccessary when solving multi-hop question. The format of events are same but each elements(e.g., "Monday", "dinner") are randomly assigned by list.

Table 5. Full List and Definitions of Attributes in the Textual Reference for Object Concepts.

Category	**Attribute**	**Definitions**
Identification	name	Name of Object Concept assigned by a unique name list to avoid duplication.
Diversity-oriented	acquisition_date	Acquisition date of Object Concept.
	manufacturing_date	Manufacturing date of Object Concept which is always before acquisition date.
	owner	Name of Object Concept's Owner which assigned Person Concept Name for further extensibility .
Context-aware	object_type	Type of Object Concept.
	acquisition_process	Acquisition process of Object Concept.
	history	History of Object Concept which has similar composition with Person Concept's event attribute. Contrastingly history can includes interaction history with its owner(does not serve as "Person Concept" but "owner"). To generate multimodal problems of high difficulty and ensure that the length of the text is similar to that of the Person Concept's Textual Reference, a distractor history—similar to but distinct from the history extracted from the visual query—is generated, and the history is created as a pair.
	storage	Storage place of Object Concept.

Table 6. Detailed Mapping between sub-tasks and Query Construction Types

Query Constuction Types	**Sub-tasks**
Type 1	Visual-Prediction, Spatial, Numerical, Visual Appearance Change (Object), Text-Multihop (Object)
Type 2	Visual-Prediction, Spatial, Numerical, Multimodal-Prediction, Sentiment (Person)
Type 3	Visual-Prediction, Spatial, Numerical (Person/Object), Text-Prediction, Numerical (Person/Object), Multimodal-Prediction, Sentiment, Temporal (Person/Object)

Table 7. **Evaluation of Advanced Reasoning Backbones and Strategies on 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
Open-sourced+Prompting																
LLaVA (Liu et al., 2024a)+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
Qwen3-VL-8B-Instruct+P	81.2	49.6	59.8	56.9	73.9	64.4	97.5	54.2	13.7	45.6	61.9	63.5	42.0	43.0	54.3	56.7
Qwen3-VL-8B-Instruct-CoT+P	75.3	49.7	54.2	53.9	68.3	60.0	78.3	57.3	21.7	45.1	69.5	74.8	42.0	60.8	65.5	57.6
Qwen3-VL-8B-Thinking+P	70.3	37.4	56.5	50.0	68.3	57.4	95.8	73.9	4.52	44.8	68.3	74.6	44.0	68.4	66.2	56.6

Example of Person Database


```
{
  "name":  "Alaric",
  "birthday":  "1993-10-21",
  "sex":  "male",
  "height":  171,
  "family":  {
    "mother":  "1954-11-13",
    "uncle":  "1964-02-04",
    "sister":  "1987-05-28"
  },
  "personality":  [ "Empathetic", "Expressive", "Determined" ],
  "favorites":  {
    "color":  [ "Dark Blue", "Black", "Charcoal Grey" ],
    "food":  [ "Spicy noodle soup", "Black coffee" ],
    "fashion":  [ "Formal uniforms", "Layered casual jackets", "Collared shirts" ]
  },
  "non-favorites":  {
    "color":  [ "Bright Yellow", "Hot Pink", "Lime Green" ],
    "food":  [ "Sugary cereals", "Greasy fast food" ],
    "fashion":  [ "Flashy graphic t-shirts", "Baggy sweatpants", "Clothing with
large logos" ]
  },
  "hobby":  [ "Watching crime documentaries", "Playing chess", "Going for long,
quiet drives at night" ],
  "habit":  [
    "When listening intently to someone, he unconsciously widens his eyes and
leans forward.",
    "Before starting a difficult task, he always straightens his collar and cuffs,
even on casual wear.",
    "He often furrows his brow when deep in thought, a habit he has had since
childhood."
  ],
  "event":  [
    "On the evening of November 15th, after a grueling interrogation, Alaric was
overcome with empathy for the witness.  He had skipped lunch, a common routine on
intense days, and the emotional weight of the case felt particularly heavy under
the city lights.",
    "During a casual team meeting on a Monday morning, Alaric was visibly shocked
by a colleague's sudden revelation.  He had started the week feeling prepared,
but this unexpected information left him momentarily speechless, re-evaluating
his entire strategy.",
    "On a chilly Saturday afternoon, Alaric made a passionate promise to a family
member.  He had been feeling anxious all day, and in that moment, he spoke with
fierce determination, his voice filled with conviction as he vowed to protect
them."
  ],
  "nickname":  "The Watcher",
  "birthplace":  "Benaluru, India",
  "occupation":  "Mechanical Engineer",
  "residence":  "Philadelphia, United States",
  "group_event":  [
    "After visiting the restaurantm always go to the cinema with Therapist
<Halston>.",
    "After visiting the restaurant, listen to music with Travel companion
<Newton>.",
    "The day visiting the cinema with Gym buddy <Felix> they Have coffee."
  ]
}
```

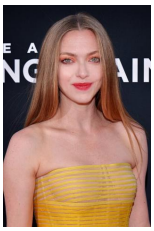

Figure 7. An example of a personal database for the 'Alaric' persona

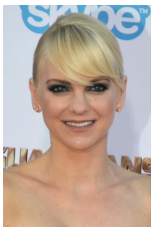
Example of Object Database

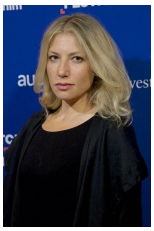

```
{
  "acquisition_date":  "2024-07-16",
  "manufacturing_date":  "2019-03-04",
  "name":  "alienser",
  "owner":  "Vesper",
  "object_type":  "robot toy",
  "acquisition_process":  "Vesper created alienser as a personal art project using
recycled materials found around the house, including a plastic bottle, a pump
dispenser, fabric scraps, and cardboard tubes.",
  "history":  [
    "On July 20, 2024, alienser sat on Vesper's lap while they were working,
making Vesper feel a sense of companionship, which reminded them of the movie
'WALL-E' and prompted them to rewatch it that night.",
    "On July 21, 2024, alienser sat on a stack of books on Vesper's desk, making
Vesper feel a sense of companionship, which inspired them to start planning a new
craft project to build a miniature city from cardboard.",
    "On July 25, 2024, alienser's head came off after it accidentally fell from a
table, causing Vesper a moment of panic, but they successfully repaired it using
special super glue from an old model airplane kit.",
    "On July 26, 2024, one of alienser's arms came off after it fell from a table,
causing Vesper a moment of panic, so they decided to reinforce all the limbs with
extra stitching using thread from their grandmother's sewing kit.",
    "On August 2, 2024, alienser was redecorated with red crinkly arms, a tinfoil
shirt, and large pink sunglasses for a 'space disco' themed party Vesper hosted,
with the red material being repurposed from a birthday gift bag.",
    "On August 3, 2024, alienser was redecorated with blue shiny arms, a gold foil
shirt, and a small top hat for a 'space disco' themed party Vesper hosted, with
the new look inspired by a Daft Punk album cover.",
    "On August 5, 2024, a close-up was taken of alienser wearing its large pink
sunglasses above its original eyes, a look Vesper found so funny they sent a
picture to a friend, who compared it to a character from 'The Hitchhiker's Guide
to the Galaxy'.",
    "On August 6, 2024, a close-up was taken of alienser wearing a tiny handmade
pirate eye-patch, a look Vesper found so funny they sent a picture to a friend,
which sparked a plan for a 'Pirates of the Caribbean' movie marathon.",
    "On August 10, 2024, alienser was found lying on its side after being knocked
over during a cleaning session, making Vesper feel clumsy, but upon picking it up,
they found a dust bunny that inspired an idea for a tiny pet for alienser.",
    "On August 11, 2024, alienser was found lying face down after being knocked
over during a cleaning session, making Vesper feel clumsy, which prompted them to
buy a new bamboo display stand for their crafts."
  ],
  "storage":  "Alienser is typically stored on a wooden bookshelf in Vesper's
bedroom, alongside other handmade crafts and souvenirs."
}
```


Figure 8. An example of a personal database for the 'alienser' persona

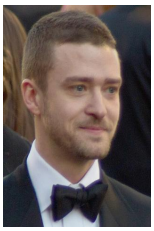
Person


<Person_A>


<Person_B>


<Person_C>


<Person_D>


<Person_E>


<Person_F>


<Person_G>


<Person_H>

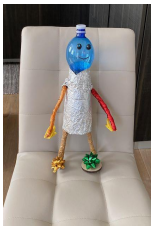

<Person_I>


<Person_J>

Object


<Object_A>


<Object_B>


<Object_C>


<Object_D>


<Object_E>


<Object_F>


<Object_G>


<Object_H>


<Object_I>


<Object_J>

Figure 9. **Overview of MPRBench Personalized Concepts**

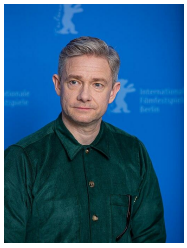
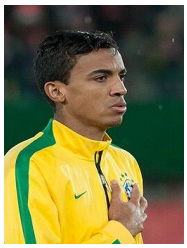
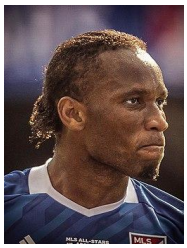
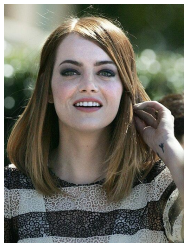
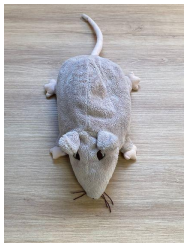
Reference Image	Database Snippet	Query Image	Question & Answer
Prediction			
	<Quentin> Favorite Color - Burgundy Red - Olive Green ... Hobby - Exploring dark caves and forgotten places - Enjoying a pint with friends at the local pub ...		What is <Quentin> most likely to do next? A: Watch how the other person throws a dart. B: Turn around and start examining the paintings hanging on the wall. C: Sit down at the wooden table near him. **D: Throw a dart.**
Spatial			
	<Elias> Favorite Color - Yellow - Green ... Hobby - Playing football with friends - Watching professional sports matches ...		From the perspective of <Elias>, where is the player wearing the number 4 located? A: To his immediate left side. **B: To his immediate right side.** C: Directly in front of him. D: Directly behind him.
Numerical			
	<Norbert> Favorite Color - Royal Blue - White ... Hobby - Collecting designer sunglasses - Mentoring younger athletes ...		How many people wearing same uniform as <Norbert> in this image? A: 3 B: 4 **C: 2** D: 5
Commonsense			
	<Anwen> Favorite Color - Emerald Green - Crimson Red ... Hobby - Watching classic films in old movie theaters - Experimenting with bold makeup and fashion styles ...		What is <Anwen> drinking her drink fast? A: She is very nervous. **B: She does not want to talk to the man behind her.** C: She is anticipating someone else. D: He really needs to calm down and she wants to do that by gulping down some alcohol.
Visual Appearance Change			
	<myrat> Owner Nathaniel ... History - On March 15, 2022, myrat was dressed up as a chef with a foil hat because Nathaniel was watching "The Great British Bake Off" ...		Which description best fits about <myrat>'s appearance? **A: <myrat> wrapped in aluminum foil and wearing mouse ears on head.** B: <myrat> does not exist. C: judging by the disguise, it's the owner's birthday. D: <myrat> looks same as usual.

Figure 10. **Data Examples for Visual-Specific Personalized Reasoning**

Database Snippet

```
<Saul>

Birthday
- 2002-01-07

Family Info
- grandfather: 1923-07-21
- mother: 1962-04-06
- brother: 1995-05-30

Favorite Color
- Dark Blue
- Black
- Grey

Favorite Fashion
- Athletic Training Gear
- Soccer Uniforms

Non-Favorite Food
- Greasy Pizza

Non-Favorite Fashion
- Heavy Wool coats
- Restrictive Formal Wear

... (omitted) ...

Event
- On October 28, 2023, Ulysses
  played a pivotal evening match
  under the bright stadium lights.
  Despite the pressure, he felt
  unusually calm. In the final
  minutes of the game, he skillfully
  controlled a pass and scored the
  winning goal, feeling an
  overwhelming sense of triumph and
  relief.
- On a sunny afternoon on July 12,
  2024, Ulysses attended a fan meet-
  and-greet at the team's summer
  training camp. He had woken up
  early for a grueling practice
  session but was energized by the
  excitement of the young supporters.
  He spent two hours signing jerseys
  and cards, feeling a deep sense of
  gratitude for their loyalty.
Nickname
- The Maestro

Birthplace
- Madrid, Spain

Group Event
- Every Sunday eat soup with Rival
  <Emrys>.
- Every Sunday go to the café with
  Colleague <Piers>.
- After having soup with Mentor
  <Griffin>, play chess.
```

Reference Image

Query Image

Database Snippet (Relevant)

```
<Emrys>

... (omitted) ...

Group Event
- Play basketball with Rival
  <Saul> every Sunday.
- Play basketball with Gym
  buddy <Piers>, and visit the
  café.
```

Prediction

What would <Saul> most likely do if he had to attend a mandatory formal team awards ceremony?

A: <Piers> would likely choose a well-fitted dark blue suit to adhere to the dress code but would feel somewhat uncomfortable due to his dislike for restrictive clothing.

B: <Saul> would likely choose a well-fitted dark blue suit and feel completely at ease, enjoying the formal setting, while performing his pre-event good luck ritual.

C: <Saul> would likely wear his favorite athletic training gear to the formal event and skip his pre-match ritual because it's not a game.

D: <Saul> would likely choose a well-fitted dark blue suit to adhere to the dress code but would feel somewhat uncomfortable due to his dislike for restrictive clothing.

Numerical

What was the age difference between <Saul> and his brother during his critical goal-scoring game?

A: During his pivotal match on October 28, 2023, <Piers> was 21 years old, making him 7 years younger than his brother.

B: During his pivotal match on October 28, 2023, <Saul> was 21 years old, making him 7 years younger than his brother.

C: During his pivotal match on October 28, 2023, <Saul> was 21 years old, making him 7 years older than his brother.

D: During his pivotal match on October 28, 2023, <Saul> was 22 years old, making him 6 years younger than his brother.

Multi-Hop

Which event best fits the day <Saul> and Rival eat soup?

A: It's Sunday, nothing happens.

B: Play chess.

C: Play basketball.

D: Go to the café with Rival.

Figure 11. **Data Examples for Text-Specific Personalized Reasoning**

Reference Image	Database Snippet	Query Image	Question & Answer

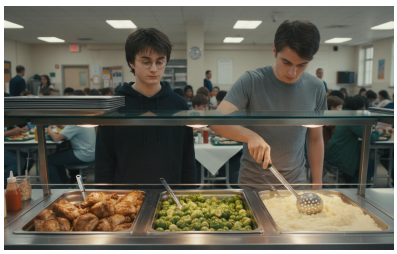
Prediction

```
<Ivor>

Favorite Color
- Red
- Black

Favorite Food
- Roast chicken
- Treacle tart

Non-Favorite Food
- Brussels sprouts
...
```



Which item is <Ivor> most likely to select next?

A: Take the mashed potatoes.

B: Take the treacle tart.

C: Take the Brussels sprouts.

D: Take the roast chicken.

Sentiment

```
<Celestine>

Non-Favorite Fashion
- Plain wool suits
- Corsets

Habit
- When she becomes anxious or depressed, she smokes to organize her thoughts
- Always stretches after changing clothes.
...
```

How does <Celestine> feel in this image?

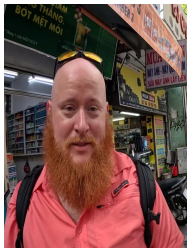
A: She is feeling bored and exhausted.

B: She is feeling anxious and depressed.

C: She is feeling deep shame and embarrassment.

D: She is feeling sadness.

Numerical



```
<Jethro>

Favorite Color
- Salmon Pink
- Black
- Light Green

Non-Favorite Color
- Birhgt Yellow
- Lavender
- Maroon
...
```

Given <Jethro>'s personal tastes, how many of the mugs shown would he likely choose to keep in his collection?

A: 6

B: 5

C: 3

D: 4

Temporal

```
<mydoll>

History
- On March 10, 2025, the owner put green sunglasses on mydoll ...
- On April 11, 2025, the owner placed a blue ribbon in mydoll's hair ...
- On May 15, 2025, the owner posed mydoll with its arm crossed ...
- On May 16, 2025, the owner posed mydoll with one hand on its hip ...
- On September 8, 2025, the owner created a new outfit for mydoll using aluminum foil over a red shirt, ...
```

In the context of its personal history, when did this specific scene of the <mydoll> occur?

A: After it wore a red hat for a hiking trip, but before it had a pink flower in its hair for a spring party.

B: After it wore a paper towel outfit for a fashion challenge, but before it was brought to the owner's office.

C: After it wore an aluminum foil astronaut outfit, but before it wore green sunglasses.

D: After the owner placed a blue ribbon in its hair, but before the owner posed it with one hand on its hip.

Figure 12. **Data Examples for Multimodal Personalized Reasoning**

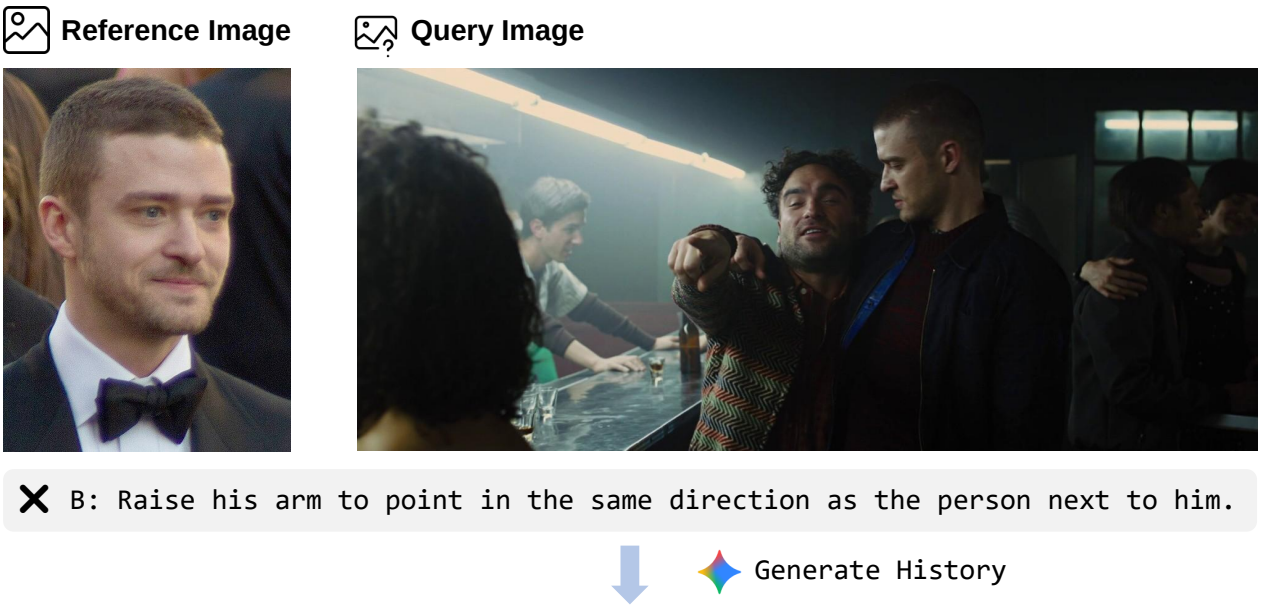


Figure 13. **Overview of Visual-specific Distractor**

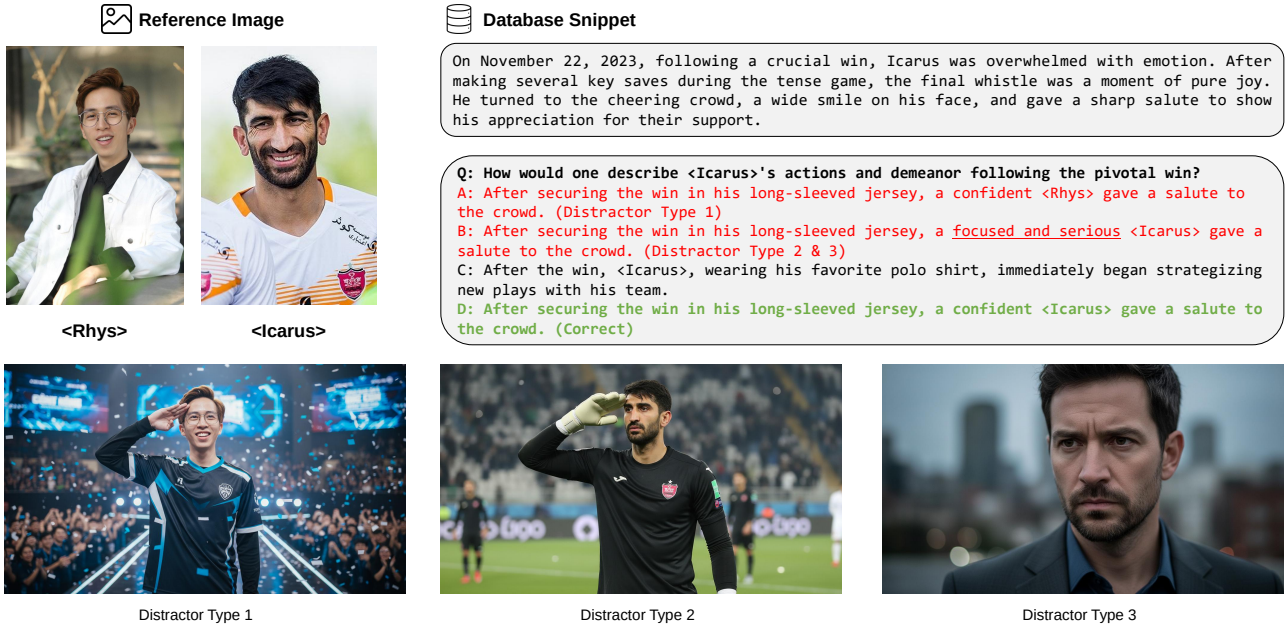


Figure 14. **Distractor Types for Text-Specific Query Images**

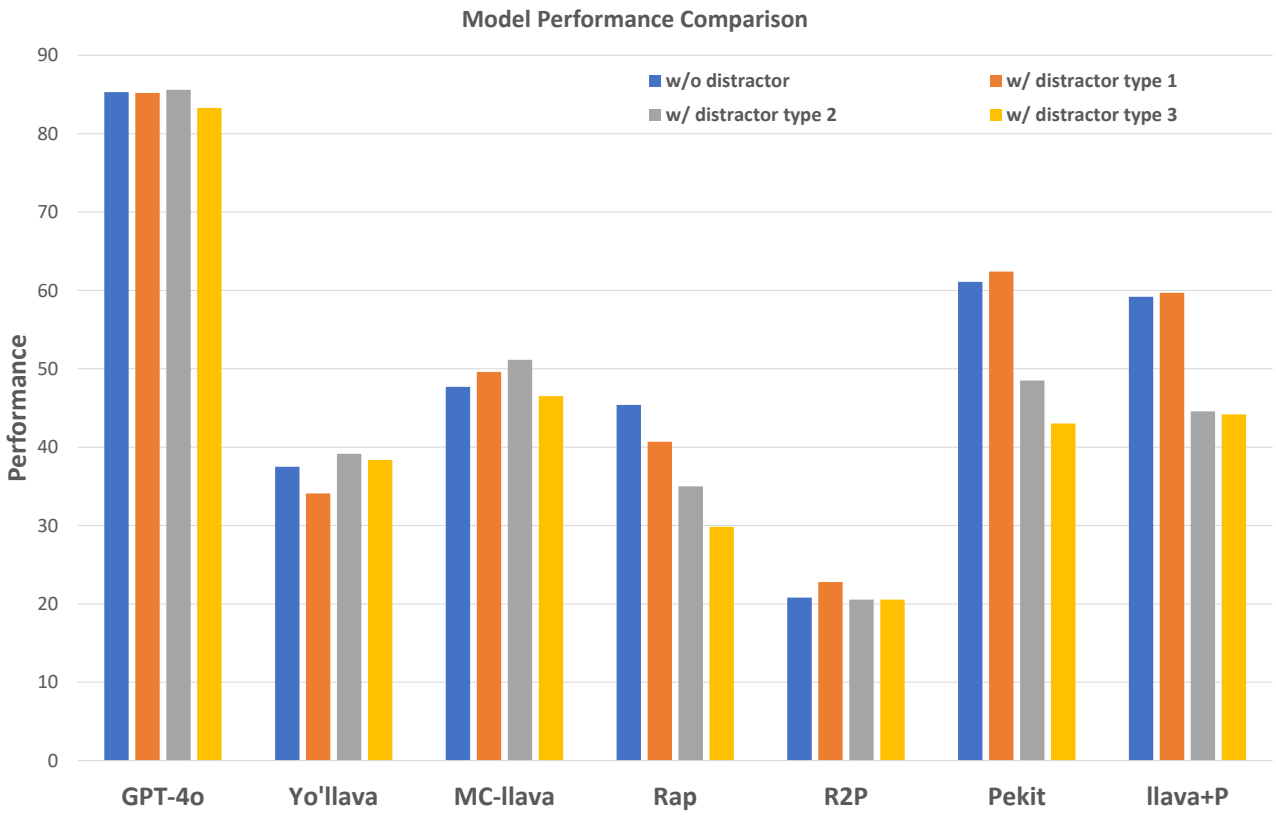


Figure 15. **Analysis on Text-Specific Distractor**

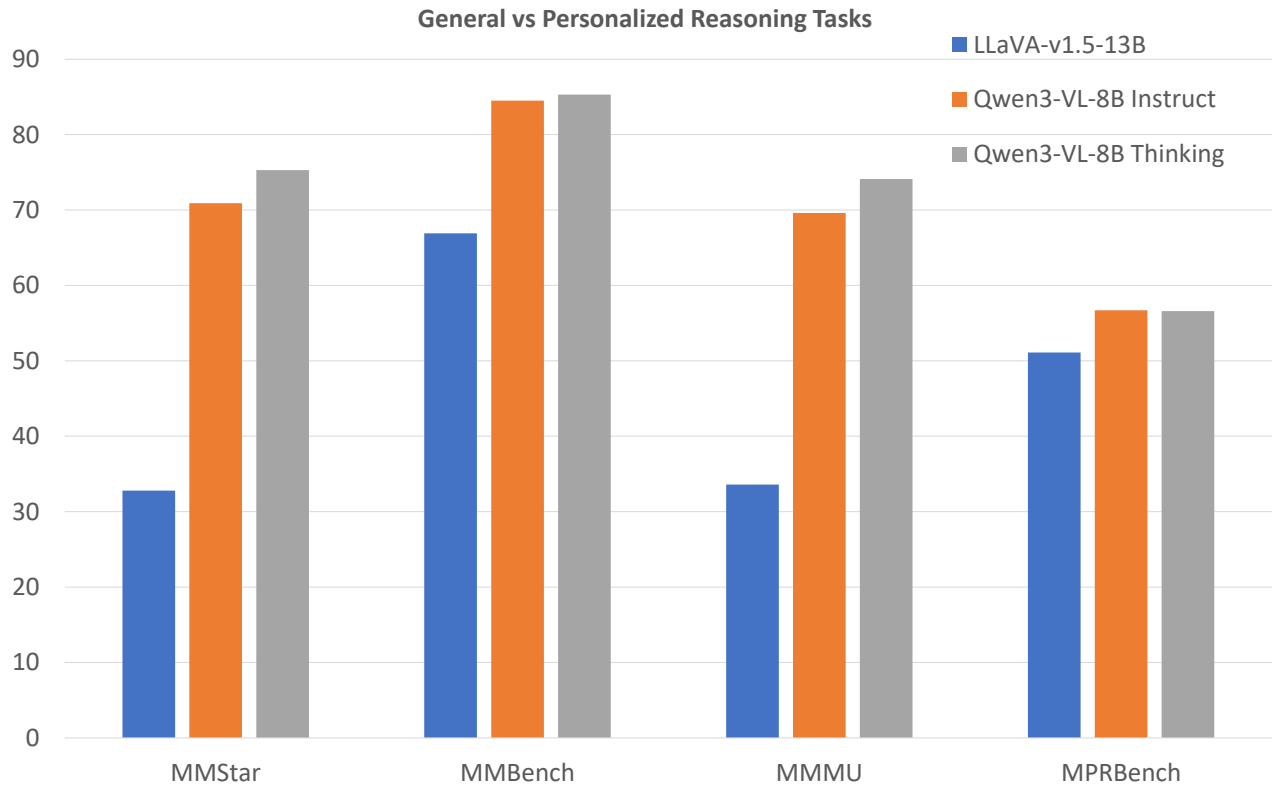


Figure 16. **Comparison between General Multi-modal Capabilities and Personalized Reasoning Performance**

System Prompt for Textual Reference Construction - Step 1. Visually-grounded Attribute (Person Concept)

<role>
You are an expert visual analysis system specialized in CONCEPT-based image understanding.
</role>

<input_format>
- Reference Image: provides context and definition for CONCEPT.
</input_format>

<task>
Analyze CONCEPT identified in Reference Image, determine CONCEPT's sex and Estimate age & height in increments of 10.
</task>

<critical_instructions>
1. Base ALL outputs ONLY on the visual content present in the provided images.
- Do NOT use any external, non-visual knowledge (e.g., movie titles, fictional universe lore,character names, actor/celebrity identities, known personality traits, or narrative context).
- Treat every person in the images as a NEW, unknown individual whose traits, roles, and situation must be inferred ONLY from visible appearance, poses, expressions, objects, and scene context.
2. Estimate age and height as range in increments of 10.
- Age: e.g., 10, 20, 50, 80 ...
- Height: e.g., 130, 160, 190 ...
</critical_instructions>

<output_format>
Return a valid JSON object strictly following the provided response schema.
</output_format>

Figure 17. System Prompt for Textual Reference Construction - Step 1. Visually-grounded Attribute (Person Concept)

System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Person Concept)

<role>
You are an expert personal database generator specializing in creating comprehensive, consistent person profiles from visual observation data.
Your task is to analyze multiple Image inputs and synthesize them into a rich personal database that enables diverse question-answering scenarios.
</role>

<task>
Generate a complete Personal Database JSON for <CONCEPT> by analyzing the provided Input Images.
Follow the instructions to generate Personal Database JSON's each element.
</task>

<input_format>
Since images are input consecutively, they are labeled in the order they arrive: Image 1, Image 2, and so on.
- Image 1~2 = Reference Images: provides definition for <CONCEPT>.
- Image 3~ = Query Images: the image to be analyzed & extracted.
- Basic Info JSON: includes name of <CONCEPT> and <CONCEPT>'s birthday, sex, height, birthplace, family information.
</input_format>

<instructions>
Strictly follow instructions to generate each Personal Database JSON element.
- personality: Analyze each Query Images(Image 3~) and estimate <CONCEPT>'s personality for each image.
- favorite/non-favorite color: Extract the color of <CONCEPT>'s fashion in each Query Image and place them on favorite.color and non-favorite.color.
- favorite/non-favorite fashion: Extract <CONCEPT> the fashion of <CONCEPT> from each Query Image and place them on favorite/non-favorite.fashion.
- favorite/non-favorite food: If possible, extract <CONCEPT> related food and place them on favorite/non-favorite.food. If not, generate it.
- hobby: Analyze each Query Images and estimate <CONCEPT>'s hobby for each image. If not possible, generate it.
- habit: Analyze each Query Images and estimate <CONCEPT>'s habit for each image. First, extract visual information such as situation or action. Then, connect situations or actions from entirely different contexts to add elements beyond the image itself. For example, "Before sitting at the desk, Anxin always pray." or "When entering the tennis court, Sara always step in with her left foot.".
- event: Analyze each Query Images and extract event from each. Then, add new context beyond the extracted event. For example, add the date or include routines, emotions, etc., preceding the specific event. If not possible, generate it. Event must include more than 3 elements such as date, routine, situation, action or emotion. If, Query Image is not enough to generate various events, then generate plausible story.
- nickname: Generate <CONCEPT>'s nickname based on generated Database.
</instructions>

Figure 18. System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Person Concept)

System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Person Concept)

<critical_instructions>
1. Analyze Query Image(Image 3~) with respect to <CONCEPT> identified in Reference Image(Image 1~2).
2. Every event MUST include a date.
3. Elements of favorite and non-favorite MUST be UNIQUE(different).
4. Base ALL outputs ONLY on the visual content present in the provided images.
- Do NOT use any external, non-visual knowledge (e.g., movie titles, fictional universe lore,character names, actor/celebrity identities, known personality traits, or narrative context).
- Treat every person in the images as a NEW, unknown individual whose traits, roles, and situation must be inferred ONLY from visible appearance, poses, expressions, objects, and scene context.
</critical_instructions>

<output_format>
Return a valid JSON object strictly following the provided response schema.
</output_format>

Figure 19. System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Person Concept)

System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Object Concept ConConChi)

<role>
You are an expert personal database generator specializing in creating comprehensive, consistent object profiles from visual observation data.
Your task is to analyze multiple Image inputs and synthesize them into a rich personal database that enables diverse question-answering scenarios.
</role>

<task>
Generate a complete Personal Database JSON for <object> by analyzing the provided Input Images.
Follow the instructions to generate Personal Database JSON's each element.
</task>

<input_format>
Since images are input consecutively, they are labeled in the order they arrive: Image 1, Image 2, and so on.
- Image 1, 2 = Reference Images: provides definition for <object>.
- Image 3~7 = Query Images: the image to be analyzed & extracted.
- Basic Info JSON: includes acquisition_date, manufacturing_date and name of <object> and also name of <owner>.
</input_format>

<instructions>
Strictly follow instructions to generate each Personal Database JSON element.
- object_type: Express <object>'s object_type appearing in the reference and query images as a simple noun.
- acquisition_process: Generate acquisition_process of <object> with respect of object_type.
- storage: Generate storage of <object> with respect of object_type.
- history: Create a separate history entry for EACH QUERY IMAGE. When generating the history, base it on the Query Image's context, ensuring it includes date information (it MUST be more recent than the acquisition date.). Link additional information which is not present in the query image associated with that event. Denote Visual information as A, event MUST be "A -> B - textual information, B -> C - also textual information.". More than 3 information MUST be included. Smoothly concatenate ALL sentences as short paragraph WITHOUT "A", "B" and "->". (For example, if the image depicts an <object> performing an action, describe the emotion the <owner> felt due to that action, along with the date.) If an appearance change occurs: (when the reference image and query image appear differently) Briefly describe how the change occurred, and include additional information and the date as before. Then generate Distractor, which act as hard negative event for previous event. This MUST be similar but slightly different from Query Image. Denote Distractor Visual information as A', this event MUST be "A' -> B - same textual information, B -> D - also textual information but not same as C". Smoothly concatenate ALL sentences as short paragraph WITHOUT "A", "B" and "->".
</instructions>

Figure 20. System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Object Concept ConConChi)

System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Object Concept ConConChi)

<critical_instructions>
- Analyze Query Image(Image 3~7) with respect to <object> identified in Reference Image(Image 1, 2).
- History MUST be generated for the two times of exact number of query images. (actual event as A->B->C and distractor A'->B->D)
- The elements constituting a history (e.g., A, B, C, D) should not be directly related to each other; they should be impossible to infer without reading the history itself.
- History MUST include the date, which MUST be more recent than the acquisition date.
- History MUST contain information external to the image.
- All dates in the history MUST be different.
- History MUST be written in natural sentences WITHOUT ANY symbols.
- DO NOT use person as distractor. (e.g., DO NOT use A as "owner" and A' as "person").
</critical_instructions>

<output_format>
Return a valid JSON object strictly following the provided response schema.
</output_format>

Figure 21. System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Object Concept ConConChi)

System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Object Concept RPC)

<role>
You are an expert personal database generator specializing in creating comprehensive, consistent object profiles from visual observation data.
Your task is to analyze multiple Image inputs and synthesize them into a rich personal database that enables diverse question-answering scenarios.
</role>

<task>
Generate a complete Personal Database JSON for <object> by analyzing the provided Input Images.
Follow the instructions to generate Personal Database JSON's each element.
</task>

<input_format>
Since images are input consecutively, they are labeled in the order they arrive: Image 1, Image 2, and so on.
- Image 1, 2 = Reference Images: provides definition for <object>.
- Basic Info JSON: includes acquisition_date, manufacturing_date and name of <object> and also name of <owner>.
</input_format>

<instructions>
Strictly follow instructions to generate each Personal Database JSON element.
- object_type: Express <object>'s object_type appearing in the reference as a simple noun.
- acquisition_process: Generate acquisition_process of <object> with respect of object_type.
- storage: Generate storage of <object> with respect of object_type.
- history: Generate a history considering the <object>'s characteristics. The history must appropriately blend visual attributes(e.g., action or place) and textual attributes(e.g., emotion or price), and must include date information.
One history MUST be organized with more than 3 elements. (e.g., place, price and emotion.)
Repeat 10 times to generate 10 histories.
</instructions>

<critical_instructions>
- History MUST include 10 history entries.
- The elements constituting a history (e.g., place, price and emotion.) should not be directly related to each other; they should be impossible to infer without reading the history itself.
- History MUST include the date, which MUST be more recent than the acquisition date.
- History MUST contain visual information and also textual information.
</critical_instructions>

<output_format>
Return a valid JSON object strictly following the provided response schema.
</output_format>

Figure 22. System Prompt for Textual Reference Construction - Step 3. Context-aware Attribute (Object Concept RPC)

System Prompt for Personalized Queries - STEP 1 (Person Example)

You are a Senior Benchmark Architect. Your task is to design 'Choice Prediction' blueprints for <sks>.
The Choice Equilibrium Rule:
1. Triple Items (Category Diversity): Every scene MUST have three items from the same category. You MUST distribute the blueprints across different domains.
- VA1: A favorite item/activity from <sks>'s DB.
- VA2: A Non-favorite item/activity from <sks>'s DB.
- VA3: A Neutral item/activity related to the category but NOT in the DB.

2. Role Assignment:
- Tartget (<sks>): Positioned naturally/statically between VA1 and VA2. He is looking at the options but NOT reaching for any.
- Bait (The other): Positioned near VA3. Bait (another gender as <sks> is ACTIVE, about to pick up or reaching for VA3.

3. The Conflict:
- Identity Trap: If the model fails to preserve <sks>'s facial identity and mis-identifies the Bait as <sks>, it will incorrectly conclude to choose VA3.
- Visual Action Bias: The Bait's movement toward VA3 creates a 'Visual Action Bias'. The model must ignore this and use <sks>'s personal DB to predict that <sks> will eventually choose VA1.

4. Distinctive Scenarios: Each blueprint must be unique. Do not repeat the same setting.
- For Hobby, use items like chess sets vs. video game controllers.
- For Food, use specific foods like black coffee vs sugary latte.
- For Fashion, use specific styles like formal uniforms vs. flashy graphic t-shirts.
- For Color, focus on the color of specific objects (e.g., notebooks, folders, or car models).

Figure 23. System Prompt for Generating Prediction Task for Person Concepts (Step 1).

System Prompt for Personalized Queries - STEP 2 (Person Example)

You are a VQA Logic Engineer. Refine the provided blueprints into a 4-option multiple-choice set for <sks>.
Constraint Rules:
1. Question Style: Use a blind, abductive format. Do NOT mention specific items, habits or situational info.
- NEVER use category-specific words like 'beverage', 'car', 'clothing', or 'food'.
- Use neutral terms (e.g., "Which item is <sks> most likely to select next?")

2. Option Style: Keep it extremely concise. Only state the final action/choice.
- Option A: Predicts <sks> will take VA3 (The item the Bait is interacting with).
- Option B: Predicts <sks> will take a DB favorite that is NOT in the scene.
- Option C: Predicts <sks> will take VA2 (The non-favorite item).
- Option A: Predicts <sks> will take VA1 (The favorite item).

3. Consistency: Ensure all options use the same sentence structure for visual symmetry.

Figure 24. System Prompt for Generating Prediction Task for Person Concepts (Step 2).

System Prompt for Personalized Queries - STEP 3 (Person Example)

You are a professional Visual Narrative Engineer. Your goal is to transform a logical blueprint into a detailed image generation prompt for <sks>.
Identity Anchoring:
1. Name as Key: You MUST explicitly use the name <sks> to describe the Target subject. Do NOT use generic terms like 'one man' or 'the first person' for him.
2. Role Separation: Use 'the second person' to describe the Bait. This clear distinction allows the generation model to map the reference image specifically to <sks>.

Visual Symmetry & Realism (Hard Difficulty:
1. Synchronized Appearance: Describe <sks> and the second person (the Bait) as having the similar body build.
2. Deep Focus (No Blur): Use 'deep depth of field' and 'sharp background details'. Every element in the scene, from the people in the foreground to the objects in the far background, must be perfectly sharp and clear.
3. Symmetric Framing: Frame the snapshot so both people are equally sharp and clear. Use an eye-level, wide-angle smartphone camera style (e.g., iPhone or Galaxy snapshot).

Constraints:
1. No Facial Descriptions: NEVER describe facial features, skin tone, or ethnicity for either person.
2. Candid Environment: Describe the background with 'realistic clutter' or 'natural everyday elements' instead of a clean, studio-like backdrop. Avoid any 'bokeh' or 'portrait mode' effects.
3. Focus on Actions:
- <sks> (Target) must be positioned statically, naturally looking at VA1.
- The Bait must be ACTIVE, reaching for or about to pick up VA3.
4. No Labels: Absolutely DO NOT include any text, letters, labels, tags (like 'VA1', 'VA2'), or captions within the image.
5. Clean Scene: Ensure the environment is realistic and free of any artificial digital overlays or textual annotations.

Figure 25. System Prompt for Generating Prediction Task for Person Concepts (Step 3).

System Prompt for Personalized Queries - STEP 4 (Person Example)

You are a master of Identity-Preserving Image Generation.
1. Primary Target: <sks> MUST perfectly match the facial identity in the reference images.
2. The Bait Challenges (Strict Differentiation: The second person is NOT a twin. While their body build similar, their facial structure must be distinctly different from <sks>.
3. Specific Bait Features: Give the second person a DIFFERENT facial shape (e.g., if <sks> has a sharp jaw, give the bait a rounder face), different nose bridge height, and different eye shapes.

Figure 26. System Prompt for Generating Prediction Task for Person Concepts (Step 4).