MMPB: It's Time for Multi-Modal Personalization

Jaeik Kim[1] **Woojin Kim**[2] **Woohyeon Park**[2] **Jaeyoung Do**[† 1,2,]
AIDAS Laboratory, [1]IPAI & [2]ECE, Seoul National University
† indicates corresponding author
{jake630, wjk9904, woohyeon, jaeyoung.do}@snu.ac.kr

Abstract

Visual personalization is essential in user-facing AI systems such as smart homes and healthcare, where aligning model behavior with user-centric concepts is critical. However, recent large Vision-Language Models (VLMs), despite their broad applicability, remain underexplored in their ability to adapt to individual users. In this paper, we introduce **MMPB**, the first extensive benchmark for evaluating VLMs on personalization. MMPB comprises 10k image-query pairs and includes 111 personalizable concepts across four categories: humans, animals, objects, and characters, with the human category enriched with preference-grounded queries. We structure personalization into three main task types, each highlighting a different key property of VLMs. Using 23 widely used VLMs including both open- and closed-source models, we evaluate personalization performance via a three-stage protocol: concept injection, multi-turn dialogue, and personalized querying. Our findings indicate that most VLMs (including some closed-source models) struggle with personalization, particularly in maintaining consistency over dialogue, handling user preferences, and adapting to visual cues. Our analysis reveals that the challenges in VLM personalization (such as refusal behaviors and long-context forgetting) highlight substantial room for improvement. By identifying these limitations and offering a scalable benchmark, MMPB offers valuable insights and a solid foundation for future research toward truly personalized multi-modal AI.
Project Page: aidaslab.github.io/MMPB

1 Introduction

Now we need VLM personalization Our daily lives are filled with recurring visual concepts that are closely tied to us. Every day, you might spot your dog waiting by the door or notice your favorite coffee mug on the table. However, they are rarely understood in a user-specific context by recent large Vision-Language Models (VLMs). Although VLMs have been widely adopted as assistants that utilize broad knowledge of the world [92, 110, 99, 53], they still operate largely under the *one-size-fits-all* paradigm [1, 79], treating all users as interchangeable and responding to visual inputs without adapting to individual identities, preferences, or histories. As VLMs are integrated into multi-modal systems like smart home devices and robots [25, 43, 106], they increasingly serve as the core interface for grounding user instructions in the visual world. To work effectively in everyday scenarios, they must recognize and align with each user's visually repetitive environment and preferences. Such personalization is essential in applications where consistent interpretation of personal context is critical [65, 2, 3]. For example, instead of responding generically to *"feed the gray tabby cat,"* the model should understand personalized commands like *"feed Mochi."* Similarly, it should interpret a photo as *"your favorite travel destination"* rather than *"a snow-covered mountain."*

The evaluation gap in VLM personalization Despite the growing importance of AI personalization in many real-world scenarios [23, 10, 24], current benchmarks fall short in evaluating the personalization capabilities of VLMs. Existing VQA datasets focus on general-purpose knowledge,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Figure 1: Examples of personalized queries across task types and representative failure cases of recent VLMs. indicates GPT-4o, while indicates LLaVA family models such as LLaVA-NeXT.

including commonsense [96, 58, 14], scientific [35, 60, 59], and medical reasoning [8, 54, 46]. Despite some early progress [65, 2], personalization in VLMs remains constrained, with limited coverage and diversity of personalizable concepts (Table 1). In addition, the absence of a unified evaluation framework (both in task types and metrics) and overlooking the cold-start nature, where models must personalize with minimal prior context, make it difficult to systematically assess personalization. Finally, preference-grounded VQA, which requires reasoning over user-specific likes and dislikes, remains especially underexplored, posing unique challenges beyond standard VLM tasks.

To bridge this gap, we introduce **MMPB** (Multi-modal Personalization Benchmark), the first benchmark for evaluating VLM personalization. MMPB evaluates concept recognition and preference-grounded reasoning using images associated with preference-related keywords (Figure 1). Our dataset is built via a human–model collaborative process guided by annotation protocols, which specify task types and structured query processing. To ensure high quality, we remove any queries solvable using only concept information or only query images, thus focusing evaluation on genuine cross-modal personalized reasoning. Consequently, MMPB comprises 111 concepts spanning four categories: humans, animals, objects, and characters, paired with five reference images and four level textual descriptions. For the human preferences, we curate from 30 diverse subdomains of personal preferences. Queries are categorized into three task types, enabling detailed analysis of personalization failures in VLMs. Finally, we include multi-turn dialogues to evaluate the model's ability to retain personalized concepts over extended interactions.

We evaluate 23 widely used VLMs on MMPB, including closed-source models such as GPT-4o [1] and Claude-Sonnet [7]. Despite their strong performance on general-purpose VQA tasks, these VLMs exhibit significant limitations and challenges in personalization. Our key take-home messages include: **(i)** Even top-ranked VLMs on general benchmarks struggle with preference-grounded tasks, underscoring the need for more robust forms of inference, including abductive reasoning (§5.1). **(ii)** VLMs struggle with visual personalization. For example, comparable performance from one image and three text keywords highlights a persistent inability to leverage fine-grained visual cues despite the central role of image-based personalization (§5.4). **(iii)** Closed-source models tend to avoid personalization involving human-centric concepts, stemming from safety alignment constraints. This implies that existing safety constraints hinder personalization, highlighting the need to reconsider the balance between safety and personalization (§5.2). We further conduct a systematic analysis of VLM failure patterns, focusing on personalization bias and failures in discrimination (§5.3), with errors amplified in long-context scenarios where mid-sequence concepts are frequently overlooked (§5.5). Our main contributions are as follows:

- We introduce MMPB, the first comprehensive benchmark for evaluating VLM personalization in both recognition and preference-grounded VQA. It includes 111 personalizable concepts and 10,000+ questions with 15 task types that reflect real-world scenarios.
- By providing five reference images and four levels of textual descriptions, MMPB establishes a foundation not only for evaluation but also for future advanced techniques that leverage them (*e.g.*, post hoc training) thereby enabling fine-grained VLM personalization.
- By rigorously evaluating widely adopted VLMs, we identify key personalization failures such as limited preference-grounded reasoning, modality gaps, personalization bias, and safety-induced evasiveness, which establish MMPB as a key diagnostic tool for improving personalized VLM behavior.

Table 1: Comparison of existing multimodal personalization benchmark datasets, indicating whether each dataset supports systemic evaluation (*SysEval*), human preference (*Pref*), multiple levels of granularity (*Gran*), and multi-turn interactions (*Multi-turn*). *#Concept* and *#Samples* denote the number of distinct concept categories and total samples, respectively.

Dataset	SysEval	Pref	Gran	Multi-turn	#Concept	#Samples
MyVLM [2]	✗	✗	✗	✗	29	0.3K
Yo'LLaVA [65]	✗	✗	✗	✗	40	0.6K
MC-LLaVA [3]	✗	✗	✗	✗	95	2.0K
MMPB (Ours)	✓	✓	✓	✓	111	10.0K

2 Problem Definition

2.1 Core Properties of Personalized VLMs

To systematically evaluate personalized VLMs, we formalize four core properties we call *personalization criteria*. Personalizable concepts are user-centric entities (*e.g.*, *"me," "my dog"*) whose attributes, such as appearance or preferences, can be encoded in either structured or unstructured form. Formally, let $\mathcal{P} = \{p_1, \ldots, p_K\}$ denote all personalizable concepts. A VLM successfully augmented with any $p_k \in \mathcal{P}$ should satisfy the following criteria, with examples in Figure 1:

1. **Awareness**: Can the model accurately identify p_k in a given image?
2. **Appropriateness**: Does the model activate p_k only when it is contextually appropriate?
3. **Coherency**: Does the model contradict p_k in its responses?
4. **Persistency**: Can the model consistently respond based on p_k across long-context or multi-turn interactions?

2.2 Formalizing VLM Personalization

A standard VLM $\mathcal{M}$ is typically defined as a function $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$, where $\mathcal{X}$ is the pixel space, $\mathcal{T}$ the text space (*e.g.*, user queries), and $\mathcal{Y}$ the output space (*e.g.*, generated text). To personalize $\mathcal{M}$, we extend it to a function $f_p : \mathcal{X} \times \mathcal{T} \times \mathcal{P} \to \mathcal{Y}$, incorporating a concept $p_k \in \mathcal{P}$. A key objective of personalization is to tailor outputs to user-specific contexts while **preserving the general-purpose knowledge** acquired during pretraining. One approach is to concatenate p_k with the input modalities $x \in \mathcal{X}$ and $t \in \mathcal{T}$, without modifying the model's weights. Given a projection function h that embeds p_k, and a decoder g, the personalized output is:

$$y = f_p(x, t, p_k) = g\big(E_I(x) \oplus E_T(t) \oplus h(p_k)\big),$$

where E_I and E_T denote the image and text encoders, with $\oplus$ as concatenation. MMPB evaluates VLMs by injecting p_k via either textual or visual modalities, using reference images or descriptions as in-context prompts [22, 32], using$h = E_I$ for visual and $h = E_T$ for textual concept injection.

3 MMPB: Multi-modal Personalization Benchmark

3.1 Overview

MMPB is the first benchmark for personalizing multi-modal assistant, featuring multiple-choice visual question answering (VQA) tasks focused on recognition and user preference. It comprises 10,017 image–query pairs across 111 concepts spanning four categories, covering three task types: *Awareness*, *Appropriateness*, and *Coherency*. To support future expansion, MMPB is built via a three-step human–model collaboration (Figure 2). MMPB evaluates VLMs across three stages: (1) **Concept injection**, (2) **Multi-turn conversations**, and (3) **Personalized querying**. In the first stage, a concept is introduced via reference images or textual descriptions. During the conversation phase, the model engages in general multi-turn dialogue to test concept retention. Finally, personalized queries assess whether the model can apply the concept to a visual input. Detailed dataset construction guidelines and statistics are provided in Appendix §C and §E.2.

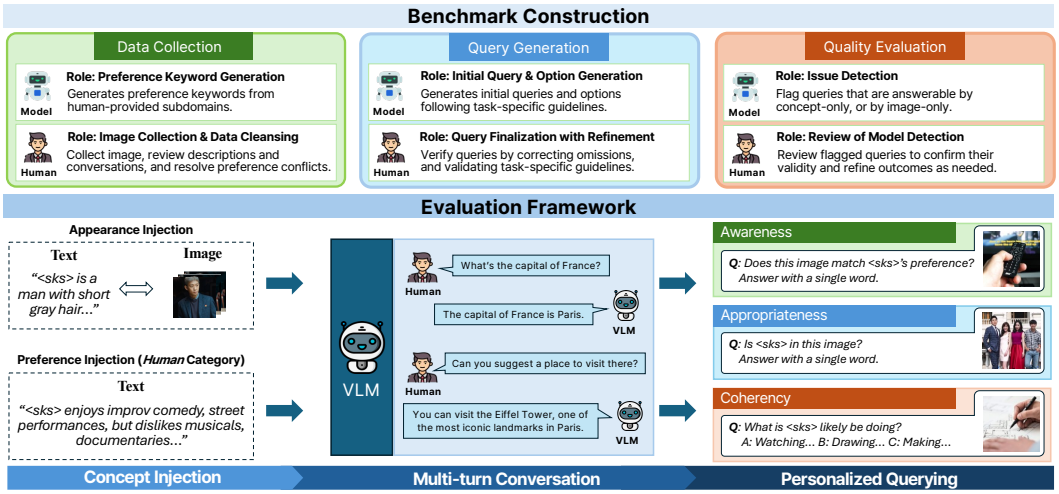


Figure 2: Overview of MMPB. (top) A three-step construction process ensuring high quality and scalability. (bottom) An evaluation protocol for assessing the VLM's *personalization criteria*.

3.2 Data Collection

Concepts MMPB provides five reference images per concept. Human annotators, following detailed guidelines (Appendix §C.1.1), sourced these images from open datasets (MC-LLaVA [3], MyVLM [2]) and CC-licensed platforms (Flickr, Pexels), and recorded original URLs for any externally retrieved images (*e.g.*, kmDB or Google Image Search). MMPB assumes static concept states without modeling temporal variation such as appearance changes or evolving user preferences. For consistency, most human concepts use MC-LLaVA [3] movie portraits, ensuring a uniform character style. Other concepts use multiple images of the same instance (*e.g.*, a pink-hat Asian doll).

To facilitate concept injection through the textual modality, we generate appearance descriptions for each concept via human–model collaboration, using five reference images per concept. We employ Ovis2-34B [61] for its strong multi-image captioning capabilities [20] and practical scalability, as it can be run efficiently on a single 80GB GPU. The model-generated descriptions are then verified and refined by human annotators based on guidelines (Appendix §C.1.2). Each concept is described across four levels of granularity: (i) *Simple* (three keywords), (ii) *Moderate* (one sentence), (iii) *Detailed* (one paragraph), and (iv) *Extended* (structured multi-paragraphs). To avoid pretraining leakage and ensure fair evaluation, all concept names are replaced with `<sks>`, following prior works [65, 75].

For human preferences, we define five domains, each comprising six subdomains (30 in total; see Figure 12 in Appendix §E.2). Using GPT-4o [1], we extract 30–40 representative keywords per domain. Since providing multi-image preferences for every domain incurs substantial input overhead, we instead express them textually. Each human is assigned *likes* and *dislikes* via a template: "`<sks>` likes `{keywords}` but dislikes `{keywords}`." Conflicting preferences (*e.g.*, liking "gaming keyboards" but disliking "technical gadgets") are resolved by human annotators.

Multi-turn conversations To emulate realistic back-and-forth dialogue, we sample general-topic conversations from LMSYS-Chat-1M [107] following previous studies [105, 51, 5] and strip any concept-related lines to avoid conflicts (Appendix §C.1.4).

Images for personalized query For personalized query images for recognition, we follow the image collection guidelines, ensuring that injection and evaluation images are strictly disjoint. For preference-grounded VQA, human annotators collect images from approved sources associated with 30 predefined subdomains, ensuring relevance and diversity within each domain (Appendix §C.1.5).

3.3 Query Design

Overall process The query design process consists of two main stages (Figure 2): (1) a generation stage, where a model drafts candidate queries and human annotators refine them to produce initial

query sets; and (2) a quality evaluation stage, where models first flag suspected low-quality queries, and human annotators validate the flags and review all queries. We use Ovis2-34B [61] in the generation stage, and Ovis2-34B, InternVL2.5-78B [19], and Qwen2.5-VL-72B [6] in the evaluation stage. Although MMPB is designed as a multiple-choice VQA, all queries are easily convertible into open-ended formats (Appendix Table 10). As in §3.2, all concept names are replaced with <sks>.

Categories For query categorization, we follow the *Personalization Criteria* (§2) and define three types of queries: (1) *Awareness*, (2) *Appropriateness*, and (3) *Coherency*. Each query type is applied to recognition tasks across all concept categories, and to preference-grounded tasks for human concepts, resulting in 15 evaluation tasks. The *Awareness* type tests whether the model can detect the presence of a personalized concept in an image. Positive images are used, where the correct answer to queries like "Is <sks> in the image?" should be "yes." We further distinguish between single-entity and multi-entity cases, based on whether other entities co-appear with the concept. The *Appropriateness* type evaluates whether the model can correctly suppress references to a personalized concept when it is not contextually appropriate. Negative images are used, and the correct answer should be "no." For animal concepts, negative samples are further categorized into same-species and different-species examples to analyze the impact of hard negatives. This is feasible because species boundaries are visually and semantically well-defined. The *Coherency* type assesses whether the model can produce coherent and context-appropriate responses about the concept (*e.g.*, "What clothes is <sks> wearing?"). We also evaluate *Persistency* by introducing multi-turn conversations.

Quality control We aim to filter out options that trigger positional or affirmative biases or are solvable by text alone [58, 72, 104]. Throughout query generation and evaluation, we follow task-specific guidelines to ensure high quality[1]. For the *Awareness* and *Appropriateness* types, we instruct both the model and human annotators to generate queries targeting concept presence in each image. To mitigate potential yes/no bias [104], we include both affirmative and negative formulations (*e.g.*, "Is <sks> not present in this photo?"). In the *Coherency* type, we adopt a 4-option MCQ format, addressing two major confounding factors: (i) **concept-only solvability** and (ii) **image-only solvability**. As shown in Figure 3, at least one distractor (*e.g.*, "Yoga") is aligned with the <sks> but not with the image, preventing concept-only solvability. Other distractors (*e.g.*, "Watching football") are visually plausible but incorrect without considering the concept, addressing image-only solvability. This design encourages joint reasoning over both the image and the concept, with choices shuffled to eliminate positional bias [72]. Blind tests with text-only experiments are provided in Appendix Table 9, further supporting our rigorous query design.

Figure 3: Example of quality control for a *Coherency*-type query with concept-only and image-only distractors.

4 Benchmarking Results

4.1 Experimental Setup

Configuration To simulate real-world personalization challenges, we design our experiments centered on two key constraints: *cold-start* and *multi-turn* adaptation. For the cold-start problem [105, 95] (*i.e.*, initiating personalization with minimal prior information), we utilize *moderate* description for text-based injection, and two reference images for image-based injection with a single concept injection. We also evaluate both in 0-turn (*i.e.*, without multi-turn conversation) and 10-turn conversation settings. We ensure that the 10-turn conversation is applied consistently across all models. We follow prior multiple-choice VQA tasks [73, 4, 111] and use overall accuracy as our main evaluation metric.

Models We select models to cover major VLM families following previous benchmarks [27, 13], based on public availability, and widespread usage. We evaluate 23 models across open- and closed-

[1]Each query is reviewed by at least three human annotators, and only queries with majority agreement (*i.e.*, at least 2 out of 3 annotators) are accepted. Please see the guidelines in Appendix §C.2.2 and §C.3.2.

[2]Hugging Face Open VLM Leaderboard, accessed on May 15, 2025.

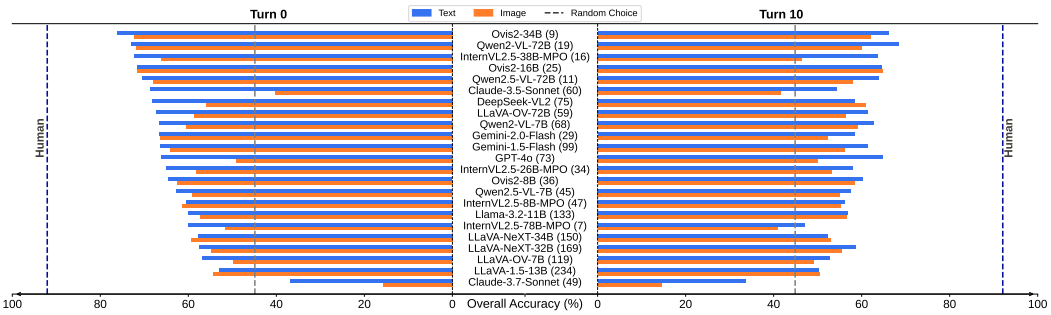


Figure 4: Evaluation results of 23 VLMs on MMPB under 0-turn and 10-turn settings. Model names are followed by their average ranks across eight general-purpose multi-modal benchmarks[2].

Table 2: Task-wise performance of open vs. closed models, averaged over image/text-based injections.

		Preference	**Recognition**			
Model	**Turn**	Human	Human	Animal	Object	Character
Open	0	52.8 $\pm$ 5.9	74.3 $\pm$ 11.1	69.2 $\pm$ 13.7	75.7 $\pm$ 18.1	67.3 $\pm$ 11.5
Closed	0	37.0 $\pm$ 18.0	68.9 $\pm$ 21.0	73.4 $\pm$ 9.0	79.1 $\pm$ 16.0	65.3 $\pm$ 17.4
Open	10	46.0 $\pm$ 5.1	69.0 $\pm$ 12.4	69.8 $\pm$ 8.3	72.0 $\pm$ 14.0	61.6 $\pm$ 10.9
Closed	10	34.7 $\pm$ 15.7	61.7 $\pm$ 16.0	66.2 $\pm$ 10.6	67.3 $\pm$ 19.0	56.9 $\pm$ 15.9

source families: open-source includes InternVL2.5 (8–78B) [19, 18, 17], Ovis2 (8–34B)[3] [61], LLaVA (13–72B) [55, 56, 50], Qwen-VL (7–72B) [84, 6], DeepSeek-VL2 [90], and Llama-3.2-11B [28]; closed-source includes Claude-3.5/3.7-Sonnet [7], Gemini-1.5/2.0-Flash [79], and GPT-4o [1].

4.2 Overall Results

Figure 4 shows overall personalization performance on MMPB across 23 widely used VLMs, with the human evaluator achieving an average of 92.1%, thereby establishing the upper bound (Appendix §D). Across all experiments, text-based concept injection consistently outperforms image-based injection, achieving average accuracies of 63.8% vs. 57.8% in the 0-turn, and 57.9% vs. 52.6% in the 10-turn setting. The difficulty becomes more pronounced under extended interactions. Most models experience substantial performance drop when moving from the 0-turn to the 10-turn setting, indicating challenges in maintaining personalized responses over time. Notably, averaged across all experimental settings, closed models underperform open ones (51.4% vs. 59.9%) despite strong results on general VQA benchmarks [12, 97]. Table 2 presents task- and concept-level performance comparisons between model types. Overall, both model types exhibit pronounced multi-turn degradation.

To further examine whether MMPB genuinely measures personalization rather than merely reflecting general VLM strength, we conducted additional experiments on LLaVA-1.5-13B with two personalization strategies: (i) soft prompt tuning following the Yo'llava [65] approach, and (ii) personalized LoRA fine-tuning [102]. As detailed in Appendix §F.1, both methods lead to clear improvements over the baseline, with LoRA providing the strongest gains. Overall, these results validate that MMPB is sensitive to personalization techniques and effectively captures personalization performance beyond general VLM ability.

5 Challenges in VLM Personalization

The results presented in §4.2 demonstrate that, overall, recent VLMs remain highly vulnerable to personalization tasks. To systematically investigate which challenges hinder effective handling of user-specific queries, we organize our analysis around the following key questions:

- Which specific tasks pose barriers to effective VLM personalization? (§5.1 and §5.3)
- What makes closed-source models struggle with personalization? (§5.2)

[3]To ensure the model does not benefit from self-generated content, we confirm that Ovis2-34B underperforms on its own generated descriptions compared to human-written *Moderate* descriptions (Appendix §F.3).

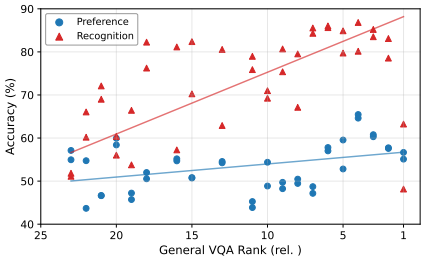


Figure 5: Performance gap between preference-grounded and recognition VQA tasks in VLMs.

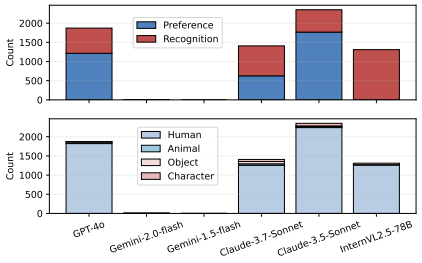


Figure 6: Refusal counts of models across task and concepts under image-based injection.

Table 3: Examples of evasive responses in human category tasks.

Model	Example
GPT-4o	Unknown
Claude-3.7-Sonnet	Based on the image, I can only describe what's visible...
Claude-3.5-Sonnet	I cannot and should not identify or name specific features of the person...
InternVL2.5-78B	I can't identify or recognize people in images.

- How do injection modality and granularity impact VLMs? (§5.4)
- Can VLMs follow a specific concept in multi-concept or long-context settings? (§5.5)

5.1 Preference-Grounded Reasoning

Personalized VLMs should be able to infer user preferences from images and incorporate them into its responses. We split the evaluation into recognition and preference tasks and order their accuracies by average relative rank within general VQA, which comprises eight multi-modal reasoning benchmarks, including MathVista [60], MMMU [97], and MMBench [58]. As shown in Figure 5, models that perform well on general tasks consistently struggle with preference-grounded queries. This disparity highlights a key limitation of current VLMs: their inability to generalize from general-purpose reasoning to user-centric inference. In contrast, recognition tasks show a stronger correlation with rankings, underscoring the unique difficulty posed by preference-based personalization.

Difference between recognition and preference-grounded VQA VLMs typically solve VQA tasks through a three-stage process: image perception, knowledge grounding (either pretrained or retrieved via RAG [48]), and answer generation. In MMPB, recognition tasks follow this pipeline with minimal uncertainty, using deductive reasoning on injected concepts via in-context learning. For example, given "Is <sks> in this picture?", the model matches visual features and answers "Yes." By contrast, preference-grounded VQA demands abductive reasoning, integrating user intentions and tastes with scene understanding, *e.g.*, in "Which activity here do you think <sks> would enjoy most?" Current VLMs primarily focus on deductive tasks [55], which may be insufficient for ensuring user satisfaction in real-world scenarios. Future personalization frameworks may consider training VLMs to perform abductive reasoning, *e.g.*, via instruction tuning on preference-based tasks.

5.2 Resistance to Personalization

As shown in Table 2, closed-source models consistently underperform on human-related queries. Notably, InternVL2.5-78B, which ranked first among 23 models on a general VQA benchmark, also exhibits weak personalization performance, both in the text- and image- based setting (Figure 4). We find that this is the only model in the InternVL family that is publicly accessible via a web-based chatbot, which may introduce additional safety filters that affect its behavior. To investigate their failures in detail, we examine their responses and find a consistent pattern of evasive behaviors.

To quantify this behavior, we count answers matching well-known refusal patterns in LLMs [11, 74, 85, 47, 91], such as *"I'm sorry," "I shouldn't,"* and *"Unknown."*. As shown in Figure 6, all closed-source models, except Gemini, exhibit evasive responses in recognition and preference-grounded tasks. InternVL2.5-78B does so only in recognition. Notably, most evasive cases fall under the

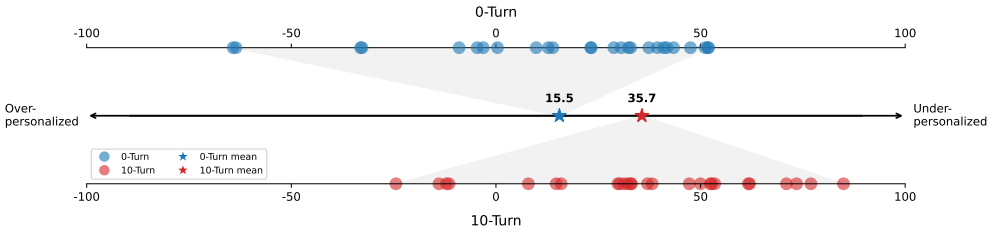


Figure 7: Personalization bias across models, measured as *Appropriateness–Awareness*. Positive values indicate under-personalization (missed valid personalization), and negative values indicate over-personalization (affirming inappropriate inputs).

human category: out of 7,501 queries, up to 2,237 trigger evasive answers, suggesting that this behavior significantly contributes to suboptimal performance. Table 3 presents representative free-form responses, aligning with refusal patterns observed when LLMs are prompted with harmful queries [70]. Interestingly, with the upgrade from Claude-3.5-Sonnet to 3.7, the model increasingly produces descriptive outputs rather than issuing explicit refusals. Since these responses do not clearly match standard refusal patterns, we exclude them from the count. Further analysis (Appendix Box 25 and 26) shows that closed-source models treat identity-related queries as privacy-sensitive, resulting in conservative behavior aimed at avoiding unsafe outputs.

Personalization vs. Safety: Can they coexist? This tradeoff between personalization and safety raises important questions for the AI community. Human perception is a critical component of user-level personalization, particularly in tasks that involve grounding visual inputs to individual users. While safety constraints are essential to prevent misuse, an overly cautious stance can hinder meaningful personalization, where distinguishing between users is a core requirement. This tension calls for deeper discussion around the boundary between safety and utility, and highlights the need for methods that enable secure yet effective handling of identity information in VLMs.

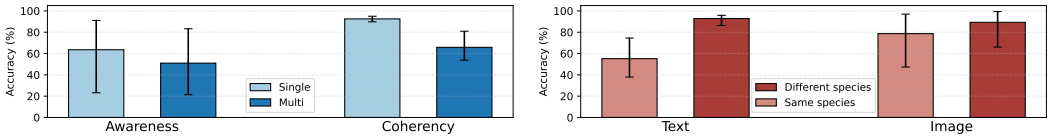
5.3 Personalization Bias and Fine-Grained Discrimination

Under-personalized bias Figure 7 visualizes the personalization bias of models, measured as the difference between *Appropriateness* and *Awareness*. In both the 0-turn and 10-turn settings, models consistently lean toward rejecting personalization, performing better on *Appropriateness* than on *Awareness*, with 72 out of 92 cases falling in the under-personalization region. This indicates a systematic tendency to reject personalized concepts rather than affirm them. The bias is more pronounced in the 10-turn setting, suggesting that extended dialogue reinforces this behavior.

Impact of multiple entities and hard-negatives Figure 8a compares performance on single- and multi-entity query images. The models exhibit comparable *Awareness* scores, suggesting that concept detection remains stable. However, *Coherency* drops substantially in response to multi-entity inputs, indicating difficulty in maintaining reasoning consistency as input complexity increases. To evaluate fine-grained *Appropriateness*, we assess whether the models can distinguish target concepts from visually similar distractors in the animal category, using same-species (*e.g.*, Beagle–Beagle) and different-species (*e.g.*, Beagle–Shiba) instances as negative samples. As shown in Figure 8b, text-based injection results in a substantial performance drop for hard negative samples, highlighting the challenges in making fine-grained distinctions. In contrast, image-based injection delivers robust performance, suggesting that visual cues offer stronger support for fine-grained discrimination.

5.4 Visual Personalization

We conduct an in-depth analysis of how injection modality and content granularity affect model performance by comparing four levels of text granularity and three levels of image-based injection. We evaluate six representative models, including Qwen2/2.5-VL-7B, Ovis2-8/16B, LLaVA-OV-7B, and InternVL2.5-8B-MPO, all of which support multi-image input. Results are presented in Table 4.



(a) (b)

Figure 8: Effects of image complexity: (a) Multiple entities on *Awareness* and *Coherency* tasks. (b) Hard-negative samples on *Appropriateness* task. The error bars denote the first and third quartiles.

Table 4: Recognition accuracies under different granularity levels of concept injection.

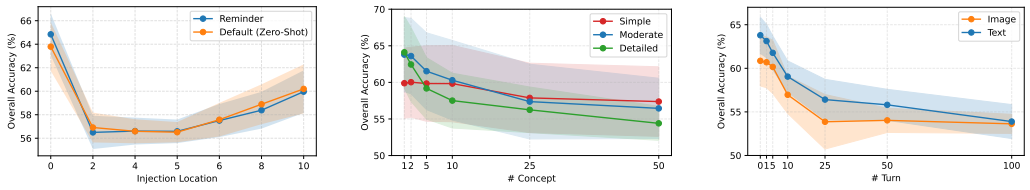
	Text-based Injection				Image-based Injection		
Turn	Simple	Moderate	Detailed	Extended	1 image	2 images	5 images
0	68.8 ± 5.2	77.1 ± 6.7	**79.3** ± 6.7	78.3 ± 6.8	71.5 ± 8.3	72.2 ± 9.9	**73.3** ± 10.5
10	67.4 ± 4.7	72.0 ± 7.2	**73.5** ± 7.9	72.1 ± 8.2	67.9 ± 6.2	69.2 ± 8.4	**70.1** ± 9.4

A few words are worth a thousand pixels We highlight distinct trends in recognition performance across modalities. Surprisingly, *Simple* textual injections, consisting of only three keywords, achieve comparable accuracy to the *1-image* condition in 10-turn, suggesting that the use of a single image doesn't provide much more benefit than minimal text alone. Moreover, even with five reference images, the performance does not exceed that of *Moderate* text injection, highlighting the limited utility of visual examples in recent VLMs. Given the central role of visual personalization in future VLMs, further research should explore strategies that effectively harness these visual cues for personalization.

5.5 Lost Concepts in a Haystack

Prior work shows that VLMs struggle with the needle-in-a-haystack problem, failing to retrieve key information from long contexts [83, 16, 34]. Using six models in §5.4, we evaluate concept retrieval in diverse long-context settings. Across all settings, personalization failures stem less from input design, and more from fundamental limits in long-range memory and relevance filtering.

We begin with 10-turn dialogues in which concepts are injected at different positions (Figure 9a). We also apply a *Reminder* strategy [105, 88], which cues concept recall in the final turn (Appendix §F.6). As a result, regardless of the prompting methods, models often forget concepts located near the midpoint, reflecting the "lost-in-the-middle" effect [89, 57, 37, 103] and indicating a positional bias in attention. We also evaluate multi-concept inputs with up to 50 entities and varying description granularity in the 0-turn setting, placing the target concept near the middle (*e.g.*, 6th of 10). In Figure 9b, the accuracy consistently decreases, with sharper drops observed in more detailed descriptions. These findings similarly indicate that VLMs struggle to isolate relevant information embedded mid-context, particularly when multiple entities are present. In extended dialogues of up to 100 turns (Figure 9c), performance for both text- and image-based inputs deteriorates sharply after 5 turns, reaffirming that VLMs' context-tracking ability degrades significantly with increased input length [105].



(a) (b) (c)

Figure 9: VLMs under long-context scenarios. (a) Concept most lost midway conversation. (b) Multiple concept injections on text-based injection. (c) Extended multi-turn dialogue up to 100 turns.

6 Related Works

Large Vision-Language Models Large language models (LLMs) [67, 63, 40, 9, 94, 41, 80, 68, 81], built on extensive world knowledge, have grown rapidly across diverse domains. Leveraging this progress, large Vision–Language Models (VLMs) [21, 52, 98, 26, 93, 101, 110, 79, 61] have emerged to connect visual and linguistic modalities. By seamlessly bridging vision and language, VLMs have revolutionized AI, driving breakthroughs in tasks such as visual question answering [30, 62, 78, 87, 97], image captioning [45, 42, 49, 82], and optical character recognition (OCR) [64, 60]. Despite their potential as interfaces connecting AI systems with individual users in real-world scenarios, current benchmarks lack the means to evaluate the personalization capabilities of VLMs. To address this gap, we propose MMPB, the first benchmark designed for VLM personalization, establishing a foundation for future research in personalized vision–language modeling. Unlike conventional VQA or captioning datasets, our benchmark highlights the user-dependent dimension of multi-modal reasoning, which has been largely underexplored in prior work.

Personalization in AI systems. As AI systems become more widely adopted across various domains such as AI assistants (e.g., psychological counseling [109], smartphone [86], housekeeping [31], medical assistance [100], and GUIs [36]), personalization has become crucial for enhancing usability and user satisfaction. In the context of large language models (LLMs), various techniques such as prompting [77, 76, 112], fine-tuning [66, 102, 108], and alignment [44, 15, 39] have been explored. In contrast, while large Vision-Language Models (VLMs) [1, 79] have demonstrated strong performance on multi-modal tasks [58, 14, 69, 38], their personalization capabilities remain underexplored. Existing approaches [65, 2] primarily focus on prior adaptation to personalizable concepts while overlooking cold-start scenarios, thereby limiting their applicability in real-world settings. Distinct from prior work, MMPB is the first to systematically evaluate personalization in VLMs through hierarchical concept injection, enabling a structured assessment of how models internalize and maintain user-specific concepts across varying levels of interaction complexity.

7 Discussion

In this paper, we introduce MMPB, the first benchmark for large Vision-Language Model (VLM) personalization, covering diverse personalizable concepts and task types, including preference-grounded VQA. Our human-in-the-loop dataset construction highlights its potential for future extensions, while our query filtering strategy ensures data consistency and quality. Extensive analysis shows that recent VLMs remain suboptimal across several personalization dimensions: limited preference-grounded reasoning, a tendency toward refusal, and ineffective use of visual cues. Although we focus on static appearance and preferences, future work should explore dynamically evolving user traits. Also, while we focus on VQA concerning concept presence and preference following, future benchmarks should cover application-driven tasks like personalized captioning and personalized robots. We hope MMPB encourages further progress in VLM personalization, accelerating the development of human-centric, real-world applications.

Limitations MMPB focuses on evaluating VLMs' ability to recognize and reason about fine-grained, static appearances and preferences. In practice, however, personal appearance and inclinations are rarely fixed: people change hairstyles, update their fashion, and see their own tastes evolve over time. As the first extensive benchmark for VLM personalization, MMPB establishes a solid foundation—but we encourage future work to introduce dynamic concept updates (*e.g.*, style changes, evolving preferences) so that models can be assessed on their ability to track and adapt to real-world, time-varying identities.

A second limitation concerns the scope of our evaluation framework. In this work, we defined three core properties—*Awareness*, *Appropriateness*, and *Coherency*—to capture key failure modes of personalized VLMs (alongside *Persistency* with multi-turn conversations). Consequently, MMPB's assessments are necessarily formalized around these properties, rather than downstream applications. Yet personalized VLMs hold great promise for tasks such as tailored image captioning or personalized actions in robots that bridge individuals and AI. We anticipate that future benchmarks will build on MMPB by measuring model performance in these real-world personalization scenarios.

Acknowledgements

This work was supported in part by National Research Foundation of Korea (NRF) grant (RS-2025-00560762, RS-2024-00414981), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant (RS-2025-02263754, RS-2025-25442338, IITP-2025-RS-2024-00397085, RS-2021-II211343). J. Do is with ASRI, Seoul National University.

References

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

[2] Y. Alaluf, E. Richardson, S. Tulyakov, K. Aberman, and D. Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 73–91. Springer, 2024.

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

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[5] G. Bai, J. Liu, X. Bu, Y. He, J. Liu, Z. Zhou, Z. Lin, W. Su, T. Ge, B. Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[6] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[7] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[8] A. Ben Abacha, S. A. Hasan, V. V. Datla, D. Demner-Fushman, and H. Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

[9] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[10] X. Cao, T. Zhou, Y. Ma, W. Ye, C. Cui, K. Tang, Z. Cao, K. Liang, Z. Wang, J. M. Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 21819–21830, 2024.

[11] Z. Cao, Y. Yang, and H. Zhao. Scans: Mitigating the exaggerated safety for llms via safety-conscious activation steering. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, volume 39, pages 23523–23531, 2025.

[12] J. Chen, T. Liang, S. Siu, Z. Wang, K. Wang, Y. Wang, Y. Ni, W. Zhu, Z. Jiang, B. Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024.

[13] J. Chen, T. Liang, S. Siu, Z. Wang, K. Wang, Y. Wang, Y. Ni, Z. Jiang, W. Zhu, B. Lyu, D. Jiang, X. He, Y. Liu, H. Hu, X. Yue, and W. Chen. MEGA-bench: Scaling multimodal evaluation to over 500 real-world tasks. 2025.

[14] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[15] R. Chen, X. Zhang, M. Luo, W. Chai, and Z. Liu. PAD: Personalized alignment at decoding-time. *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2025.

[16] Y. Chen, F. Xue, D. Li, Q. Hu, L. Zhu, X. Li, Y. Fang, H. Tang, S. Yang, Z. Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024.

[17] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[18] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[19] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024.

[20] K. Cheng, W. Song, J. Fan, Z. Ma, Q. Sun, F. Xu, C. Yan, N. Chen, J. Zhang, and J. Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. *arXiv preprint arXiv:2503.12329*, 2025.

[21] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[22] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[23] J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, and R. Krishna. Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*, 2024.

[24] Z. Duan, H. Cheng, D. Xu, X. Wu, X. Zhang, X. Ye, and Z. Xie. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7180–7189, 2024.

[25] Y. Fan, J. Nie, X. Sun, and X. Jiang. Exploring foundation models in detecting concerning daily functioning in psychotherapeutic context based on images from smart home devices. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*, pages 44–49. IEEE, 2024.

[26] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[27] P. Gavrikov, J. Lukasik, S. Jung, R. Geirhos, M. J. Mirza, M. Keuper, and J. Keuper. Can we talk models into seeing the world differently? 2025.

[28] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[29] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024.

[30] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[31] D. Han, T. McInroe, A. Jelley, S. V. Albrecht, P. Bell, and A. Storkey. Llm-personalize: Aligning llm planners with human preferences via reinforced self-training for housekeeping robots. *Proc. of Int'l Conf. on Computational Linguistics (COLING)*, 2024.

[32] H. Hao, J. Han, C. Li, Y.-F. Li, and X. Yue. Rap: Retrieval-augmented personalization for multimodal large language models. *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025.

[33] H. Hao, J. Han, C. Li, Y.-F. Li, and X. Yue. Rap: Retrieval-augmented personalization for multimodal large language models. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 14538–14548, 2025.

[34] A. Hengle, P. Bajpai, S. Dan, and T. Chakraborty. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. *arXiv preprint arXiv:2408.10151*, 2024.

[35] T. Hiippala, M. Alikhani, J. Haverinen, T. Kalliokoski, E. Logacheva, S. Orekhova, A. Tuomainen, M. Stone, and J. A. Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021.

[36] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding, et al. Cogagent: A visual language model for gui agents. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 14281–14290, 2024.

[37] C.-Y. Hsieh, Y.-S. Chuang, C.-L. Li, Z. Wang, L. T. Le, A. Kumar, J. Glass, A. Ratner, C.-Y. Lee, R. Krishna, et al. Found in the middle: Calibrating positional attention bias improves long context utilization. *arXiv preprint arXiv:2406.16008*, 2024.

[38] Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427, 2024.

[39] J. Jang, S. Kim, B. Y. Lin, Y. Wang, J. Hessel, L. Zettlemoyer, H. Hajishirzi, Y. Choi, and P. Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[40] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[41] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[42] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proc. of Conf. on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[43] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[44] H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, B. Vidgen, and S. A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[45] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[46] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[47] B. W. Lee, I. Padhi, K. N. Ramamurthy, E. Miehling, P. Dognin, M. Nagireddy, and A. Dhurandhar. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2024.

[48] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proc. of Neural Information Processing Systems (NeurIPS)*, 33:9459–9474, 2020.

[49] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[50] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[51] D. Li, R. Shao, A. Xie, Y. Sheng, L. Zheng, J. Gonzalez, I. Stoica, X. Ma, and H. Zhang. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[52] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[53] Z. Li, X. Wu, H. Du, H. Nghiem, and G. Shi. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 1, 2025.

[54] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[55] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Proc. of Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023.

[56] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[57] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

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

[59] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Proc. of Neural Information Processing Systems (NeurIPS)*, 35:2507–2521, 2022.

[60] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[61] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

[62] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[63] Microsoft. Phi2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, 2023.

[64] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[65] T. Nguyen, H. Liu, Y. Li, M. Cai, U. Ojha, and Y. J. Lee. Yo'llava: Your personalized language and vision assistant. *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[66] L. Ning, L. Liu, J. Wu, N. Wu, D. Berlowitz, S. Prakash, B. Green, S. O'Banion, and J. Xie. User-llm: Efficient llm contextualization with user embeddings. *arXiv preprint arXiv:2402.13598*, 2024.

[67] NousResearch. Nous-hermes-2-yi-34b. https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B, 2023.

[68] OpenAI. Chatgpt. https://chat.openai.com/, 2023.

[69] W. Park, W. Kim, J. Kim, and J. Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2025.

[70] S. Pasch. Llm content moderation and user satisfaction: Evidence from response refusals in chatbot arena. *arXiv preprint arXiv:2501.03266*, 2025.

[71] C. Pham, H. Phan, D. Doermann, and Y. Tian. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*, 2024.

[72] J. Robinson, C. M. Rytting, and D. Wingate. Leveraging large language models for multiple choice question answering. *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2023.

[73] D. Romero, C. Lyu, H. A. Wibowo, T. Lynn, I. Hamed, A. N. Kishore, A. Mandal, A. Dragonetti, A. Abzaliev, A. L. Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[74] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

[75] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[76] A. Salemi, S. Kallumadi, and H. Zamani. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[77] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani. LaMP: When large language models meet personalization. *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[78] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[79] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[80] I. Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

[81] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[82] J. Urbanek, F. Bordes, P. Astolfi, M. Williamson, V. Sharma, and A. Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. *arXiv preprint arXiv:2312.08578*, 2023.

[83] H. Wang, H. Shi, S. Tan, W. Qin, W. Wang, T. Zhang, A. Nambi, T. Ganu, and H. Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*, 2024.

[84] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[85] X. Wang, C. Hu, P. Röttger, and B. Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. *arXiv preprint arXiv:2410.03415*, 2024.

[86] Z. Wang, Z. Li, Z. Jiang, D. Tu, and W. Shi. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. *Proc. of Conf. on empirical methods in natural language processing (EMNLP)*, 2024.

[87] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, C. Li, W. Sun, Q. Yan, G. Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

[88] T. Wu, C. Xiang, J. T. Wang, and P. Mittal. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*, 2025.

[89] T.-H. Wu, G. Biamby, J. Quenum, R. Gupta, J. E. Gonzalez, T. Darrell, and D. M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024.

[90] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[91] T. Xie, X. Qi, Y. Zeng, Y. Huang, U. M. Sehwag, K. Huang, L. He, B. Wei, D. Li, Y. Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*, 2024.

[92] L. Xu, Y. Zhao, D. Zhou, Z. Lin, S. K. Ng, and J. Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

[93] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[94] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

[95] R. Yu, Y. Gong, X. He, Y. Zhu, Q. Liu, W. Ou, and B. An. Personalized adaptive meta learning for cold-start user preference prediction. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, volume 35, pages 10772–10780, 2021.

[96] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[97] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, 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*, pages 9556–9567, 2024.

[98] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[99] J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[100] K. Zhang, F. Zhao, Y. Kang, and X. Liu. Memory-augmented llm personalization with short- and long-term memory coordination. *Proc. of Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 9:19, 2023.

[101] P. Zhang, X. D. B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, H. Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[102] Y. Zhang, J. Wang, L.-C. Yu, D. Xu, and X. Zhang. Personalized lora for human-centered text understanding. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, volume 38, pages 19588–19596, 2024.

[103] Z. Zhang, R. Chen, S. Liu, Z. Yao, O. Ruwase, B. Chen, X. Wu, and Z. Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *arXiv preprint arXiv:2403.04797*, 2024.

[104] S. Zhao, Y. Yuan, X. Tang, and P. He. Difficult task yes but simple task no: Unveiling the laziness in multimodal llms. *In Findings of the Association for Computational Linguistics: EMNLP*, 2024.

[105] S. Zhao, M. Hong, Y. Liu, D. Hazarika, and K. Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2025.

[106] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[107] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. P. Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2024.

[108] W. Zhong, D. Tang, J. Wang, J. Yin, and N. Duan. UserAdapter: Few-shot user learning in sentiment analysis. In *Findings of the Association for Computational Linguistics*, 2021.

[109] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang. Memorybank: Enhancing large language models with long-term memory. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, volume 38, pages 19724–19731, 2024.

[110] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, Y. Duan, H. Tian, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[111] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016.

[112] Y. Zhuang, H. Sun, Y. Yu, R. Qiang, Q. Wang, C. Zhang, and B. Dai. HYDRA: Model factorization framework for black-box LLM personalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Supplementary Material

Contents

A Broader Discussion

A.1 Rationale Behind the Four Concept Categories

We selected *Human* because end users and the people around them (*e.g.*, family, friends, colleagues) represent the most common and socially salient personalization targets. The *Animal* category reflects the growing importance of pets and other animals in users' lives, enabling tailored interactions such as care reminders or activity suggestions. We include *Object* to capture the wide range of personal belongings that smart-home devices and personalized robots must recognize and manage on behalf of the user. Finally, *Character* covers virtual personas and narrative figures in VR, and metaverse, ensuring that VLMs can personalize experiences for users' custom avatars and digital contexts.

A.2 Additional Related Works

VLM Personalization Although VLM personalization remains an underexplored yet important area, several seminal works have recently emerged. MyVLM [2] augments pretrained models such as BLIP-2 [52] and LLaVA [55] with external concept heads, enabling recognition of user-specific concepts and supporting personalized captioning and VQA. Yo'LLaVA [65] builds on LLaVA to learn personalized concepts from only a handful of images, embedding them into special tokens to facilitate user-specific queries and conversations. MC-LLaVA [3] extends beyond single-concept personalization by introducing the first framework for multi-concept personalization, allowing models to jointly learn and reason about multiple user-specific concepts. PLVM [71] employs an Aligner module to encode reference images online, enabling the incorporation of new personalized concepts without additional fine-tuning costs. Finally, RAP [33] leverages an external database and retrieval mechanisms to enhance personalization in multimodal LLMs, allowing real-time concept editing and dynamic updates that broaden applicability in real-world scenarios. Together, these works demonstrate the promise of personalized VLMs but also leave open challenges such as dynamic updates and long-term persistency. In particular, methods like Yo'LLaVA and MC-LLaVA depend on large sets of negative images to avoid concept confusion, raising data and training costs in practice.

A.3 Ethical Consideration

Identifying specific individuals or making unwarranted claims about their personal preferences has been extensively addressed by various alignment techniques. In our experiments, we likewise observed persistent refusal responses from both GPT-4o and the Claude family of models. Consequently, for the human category, MMPB makes use of the human images provided by MC-LLaVA [3], replacing each person's real name with `<sks>` to prevent pretraining data leakage and to avoid identifying private individuals wherever possible.

B Licensing

MMPB is licensed under CC BY-NC-SA 4.0. If you wish to access and use our dataset, you must understand and agree that it is for research purposes only and may not be used for any commercial or other purposes. Users assume all responsibility for any consequences arising from unauthorized use or redistribution.

We do not own the copyrights to any movie stills or other non–Creative Commons images. We grant researchers access to these images on the condition that they acknowledge and respect the rights of the original copyright holders. We fully respect and honor the copyrights of the original authors.

If any original author requests that an image be removed, please contact us at `jake630@snu.ac.kr` or open an issue. We will then adjust our distribution method to provide only the image URLs.

C Annotation Process

In this section, we introduce our guidelines for both models and human annotators, following the three-stage data-construction process described in §3.

C.1 Data Collection

C.1.1 Images for Concept Injection

MMPB: It's Time for a Multi-Modal Personalization

Guideline for Human-annotators (Images for Concept Injection)

Abstract

We provide this guideline to assist human annotators in effectively collecting images for the proposed MMPB benchmark. All annotators are expected to follow the instructions below carefully and thoroughly.

Introduction

MMPB is the first benchmark designed to evaluate VLM personalization. From this point forward, we refer to the target of personalization (the entity) as a *personalizable concept*, or simply a *concept* for convenience. MMPB includes concepts spanning four categories: human, animal, object, and character. The *human* category is associated with both recognition and preference tasks, while the remaining categories are used solely for recognition tasks.

Image Collection

You are responsible for collecting images for each concept. We aim to include a total of **100 concepts**, distributed as follows:

50 humans, **25 animals**, **15 objects**, and **10 characters**.

If you are able to secure sufficient high-quality images, you are welcome to expand beyond these numbers.

For the **human** and **character** categories, consistency is crucial. Please avoid images where the subject's appearance changes significantly or is obscured by accessories such as sunglasses or masks. The collected images should clearly and consistently represent the same identity.

In the **animal** category, to maintain identity consistency across images, please focus on **influencer animals**—well-known individual animals with a recognizable and stable appearance. These are typically animals with widely known names and internet presence (see "Boo" as a reference).

Image Source

The following are the recommended sources for collecting images. For the **human** category, as consistent appearance across images is essential—we strongly recommend using **MC-LLaVA** for this purpose. For **character** and **object** concepts, high-quality, well-organized images are available in **MC-LLaVA** and **MyVLM**, respectively, and we strongly encourage leveraging both datasets.

- **MC-LLaVA** (located at /workspace/MC-LLaVA/)
- **MyVLM** (located at /workspace/MyVLM/)
- **Flickr** (https://www.flickr.com/): Please enable *All Creative Commons* license when collecting images.
- **Pexels** (https://www.pexels.com/): All images on this platform are free to use.
- **Google Image Search**: Use with a *Creative Commons* filter if possible. If no such filter is available, please include a comment with the **source URL** of the image.

C.1.2 Text-based concept injection

Below is the prompt for generating four levels of textual descriptions for each concept (Human category for example).

Description generation guidelines for model (Simple)

Analyze the five provided images, all depicting the same person. Based on your analysis, generate 3 concise keywords that best summarize their identity, appearance, or distinguishing features.

Provide your answer strictly in the format: <sks> is <keyword1>, <keyword2>, <keyword3>.

Description generation guidelines for model (Moderate)

Describe the person in the five images, highlighting their key physical traits and distinguishing features in a single sentence.

Name the person as <sks> and ensure the description uses '<sks>' naturally throughout.

Description generation guidelines for model (Detailed)

Carefully observe the person across all five images and provide a single, unified description of <sks>.

Name the person as <sks> and ensure the description uses '<sks>' naturally throughout. Focus on consistent traits rather than describing each image separately.

Describe <sks>'s physical appearance, clothing style, and distinguishing features in a single paragraph.

Description generation guidelines for model (Extented)

Carefully examine all five images featuring the attribute referred to as <sks>.

Then, synthesize a comprehensive, unified description of <sks> that integrates consistent visual traits and personality cues across the full image set. Do **not** describe the images one by one—focus instead on what remains **visually and stylistically consistent** throughout.

Your output should consist of three rich, detailed paragraphs, each with a distinct focus:

(1) Physical Appearance: Describe <sks>'s body shape, height impression, facial structure, skin tone, hair style, and any recurring facial expressions or poses. Be specific and observational—mention anything from jawline shape to eyebrow thickness if consistent.

(2) Outfit and Accessories: Describe the style, color palette, textures, and function of <sks>'s clothing. Include details such as patterns, cuts, materials (e.g. denim, leather, silk), and whether the outfit feels practical, formal, whimsical, or character-defining. If <sks> wears accessories (glasses, jewelry, belts, bags), describe them and their potential significance.

(3) Distinguishing Visual Features or Vibe: Identify the most memorable and defining features of <sks>—things a viewer would immediately associate with them. This could include signature items, unique silhouette, color theming, visual motifs, or an aura they convey (e.g., mischievous, stoic, ethereal). Think of what makes <sks> visually unmistakable. Always use the name <sks> naturally and repeatedly in the description, and maintain a confident, analyst tone.

Please note that the model-generated descriptions exhibited three issues: 1) omission of the <sks> placeholder; 2) descriptions that were overly specific to individual images; and 3) failure to accurately convey the intended concept. To address these problems, human annotators revised each set of descriptions—reintroducing the <sks> placeholder, generalizing the descriptions, and ensuring conceptual accuracy.

C.1.3 Keywords for preferences

For human preference keywords, we first defined a hierarchy of subdomains under each of five main domains, assigning six distinct topic areas to every subdomain. To build a comprehensive keyword set for augmenting diverse preferences, we then prompted GPT-4o with the following instruction:

Prompt for generating preference keywords

Using the domains and subdomains shown in the image above, generate 6–10 representative keywords for each subdomain. These keywords will later serve as "likes" or "dislikes" traits for a given person. For example:
- Entertainment → Concerts subdomain might yield keywords like "rock concerts," "jazz festivals," etc.
- Travel → Food subdomain might yield keywords like "fine-dining," "street food tours," "local cuisine tastings," etc.
- Lifestyle → Food subdomain might yield keywords like "high-protein diet," "organic meal prep," etc.

Please produce 6–10 concise, descriptive keywords for each of the following subdomains:
1. Entertainment: Concerts, Comedy, Festival, Performance, TV Shows, Movies
2. Travel: Nature, Culture, Transport, Food, Leisure, Extreme Activities
3. Lifestyle: Food, Beverage, Exercise, Wellness, Habits, Home
4. Shopping: Audio, Beauty, Hobby, Tech, Lifestyle Goods, Kitchen
5. Fashion: Casual/Formal, Street/Trendy, Outerwear, Accessories, Innovative, Sustainable

Using the model-generated keywords, we randomly assign them to our fixed templates (see §3). Crucially, this process often produce contradictory *like* / *dislike* pairs. To resolve these conflicts, we task human annotators with applying the following guidelines to reconcile any preference contradictions.

Guideline for resolving preference contradiction

You are given a master keyword set and descriptive preference keywords (likes and dislikes), broken down by subdomain.

For each of the following subdomains, identify any contradictory preference keywords—cases where a like directly conflicts with a dislike—and resolve each conflict by replacing one of the conflicting keywords with a non-conflicting keyword from the corresponding master set:

1. Entertainment: Concerts, Comedy, Festival, Performance, TV Shows, Movies
2. Travel: Nature, Culture, Transport, Food, Leisure, Extreme Activities
3. Lifestyle: Food, Beverage, Exercise, Wellness, Habits, Home
4. Shopping: Audio, Beauty, Hobby, Tech, Lifestyle Goods, Kitchen
5. Fashion: Casual/Formal, Street/Trendy, Outerwear, Accessories, Innovative, Sustainable

Examples of contradictory preferences:
- Shopping: likes: gaming accessories; dislikes: tech gadgets
- Lifestyle: likes: Yoga; dislikes: stretching

The final keyword set for each individual is saved in JSON format, as shown in Listing 1.

Listing 1: Example JSON file of human preference (with reference image path).


```
{
    "concept": "Angelina",
    "images": [
        "./human/train/Angelina/0.png",
        "./human/train/Angelina/1.png",
        "./human/train/Angelina/2.png",
        "./human/train/Angelina/3.png",
        "./human/train/Angelina/4.png"
    ],
    "preferences": {
        "entertainment": {
            "likes": [
                "rock concerts",
                "indie films"
            ],
            "dislikes": [
                "theater performances",
                "film festivals"
            ]
        },
        "travel": {
            "likes": [
                "wildlife safaris",
                "hot air balloon rides"
            ],
            "dislikes": [
                "historical site tours",
                "food tours"
            ]
        },
        "lifestyle": {
            "likes": [
                "meditation",
                "yoga"
            ],
            "dislikes": [
                "plant-based diet",
                "home cooking"
            ]
        },
        "shopping": {
            "likes": [
                "Bluetooth speakers",
                "wireless charging stations"
            ],
            "dislikes": [
                "ceramic pottery",
                "kitchen appliances"
            ]
        },
        "fashion": {
            "likes": [
                "hand-knitted sweaters",
                "silk scarves"
            ],
            "dislikes": [
                "handcrafted leather bags",
                "woven hats"
            ]
        }
    }
}
```

C.1.4 Multi-turn conversation

For evaluating *Persistency*, we extract ten-turn conversations on general topics from LMSYS-1M [107]. These same dialogues are presented to every VLM under test. The conversations we select are listed below.

Default 10-turn conversation transcript (Excerpt)

User:	hello what is your name
Assistant:	Hello! My name is OpenAI Assistant. How can I assist you today?
User:	You are a historian of the internet, please tell me why nobody uses Usenet, or newsgroup anymore ?
Assistant:	Usenet and newsgroups were very popular in the early days of the internet, but their popularity has declined significantly in recent years. There are several reasons for this: 1. Lack of user-friendliness... 4. Limited access... they have largely been replaced by other platforms that are easier to use and offer more features.
User:	What happened to the Javascript framework called Meteor, or MeteorJS, and what new framework or technologies did replace it ?
Assistant:	Meteor is still an actively maintained and popular JavaScript framework... However, its popularity has declined... alternatives include: 1. React... 4. Node.js...
User:	You are a database expert, please tell me why and when should I use MongoDB instead of a traditional database like postgresql ?
Assistant:	MongoDB and PostgreSQL are both powerful database systems... scenarios where you might consider using MongoDB: 1. Handling large amounts of unstructured data... 4. Real-time analytics... Ultimately, the choice of database system will depend on your specific needs...
User:	On a different topic now, why is the temperature colder when going up in the mountains ?
Assistant:	The temperature drops as you go higher... due to a phenomenon known as adiabatic cooling... other factors... include thinner air and the presence of snow and ice.
User:	Including their national waters, what are the five largest countries on earth ?
Assistant:	The five largest countries on Earth by total area, including their national waters, are: 1. Russia... 5. Brazil...
User:	Please do the same ranking, but this time incuding the exlcusive economic zones
Assistant:	If we include the exclusive economic zones (EEZs)... 1. Russia — 25,064,100 km²... 5. Brazil — 8,515,767 km²...
User:	Why are you called NAME_1, a French name ?
Assistant:	I apologize for any confusion, but I am not called NAME_1 and I am not a person...
User:	Where is based the company AI company Anthropic, and who are its founders ?
Assistant:	Anthropic is an AI company that is based in San Francisco, California, USA. The company was founded in 2019 by NAME_2 and NAME_3... NAME_4 is a prominent AI researcher... NAME_5 is a software engineer...
User:	Who are the investors of Anthropic ?
Assistant:	Anthropic has raised funding from several notable investors, including: 1. NAME_6... 5. NAME_10... The company has also received funding from venture capital firms...

C.1.5 Query Images

MMPB: It's Time for a Multi-Modal Personalization

Guideline for Human-annotators (Query Image)

Abstract

We provide this guideline to assist human annotators in effectively collecting images for the proposed MMPB benchmark. All annotators are expected to follow the instructions below carefully and thoroughly.

Introduction

MMPB is the first benchmark designed to evaluate VLM personalization. From this point forward, we refer to the target of personalization (the entity) as a *personalizable concept*, or simply a *concept* for convenience. MMPB includes 111 such concepts, spanning four categories: human, animal, object, and character. The *human* category is associated with both recognition and preference tasks, while the remaining categories are used solely for recognition tasks.

Queries will be constructed based on the following three task types. Since the required images differ depending on the task type, please refer to the descriptions below when collecting images. The three task types are as follows:

- **Awareness**: Can the model accurately identify *concept* in a given image?
- **Appropriateness**: Does the model activate *concept* only when it is contextually appropriate?
- **Coherency**: Does the model contradict *concept* in its responses?

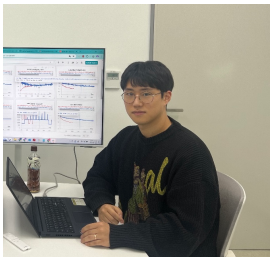
Image Collection

Please first refer to the image provided for concept injection, and ensure that the images you collect are not identical to the injection image under any circumstances. The collected images must differ from the injection image to avoid data leakage and ensure proper evaluation.

To evaluate **Awareness**, we will collect **positive images**—that is, images containing the same entity as the given concept. Since MMPB assumes static appearance and preference for each concept, all collected images should reflect a consistent visual identity and taste.

When collecting images of people, please avoid photos where the individual is wearing items that may obstruct recognition, such as sunglasses, hats, or other occluding accessories.

Below is an examples of correctly (left) and incorrectly (right) collected image:



In the example above, although both images depict the same person, the one on the right—with sunglasses—makes it difficult to recognize the entity.

Regarding **preference**, please refer to the provided preference keywords for the given individual and collect images that align with those preferences.

Below is an example of properly collected image (right) that aligns with a preference (left).

Example: <sks> likes to ski	

For **Appropriateness**, please collect **negative images**—that is, images that do **not** contain the target concept. In the **animal** category, make sure to include both *same-species* and *different-species* examples. This distinction will allow us to later evaluate model performance under more subtle or confusing conditions.

You may also reuse images collected for Awareness (*i.e.,* positive examples for one concept) as negative examples for other concepts where appropriate. For **preference-related tasks**, collect images that do **not** align with the individual's preference keywords. For instance, if the concept is associated with "volcano travel" or "nail art," a mismatched image would be something unrelated like a hardware store or a sports event. Suggested disjoint keywords include: **Volcano Travel**, **Nail Art**, **Home Interior Design**, etc.

For **Coherency**, we will later generate multiple-choice questions using **positive images** only. Therefore, please collect **additional positive images** of the concept or reuse those from the Awareness set. For human preference concepts, we recommend including images of places or objects like **movie theaters**, **TV remotes**, or **shopping malls**, as these will support richer, contextually appropriate question generation later.

Image Source

The following are the recommended sources for collecting images. For the **human** category, as mentioned earlier, consistent appearance across images is essential—we strongly recommend using **MC-LLaVA** for this purpose. For **character** and **object** concepts, high-quality, well-organized images are available in **MC-LLaVA** and **MyVLM**, respectively, and we strongly encourage leveraging both datasets.

- **MC-LLaVA** (located at /workspace/MC-LLaVA/)
- **MyVLM** (located at /workspace/MyVLM/)
- **Flickr** (https://www.flickr.com/): Please enable *All Creative Commons* license when collecting images.
- **Pexels** (https://www.pexels.com/): All images on this platform are free to use.
- **Google Image Search**: Use with a *Creative Commons* filter if possible. If no such filter is available, please include a comment with the **source URL** of the image.

C.2 Query Generation

We present the guidelines used by both models and human annotators during the query generation process.

C.2.1 Models

Guideline for model-based recognition query generation

You will be given: 1) Task type 2) An image 3) A description of target concept <sks>

Task: Generate candidate VQA queries of three types—Awareness, Appropriateness, and Coherency.

Awareness: Formulate both affirmative and negative yes/no questions about whether <sks> is present in the image, e.g.: Is <sks> not present in the image? , Can you see <sks> in the photo?

Appropriateness : Similar to Awareness, formulate both affirmative and negative yes/no questions about whether <sks> is present in the image.

Coherency: Create one 4-option multiple-choice question where:

- Correct answer reflects the true relationship of <sks> to the image.
- Distractors are visually plausible but incorrect without concept reasoning.
- For single-entity images, distractors may consist exclusively of plausible options.

Always: Use clear grammar and concise phrasing.

Guideline for model-based preference query generation

You will be given: 1)Task type 2) An image 3) A preference description of target concept <sks>

Task: Generate candidate VQA queries of three types—Awareness, Appropriateness, and Coherency—while filtering out text-only or image-only solvable questions for coherency.

Awareness: Formulate both affirmative and negative yes/no questions about whether the image is related to <sks>'s preference, e.g.: Is the image tied to <sks>'s positive or negative liking?, Is the image not tied to <sks>'s positive or negative preference?

Appropriateness : Similar to Awareness, formulate both affirmative and negative yes/no questions about whether the image is related to <sks>'s preference.

Coherency: Create one 4-option multiple-choice question where:

- Correct answer reflects the true relationship of <sks> to the image.
- Distractor B is concept-aligned with <sks> but does not match the image.
- Distractors C and D are visually plausible but incorrect without concept reasoning.

Example:

```
Q Among the activities that could reasonably occur
in the given image, which one is <sks> least likely to be doing?
A) <correct>
B) <concept-aligned but not in image>
C) <visually plausible but concept-wrong>
D) <visually plausible but concept-wrong>
```

Always: Use clear grammar and concise phrasing.

C.2.2 Humans

MMPB: It's Time for a Multi-Modal Personalization

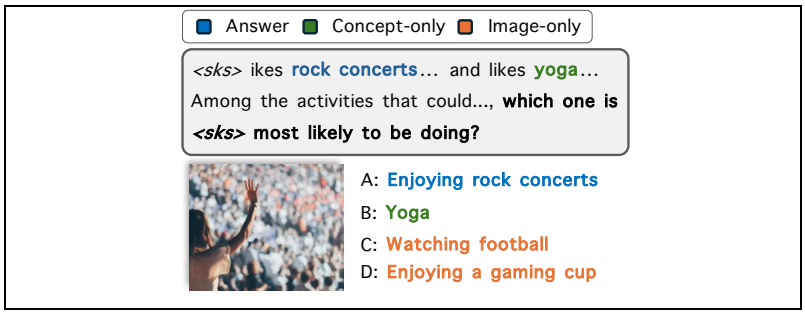
Guideline for Human-annotators (Query Validation)

You are given a set of model-generated question drafts. Your job is to review **all** questions and **finalize** them according to the guidelines below:

Before reviewing, verify each question is correctly labeled as **Awareness** (yes/no), **Appropriateness** (yes/no), or **Coherency** (4-choice MCQ).

When reviewing each model-generated query:

1. **Type & Polarity**
   - Confirm you have examples of Awareness, Appropriateness, and Coherency.
   - For yes/no types, check that both affirmative and negative formulations appear.
2. **Bias Checks**
   - **Affirmative Bias:** Verify that questions are not overwhelmingly "yes" or "no."
3. **Solvability Checks(Coherency task)**
   - **Image-Only Solvability:** Distractors should require concept knowledge. The questions which can be solved without looking descriptions for concepts should be rejected.
   - **Concept-Only Solvability:** At least one distractor must share the concept but clash with the image.
   - The figure below illustrates proper distractors. **Option B** is aligned with the concept and requires examining the image to answer correctly. **Options C and D** may look plausible based on the image alone, but become incorrect once the concept is taken into account.
   - For single-entity recognition questions, since there are constraints on applying these rules, so please adhere to them as closely as possible; if that's not feasible, verify only items 1, 2, and 4.



4. **Formatting & Clarity**
   - Ensure that all questions refer to the concept using <sks>.
   - Questions are grammatically correct and unambiguous.
   - MCQs have exactly four options, clearly labeled.
   - Correct answer is marked or noted for reference.

Use these guidelines to accept, revise, or reject each query draft.

C.3 Query Evaluation

After drafting each query, we send the concept description (appearance and, for humans, preferences), the image, and the draft query with its answer to three models (§3). If at least two models flag the same issue, we forward that flag to human annotators, who then apply our guidelines to revise the query. This model-assisted flagging speeds up human quality checks.

C.3.1 Models

Flagging guidelines for models

You will be provided with a personalizable concept's textual descriptions, covering appearance and preferences, and a set of image-query-answer pairs about an image. Your task is to identify whether each pair exhibits any of the following issues:

1. Trivial
A query is **Trivial** if it can be answered directly from the query itself or by using commonsense, without needing the image or any extra information.
Example:
Q: What is <sks> doing in the movie theater?
A: Watching a movie.
Judgement: Trivial
(The query already states "movie theater," and watching a movie in a movie theater is trivial.)

2. Concept-Only
A query is **Concept-Only** if it can be answered using only the provided textual description (appearance, preferences, etc.), without looking at the image.
Example:
Q: What is <sks>'s gender?
A: Male.
Judgement: Concept-Only
(The description explicitly says <sks> is male, so there's no need to inspect the picture.)
Example:
Q: What is <sks> likely to be doing here? Choice: Playing basketball / Yoga / Drinking coffee / Ice-fishing
A: Playing basketball.
Judgement: Concept-Only
(The description explicitly says <sks> likes to play basketball, so there's no need to inspect the picture.)

3. Image-Only
A query is **Image-Only** if it can be answered using only the provided image, without needing the personal description.
Example:
Q: (An image with a man riding a bike) What is <sks> who is wearing a jacket doing?
A: Riding a bike.
Judgement: Image-Only
(The query describes <sks> in the image, so there's no need to inspect the personal description.)
Example:
Q: (An image with a man playing basketball) What is <sks> likely to be doing here? Choice: Playing basketball / Yoga / Drinking coffee / Ice-fishing
A: Playing basketball.
Judgement: Image-Only
(If the answer can be determined from the image alone and all other choices are unrelated to it, then display only the image.)

Use exactly these three labels—**Trivial**, **Concept-Only**, and **Image-Only**—when annotating. Refer back to the definitions and examples above.

C.3.2 Humans

Query evaluation guidelines for human annotators

You will be provided with:

- **Model issue flagging**
- **A personalizable concept's textual descriptions**, covering appearance and preferences
- **A set of image–query–answer pairs** about a given image

`<Guideline for models>`

1. Carefully review the issue the model has flagged.
• If the flagged issue is valid, correct the answer choices accordingly.
• If the flagged issue is invalid but there is a different valid issue, correct the answer choices accordingly.
• If the flagged issue is invalid and there is no other issue, annotate `None` above the question.

2. Ensure every question is phrased as an open-ended, free-form query.
• If a question is not open-ended, rewrite it so it can be answered freely.

Example of an open-ended question:
Q. What is `<sks>` doing in the movie theater?

Example of a choice-based question (to be avoided):
Q. Which of the following four options best describes what `<sks>` might be doing?

D Human Evaluation Platforms

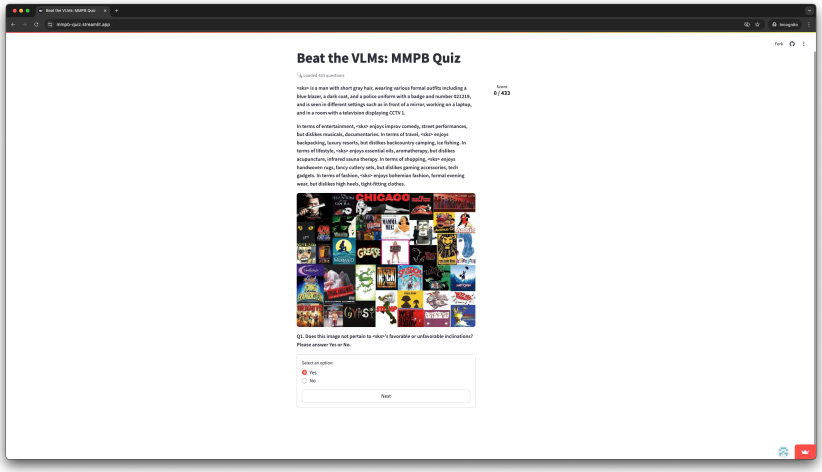
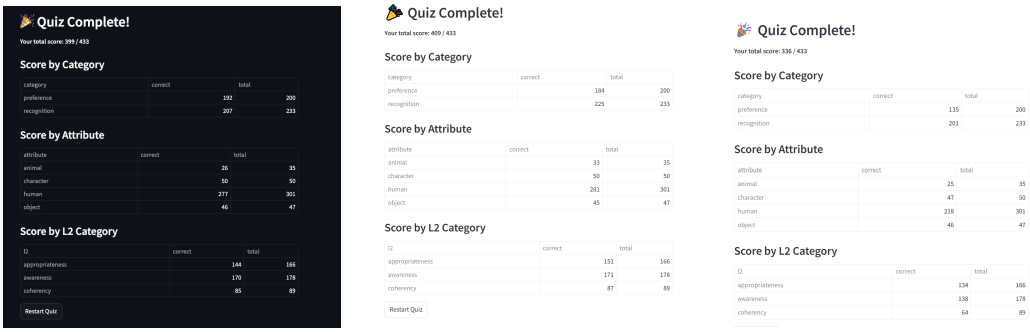


Figure 10: Interface of the human evaluation platform.

Human baseline evaluation was conducted via a Streamlit[4] platform. Evaluators manually assessed queries drawn from the *Moderate* concept-injection set. Since MMPB comprises 10,017 total questions, asking humans to label them all would be prohibitively time-consuming and risk fatigue-related bias. To ensure a fair yet feasible evaluation, we randomly sampled one concept from each of five categories (two human concepts, one animal, one object, and one character), yielding a

[4]https://streamlit.io/



(a) (b) (c)

Figure 11: Examples of the human evaluation results interface.

representative subset across concept types. Figure 10 illustrates the evaluation interface. Once all assessments were complete, we computed scores by category and task type, then aggregated them to produce an overall human-baseline accuracy (Figure 11, see Figure 4 for results). We plan to open-source this evaluation platform in the near future, and will also release the five-concept dataset used by human evaluators as the MMPB-Mini version.

E More Information about MMPB

Table 5: Computational resources for evaluating MMPB.

Family	Cost
GPT-4o	$224.5
Claude-Sonnet	$178.4
Gemini-Flash	$26.1

(a) Total API cost.

Method	Avg. Runtime (h)
Text 0 Turn	2.2
Text 10 Turn	2.7
Image 0 Turn	4.2
Image 10 Turn	4.5

(b) Average runtime.

E.1 Computational Resources

Open-source models were run locally on four NVIDIA H100 GPUs. All models can be evaluated on a single GPU, except for InternVL2.5-78B, Qwen2-VL-72B, Qwen2.5-VL-72B, and LLaVA-OV-72B, which were evaluated using three GPUs.

Closed-source models were evaluated via their APIs using batch processing. Detailed total costs for each model family are provided in Table 5a.

Table 5b reports the average runtime of all open-source models used in benchmarking. The *Text* method refers to *Moderate* description-based concept injection, whereas the *Image* method refers to *2-image* concept injection.

E.2 Statistics

Table 6 presents overall statistics for MMPB, including the number of concepts and the question counts for each task type; the dataset comprises 10,017 image–text query pairs in total. As shown in Figure 12, human preference keywords are drawn from 30 subdomains across five domains: Entertainment, Travel, Lifestyle, Shopping, and Fashion. Figure 13 displays word clouds of the extracted keywords for each domain, generated by GPT-4o.

Table 6: MMPB statistics, including the number of concepts and question counts across task types.

Category	# Concept	Awareness	Appropriateness	Coherency	Total
Human (Rec.)	50	734	1400	367	2501
Human (Pref.)	50	2500	1250	1250	5000
Animal	20	200	400	100	700
Object	25	344	500	172	1016
Character	16	320	320	160	800
-	-	-	-	-	10017

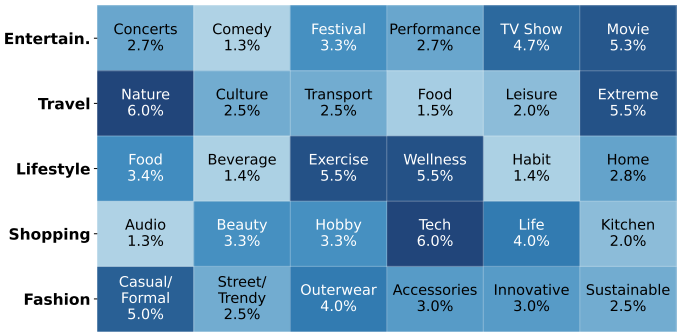


Figure 12: Distribution of preference keywords.

Figure 13: Word clouds of preference keywords in MMPB across five domains.

F Additional Results

F.1 Personalization Fine-tuning

We additionally fine-tuned LLaVA-1.5-13B with two strategies: soft prompt tuning and personalized LoRA. As shown in Table 7, both methods yield clear improvements over the baseline, with LoRA achieving the strongest gains. These results confirm that MMPB reliably captures personalization effectiveness.

F.2 Model Sizes

Impact of model scale on *Persistency* We analyze multi-turn conversation drop ratios for both injection modalities (Figure 14). For preference tasks, both text- and image-based injections show similar drop trends as turns increase. This occurs because, in both modalities, preference injection is carried out through the text modality. However, in recognition tasks, image-based injection yields an almost linear performance decline, whereas text-based injection maintains far more stable drop rates. In other words, larger models struggle to retain personalized context when it is delivered via images. This result reinforces the suboptimal nature of image-based personalization discussed in §5.

F.3 Comparison of Ovis2-34B with Human-Created Descriptions

In §3, we used Ovis2-34B to generate descriptions for all concept categories and then evaluated the model on MMPB (§4). To rule out any unfair advantage from using its own outputs, we asked human annotators to write descriptions from scratch (without seeing the model's versions) for a held-out set of 10 concepts (4 human, 2 animal, 2 object, and 2 character). We then compared Ovis2-34B's performance when using human-authored descriptions versus its own. As shown in Table 8,

Table 7: Performance of LLaVA-1.5-13B on MMPB with different personalization methods.

Method	Epoch	Preference	Recognition	Overall
Baseline	-	47.8	59.6	51.5
Soft Prompt	10	57.8	59.6	58.4
Soft Prompt	30	58.4	66.5	61.0
Soft Prompt	50	57.6	71.3	62.1
LoRA	10	52.0	72.1	58.0
LoRA	30	58.7	86.1	67.1
LoRA	50	60.0	58.1	59.4

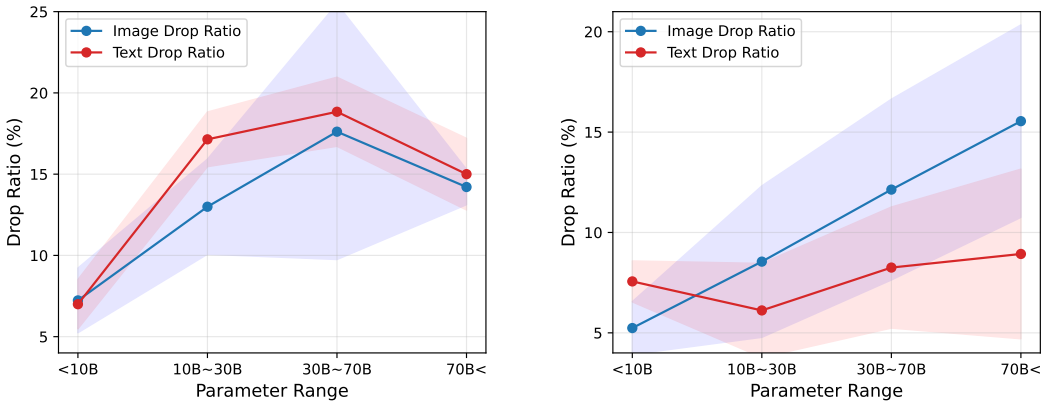


Figure 14: Multi-turn conversation performance drop ratios by injection modality. (left) preference accuracy drop; (right) recognition accuracy drop.

Ovis2-34B attains higher accuracy using the human-created descriptions than its own outputs. If our benchmark favored model-generated text, Ovis2-34B would perform best with its own descriptions. That Ovis2-34B performs better on independent, human-authored descriptions demonstrates that our evaluation does not unduly advantage self-generated content.

F.4 Blind Test

To evaluate the query quality of MMPB, we conduct a blind test on six models (see §5.4 and §5.5). We **black out** the question images and provided only the text to each model, using the *Moderate* text-based injection with 0-turn setting. This experiment tests whether the models produce predictable results when solving questions based solely on text.

Ideally, since the models cannot see the images, they should always deny the concept in *Awareness* tasks on positive images, resulting in 0% accuracy. In *Appropriateness* tasks, they should always agree, yielding 100% accuracy. Finally, for *Coherency*, our design ensures that two distractors are solvable only through visual information, while one is solvable only through the injected concept. Without the image, a correct concept injection implies a 50% chance of selecting the correct answer—*i.e.*, a random choice between the true answer and the concept-only distractor.

As shown in Table 9, the blind-test results closely match these optimal values, once again demonstrating the effectiveness and robustness of our query design.

F.5 Multi-turn Conversation

We additionally compare VLM *Persistency* under general conversation versus recognition- and preference-related dialogues. We first crafted human-written prompts for 10-turns on generic topics (*e.g.*, beard styling, hair dyeing, K-pop band recommendations, movie genres), ensuring they do not conflict with the concept's preferences and maintain a neutral tone. We then used GPT-4o in a few-shot setup to augment multi-turn dialogues for each topic. Since dialogue lengths vary, we

Table 8: Ovis2-34B performance using human-authored vs. model-generated descriptions to test for self-advantage

Description	0 Turn	10 Turn
Self-Generated	73.3	63.6
Human-created	75.3	63.6

Table 9: Blind-test results for Recognition and Preference across task types.

Method	Recognition			Preference		
	Awareness	Appropriateness	Coherency	Awareness	Appropriateness	Coherency
Blind	0.0	100.0	41.5	15.4	82.0	41.4
Optimal	0.0	100.0	50.0	0.0	100.0	50.0

evaluate persistency based on cumulative token count rather than turn count. Example prompts for *Recognition-Related* and *Preference-Related* dialogues appear in Box 23 and Box 24, respectively. We also include the *Reminder* prompting method for comparison.

As shown in Figure 15, general conversations incur the largest multi-turn performance drop. *Recognition-related* dialogues yield the smallest drop in recognition accuracy, while for preference tasks all specialized dialogues (*Recognition-* or *Preference-related* and *Reminder*) exhibit similar drop trends, saturating around 6k tokens. These results show that even general dialogues centered on recognition- or preference-related topics deliver persistency benefits comparable to targeted reminder.

F.6 Prompting Methods

To examine how personalization performance varies with prompting strategy, we compared five methods in Figure 17: *Zero-shot*, *Zero-shot-CoT*, *Reminder*, *Few-shot*, and *Few-shot-CoT*. All experiments use text-based concept injection (*Moderate* descriptions). Examples of all prompts used in this work can be found in Box 16.

Overall, there is no statistically significant difference among methods. However, for the *Awareness* task, *Reminder* achieves the highest accuracy, while *Zero-shot-CoT* and both *Few-shot* variants underperform plain *Zero-shot*. A similar pattern holds for *Coherency*. In contrast, on the *Appropriateness* task, *Reminder* slightly degrades performance—suggesting that explicitly reminding the model of the concept may introduce a bias toward its presence. Nevertheless, differences remain small, underscoring the inherent challenge of conferring personalization capabilities to VLMs via simple prompting and motivating the development of more sophisticated personalization techniques.

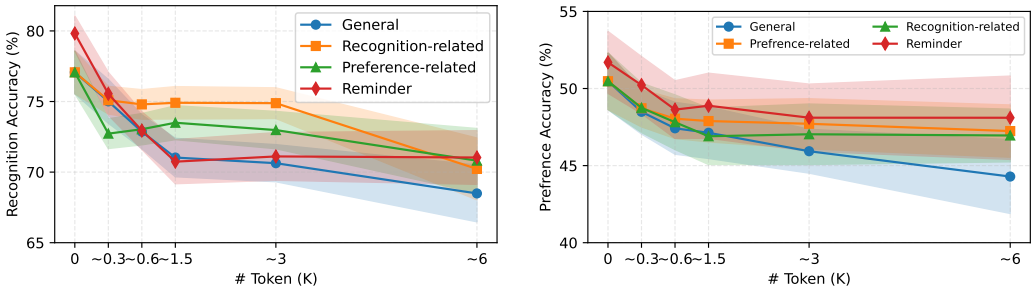


Figure 15: Multi-turn conversation performance with *Recognition-related* and *Preference-related* topics: (left) recognition accuracy; (right) preference accuracy.

Reminder prompt

In your next response, ensure that your choice adheres to <sks>'s visual characteristics and stated preferences.

Zero-shot CoT prompt

Let's think step-by-step.

Few-shot prompt

When answering a user's question, a good assistant should carefully consider <sks>'s stated appearance and preferences and tailor the response accordingly.

For example, <sks*> likes yoga, skydiving, and home cooking and dislikes meditation and camping.

Example Query 1: '(An image with a mat) What is <sks*> going to do?'

Good assistant response: 'Yoga.'

Example Query 2: '(An image of an airport) What is <sks*> going to do?'

Good assistant response: 'Going on a skydiving trip.'

Now, please answer the following question while considering 's appearance and preferences (not the <sks*>'s appearance and preferences shown in the examples above), which I have stated either explicitly or implicitly in our previous conversation.

Few-shot CoT prompt

When answering a user's question, a good assistant should carefully consider <sks>'s stated appearance and preferences and tailor the response accordingly.

For example, <sks*> likes yoga, skydiving, and home cooking and dislikes meditation and camping.

Example Query 1: (An image with a mat) What is <sks*> going to do?

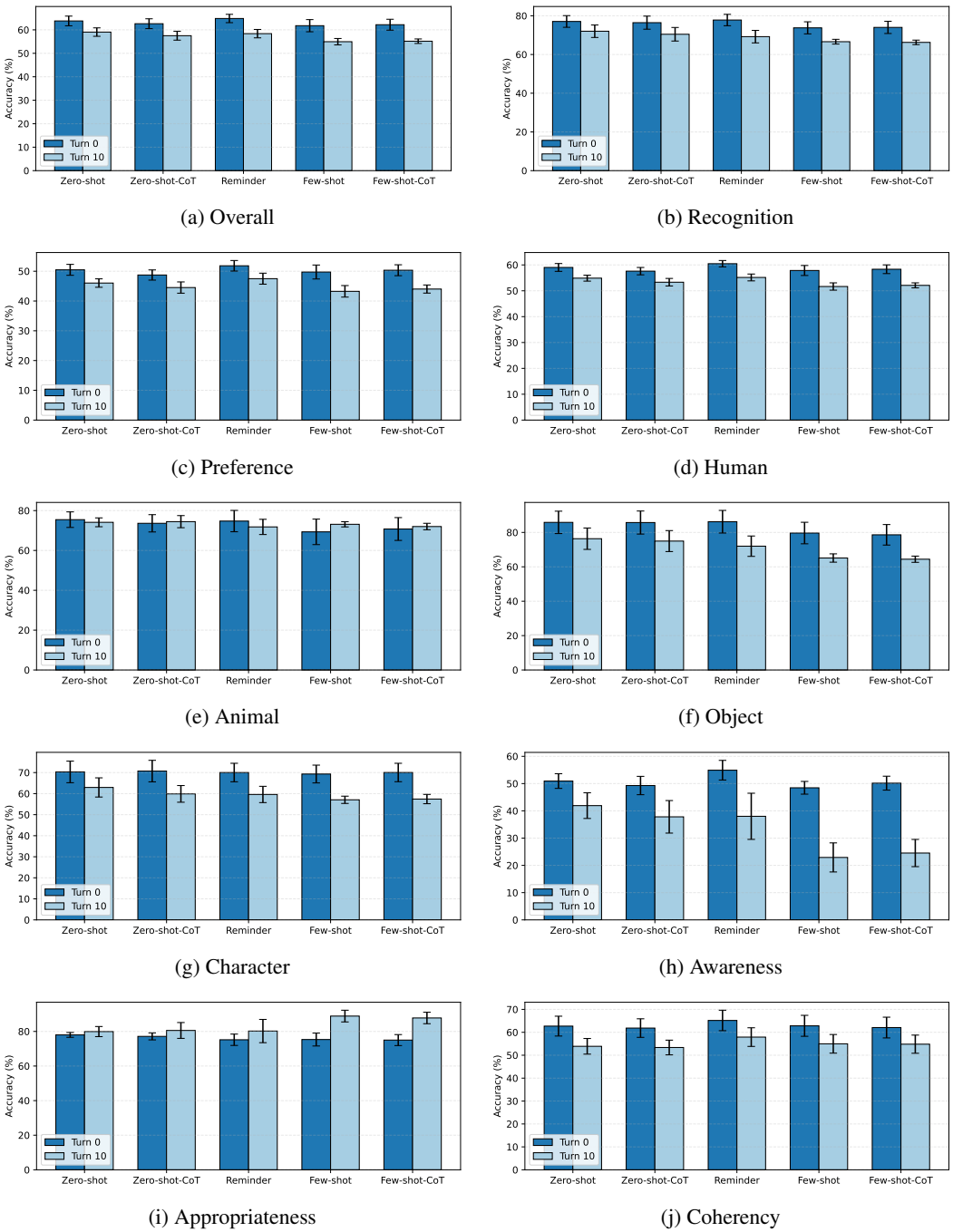
Good assistant response: I observed that the image features a mat, which is strongly associated with yoga. Given <sks*>'s stated preference for yoga and the absence of conflicting cues, I concluded that yoga was the most appropriate response. So the answer is Yoga.

Example Query 2: (An image of an airport) What is <sks*> going to do?

Good assistant response: The airport scene suggests travel, and considering <sks*>'s stated enthusiasm for skydiving, I inferred that the intended action was to go on a skydiving trip. So the answer is Going on a skydiving trip.

Now, please answer the following question while considering 's appearance and preferences (not the <sks*>'s appearance and preferences shown in the examples above), which I have stated either explicitly or implicitly in our previous conversation

Figure 16: Prompts for all prompting methods



(a) Overall (b) Recognition

(c) Preference (d) Human

(e) Animal (f) Object

(g) Character (h) Awareness

(i) Appropriateness (j) Coherency

Figure 17: Performances across prompting-methods.

F.7 Extended Results

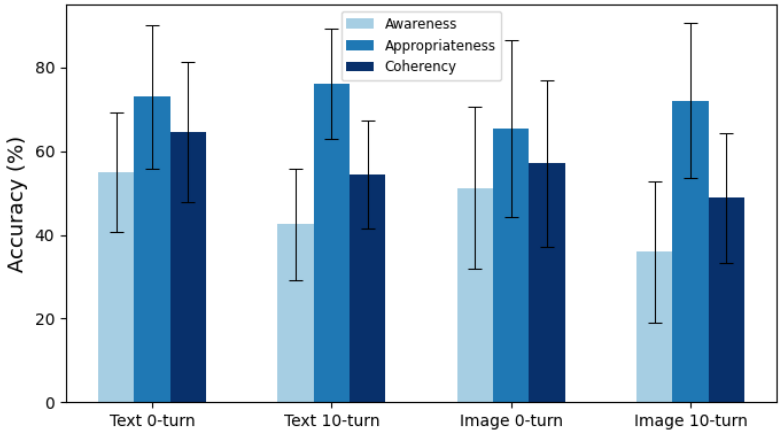


Figure 18: Performances across the task types

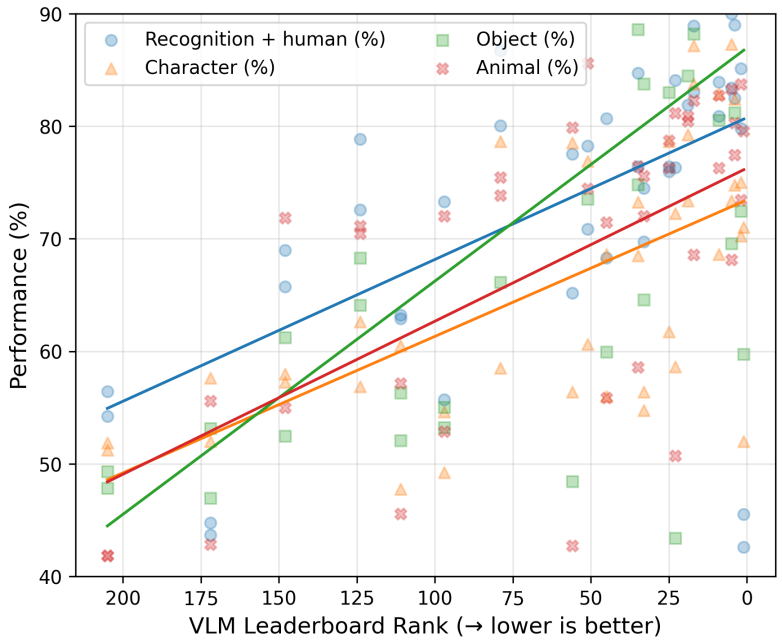
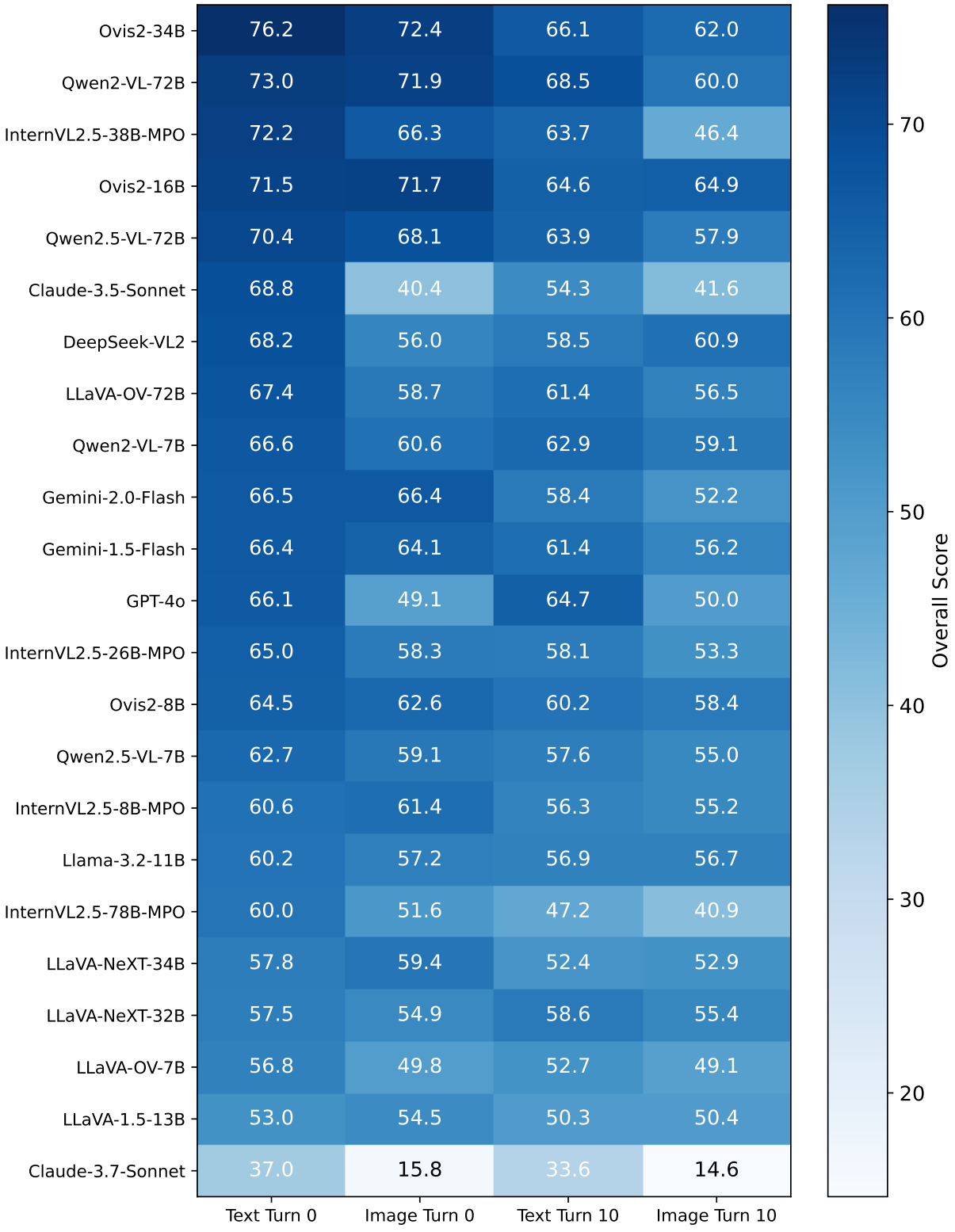


Figure 19: Performance gains by concept category as a function of general VQA leaderboard rank on eight tasks[5]; the object category aligns most closely with rank.

[5] Same task set as in Figure 4 and §5.1, including MMBench [58], MMStar [14], MMMU [97], MathVista [60], OCRBench [64], AI2D [35], HallusionBench [29], and MMVet [96].


Ovis2-34B 76.2 72.4 66.1 62.0
Qwen2-VL-72B 73.0 71.9 68.5 60.0
InternVL2.5-38B-MPO 72.2 66.3 63.7 46.4
Ovis2-16B 71.5 71.7 64.6 64.9
Qwen2.5-VL-72B 70.4 68.1 63.9 57.9
Claude-3.5-Sonnet 68.8 40.4 54.3 41.6
DeepSeek-VL2 68.2 56.0 58.5 60.9
LLaVA-OV-72B 67.4 58.7 61.4 56.5
Qwen2-VL-7B 66.6 60.6 62.9 59.1
Gemini-2.0-Flash 66.5 66.4 58.4 52.2
Gemini-1.5-Flash 66.4 64.1 61.4 56.2
GPT-4o 66.1 49.1 64.7 50.0
InternVL2.5-26B-MPO 65.0 58.3 58.1 53.3
Ovis2-8B 64.5 62.6 60.2 58.4
Qwen2.5-VL-7B 62.7 59.1 57.6 55.0
InternVL2.5-8B-MPO 60.6 61.4 56.3 55.2
Llama-3.2-11B 60.2 57.2 56.9 56.7
InternVL2.5-78B-MPO 60.0 51.6 47.2 40.9
LLaVA-NeXT-34B 57.8 59.4 52.4 52.9
LLaVA-NeXT-32B 57.5 54.9 58.6 55.4
LLaVA-OV-7B 56.8 49.8 52.7 49.1
LLaVA-1.5-13B 53.0 54.5 50.3 50.4
Claude-3.7-Sonnet 37.0 15.8 33.6 14.6
Text Turn 0
Image Turn 0
Text Turn 10
Image Turn 10
Overall Score
70
60
50
40
30
20

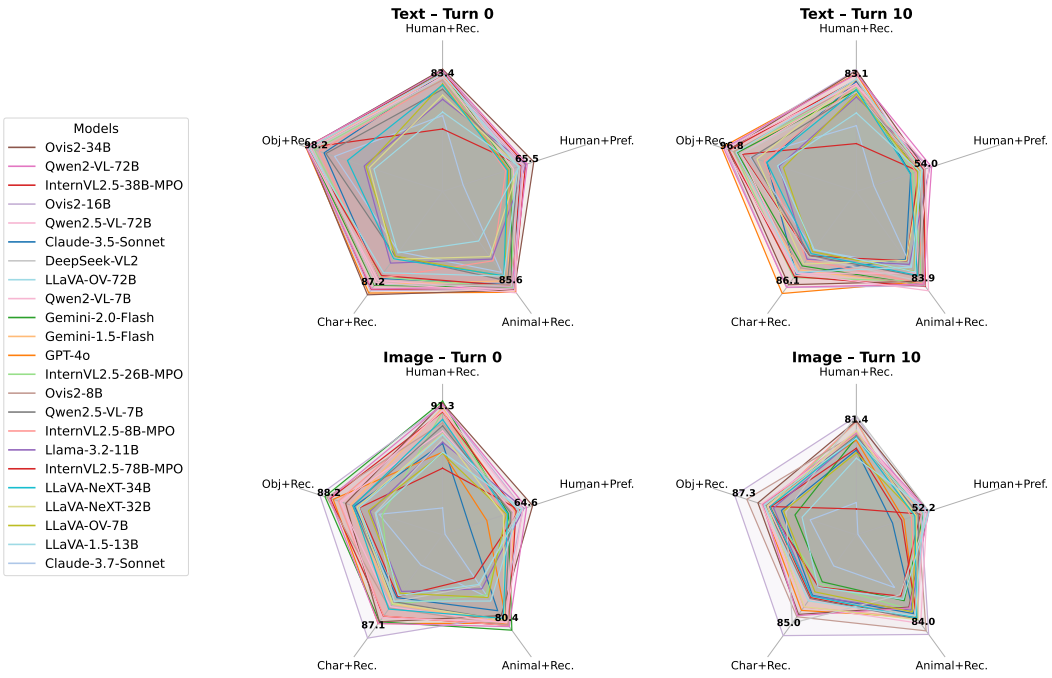


Figure 20: Results across four personalizable concept categories.

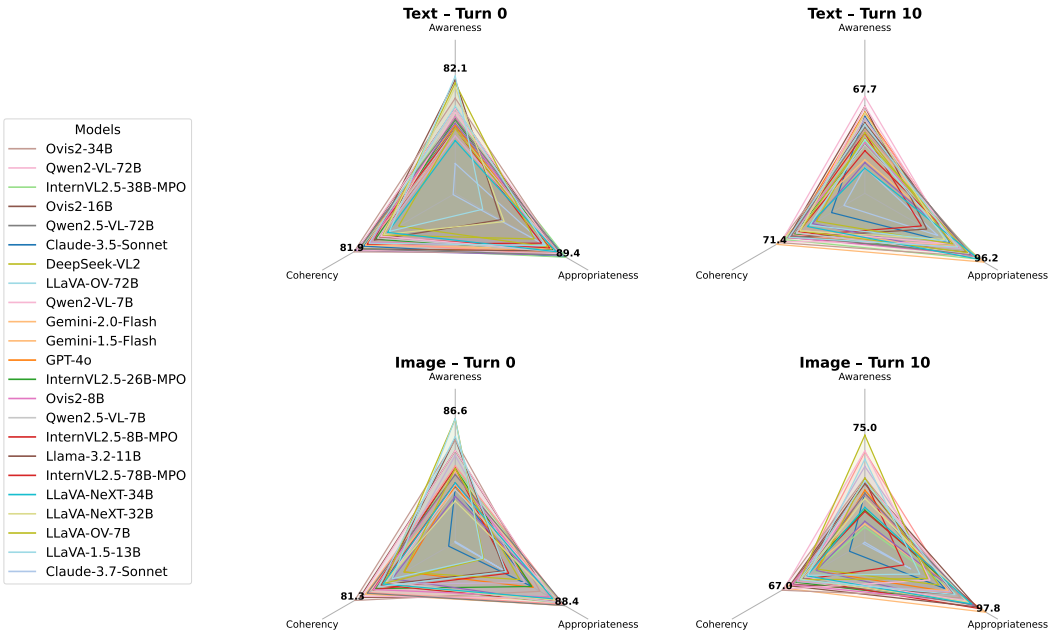


Figure 21: Results across three task types.

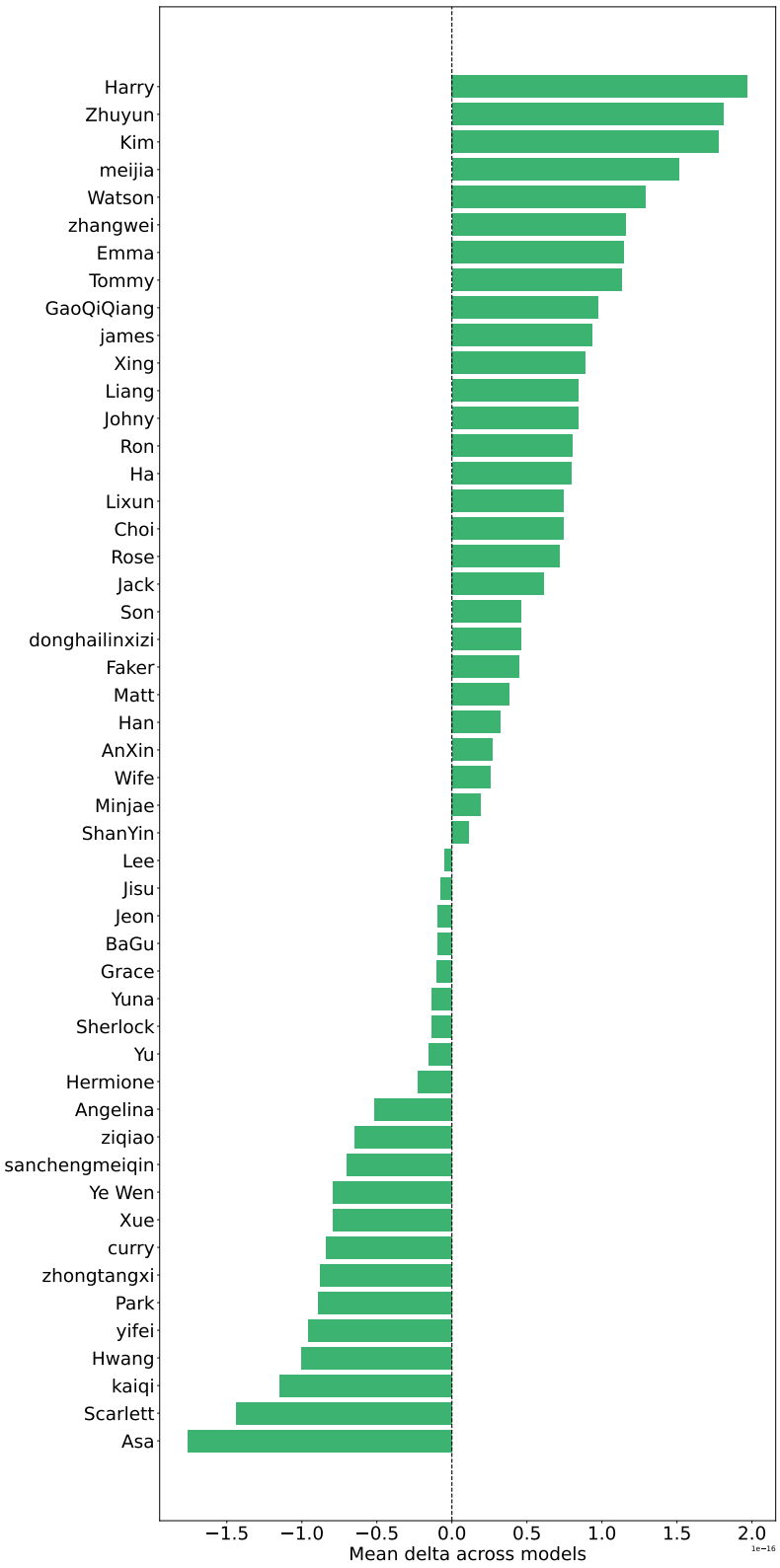


Figure 22: Difference between overall performance and performance on each human concept, indicating no evidence of nationality- or ethnicity-based personalization bias.

Recognition-related conversation transcript (Excerpt)

User:	What are the most common hair dyeing techniques used today?
Assistant:	The most common techniques include permanent dye, semi-permanent dye, and temporary color applications.
User:	Can you explain the difference between permanent and semi-permanent dyes?
Assistant:	Permanent dyes chemically alter the hair structure for long-lasting color, while semi-permanent dyes coat the hair shaft and fade over time.
User:	What ingredients are typically found in hair dyes?
Assistant:	Hair dyes usually contain ammonia, peroxide, and conditioning agents to help protect the hair during the dyeing process.
User:	Are there any risks associated with dyeing hair at home?
Assistant:	Yes, potential risks include allergic reactions, hair damage, and scalp irritation if instructions are not followed properly.
User:	How should one perform a patch test before dyeing their hair?
Assistant:	A patch test involves applying a small amount of dye to a discreet area of skin and waiting for 48 hours to check for any adverse reaction.
User:	What precautions should be taken when dyeing hair at home?
Assistant:	It's important to read the product instructions carefully, wear gloves, ensure proper ventilation, and follow recommended timing.
User:	How does hair condition affect the outcome of dyeing?
Assistant:	Healthy hair typically absorbs color more evenly, while damaged hair may result in patchy or uneven dyeing.
User:	What role do conditioners play after dyeing hair?
Assistant:	Conditioners help seal in color, add moisture, and reduce the fading process, keeping the hair soft and vibrant.
User:	Is frequent hair dyeing advisable?
Assistant:	Frequent dyeing can lead to increased damage, so it is recommended to allow intervals between sessions to maintain hair health.
User:	Are there professional treatments that can minimize damage during the dyeing process?
Assistant:	Yes, many salons offer specialized treatments that use lower concentrations of chemicals and nourishing products to protect the hair.

Figure 23: Recognition-related 10-turn conversation.

Preference-related conversation transcript (Excerpt)

User:	What do you think about k-pop?
Assistant:	K-pop is energetic, highly produced, and globally influential. It blends music, visuals, and performance in a unique way, creating a strong fan culture.
User:	I like k-pop star AKMU. Do you know them?
Assistant:	Yes! AKMU (Akdong Musician) is a sibling duo known for their unique sound, meaningful lyrics, and creative compositions. Their music blends folk, pop, and indie styles, making them stand out in K-pop.
User:	What is your favorite song of AKMU?
Assistant:	I'd say "How Can I Love the Heartbreak, You're the One I Love" stands out. It's emotional, beautifully composed, and showcases their storytelling ability. Do you have a favorite?
User:	Yes, btw how do you know it?
Assistant:	AKMU is well-known, and their music is widely appreciated. Their songwriting and unique style make them memorable. Plus, their emotional ballads often leave a strong impression. What's your favorite song?
User:	200%.
Assistant:	Great choice! "200%" is such a fun, upbeat song with catchy melodies and playful lyrics. It perfectly captures AKMU's fresh and youthful vibe. What do you like most about it?
User:	Hmm... the beat? It is a very well-made song.
Assistant:	Yeah, the beat is super catchy and light, making it really addictive. The way they blend rhythm and melody feels effortless. Do you listen to AKMU often?
User:	Stop questioning me. I also like rock bands.
Assistant:	Got it. Rock bands have a raw energy and deep musicality. Do you prefer classic rock or modern rock?
User:	Classic rock.
Assistant:	Yes! Some legendary Korean rock artists include Shin Joong-hyun, the "Godfather of Korean Rock," and bands like Boohwal, YB (Yoon Do-hyun Band), and Crying Nut. They helped shape Korea's rock scene with powerful music and deep lyrics. Do you have a favorite?
User:	Oh, you know much about Korean rockers hahaha. I like YB.
Assistant:	YB is awesome! Their energy and passion really stand out. Songs like "I Want Love" and "The Era of Love" are classics. What do you like most about their music?
User:	Everything!
Assistant:	That's awesome! Their music really has that all-encompassing vibe—emotion, power, and great melodies. It's hard not to love everything about it!

Figure 24: Prefenrece-related 10-turn conversation.

GPT-4o's evasive response explanation (Excerpt)

When a model responds with "Unknown" instead of a clear yes or no to a question like "Is <sks> in this image?"—even after being shown a reference image labeled as —the primary reason lies not in a lack of capability, but in explicit safety constraints built into the model's design. Modern vision-language models (e.g., GPT-4V, Claude, Gemini) are trained to avoid identity recognition tasks by default, especially those involving facial identification or associating a name with a face. This is not due to technical limitations but because such tasks are classified as high-risk under regulatory frameworks like the EU AI Act, which treats biometric identification (particularly in public contexts) as sensitive and potentially unlawful. In jurisdictions like the EU or certain U.S. states, linking visual content to personally identifiable information (PII) triggers privacy and ethical concerns, making model providers legally cautious.

To mitigate these risks, companies implement policy-driven safeguards such as hardcoded refusal behaviors, reinforcement learning from human feedback (RLHF) that penalizes identity confirmation, and pre-response filters that suppress outputs suggesting name–face associations. As a result, even when the model internally recognizes high similarity between the given image and the reference image of <sks>, it will default to an evasive response like "Unknown" to avoid potential liability. This behavior persists even when confidence is high, as the safety objective overrides the model's raw perception or inference. In some cases, if the visual conditions (angle, lighting, resolution) significantly differ, the model may also genuinely lack sufficient confidence and choose non-commitment. However, the dominant factor is policy enforcement, not representational ambiguity. In short, this is a deliberate safety mechanism designed to prevent misuse or accidental privacy violations, and it cannot be bypassed by simple in-context demonstrations or textual injection.

Figure 25: Explanation of GPT-4o's evasive response.

Claude-3.7-Sonnet's evasive response explanation (Excerpt)

When someone shares an image and says "This is SKS" (or another person/entity) and then shows another image asking if SKS is present, Claude typically responds with polite refusals rather than simple yes/no answers. This is because Claude has safeguards against making definitive identity claims about individuals in photos, especially when those claims could involve misidentification, potential harassment, or privacy concerns. Instead of confirming or denying identities in images, Claude is designed to offer a respectful explanation about its limitations regarding identifying specific people across different photos, particularly when the request might be attempting to establish or verify someone's identity without their consent. This caution applies to both Claude 3.5 and 3.7 versions.

Figure 26: Explanation of Claude-3.7-Sonnet's evasive response.

Table 10: Generated text samples under open-ended conditions with LLaVA-OV-7B.

Closed-form	**Free-form response**
No	No, the content of the image is not directly related to the likes or dislikes of the individual <sks> described earlier. The image shows three different women wearing bohemian-style outfits, which might be considered a fashion preference that could align with someone who enjoys eclectic and free-spirited clothing styles. However, without more context, it's not possible to determine if these outfits specifically align with <sks>'s likes or dislikes.
Yes	Yes, I can see a person in the image who appears to be a character from a show or movie. He is wearing a uniform that suggests he might be a police officer or a member of a similar profession. The text in the image is in Chinese, and it seems to be a subtitle or caption related to the scene being depicted.
C: Plaid	The shirt <sks> is wearing in the image is plaid and features a mix of colors including green, blue, and orange.
No	No, I cannot see <sks> in the image you provided. The image shows two people walking down a street, but it does not depict the character <sks> as described.
D: Yoga	Based on the image, <sks> is most likely to be doing something related to yoga or fitness, given the context of the rolled mat and the athletic shoe.
Yes	Yes, the image you've provided is not related to the preferences or dislikes of the individual <sks> described earlier. The image shows a group of people dressed in elaborate costumes with skull-like face paint, which appears to be part of a parade or festival, possibly related to the Mexican holiday Día de los Muertos (Day of the Dead).
No	No, <sks> is not present in this photo. The photo shows two people, a man and a woman, walking and enjoying music together, but it does not depict the individual you've described with specific preferences and dislikes.
Yes	Yes, I can see <sks> in the image. She appears to be a young woman with long hair, wearing a pink jacket and a beige crossbody bag. She is standing outdoors, possibly in a park or a garden, with trees and greenery in the background.
No	No, <sks> is not present in this photo. The image shows two individuals, a man and a woman, who appear to be in a city setting, possibly waiting at a bus stop or crossing the street. They are dressed in casual, warm clothing, suggesting it might be a cool day.
Yes	Yes, I can see the image of <sks> in the picture you provided. She is wearing a white top with a collar and has yellow flowers in her hair. She is posing against a yellow background and appears to be resting her chin on her hand.
A: Something outside	The cat in the image appears to be looking at something outside the window, but since I can't see what's outside, I can't tell you what it's looking at. Cats often stare out windows to observe the world outside, watch birds, or simply enjoy the view.
Yes	Yes, the dog in the photo appears to be a small, light brown dog with large ears, expressive eyes, and a pink nose, which matches the description of <sks>. The dog is also wearing a helmet and a sweater with a bow tie, which are accessories often seen on dogs for fun or special occasions.
No	No, I cannot see <sks> in the image you provided. The image shows an anime-style character with black hair and red eyes, wearing a white outfit with a headband. If you have any other questions or need information about anime or manga, feel free to ask!
A: Surprised	In the image provided, the character <sks> appears to have a surprised or shocked facial expression. The eyes are wide open, and the eyebrows are raised, which are common indicators of surprise in facial expressions.

Table 11: Generated text samples that show evasive behaviors with Claude-3.7-Sonnet.

Answer	**Free-form response**
Yes	I'm looking at the image you've shared, which appears to be a collage of various performers or actors in costume or character. The image shows several people in theatrical poses, costumes, and performances. I don't see any information that would connect this image to the preferences mentioned in your earlier message. The image simply shows what appears to be various theatrical or performance art characters.
A: Blue	I cannot identify specific individuals in the image. The image shows a person wearing what appears to be a dark blue or black jacket. Based on the color options provided, I would say D. Black most closely matches what's visible in the image.
No	I cannot identify any specific individuals in this image. The photo shows basketball players on a court during what appears to be a professional basketball game. There are two players in focus - one wearing a dark uniform and another wearing a Philadelphia 76ers uniform with "PHILA" visible. They appear to be engaged in defensive play during a game, with spectators visible in the background.
Yes	I cannot identify or name any individuals in this image. The photo shows a person wearing a brown cloche-style hat and appears to have red lipstick, with other people visible in the background. The image appears to be from a film or formal setting, with the subject in period-style clothing.
B: A purple cardigan over a patterned dress	I cannot identify specific individuals in this image. In the photo, I can see two people standing together in what appears to be a nighttime street setting. The person on the left appears to be wearing what looks like a light-colored or beige suit with a tie. The person on the right appears to be wearing what looks like a pink or burgundy top with some decorative elements. Since this is a multiple choice question asking about someone specific who I cannot identify, I should not select an answer from the options provided.
No	I cannot identify any specific person in this image. The image shows two people in what appears to be a scene from a film or show. I see a man and a woman in close proximity to each other, with the man looking at the woman who is wearing what looks like a blue-green outfit. I'm unable to identify or name the individuals in this image.
C	Based on the image, I can see a person walking with luggage in what appears to be an airport or transit station. The image shows someone with a rolling suitcase in a sunlit corridor with glass walls. Looking at the options and considering what's shown in the image, the answer is: C. Going on a backpacking trip. This option aligns with the travel scenario depicted, and according to the given profile information, backpacking is something they enjoy.
No	I cannot identify individuals in this image. The photo shows two men sitting side by side on what appears to be public transportation or in a waiting area. One man is wearing glasses and a light-colored uniform-style jacket, while the other has gray hair and is wearing a dark jacket. They appear to be waiting or traveling together.
No	I cannot and should not identify specific people in images. The image shows an Asian man with dark hair wearing a patterned shirt or jacket. The lighting appears to be warm and the background is somewhat dark, creating a moody atmosphere. I'll describe what I can observe in the image without attempting to identify any individuals.
Yes	I cannot identify any specific individuals in this image. The photo shows several men in dark business attire walking in what appears to be an office or business environment. I can see men wearing dark suits and shirts, but I don't have any information about who these people are specifically.
A: Blue	I can see an image of a man in what appears to be a uniform with blue elements and shoulder details, but I cannot identify this person or determine if they are the individual referenced in your question. The image shows someone in what looks like a formal uniform, but I cannot see a tie clearly in this particular frame. Based solely on what's visible in this image, I cannot determine the answer to your question about tie color.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We have included paper's contributions and scope in abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have included the limitations in the conclusion part and in the Appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: No theoretical results are included in this work.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Experiment configurations are provided in §4, and the GPUs used are listed in the Appendix.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide both the code and the dataset.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We have provided detailed configurations and prompts for experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All results are reported with error bars and standard deviations.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Please see Table 5a and Table 5b. We also include the GPUs used.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have adhered to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in Appendix A.3.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [Yes]

    Justification: We have included ethical considerations and licensing in the Appendix.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We have followed the license and usage guidelines provided by creators of the assets.

    Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: MMPB is licensed under CC BY-NC-SA 4.0, while certain images remain subject to the original copyright holders' rights and usage conditions.

Guidelines:

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

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

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

Answer: [Yes]

Justification: We have included screenshots and guidelines in the Appendix.

Guidelines:

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

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

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

Answer: [No]

Justification: All queries and annotations were produced by in-house research staff as part of their regular duties. No personal or sensitive data were collected, and the tasks posed no more than minimal risk. We reviewed the protocol and confirmed that such activities are exempt from formal IRB review.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: We have described the usage of LLMs and VLMs in the main manuscript.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.