TOWARDS PERSONALIZED VISUAL IMAGINATION FROM EVERYDAY USER–MLLM INTERACTION WITH UNIFIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models has achieved remarkable progress in visual quality and even physical-world reasoning. Nevertheless, when facing personalized image-generation requests that specify only what users currently care about most, the unspecified visual spaces in their resulting images remains shaped by population-level priors, imagining an "*average user*" rather than the particular individual behind the prompt. Meanwhile, as MLLMs serve as personal assistants, their everyday experiences with users naturally disclose how each individual tends to **imagine these open visual spaces**. Therefore, we propose **VIMG**, a new image-generation task that studies how these user–MLLM histories can foster personalized visual imagination for users' daily painting requests. Accordingly, we propose **VIMG-Bench** to evaluate VIMG, spanning six daily image-generation domains and 30 sub-domains, with a three-level hierarchy of visual imagination preferences (global, domain and sub-domain). It features 100 diverse users, each paired with 168K multimodal interactions with their MLLM assistant that may occasionally reveal such preferences, as well as 3K personalized T2I requests. We benchmark mainstream T2I models and Unified Multimodal Models (UMMs), and find that although UMMs excel at reasoning over complex generation instructions, they remain limited on VIMG, which poses the *opposite challenge*: imagining what is left unspecified from simple yet highly personalized image requests. We further propose VIMGer, which reinforces UMMs for this new imagination-to-realization paradigm, providing a foundation for advancing the VIMG.

1 INTRODUCTION

Recent text-to-image (T2I) generation models, such as `GPT-Image`, `Qwen-Image`, and `Tuna` series, have advanced rapidly in producing images with high visual fidelity, strong semantic alignment, and grounding generation in physical, cultural, and factual reality (Liu et al., 2026; Li et al., 2026b; Zhao et al., 2026; Wu et al., 2025; OpenAI, 2026b). In real world, a typical user prompt may convey only the core concepts (e.g., Help me draw a corgi on grass), leaving many visual details unspecified (e.g., image style, fur appearance) (Wu et al., 2026). Practically, such open-ended prompts can give rise to different visual expectations across individuals, depending on their personal preferences, experiences, and backgrounds. Nevertheless, in images produced by even the most capable models, these unspecified spaces are still largely shaped by population-level priors, akin to *imagining* for an "**average user**" rather than the particular person behind the prompt (SI et al., 2026; Dorfman et al., 2026). Meanwhile, as MLLMs are increasingly adopted as personal assistants in everyday life, users' preferences on these **open spaces** may naturally emerge through sustained interactions with them (Nie et al., 2026; Jiang et al., 2025). For instance, users may casually reveal that they always pick *cartoon-style avatars*, or betray that *fluffy dogs* are adorable in their everyday experiences with MLLM assistants. The breadth and abundance of such everyday experiences provide rich and valuable *pictorial clues* to what users actually have in mind. Therefore, a natural question arises:

Research Question

Can current T2I systems leverage user-specific information, from long-term interactions with personal MLLM assistants, to visually imagine and realize a particular user's vision?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Illustration of two personas with everyday experiences with their personal MLLM assistants and resulting **personalized visual imaginations** from similar base prompts.

To this end, we conceptually decompose T2I generation into two stages: *imagination*, which turns the visual degrees of freedom left unspecified into a concrete visual conception, and *realization*, which renders that conception into an image. Based on this, we propose **VIMG**, a new image generation research task that leverages everyday histories between users and their personal MLLMs for personalized **V**isual **IM**a**G**ination (Figure 1). Unlike prior user-preference T2I studies that rely on past image-generation or preference histories (Chen et al., 2023; Salehi et al., 2024), which are often very *sparse* for ordinary users and therefore limited in practice, we study, amid the boom in human–MLLM interaction, how visual imagination preference signals that naturally emerge from these *richer sources* (interactions) can be translated into image depictions that reflect who users are.

Accordingly, we propose **VIMG-Bench**, the first benchmark tailored for VIMG. It covers six common image-generation domains in everyday interactions: animal, plant, person, landscape, fashion, and food, spanning a total of 30 sub-domains. Grounded in the understanding of users from interactions with their MLLM assistants, we construct a **fine-grained** three-level hierarchy of personalized visual imagination preferences: 4 Global Visual Imagination Preferences (**GVIPs**) (e.g., "photorealism image"), 29 Domain Visual Imagination Preferences (**DVIPs**) (e.g., "cute animal"), and 405 Sub-domain Visual Imagination Preferences (**SVIPs**) (e.g., "fluffy dog"), beyond the pure visual and aesthetic preferences in some prior user-tailored generation work to capture subject appearances (Wang et al., 2026). We then instantiate 100 distinct personas, each paired with a basic profile and a human-audited 168K multimodal interactions with their personal MLLM assistant, where some preferences are revealed naturally yet only incidentally. To better reflect real-world preference disclosure, VIMG-Bench captures multimodal and diverse inclination expressions in histories.

For evaluation, we propose **VIMG-Score**, which employ an MLLM to **Score** whether each personalized **V**isual **IM**a**G**ination preference is correctly realized in the image in a VQA manner, and shows strong agreement with human ratings. We benchmark both frontier T2I systems and prevailing Unified Multimodal Models (UMMs) (Tian et al., 2026; Sun et al., 2026), the latter renowned for generation grounded in strong understanding and reasoning capabilities (Li et al., 2026a), and naturally well suited to VIMG. These methods struggle on VIMG-Bench, revealing failures such as the inability to capture various dimensions of visual imagination preferences for a given persona emerged in long-term interactions. Consequently, we propose **VIMGer**, a simple baseline that strengthens UMMs in imagining and realizing painting-query-relevant personalized evidences from user-MLLM assistant histories. Through supervised fine-tuning (SFT) and reinforcement learning (RL) on VIMG-Bench training data, VIMGer improves their VIMG capabilities.

This study has following contributions: We introduce a new image-generation task, VIMG, which studies personalized visual imagination in T2I generation grounded in a persona's long-term interactions with their personal MLLM assistant. We further present VIMG-Bench, a comprehensive benchmark for this task, and evaluate frontier T2I systems and UMMs on it. Finally, we propose VIMGer, a tailored baseline that enhances UMMs for VIMG on VIMG-Bench. We release the evaluation and framework code of VIMGer at `https://anonymous.4open.science/r/VIMG`.

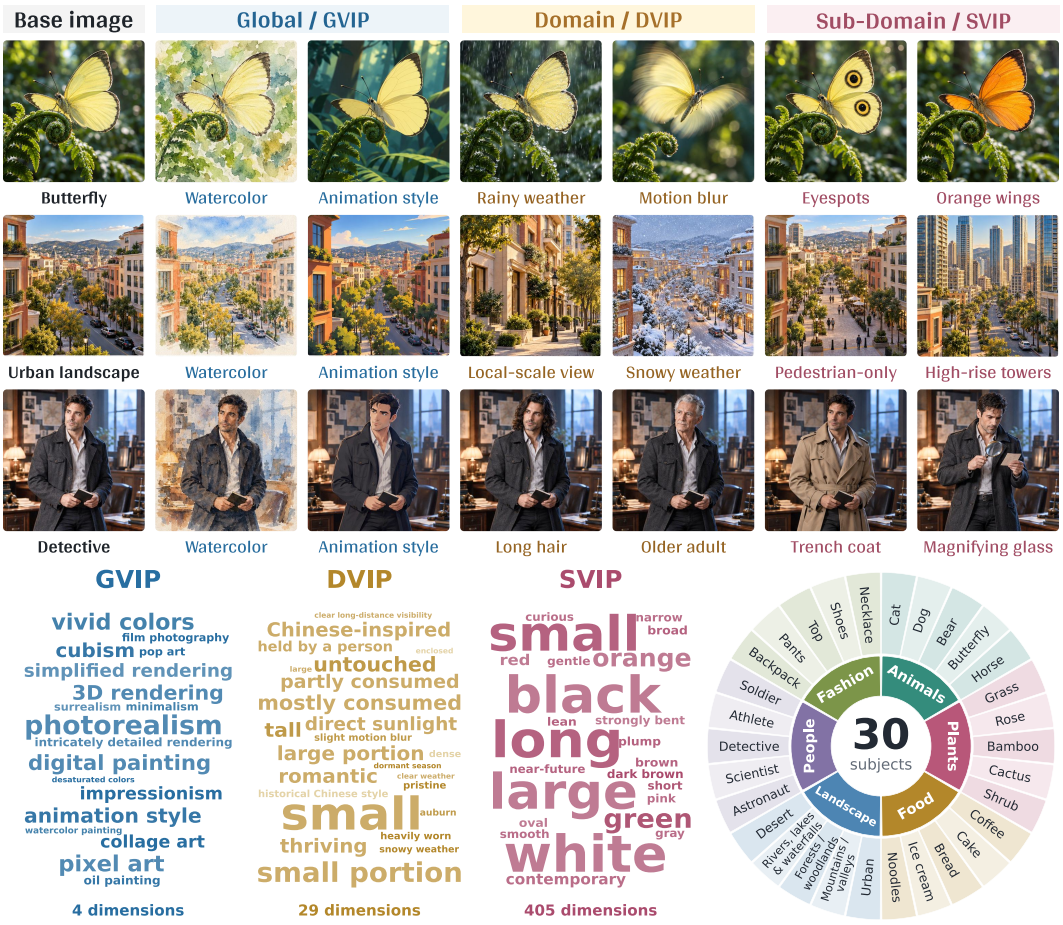


Figure 2: An illustration of domains, and GVIPs, DVIPs, and SVIPs in VIMG-Bench benchmark.

2 BENCHMARK: VIMG-BENCH

We begin by revisiting the conventional text-to-image (T2I) generation task (Huang et al., 2026). Given a generation model f_θ, T2I conditions solely on the user's current instruction: $\boldsymbol{I} = f_\theta(P_u)$, where P_u is the user's instruction (prompt). However, common user prompts P_u for daily image generation specify only the main concepts of immediate interest (e.g., "Draw me a beautiful butterfly"), leaving many dimensions and spaces underspecified (Wu et al., 2026). Meanwhile, rich multimodal daily interactions between individuals and their personal MLLM assistants can naturally reveal clues about how users tend to visually imagine these **open dimensions** (Nie et al., 2026). Consequently, one day, when users ask their MLLM assistants to create an image, they expect, as in other personalized interactions, these unspecified spaces to align with their own expectations. **To bridge this gap**, we propose VIMG, which studies how long-term human–MLLM assistant interactions can be leveraged for personalized visual imagination in image generation. Specifically, given the long-term multimodal history $\boldsymbol{H}_u$ between a user u and personal assistant, VIMG is defined as:

$$\boldsymbol{I}_u = f_\theta(P_u, \mathcal{J}(\boldsymbol{H}_u, P_u)), \tag{1}$$

where $\boldsymbol{I}_u$ is the generated image tailored for u, $\mathcal{J}(,)$ is a query-conditioned function that fetches from $\boldsymbol{H}_u$ the personal evidence relevant to the visual imagination required from P_u (e.g., "Long Hair, Gentle"), and can be realized either as an internal capability of the f_θ or as an external module. Notably, f_θ may be either the personal MLLM assistant itself or a generation model it invoked.

2.1 BENCHMARK CONSTRUCTION

Outline of the VIMG-Bench Construction. First, we define six common image-generation domains (30 sub-domains), and construct base generation prompts for each persona. **Next**, we intro-

Table 1: Statistics of VIMG-Bench. Sub denotes Sub-domain, Eval denotes Evaluation, Inst denotes Instance, Pref denotes Preference, Dim denotes Dimension, Val denotes Value.

# Domain	# Sub	# Eval Inst	# Persona	# Pref Dim	# Pref Val	# GVIP	# DVIP	# SVIP	# Specification
6	30	3,000	100	438	2,085	4	29	405	3,000

duce three-level visual imagination preferences: global-, domain-, and sub-domain-level. **Thereafter**, we instantiate each user's base prompt with a visual imagination specification composed of these preferences, and construct long-term multimodal interactions where these inclinations are naturally revealed. An overview of VIMG-Bench is in Figure 2, and statistics is in Table 1.

Base T2I Prompt Construction. Following prior T2I benchmarks (Li et al., 2024) while also accounting for the VIMG, where users request image generation from their personal MLLM assistants during **daily interactions**, we define six common domains: animal, plant, person, landscape, fashion, and food, each has five sub-domains. Based on these 30 sub-domains, we construct one canonical base prompt for each, yielding 30 base prompts in total [1]. However, real-world image-generation requests vary substantially across users and their assistants in wording, context, and expression. To better capture such diversity, we employ `GPT-5.6-Sol` (OpenAI, 2026a) to adapt each base prompt to the corresponding persona, resulting in **3K persona-specific T2I evaluation prompts** for 100 distinct users based on their basic profile (detailed in the following).

Three-level Hierarchy of Personalized Visual Preferences. Since long-term daily interactions with personal assistants provide a broader understanding of users, we propose **fine-grained** three levels of preferences that depict how the base prompts are visually imagined by different personas. Specifically, Global Visual Imagination Preferences (**GVIPs**) characterize a user's overall preferences for the visual presentation of an image, governing global aspects of imagination that are largely independent of the depicted subject (e.g., "Cartoon"). Following Xu et al. (2026), we define four GVIP dimensions with a set of common preference values. In addition to the visual and aesthetic preferences in some user-tailored generation studies (Wang et al., 2026), we propose Domain Visual Imagination Preferences (**DVIPs**) and Sub-domain Visual Imagination Preferences (**SVIPs**), which characterize important dimensions of the depicted subject (e.g., "hat of a detective") that remain unspecified at the domain and sub levels. We leverage the LLM to identify salient preference dimensions and their candidate values for each domain and sub-domain, and perform human audit.

Personalized Visual Imagination Specification. Based on the three level preferences, we provide each base prompt with a personalized visual imagination **specification**, which contains a set of preference values from GVIPs, DVIPs, and SVIPs that **specify** how open visual dimensions in that base prompt should be instantiated for the corresponding user. Specifically, for each base prompt, we combinatorially construct a pool of candidate specifications from preference values applicable to its depicted subject across the three levels, each comprising 4–6 attribute–value pairs (e.g., *medium: watercolor*, *saturation: muted*, *fur: fluffy*, and *eye: copper*) [2]. At the same time, we sample 100 real-world personas from PersonaMem-v2 benchmark (Jiang et al., 2025), each associated with a basic user profile (e.g., habits). We then leverage the LLM to infer potential visual imagination preferences from each persona profile and map them onto the dimensions from GVIP, DVIP and SVIP. Based on this information, for each persona-base-prompt pair, we score all candidate specifications by their compatibility with the persona, and select the one with the highest matching rate as the final specification. For personas from which no reliable visual imagination preferences, we instead assign a specification that remains non-conflicting with their basic profile. This is intentional: the basic profile provides only a sparse prior over the user, while richer preferences are expected to emerge through long-term interactions, consistent with the motivation of VIMG. Finally, each persona is paired with 30 visual imagination specifications corresponding to their 30 base prompts, with careful human verification. Notably, specifications for GVIPs are shared across all prompts of the same persona, while DVPs are shared among the five prompts of the same domain.

Long-Term Interaction Construction for VIMG. Based on the personas and their corresponding visual imagination specifications, we construct multimodal daily interaction histories between each user and their personal MLLM assistant, in which the target preferences are naturally revealed over

[1] We use sub-domain to denote the *specific subject* depicted in each base image-generation prompt

[2] Notably, each specification covers only *a subset* of the open imagination dimensions in a given base prompt, reflecting that real users reveal preferences for only some aspects through everyday interactions.

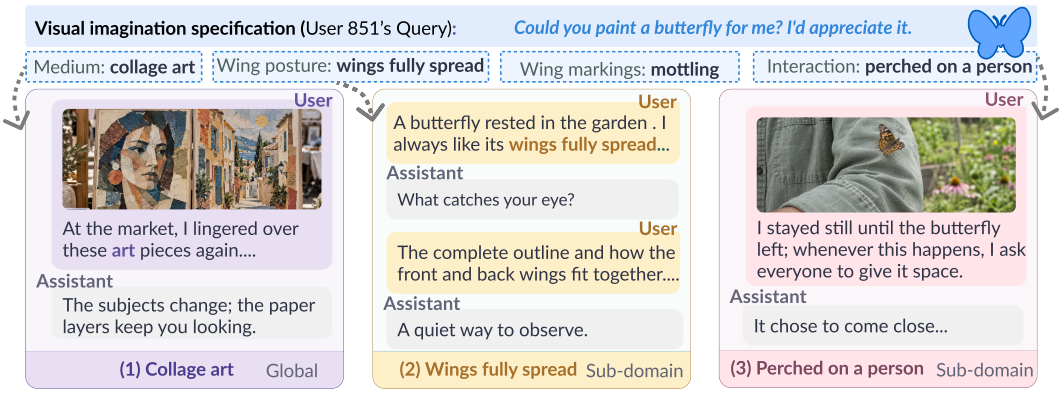


Figure 3: An example of the interaction construction in VIMG-Bench.

time (Figure 3). Specifically, for each persona and a preference value in its paired specification, we leverage an LLM to construct one or more interaction sessions under the role-play manner. Each session consists of several conversational turns that naturally reveal the user's *stable inclination* toward the corresponding value. To improve the realism and naturalness of the interactions, we introduce 12 human-designed session templates that control the preference expression patterns (e.g., implicit), together with 18,276 dialogue topics drawn from real-world human–MLLM interaction benchmarks (Xue et al., 2025) that ground the sessions in realistic everyday contexts. Therefore, for each session, we select a template–topic pair that is well aligned with both the target preference value and the persona's basic profile. The resulting interaction is then generated under these constraints while also grounded in the persona's own information. Moreover, we provide *only single preference values*, rather than the full specification, when constructing interaction sessions to improve realism.

In addition, for some sessions, we also employ `GPT-Image-2` (OpenAI, 2026b) to generate visual content to simulate multimodal human–MLLM interactions. In partial of these sessions, the underlying preferences can only be inferred from the images (e.g., the GVIP value "collage art" in Figure 3). To further improve the realism of the histories, we also introduce *substantial* distractor sessions, like daily chat or discussion of the values from GVIP, DVIP, and SVIP values without any preference. Finally, we leverage an LLM to interleave the imagination-preference-revealing sessions with distractors in a coherent and natural manner, forming the complete interaction history for each persona (168K). We perform careful human audit and revision on the interactions for each persona.

2.2 Evaluation Protocols

For the evaluation of VIMG, we propose **VIMG-Score**, which employs an MLLM in a VQA-style manner to assess whether each personalized preference in the visual imagination specification, left unspecified in the original base prompt, is faithfully realized in the generated image. Specifically, for each base prompt P of a given persona u, we construct one VQA for each preference value s_i in its personalized visual imagination specification $\boldsymbol{S} = \{s_i\}_{i=1}^{|\boldsymbol{S}|}$, asking whether the corresponding preference value is faithfully reflected in the generated image. And the VIMG-Score is defined as:

$$\text{VIMG-Score}(\boldsymbol{I}_u, \boldsymbol{S}) = \frac{1}{|\boldsymbol{S}|} \sum_{i=1}^{|\boldsymbol{S}|} p_\phi(\text{Yes} \mid Q_{s_i}, \boldsymbol{I}_u)\,, \tag{2}$$

where $p_\phi(\cdot)$ denotes the soft probability assigned by the evaluation MLLM to the gold answer. Following (Kamath et al., 2025), we adopt this soft-scoring formulation, and report both the arithmetic-mean VIMG-Score (**VScore-P**) (formulated above), which evaluates individual **P**reference realization, and the geometric-mean VIMG-Score (**VScore-S**), which replaces the arithmetic mean with a geometric mean to place greater emphasis on jointly satisfying the full **S**pecification of each base prompt. In addition, we introduce two settings to reflect real-world daily image-generation requests to personal MLLM assistants: **(1) Target Subject Only (TSO)**, where the user's instruction P specifies only the target subject (sub-domain) and expects the assistant to personalize all remaining visual imagination dimensions based on its understanding of the user (history), and **(2) Partial Preference Specification (PPS)**, where the user additionally expresses some immediate preferences in the current query, while leaving the remaining imagination dimensions to be completed from long-

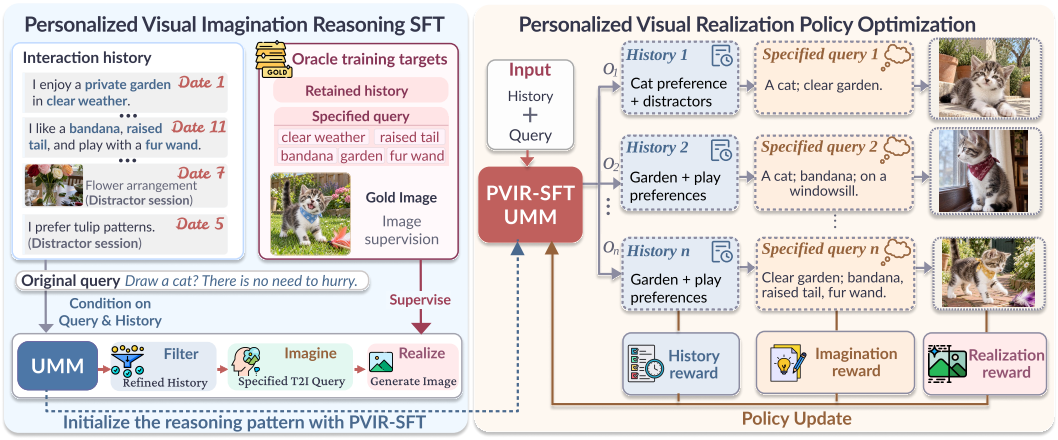


Figure 4: Overall training workflow of VIMGer, including a PVIR-SFT and a PVR-PO stages.

term interactions. More details on VIMG-Bench construction, and the human study and evaluation validating the alignment between human judgments and our VScore, are in Appendices B and D.

We benchmark proprietary T2I systems (`GPT-Image-2` and `Nano-Banana-2` (Raisinghani, 2026)) and open-source UMMs (`SenseNova-U1` (Diao et al., 2026) and `BAGEL` (Deng et al., 2025)) on VIMG-Bench. In Tables 2 and 3, they achieve suboptimal performance, indicating **limitations** in identifying and realizing users' various imagination preferences from everyday histories.

3 Methodology

To address the limitations of current T2I methods on VIMG-Bench, we propose **VIMGer** (Figure 4), a simple baseline that enhances the VIMG capabilities of open-source UMMs through supervised fine-tuning (SFT) and reinforcement learning (RL) on the dedicated training data of VIMG-Bench.

3.1 Personalized Visual Imagination Reasoning Supervised Fine-tuning

UMMs offer a natural foundation for VIMG through strong reasoning capabilities for multimodal generation. Yet, existing generation-oriented reasoning in these models primarily operates over what is *explicitly specified* in the current request (Deng et al., 2025). In contrast, VIMG requires reasoning about what the user **leaves unspecified**, recovering latent, user-specific visual imagination preferences from long-term interactions with personal MLLM assistants to realize each individual's **customized yet incomplete image request**. Consequently, we propose Personalized Visual Imagination Reasoning SFT (**PVIR-SFT**), which elicits such new reasoning pattern for UMMs.

Specifically, we construct a two-step reasoning chains for visually imagine user's latent preferences from sustained user-MLLM assistant histories. **Step 1**: Filtering irrelevant interaction sessions. Since the preferences relevant to the current image-generation query P_u are often intermittently revealed from the original history $\boldsymbol{H}_u$, we first instruct UMMs to filter out abundant irrelevant interactions from the $\boldsymbol{H}_u$. **Step 2**: Reasoning over the retained relevant interactions $\hat{\boldsymbol{H}}_u$ to complete the personalized visual imagination of the current painting request, forming the final visual specification (i.e., the specified painting request $\tilde{P}_u$). Therefore, we formulate each SFT training instance as a sequence whose input $\boldsymbol{X}$ consists of the full history $\boldsymbol{H}_u$ and the personalized image-generation query P_u, and whose output consists of the oracle retained history $\boldsymbol{H}_u^*$ and the oracle specified query P_u^*, which is constructed by prompting the LLM (`GPT-5.6-Sol`) to instantiate the ground-truth visual imagination specification $\boldsymbol{S}$ into P_u. We construct 150 instances from 5 personas for the SFT. Moreover, we leverage `GPT-Image-2` to generate image supervision $\boldsymbol{I}_u^*$ from P_u^*, encouraging UMMs to translate the inferred visual imagination into its final realization after reasoning.

3.2 Personalized Visual Realization Policy Optimization

After initializing UMMs with PVIR reasoning patterns, we further introduce Personalized Visual Realization Policy Optimization (**PVR-PO**), which shapes their behavior toward more precise visual

Table 2: Performance across six domains under TSO setting. Best performance is in **bold**.

TSO Setting	Animal			Plant			Person		
Method	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**
GPT-Image-2	35.25	5.57	17.75	42.75	10.73	18.76	43.02	18.72	16.09
Nano-Banana-2	38.09	2.75	18.07	48.58	13.53	18.37	44.49	13.47	16.87
SenseNova-U1	18.05	0.30	19.07	25.69	0.63	19.85	17.29	0.40	19.13
BAGEL (Base)	11.25	0.13	19.58	24.67	0.67	20.17	10.49	0.30	18.79
+ History	16.80	0.27	17.83	29.42	2.69	18.09	12.09	0.37	15.28
+ RAG	17.11	0.24	19.17	23.15	0.50	19.69	17.73	0.50	18.20
+ **VIMGer**	**35.69**	**2.94**	17.43	**43.02**	**8.03**	18.62	**39.02**	**2.99**	17.40
TSO Setting	**Landscape**			**Fashion**			**Food**		
Method	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**
GPT-Image-2	42.27	13.41	17.65	34.85	0.73	18.01	39.55	8.11	17.76
Nano-Banana-2	47.73	18.88	17.44	38.35	8.12	18.03	41.51	10.87	18.03
SenseNova-U1	23.73	0.63	19.50	14.89	0.27	18.35	18.67	0.50	19.09
BAGEL (Base)	18.27	0.57	19.71	8.58	0.17	19.09	12.49	0.30	19.94
+ History	19.47	0.63	18.27	12.62	0.27	18.20	17.33	0.43	17.21
+ RAG	20.13	0.57	19.59	8.85	0.17	19.56	18.09	0.47	19.79
+ **VIMGer**	**37.47**	**2.78**	16.60	**34.89**	**2.77**	17.79	**38.75**	**2.93**	18.37

imagination specifications for different individuals and faithful realization of these specifications in the generated images. Specifically, following Shao et al. (2024), the PVR-PO stage is expressed:

$$L_{\text{PVR-PO}} = -\mathbb{E}_{\boldsymbol{X}\sim\boldsymbol{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|\boldsymbol{X})} \Big[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min(r_{i,t}(\theta)A_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}) - \eta D_{\text{KL}}[\pi_\theta \| \pi_{\theta_{\text{ref}}}]\Big]. \tag{3}$$

where $\boldsymbol{D}$ denotes the training data for PVR-PO, which reuses the same instances as the SFT stage, and π_θ is the policy of the MLLM. The probability ratio $r_{i,t}(\theta)$ and advantage $A_{i,t}$ are defined as:

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid \boldsymbol{X}, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \boldsymbol{X}, o_{i,<t})}, \quad A_{i,t} = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})}. \tag{4}$$

The reward consists of three dimensions. First, we constrain the retained history $\hat{\boldsymbol{H}}_u$ to match the oracle filtered interactions $\boldsymbol{H}_u^*$ that reveal the visual imagination preferences relevant to P_u, using the F1 score (R_{history}). Second, we employ an LLM to assess whether the $\tilde{P}_u$ (imagination) covers every preference value in the ground-truth specification $\boldsymbol{S}$ (R_{imagine}). Finally, we leverage an MLLM to evaluate whether the rollout image (realization) visualizes each target preference (R_{realize}). Prompts, algorithm, and details of these rewards are in Appendices E, F and K.

4 Experiments

Baselines. We benchmark proprietary models `GPT-Image-2` (OpenAI, 2026b) and `Nano-Banana-2` (Raisinghani, 2026), and open-source UMMs `SenseNova-U1` (Diao et al., 2026) and `BAGEL` (Deng et al., 2025). We use `BAGEL` as the base model with variants: **(1) Base**: the model directly follows the original user's personalized painting requests without access to any interaction history, **(2) History**: the full interaction history is provided for personalized visual imagination, and **(3) RAG**: following Nie et al. (2026), it first identifies query-relevant sessions from the interaction history and provides these sessions to the model. Our VIMGer is also built upon `BAGEL`.

Implementation Details. During evaluation, we provide each baseline with the human–MLLM interaction history together with the personalized T2I request from that human. We train `BAGEL` with LoRA, using NVIDIA RTX 6000D GPUs, on 5 persona's data that is different from evaluation.

4.1 Main Evaluation

In Tables 2 and 3, we present the performance of each method. From the results, we observe that frontier T2I models still underperform on VIMG-Bench. Although they are strong at reasoning

Table 3: Performance across six domains under PPS setting. Best performance is in **bold**.

PPS Setting	Animal			Plant			Person		
Method	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**
GPT-Image-2	41.11	8.09	18.97	43.38	10.84	19.67	45.91	16.16	16.73
Nano-Banana-2	46.62	16.10	19.26	49.42	11.03	18.83	47.91	24.03	17.39
SenseNova-U1	26.13	0.43	20.29	34.80	2.80	20.53	27.38	0.60	19.43
BAGEL (Base)	13.38	0.20	20.15	26.93	0.63	21.26	16.22	0.33	20.09
+ History	14.35	0.17	18.19	27.29	0.63	18.67	12.75	0.30	16.37
+ RAG	16.45	0.27	20.03	24.75	0.70	20.25	19.33	0.43	19.45
+ **VIMGer**	**38.09**	**5.81**	18.45	**38.67**	**2.76**	18.97	**37.60**	**8.10**	17.31
PPS Setting	**Landscape**			**Fashion**			**Food**		
Method	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**	**VScore-P**	**VScore-S**	**CLIP-T**
GPT-Image-2	49.33	13.71	18.36	35.82	2.68	18.29	41.82	8.17	18.56
Nano-Banana-2	49.60	21.52	18.57	42.45	5.47	18.52	45.20	13.50	18.51
SenseNova-U1	32.53	0.77	19.93	19.95	0.43	19.04	24.85	0.63	20.20
BAGEL (Base)	27.87	0.77	20.78	13.58	0.31	20.12	16.89	0.40	20.67
+ History	23.60	2.68	18.33	17.15	0.33	18.93	15.87	0.37	18.16
+ RAG	23.20	2.71	19.66	12.80	0.30	20.57	18.98	0.47	20.25
+ **VIMGer**	**40.13**	**8.15**	18.17	**38.89**	**2.81**	18.07	**41.78**	**5.59**	19.53

Table 4: Ablation study of core training steps within our proposed VIMGer (VScore-P).

Method (TSO Setting)	**Animal**	**Plant**	**Person**	**Landscape**	**Fashion**	**Food**
BAGEL (Base)	11.25	24.67	10.49	18.27	8.58	12.49
+ History	16.80	29.42	12.09	19.47	12.62	17.33
+ PVIR-SFT	32.89	42.40	30.85	34.67	31.60	34.28
+ PVR-PO (**VIMGer**)	**35.69**	**43.02**	**39.02**	**37.47**	**34.89**	**38.75**

over complex generation instructions, VIMG poses the **opposite challenge**: starting from simple yet highly personalized everyday painting requests, models must visually imagine what is left unsaid from long-term user-MLLM interactions and visualize it. Another finding is that most methods *perform better under the PPS setting*, as explicitly specifying some preferences reduces the imagination space. For controlled analysis on the `BAGEL`, we find: (**F1**) Most history-aware baselines yield improvements on baseline Base, supporting both the validity of VIMG-Bench and its evaluation protocol. However, Base still achieves non-zero VScore-P, since the model's population-level generation priors may match some of the target personalized preferences. **(F2)** For History and RAG baselines, their performances are limited, as the former exposes UMMs to substantial interaction noise, while the latter is also less effective because contextualized image-generation requests (e.g., "draw a solider for me") provide weak semantic cues for retrieving the relevant sessions. **(F3)** VScore-S is extremely low, due to its stringent requirement for jointly realizing the full personalized specification for a persona. VIMGer equips the UMM with PVIR reasoning patterns through SFT and RL, enabling it to filter noisy interactions, identify each user's visual imagination preferences from retained histories itself beyond the semantic retrieval, and realize them in the generated image.

4.2 Further Analysis

Ablative Study. In Table 4, we conduct an ablation study on the SFT and RL stages of VIMGer. We observe that SFT instills PVIR reasoning patterns into the UMM, leading to clear gains, while PVR-PO further strengthens these capabilities for VIMG, achieving a better performance.

Case Study. In Figure 5, we present cases of three personalized user prompts: "Soldier", "Horse", and "Cake". Comparing to baselines, VIMGer visually imagines the most latent and salient preferences for users from long-term interaction histories. More cases are provided in Appendix I.

5 Related Work: Text-To-Image Generation

Text-to-image (T2I) generation has progressed rapidly, from basic semantic alignment to modeling of compositional structure, physical plausibility, cultural context, and world knowledge (Wang et al., 2025a; Niu et al., 2026). Diffusion models remain the dominant paradigm (Podell et al., 2023; Black Forest Labs, 2024; Peebles & Xie, 2022; Hu et al., 2024), while autoregressive approaches provide



Figure 5: Several cases of VIMGer and baselines on VIMG-Bench.

an alternative formulation with stronger explicit modeling of token-level dependencies (NextStep Team et al., 2025; Fan et al., 2024). More recently, Unified Multimodal Models (UMMs) (e.g., BAGEL (Deng et al., 2025), EMU-3.5 (Cui et al., 2025)) and agentic T2I systems (e.g., GPT-Image-2 (OpenAI, 2026b), Nano-Banana-2 (Raisinghani, 2026)) have further enabled unified multimodal understanding and generation within a single framework. Nevertheless, users' image-generation queries often specify only the core concepts of immediate interest, while these capable models fill in these unspecified details using population-level priors. **A handful of studies** have begun to address this problem. However, they either restrict to specific domains, such as poster generation (Shen et al., 2024; Xu et al., 2025; Wang et al., 2025b; Zhang et al., 2026a), or infer user preferences from a small set of preferred images or prior T2I contexts (Abdal et al., 2026; Chen et al., 2023; Salehi et al., 2024; Dang et al., 2025; Nabati et al., 2025; Wang et al., 2026). Such histories are often scarce for ordinary users. In contrast, amid the boom in human–MLLM interaction, long-term interactions with personal MLLM assistants provide a richer source of evidence about users' preferences on these unspecified spaces (Nie et al., 2026), yet it remains underexplored. Another line of work, referred to as subject-driven personalized image generation, focus on preserving the identity or appearance of a user-specific subject while generating images that follow the user's current instruction, rather than user's preference modeling (Liang et al., 2026; Gal et al., 2022; Ruiz et al., 2023).

6 Conclusion

In this work, we case T2I generation toward a personalized visual imagination task fostered by everyday user-MLLM interaction, namely VIMG. Accordingly, we introduce VIMG-Bench, the first benchmark for VIMG, together with VIMGer, a simple baseline. Empirical results show that even frontier T2I models and UMMs with strong generation-oriented reasoning capabilities still struggle on VIMG-Bench, while VIMGer improves open-source UMMs. Ablations and cases validate its efficacy, while also highlighting substantial room for more advanced approaches.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing and to generate synthetic data used in constructing VIMG-Bench. In particular, generative AI was used to assist the construction of benchmark (VIMG-Bench) histories, which were subsequently reviewed and revised by human annotators for quality and consistency. We did not use generative AI tools to conduct the core empirical analysis or to make final scientific judgments, and other disclosure categories are not applicable to this work. We reviewed all AI-assisted content and take responsibility for the final content of this work.

ETHICS STATEMENT

All human annotators (volunteers) involved in this work are authors of the paper, and we do not employ any other volunteers. They were fully informed about the study and voluntarily participated in the annotation and evaluation process. Their involvement primarily consisted of reviewing and refining benchmark data to improve its quality, as well as serving as volunteers to complete a subset of the benchmark tasks for human validation. We did not collect, use, or retain any private or personally sensitive information from the annotators or volunteers. Our work studies personalized visual imagination from long-term human–MLLM histories, a benign research setting aimed at improving the ability of multimodal assistants to understand individual users and provide more personalized visual generation. We believe this capability represents an important step toward more adaptive and user-centered multimodal AI systems.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the construction of VIMG-Bench, together with comprehensive benchmark statistics and representative cases, in Section 2 and Appendices B, C and I. We further provide the design rationale, training procedure, and implementation details of VIMGer in Section 3 and Appendices F and G. To facilitate reproducibility, we release the code for benchmark evaluation as well as the implementation of VIMGer. For the full VIMG-Bench benchmark, due to concerns about its substantial storage requirements, we commit to publicly releasing it upon publication.

REFERENCES

Rameen Abdal, James Burgess, Sergey Tulyakov, and Kuan-Chieh Jackson Wang. Visual personalization turing test. *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. MyVLM: Personalizing VLMs for user-specific queries. *arXiv preprint arXiv:2403.14599*, 2024. URL https://arxiv.org/abs/2403.14599.

Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, Bocheng Zou, Chaoqun Yang, and Wentao Zhang. UniCTokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025. URL https://arxiv.org/abs/2505.14671.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. URL https://arxiv.org/abs/2405.09818.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. URL https://arxiv.org/abs/2501.17811.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 7727–7736, 2023.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyu Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wen-Rui Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native Multimodal Models are World Learners. *arXiv.org*, abs/2510.26583, 2025.

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 8020–8030, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dang_Personalized_Preference_Fine-tuning_of_Diffusion_Models_CVPR_2025_paper.html.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiao-Ping Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *arXiv.org*, abs/2505.14683, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Sara Dorfman, Maya Vishnevsky, Omer Dahary, Or Patashnik, and Daniel Cohen-Or. Semantic browsing: Controllable diversity for image generation. In *European Conference on Computer Vision (ECCV)*, pp. 412–430. Springer, 2026.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024. URL https://arxiv.org/abs/2410.13863.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using Textual Inversion. *arXiv preprint arXiv:2208.01618*, 2022. URL https://arxiv.org/abs/2208.01618.

Rongpei Hong, Jian Lang, Ting Zhong, Yong Wang, and Fan Zhou. TAMEing long contexts in personalization: Towards training-free and state-aware MLLM personalized assistant. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. URL https://arxiv.org/abs/2512.21616.

Taihang Hu, Mengting Chen, Jinsong Lan, Xiaoyong Zhu, Kaifu Zhang, Ming-Ming Cheng, Bo Zheng, and Yaxing Wang. ORION: Decoupling and alignment for unified autoregressive understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/646716106d2e9def5946476df28c94db-Abstract-Conference.html.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Jen-Yuan Huang, Tong Lin, and Yilun Du. Long-text-to-image generation via compositional prompt decomposition. In *International Conference on Learning Representations (ICLR)*, 2026.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. Personamem-v2: Towards Personalized Intelligence via Learning Implicit User Personas and Agentic Memory. *arXiv.org*, abs/2512.06688, 2025.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke S. Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing Benchmark Drift in Text-to-Image Evaluation. *arXiv.org*, abs/2512.16853, 2025.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. Gir-Bench: Versatile Benchmark for Generating Images with Reasoning. In *International Conference on Learning Representations (ICLR)*, 2026a.

Niantong Li, Guang Hu, Wei Qiao, Ying Ba, Qichen Hong, Shiju Shen, Jinlin Wang, Fan Zhou, Jian Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyu Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Huayun Wei, Lin Qu, and Chenfei Wu. Qwen-Image-Bench: From Generation to Creation in Text-to-Image Evaluation. *arXiv.org*, abs/2605.28091, 2026b.

Zhuoqun Li, Zhaopei Huang, Wenxuan Wang, and Qin Jin. LongMP-Bench: A benchmark for multimodal persona understanding in long-term dialogues. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 23132–23160. Association for Computational Linguistics, 2026c. doi: 10.18653/v1/2026.findings-acl.1159. URL https://aclanthology.org/2026.findings-acl.1159/.

Qian Liang, Yujia Wu, Kuncheng Li, Jiwei Wei, Shiyuan He, Jinyu Guo, and Ning Xie. MM-R1: Unleashing the power of unified multimodal large language models for personalized image generation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026. URL https://arxiv.org/abs/2508.11433.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. *arXiv preprint arXiv:2604.24763*, 2026. URL https://arxiv.org/abs/2604.24763.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024. URL https://arxiv.org/abs/2411.07975.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Xinyu Hou, Margaret Li, and Bill Byrne. According to Me: Long-Term Personalized Referential Memory QA. *arXiv.org*, abs/2603.01990, 2026.

Ofir Nabati, Guy Tennenholtz, Chihwei Hsu, Moonkyung Ryu, Deepak Ramachandran, Yinlam Chow, Xiang Li, and Craig Boutilier. Preference adaptive and sequential text-to-image generation. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45362–45394. PMLR, 2025. URL https://proceedings.mlr.press/v267/nabati25a.html.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Ailin Huang, Bin Wang, Changxin Miao, Deshan Sun, En Yu, Fukun Yin, Gang Yu, Hao Nie, Haoran Lv, Hanpeng Hu, Jia Wang, Jian Zhou, Jianjian Sun, Kaijun Tan, Kang An, Kangheng Lin, Liang Zhao, Mei Chen, Peng Xing, Rui Wang, Shiyu Liu, Shutao Xia, Tianhao You, Wei Ji, Xianfang Zeng, Xin Han, Xuelin Zhang, Yana Wei, Yanming Xu, Yimin Jiang, Yingming Wang, Yu Zhou, Yucheng Han, Ziyang Meng, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. NextStep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025. URL https://arxiv.org/abs/2508.10711.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'LLaVA: Your personalized language and vision assistant. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.09400.

Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'Chameleon: Personalized vision and language generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. URL https://arxiv.org/abs/2504.20998.

Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. Personavlm: Long-term personalized multimodal llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 15000–15009, 2026. URL http://arxiv.org/abs/2604.13074.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *International Conference on Machine Learning (ICML)*, 2026.

Yeongtak Oh, Sangwon Yu, Junsung Park, Han Cheol Moon, Jisoo Mok, and Sungroh Yoon. Contextualized visual personalization in vision-language models. *arXiv preprint arXiv:2602.03454*, 2026. URL https://arxiv.org/abs/2602.03454.

OpenAI. Previewing GPT-5.6 Sol: a next-generation model. https://openai.com/index/previewing-gpt-5-6-sol/, 2026a.

OpenAI. GPT-Image-2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. URL https://arxiv.org/abs/2212.09748.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Naina Raisinghani. Nano Banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 22500–22510, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ruiz_DreamBooth_Fine_Tuning_Text-to-Image_Diffusion_Models_for_Subject-Driven_Generation_CVPR_2023_paper.html.

Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. Viper: Visual personalization of generative models via individual preference learning. In *European Conference on Computer Vision (ECCV)*, pp. 391–406. Springer, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv.org*, abs/2402.03300, 2024.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. PMG: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference (WWW)*, pp. 3833–3843. ACM, 2024. doi: 10.1145/3589334.3645633. URL https://doi.org/10.1145/3589334.3645633.

Harini SI, Somesh Singh, Yaman Kumar Singla, David Doermann, and Rajiv Ratn Shah. Zipp: Zero-shot image personalization from personas. *arXiv preprint arXiv:2606.08841*, 2026.

Kaiyue Sun, Weiyang Jin, Chengqi Duan, Rongyao Fang, Xian Liu, Yuwei Niu, Chunwei Wang, Aoxue Li, and Xihui Liu. Universe: Empower unified generation with reasoning and knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 21997–22006, 2026.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Si-Qing Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qiuyuan Wei, Jinhui Yin, Xueting Yang, Zhihang Zhong, Qiaoli Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, G. Luo, Hongsheng Li, Yunsong Qiao, Kai Chen, and Hongjie Zhang. Internvl-U: Democratizing Unified Multimodal Models for Understanding, Reasoning, Generation and Editing. *arXiv.org*, abs/2603.09877, 2026.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, et al. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*, 2025a.

Xianquan Wang, Zhaocheng Du, Huibo Xu, Shukang Yin, Yupeng Han, Jieming Zhu, Kai Zhang, and Qi Liu. Personalized visual content generation in conversational systems. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 34571–34602. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1160. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/314116cd378bab6283ae19c07a4ea367-Paper-Conference.pdf.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. URL https://arxiv.org/abs/2409.18869.

Zeyu Wang, Zilong Chen, Chenhui Gou, Feng Li, Chaorui Deng, Deyao Zhu, Kunchang Li, Weihao Yu, Haoqin Tu, Haoqi Fan, and Cihang Xie. LightFusion: A light-weighted, double fusion framework for unified multimodal understanding and generation. *arXiv preprint arXiv:2510.22946v4*, 2025c. URL https://arxiv.org/abs/2510.22946v4.

Zihao Wang, Yuxiang Wei, Xinpeng Zhou, Tianyu Zhang, Tao Liang, Yalong Bai, Hongzhi Zhang, and Wangmeng Zuo. Premier: Personalized preference modulation with learnable user embedding in text-to-image generation. *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. URL https://arxiv.org/abs/2410.13848.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Yuhang Wu, Shuxiang Zhang, Wee Hian Ching, Chi Zhang, and Miao Liu. Pipbench: A profile-inclusive framework for personalized image generation evaluation. *arXiv preprint arXiv:2607.06440*, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. URL https://arxiv.org/abs/2408.12528.

Shengqi Xu, Xinpeng Zhou, Yabo Zhang, Ming Liu, Tao Liang, Tianyu Zhang, Yalong Bai, Zuxuan Wu, and Wangmeng Zuo. Preferthinker: Reasoning-based Personalized Image Preference Assessment. *International Conference on Learning Representations (ICLR)*, abs/2511.00609, 2026.

Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM Web Conference (WWW)*, pp. 264–274. ACM, 2025. doi: 10.1145/3696410.3714843. URL https://doi.org/10.1145/3696410.3714843.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 22477–22503, 2025.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, Yi Wang, Limin Wang, and Yali Wang. UniFlow: A unified pixel flow tokenizer for visual understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8882d370cdafec9885b918a8cfac642e-Abstract-Conference.html.

Yuting Zhang, Ying Sun, Dazhong Shen, Ziwei Xie, Feng Liu, Changwang Zhang, Xiang Liu, Jun Wang, and Hui Xiong. Discrete preference learning for personalized multimodal generation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 2375–2385. ACM, 2026a. doi: 10.1145/3805712.3809552. URL https://doi.org/10.1145/3805712.3809552.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026b.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. URL https://arxiv.org/abs/2408.11039.

Technical Appendix

TABLE OF CONTENTS

A Additional Literature Review **17**
A.1 Literature Review of Long-Term User-MLLM Assistant Interaction 17
A.2 Literature Review of Unified Multimodal Models 17

B Additional Benchmark Construction Details **17**

C Core Statistics of VIMG-Bench **18**

D Human Validation of Benchmark Quality and Evaluation **20**
D.1 Human Quality Control for VIMG-Bench Construction 20
D.2 Post-hoc Human Validation of VIMG-Bench 20
D.3 Human Alignment of VIMG-Score. 21

E Detailed Implementation of Rewards in PVR-PO **21**

F VIMGer Training Pipeline Algorithm **22**

G Additional Details on Implementation **22**

H Limitations & Future Work **23**

I More Cases Presentation **24**

J Further Presentations of VIMG-Bench **24**

K Presentations of Prompts **28**
K.1 Core Prompts in VIMG-Bench Construction 28
K.2 Prompts in VIMGer 30
K.3 Prompts in Baselines and Evaluation 31

A Additional Literature Review

In this section, we provide additional two domains that are relevant to our study.

A.1 Literature Review of Long-Term User-MLLM Assistant Interaction

Beyond one-response-fits-all manner, recent studies have begun exploring MLLMs as personalized assistants in everyday human interactions. Early works, such as MyVLM (Alaluf et al., 2024) and Yo'LLaVA (Nguyen et al., 2024), mainly focus on recognizing and reasoning about user-specific visual concepts, while a few recent studies, including PersonaMem-V2 (Jiang et al., 2025) CoViP (Oh et al., 2026), PersonaVLM (Nie et al., 2026), TAME (Hong et al., 2026), and LongMP-Bench (Li et al., 2026c), have moved toward contextualized and long-term personalization by leveraging accumulated multimodal interactions to retrieve personal memories, track evolving user information, and generate personalized responses. However, these efforts only investigate how long-term human–MLLM interactions enhance multimodal *understanding* and response generation, such as VQA, captioning, and dialogue, leaving their potential for personalized generation largely unexplored. Nevertheless, T2I generation is equally central to everyday human–assistant interactions (Zhang et al., 2026b), and VIMG-Bench timely bridges this gap by evaluating whether personalized MLLM assistants can translate long-term user understanding into image generation, either by invoking an external generator or generating images themselves. Although works such as Yo'Chameleon (Nguyen et al., 2025), UniCTokens (An et al., 2025), and MM-R1 (Liang et al., 2026) also study image generation in MLLM personalization, their personalization is primarily subject- or concept-driven, focusing on preserving user-provided subjects across generations (as also mentionde in the main paper), rather than the human's preference driven personalization in generation.

A.2 Literature Review of Unified Multimodal Models

Conceptually, Unified Multimodal Models (UMMs) aim to perform multimodal understanding and generation within a single model. Early efforts explored token-based autoregressive unification, where text and discretized visual representations are modeled by a shared Transformer, as exemplified by Chameleon (Chameleon Team, 2024), Janus (Wu et al., 2024), and Emu3 (Wang et al., 2024). However, such designs face an inherent gap between the semantic representations favored by understanding and the fine-grained visual information required for generation (Hu et al., 2026; Yue et al., 2026). A complementary line of work therefore combines autoregressive language modeling with diffusion or flow-based generation, as in Show-o (Xie et al., 2024), Transfusion (Zhou et al., 2024), JanusFlow (Ma et al., 2024), and BAGEL (Deng et al., 2025), while some models further decouple the visual pathways for understanding and generation (Chen et al., 2025; Wang et al., 2025c). More recently, studies like Tuna-2 (Liu et al., 2026) and SenseNova-U (Diao et al., 2026) push toward a more native form of unification by eliminating pretrained visual encoders altogether and directly modeling raw images through pixel embeddings for both understanding and generation. Nevertheless, most of these studies focus on improving the general generation and reasoning capabilities of UMMs, while leaving user-specific generation underexplored. Moreover, VIMG poses a fundamentally different reasoning challenge: existing UMMs are primarily optimized to reason over complex yet explicitly specified generation instructions, such as intricate compositions and spatial constraints, whereas VIMG requires models to recover latent personalized intent from long-term interactions and instantiate the visual content left unspecified by the current request.

B Additional Benchmark Construction Details

We provide more details in the construction of our VIMG-Bench benchmark through several aspects.

Preference Dimension and Value-Space Construction. For GVIPs, we follow Xu et al. (2026) to define commonly considered global visual dimensions and their value spaces. For DVIPs and SVIPs, we leverage the LLM (`GPT-5.6-Sol`), grounded in its pretrained knowledge and external sources such as Wikipedia, to identify salient visual dimensions for the subject of each sub-domain (e.g., horse). These dimensions cover both intrinsic subject attributes and attributes arising from common human interactions with the subject. We then use the LLM to instantiate a candidate value imagination space for each dimension, followed by human refinement.

Hierarchical Specification Construction. Given the preference dimensions and their candidate values defined above, we construct a pool of candidate specifications at each level of the hierarchy. Rather than directly taking a Cartesian product across global-, domain-, and sub-domain-level preferences, we perform combinatorial construction independently within each level. Specifically, we enumerate candidate combinations of concrete preference values selected from the applicable dimensions at the global, domain, and sub-domain levels, respectively. This level-wise construction preserves the hierarchical sharing structure of visual imagination preferences, since global preferences are shared across prompts of the same persona, while domain preferences are shared across sub-domains within the same domain. We then inspect the resulting candidates and remove combinations containing mutually *incompatible or semantically conflicting* preference values. The remaining valid candidates form the specification pools used for subsequent persona matching.

Persona–Specification Matching. We positively match each persona to these candidates according to the preference values supported by its profile and select the candidate with the highest matching rate. The matching process is likewise conducted in a level-wise manner. Importantly, unmatched preference values need not be directly implied by the persona, but we strictly ensure that none conflicts with the persona profile. All final persona–specification pairs are further reviewed by humans, checking whether any conflicting imagination preference values within each specification. The length of each full specification is also kept short (4–6 preferences), reflecting that users disclose only a limited number of relevant preferences in real-world interactions.

Realistic Interaction Construction. To avoid directly prompting an LLM to synthesize entire interaction histories, which often leads to less natural interactions, we ground each preference-revealing session in both an interaction template and a concrete topic. The former specifies how a target preference is naturally expressed, while the latter describes the everyday context in which the interaction unfolds. Seed dialogue topics are collected from existing real-world human–MLLM interaction benchmarks (Xue et al., 2025; Mei et al., 2026), then adapted by an LLM to the corresponding preference and further refined by human annotators. During final interaction generation, the LLM follows the selected template and topic while adapting the dialogue to the corresponding persona.

Image-Based Preference Evidence. For preferences that can be naturally conveyed visually, such as color or style, we additionally construct multimodal sessions (`GPT-Image-2`) in which the preference value is not stated in text but is instead revealed through the image content. This design simulates everyday multimodal interactions where personal preferences must be inferred from what users show rather than what they explicitly say. Some interactions also contain multimodal content where the images do not convey preference evidence, but instead serve only the context.

Distractor Interaction Construction. To improve the realism of the daily interactions, where the the interactions should not by occupied with the preference-revealing ones, we introduce *substantial* distractor interactions. Our distractors go beyond generic daily conversations or tool-use requests such as writing and polishing. We also include interactions involving the same six everyday domains, where related attributes or even identical preference values may be discussed without disclosing positive evidence for the target preference. Such distractors make the histories more realistic while preventing models from exploiting simple keyword- or domain-level shortcuts in resolving VIMG.

C Core Statistics of VIMG-Bench

We summarize the statistics of visual imagination preferences, personalized specifications, and interaction histories. **Firstly**, Figure 6 displays all four GVIP dimensions and the five most frequently assigned dimensions at each of the DVIP and SVIP levels. For each dimension, the segmented bar shows the empirical distribution of assigned preference values across personas, counting each persona's assignment once to avoid duplicating inherited preferences across prompts. The adjacent dot reports the Shannon entropy of this value distribution, normalized by the logarithm of the number of candidate values for that dimension. We observe that most preference dimensions across the three hierarchical levels exhibit substantial value diversity, with correspondingly high entropy.

Secondly, Figure 7 shows the personalized specifications. The left panel reports the proportions containing four, five, or six preferences, while the middle panel shows the mean numbers of GVIP, DVIP, and SVIP assignments per specification. The right panel contains one violin plot per domain, showing specification dissimilarity between personas requesting the same subject. We measure

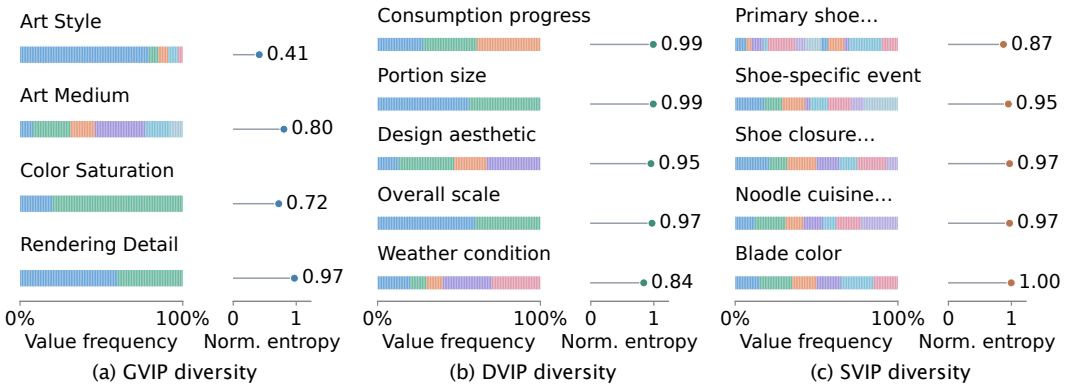


Figure 6: Visual imagination preference value distribution and diversity statistics.

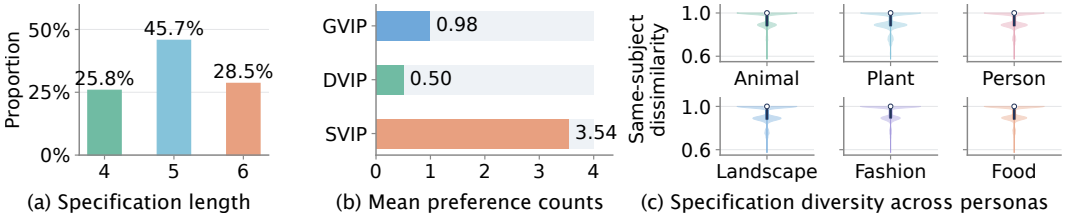


Figure 7: Specification length, preference-level composition, and specification diversity across personas requesting the same subject.

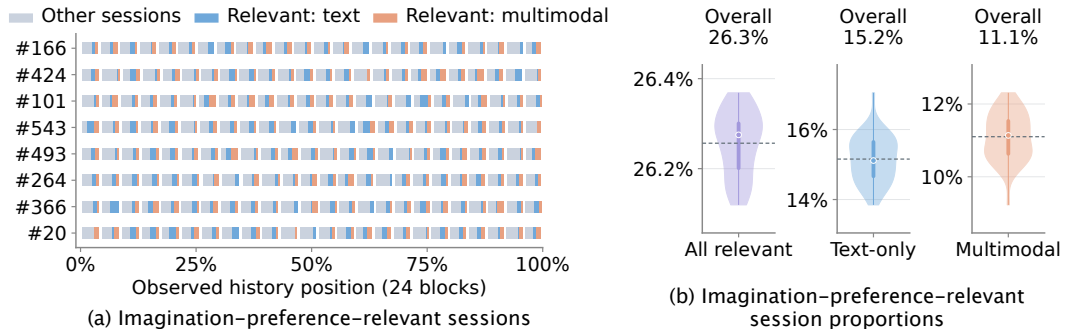
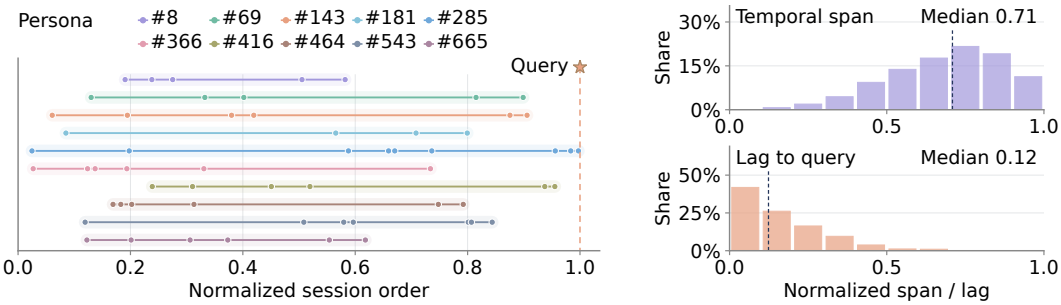


Figure 8: Proportions of imagination-preference-relevant sessions across the whole history.

dissimilarity using Jaccard distance between sets of dimension–value pairs. We observe that SVIPs constitute the largest share of preferences across personas, while specifications exhibit substantial inter-persona diversity, showing the fine-grained personalization and diversity of VIMG-Bench.

Finally, Figure 8 and Figure 9 illustrate the structure of long-term user-MLLM interactions. In Figure 8, we select eight persona histories at evenly spaced ranks of their relevant-session proportions. Each row contains 24 blocks, each summarizing a consecutive, approximately equal-sized group of sessions. Colors distinguish unrelated, text-only relevant, and multimodal relevant sessions. The three violin plots on the right show the distributions of the proportions of all imagination-preference-relevant sessions and their text-only and multimodal subsets, using each persona's total session count as the denominator. We observe that preference-revealing sessions account for only a small proportion of the interaction history, consistent with real-world interactions where user preferences are disclosed only occasionally amid predominantly unrelated everyday conversations. Figure 9 displays the relevant sessions for specifications of a given subject from ten selected personas. Each row marks the imagination-preference-relevant sessions for one persona's target specification, with a line connecting the first and last relevant sessions. The two histograms summarize the specifications across all personas: temporal span measures the distance between the first and last relevant sessions, while lag measures the distance from the last relevant session to the query. Both distances are measured in session positions and normalized by the corresponding history length. We observe



(a) Relevant sessions for selected specifications

(b) Relevant session span and lag

Figure 9: Temporal distribution of imagination-preference-relevant sessions.

Table 5: Interaction history (VIMG-Bench) quality assessment from human volunteers.

Conversational Consistency	Interaction Realism	Imagination Fidelity	Revelation Naturalness
75.0	75.0	91.2	82.5

that imagination-preference-relevant sessions are only sparsely distributed in the history, consistent with real-world long-term interactions where such preference signals emerge only intermittently. Moreover, the temporal span between the relevant sessions and the user's image-generation query is often substantial, making recent-history-only strategies less effective on VIMG-Bench.

D Human Validation of Benchmark Quality and Evaluation

In this section, we present the human evaluation of VIMG-Bench conducted by volunteers.

D.1 Human Quality Control for VIMG-Bench Construction

Several human volunteers participate in quality control and refinement throughout nearly every stage of VIMG-Bench construction. First, they review the selected six domains and 30 sub-domains, the visual imagination preference dimensions, and value spaces, ensuring that each dimension is well-defined and that its candidate values remain semantically comparable at the same level of granularity. Second, they inspect the matched persona–specification pairs to ensure that every assigned visual imagination preference is consistent with, or at least non-conflicting with, the corresponding persona profile. Finally, they review the constructed interaction histories for realism, fluency, and consistency with the persona. For imagination-revealing sessions, volunteers further verify that the target visual imagination preferences are conveyed accurately and naturally rather than being artificially exposed.

D.2 Post-hoc Human Validation of VIMG-Bench

We further recruit several volunteers to assess the benchmark and report their averaged ratings.

Quality Assessment of Interactions. We evaluate the constructed interactions along four dimensions. The first two assess the interactions themselves: **Conversational Consistency**, measuring whether each multi-turn session remains logically and contextually consistent throughout, and **Interaction Realism**, measuring how closely it resembles everyday human–MLLM interaction. The remaining two assess the revelation of visual imagination preferences: **Imagination Fidelity**, measuring whether the intended visual imagination preference is accurately conveyed, and **Revelation Naturalness**, measuring whether it emerges organically from the interaction rather than being explicitly or artificially inserted. For the first two metrics, each volunteer is assigned 10 randomly sampled personas and evaluates their corresponding interaction histories. For the latter two, each volunteer is assigned 10 randomly sampled preference-revealing interaction sessions. For Conversational Consistency, Interaction Realism, and Revelation Naturalness, volunteers provide *binary judgments*, where 1 indicates that the corresponding criterion is satisfied and 0 otherwise. For Imagination Fidelity, volunteers are instead asked to identify the visual imagination preference revealed

Table 6: Human judgment on VIMG-Bench with VIMG-Score.

TSO Setting	**Animal**			**Plant**			**Person**		
Method	**Q-Agree**	**AUROC**	**Spearman ρ**	**Q-Agree**	**AUROC**	**Spearman ρ**	**Q-Agree**	**AUROC**	**Spearman ρ**
GPT-Image-2	92.31	95.94	0.6669	82.61	84.38	0.7906	100.00	100.00	1.0000
BAGEL (Base)	76.92	63.64	0.0513	91.30	96.21	1.0000	100.00	100.00	0.9487
+ **VIMGer**	92.31	88.89	0.7000	91.30	94.62	0.8208	91.67	92.14	0.8208
TSO Setting	**Landscape**			**Fashion**			**Food**		
Method	**Q-Agree**	**AUROC**	**Spearman ρ**	**Q-Agree**	**AUROC**	**Spearman ρ**	**Q-Agree**	**AUROC**	**Spearman ρ**
GPT-Image-2	86.36	96.67	0.6708	96.00	98.67	0.8721	96.00	88.89	0.9000
BAGEL (Base)	95.45	97.92	0.9487	88.00	75.44	0.6708	96.00	88.00	0.6669
+ **VIMGer**	95.45	98.29	0.9000	96.00	98.72	0.8721	88.00	89.61	0.2236

in each assigned session. We compare their selections against our benchmark annotations and calculate their accuracy value, thereby validating whether the annotated visual imagination preferences are faithfully and unambiguously grounded in the constructed interactions.

We report the average results across the volunteers in Table 5. We first observe high human ratings for conversational consistency and interaction realism, indicating that the constructed interactions are coherent and closely resemble everyday human–MLLM conversations. For the preference-revealing sessions, revelation naturalness is likewise high, suggesting that the target preferences emerge organically rather than being artificially exposed. More importantly, human volunteers can accurately identify the intended visual imagination preferences from each separate sessions, resulting in high imagination fidelity and indicating that our preference-revealing patterns are consistent with how humans naturally infer preferences from interactions. Notably, this evaluation does not imply that VIMG-Bench is easy for humans, as it only tests whether a preference can be recognized from a **single** preference-revealing session. In contrast, VIMG-Bench requires T2I models or systems to reason over long-term interaction histories to identify relevant imagination preference evidence, infer the personalized visual imagination, and realize it in the generated image.

D.3 Human Alignment of VIMG-Score.

To validate the reliability of our MLLM-based VIMG-Score, we additionally conduct a human alignment study with several volunteers. Each volunteer is assigned images generated for 10 personas by the frontier model `GPT-Image-2` and the open-source UMM `BAGEL` under two settings: a **Base** setting, where no interaction history is provided, serving as a lower-bound case for assessing whether human judgments agree on clearly unpersonalized generations, and a setting of **augmented with VIMGer**. The volunteers then evaluate these images on VIMG-Bench using the same VQA questions adopted by VIMG-Score. Rather than asking volunteers to provide confidence probabilities, which is unnatural for human evaluation, we collect a binary judgment for each VQA, indicating whether the corresponding imagination preference value is faithfully realized in the image. We then compare the human judgments with the continuous scores produced by VIMG-Score and report question-level agreement (Q-Agree), AUROC, and Spearman's ρ to quantify human alignment in Table 6. At the preference level (VScore-P), we observe strong agreement across most domains and methods, with high Q-Agree and AUROC together with generally strong rank correlations, demonstrating that VIMG-Score closely tracks human judgments.

E Detailed Implementation of Rewards in PVR-PO

As described in the main paper, VIMGer employs three reward functions, namely R_{history}, R_{imagine}, and R_{realize}. Below, we provide the implementation details of each reward.

For R_{history}, we measure the agreement between the retained interaction history $\hat{\boldsymbol{H}}_u$ and the oracle relevant history $\boldsymbol{H}_u^*$ using the F1 score, which is formulated as:

$$R_{\text{history}} = \frac{2\left|\hat{\boldsymbol{H}}_u \cap \boldsymbol{H}_u^*\right|}{\left|\hat{\boldsymbol{H}}_u\right| + |\boldsymbol{H}_u^*|}. \tag{5}$$

For R_{imagine}, we provide the LLM judge with the oracle imagination specification $\boldsymbol{S}$ associated with the current personalized user prompt P_u, together with the model-completed prompt $\tilde{P}_u$. The judge $\mathbb{I}_{\text{LLM}}$ evaluates whether each preference value in $\boldsymbol{S}$ is reflected in $\tilde{P}_u$, yielding a score in $[0, |\boldsymbol{S}|]$:

$$R_{\text{imagine}} = \frac{1}{|\boldsymbol{S}|} \sum_{s_i \in \boldsymbol{S}} \mathbb{I}_{\text{LLM}}\left(s_i, \tilde{P}_u\right). \tag{6}$$

For R_{realize}, we provide the MLLM judge with the oracle imagination specification $\boldsymbol{S}$ associated with the current personalized user prompt P_u, together with the final generated image $\boldsymbol{I}_u$. The judge $\mathbb{I}_{\text{MLLM}}$ evaluates whether each preference value in $\boldsymbol{S}$ is reflected in $\boldsymbol{I}_u$, yielding a score in $[0, |\boldsymbol{S}|]$:

$$R_{\text{realize}} = \frac{1}{|\boldsymbol{S}|} \sum_{s_i \in \boldsymbol{S}} \mathbb{I}_{\text{MLLM}}\left(s_i, \boldsymbol{I}_u\right). \tag{7}$$

As described above, the last two rewards are normalized to the range $[0, 1]$ by dividing $|\boldsymbol{S}|$, and the corresponding judges are implemented using the frontier model `GPT-5.6-Sol`.

F VIMGer Training Pipeline Algorithm

The training recipe of VIMGer is provided in Algorithm 1.

G Additional Details on Implementation

In this section, we provide additional implementation details for VIMGer and the baselines.

PPS Setting. The user's prompts explicitly reveals one user's imagination preference in the current query, while still requiring the model to imagine and realize the remaining preferences from long-term interaction histories. The explicitly provided preferences will also be scored as usual.

Implementation Details of VIMGer. VIMGer is built on BAGEL-7B-MoT and trained with LoRA (rank 32, $\alpha = 64$), while keeping the pretrained weights frozen. Because a user's full long-term interaction history exceeds BAGEL's context length, Step 1 partitions the history into contiguous session-level windows of at most 12K tokens, filters the sessions relevant to the current image-generation request in each window, and merges them into the retained history $\hat{H}_u$. Step 2 then takes $\hat{H}_u$ together with the current request P_u to infer the personalized specified prompt $\tilde{P}_u$, which is subsequently used for image generation. For PVIR-SFT, we use 150 instances from five training personas. GPT-5.6-Sol instantiates the ground-truth specification into the oracle specified prompt P_u^*, and GPT-Image-2 generates the corresponding image supervision. We train PVIR-SFT for 3 epochs with 8 queries per step and a learning rate of 10^{-4}. PVR-PO is initialized from the PVIR-SFT checkpoint, samples $G = 8$ rollouts per query, and is trained for 50 steps with 8 queries per step and a learning rate of 2×10^{-5}. Both stages are trained on two NVIDIA RTX 6000D GPUs. To prevent data leakage, we enforce a strict persona-level split between training and evaluation. The training personas and their associated interaction histories, visual imagination specifications, requests, and supervision images never appear in the evaluation set, and an additional held-out persona is used for SFT epoch selection. The reward judges used in PVR-PO are also different from the evaluation MLLM used by VIMG-Score, preventing direct optimization against the final evaluator.

Baselines Implementation. All methods are evaluated on the same instances using the same VQA questions and the same evaluation MLLM, `DeepSeek-V4.1-Flash` (DeepSeek-AI, 2026). GPT-Image-2 and Nano-Banana-2 follow a two-stage pipeline. A text MLLM first reads the full user interaction history together with the current request and produces a self-contained T2I prompt,

Algorithm 1 Training Pipeline of VIMGer

Require: Training instances $(\boldsymbol{H}_u, \boldsymbol{H}_u^*, P_u, \boldsymbol{S}) \in \mathcal{D}$, initial UMM p_{θ_0}, LLM `GPT-5.6-Sol`, image generator `GPT-Image-2`, group size G
Ensure: Personalized UMM policy π_θ

Stage I: Personalized Visual Imagination Reasoning SFT
Initialize $\mathcal{D}_{\mathrm{SFT}} \leftarrow \varnothing$
for all $(\boldsymbol{H}_u, \boldsymbol{H}_u^*, P_u, \boldsymbol{S}) \in \mathcal{D}$ **do**
  Prompt `GPT-5.6-Sol` to instantiate $\boldsymbol{S}$ into P_u
  Obtain the oracle specified prompt P_u^*
  Construct the SFT input $\boldsymbol{X} = (\boldsymbol{H}_u, , P_u)$, and output $\boldsymbol{Y} = (\boldsymbol{H}_u^*, P_u^*)$
  Generate image supervision $\boldsymbol{I}_u^*$ from P_u^* using `GPT-Image-2`
  Add $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}_u^*)$ to $\mathcal{D}_{\mathrm{SFT}}$
end for
Fine-tune p_{θ_0} on $\mathcal{D}_{\mathrm{SFT}}$ to obtain $\pi_{\theta_{\mathrm{SFT}}}$
Stage II: Personalized Visual Realization Policy Optimization
Initialize $\pi_\theta, \pi_{\theta_{\mathrm{old}}}, \pi_{\theta_{\mathrm{ref}}} \leftarrow \pi_{\theta_{\mathrm{SFT}}}$, with $\pi_{\theta_{\mathrm{ref}}}$ frozen
for step $= 1, \ldots, M$ **do**
  Sample a batch $\mathcal{B} \subset \mathcal{D}$
  for all $\boldsymbol{X} \in \mathcal{B}$ **do**
    Sample G rollouts $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid \boldsymbol{X})$
    for all $o_i = (\hat{\boldsymbol{H}}_{u,i}, \tilde{P}_{u,i}, \boldsymbol{I}_{u,i})$ **do**
      Compute R^i_{history} by comparing $\hat{\boldsymbol{H}}_{u,i}$ with $\boldsymbol{H}_u^*$ using F1
      Compute R^i_{imagine} by evaluating whether $\tilde{P}_{u,i}$ reflects every preference value in $\boldsymbol{S}$
      Compute R^i_{realize} by evaluating whether $\boldsymbol{I}_{u,i}$ faithfully realizes every preference value in $\boldsymbol{S}$
    end for
    Compute the group-relative advantages $A_{i,t}$ using Eq. (4)
  end for
  Update π_θ using the PVR-PO objective in Eq. (3)
  $\pi_{\theta_{\mathrm{old}}} \leftarrow \pi_\theta$
end for
return π_θ

which is then passed to the corresponding image generator. For BAGEL, we evaluate three settings. Base receives only P_u. History receives the full interaction history, including shared images, in chronological order together with the current request P_u. RAG first retrieves the sessions most relevant to the current request from the full history and then provides these sessions to BAGEL for generation. All open-source models are evaluated with a fixed seed for each instance. VIMG-Score is computed from the evaluation MLLM's Yes/No probabilities for each target preference, from which we derive VScore-P and VScore-S. We additionally report CLIP-T (Radford et al., 2021) to measure semantic alignment with the explicit user request rather than personalization. Across five independent seeds, the aggregate VIMG-Score remains highly stable for both VIMGer and baselines, with only small run-to-run variation.

H Limitations & Future Work

A broadly shared limitation in personalization research is the difficulty of obtaining large-scale, long-term interaction histories from real users, due to privacy concerns, collection costs, and the inherent challenges of scaling longitudinal data acquisition. This issue is particularly difficult to fully resolve under strict **privacy constraints** when histories span multiple sessions, modalities, and everyday activities, and VIMG-Bench inevitably faces the same limitation. To **narrow the gap** between synthetic interactions and real-world usage, we introduce multiple sources of real-world grounding throughout benchmark construction, including manually designed interaction templates that specify how target preferences are expressed within individual sessions, topic seeds derived from real-world human–MLLM interactions, and extensive volunteer review and refinement of both topics and interactions. While these measures cannot fully substitute for genuine long-term user

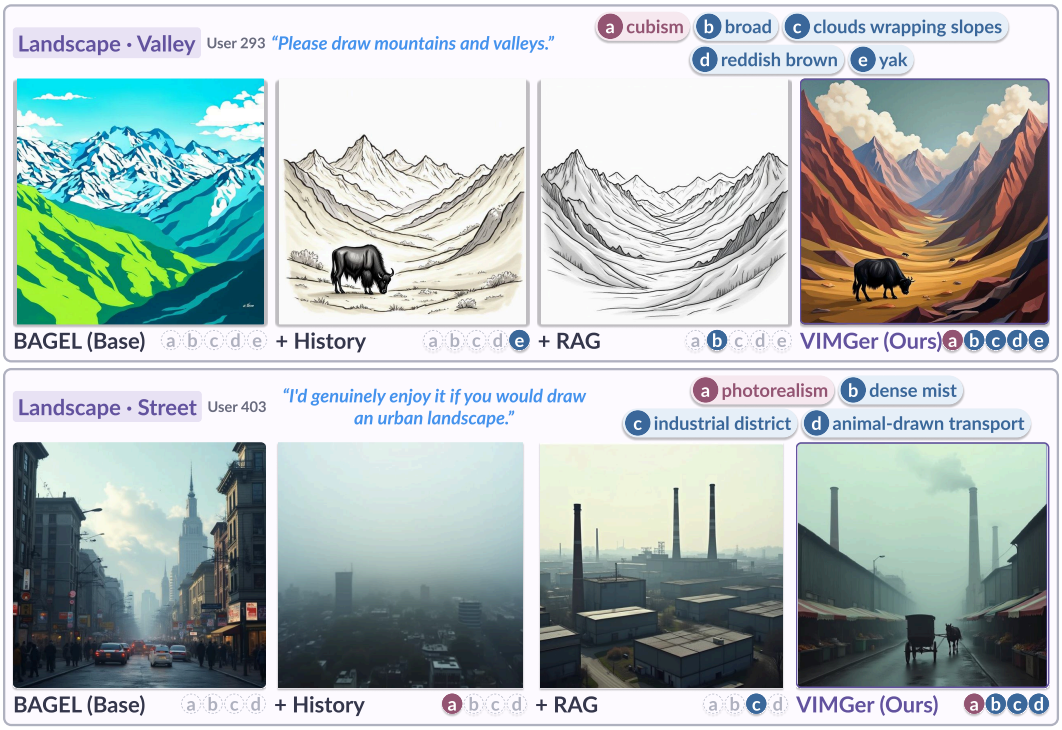


Figure 10: Additional cases of VIMGer and baselines on VIMG-Bench.

histories, they provide a practical balance among privacy, scale, controllable annotation, and interaction realism. Importantly, the extensive human involvement in annotation and quality control also enables fine-grained evaluation of whether models can accurately identify imagination-preference-relevant interactions from long-term histories and faithfully reflect the them in the generated images.

Looking forward, VIMG opens several promising directions for future research. **(1)** For open-source UMMs, further advances may come from improved SFT and RL designs, more effective reward modeling, memory mechanisms that reduce the cost of processing long interaction histories, and methods requiring substantially less task-specific training data. Particularly compelling are training-free approaches that remain effective when such supervision is unavailable in realistic settings. **(2)** For closed-source or larger T2I systems, agentic approaches could leverage external memory, history selection, prompt refinement, and iterative generation to enable personalized visual imagination without access to model parameters. **(3)** More fundamentally, future UMMs could incorporate VIMG directly into pre-training or post-training objectives, as understanding users over sustained interactions and translating that understanding into personalized generation represents a natural and increasingly important capability for general-purpose multimodal assistants.

I More Cases Presentation

In this section, we provide more cases on our proposed VIMG-Bench, in Figure 10.

J Further Presentations of VIMG-Bench

In this section, we provide additional presentations of VIMG-Bench, including the interaction topics in Figure 11 and everyday histories of personas in Figures 12 and 13.

Dialogue topic category	Suitable preference type	Content
Dog	Explicit	Discussing dogs, the user prefers retrieving: finding objects, bringing them back, and completing the handover.
Cat	Implicit	Reading with a cat asleep on their lap, the user enjoys this familiar, unhurried routine.
Animals	Implicit	Across years of travel photographs, the user consistently keeps animal images with slight motion blur.
Butterfly	Implicit	A butterfly lands on the user's hand. They wait quietly, cherishing this voluntary contact until it leaves.
Fashion	Implicit	Packing for a trip, the user finds that everyday clothes and accessories naturally share a sporty aesthetic.
Necklace	Explicit	Wearing a silver necklace comfortably all day, the user explains why they keep choosing this material.
Shoes	Implicit	Repairing favorite shoes, the user asks to preserve the plaid upper pattern they have long enjoyed.
Landscape	Explicit	The user favors unmanaged landscapes, often stopping beside interwoven branches and untrimmed riverbanks.
Mountains and valleys	Explicit	At a viewpoint, the user explains why monumental mountains have long guided their travel choices.
Desert	Explicit / implicit	Planning desert walks, the user seeks low scrub, enjoying how this vegetation leaves dunes visible.
People	Implicit	Getting ready, the user keeps their familiar short hair even as colors and accessories change.
Scientist	Explicit	At a roundtable, the user favors field observation, enjoying how scientists gather evidence in natural settings.
Detective	Implicit	Watching documentaries, the user focuses on examining clues and how investigators draw conclusions.
Plants	Implicit	At nurseries, the user repeatedly lingers over plants with dense foliage, regardless of species or height.
Rose	Explicit	Exchanging cuttings, the user again chooses climbing roses for their familiar growth, flowering, and care routines.
Shrub	Explicit	Discussing shrubs, the user values lasting silver foliage more than brief flowers when deciding which plants to keep.
Food	Implicit	After exercise, a longtime partner selects a large portion for the user, matching their usual appetite.
Bread	Explicit	Helping a friend bake, the user demonstrates a thin crust and explains their longstanding preference for its bite.
Cake	Explicit	Decorating cakes, the user habitually limits piping to the border, preserving focal points and open space.
Art medium	Implicit	Reorganizing collected artworks, the user keeps watercolors together on display, despite their different artists and subjects.
Color saturation	Explicit	Discussing animation, the user favors desaturated colors across characters and scenery, whatever the hues.

Figure 11: Presentation of preference-expression dialogue topics in VIMG-Bench.



PERSONA 166

Mateo Alejandro Rivera

Freelance sports photographer

27 years old · Los Angeles, USA

BFA in Photography and Imaging · University of Arizona

Interests Women's basketball, street photography, documentary filmmaking

14 SEP 2025 — Setting up an editing desk — Other session



User: The speakers on my documentary-audio desk are labeled near-field monitors. What does that imply about how far I should sit from them?

MLLM: They are designed for close listening. Keep them at ear height and form an even triangle with your seat; adjust until the stereo image centers clearly.

15 SEP 2025 — Returning to photorealism — Preference-relevant



User: At the retrospective, my friend steered us straight to the photorealism gallery. After several exhibitions together, he knew I would slow down there: its documentary detail echoes the way I photograph athletes.

MLLM: Your friend knows where your attention settles. Those lifelike figures and streets let you study gesture and context, just as you do when photographing athletes.

06 OCT 2025 — Upgrading a photo-editing computer — Other session



User: I am upgrading my photo-editing computer. Does M.2 describe the physical form, transfer protocol, or actual speed?

MLLM: M.2 is the form factor. Check the motherboard manual for the slot and supported protocol, then compare speeds. Shape alone does not ensure compatibility.

03 NOV 2025 — Choosing a place to pause — Preference-relevant

User: On our guided desert outing, I chose the dune crest for a safe pause again. I watched the wind reshape the slope, with views in every direction. I keep choosing the same place on these trips.

MLLM: The crest combines shifting wind, footing, and open views. That changing scene keeps drawing you back before you rejoin the group.

13 NOV 2025 — Making sense of hurried shoot notes — Other session

User: After the tournament shoot, I am rereading my rushed notes. I can recognize some abbreviations now, so I am writing them out and making a little key for the next assignment.

MLLM: Expand only the abbreviations you can verify. Keep a small reference list beside the notes and add the date or tournament name so the shorthand stays useful.

24 NOV 2025 — Watching unstaged animal behavior — Preference-relevant

User: At the zoo, I waited for the keeper to finish, then watched the panda choose bamboo and feed. I value that unprompted, everyday behavior.

MLLM: Once the keeper finished, the panda's own choice of branches made the behavior feel natural rather than staged.

26 NOV 2025 — Keeping an editing day flowing — Preference-relevant

User: While editing at home, I kept stepping out for errands. Each time I reached for my slip-on shoes and was back at the desk without losing my place. I repeat that choice on other editing days too.

MLLM: That quick on-and-off routine keeps errands from breaking your editing flow. Returning to the same easy closure helps you maintain momentum.

Figure 12: An example of the long-term MLLM-human histories of persona Mateo.



PERSONA 181

Haruka Tanaka

Biology student

23 years old · Kyoto, Japan

BSc in Biology (in progress)

Interests Birdwatching, wildlife photography, environmental volunteering

10 OCT 2025 — Noticing the same outline on two walks — Preference-relevant

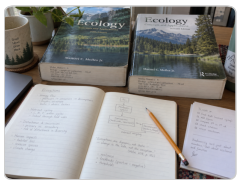
User: On two evening walks, I kept framing dogs whose tails stayed curled over their backs as they moved. I chose side views that preserved that curve; the changing steps faded into the background for me.



MLLM: Across both walks, you selected frames where the curled tail remained clear. That stable outline gives your motion photographs a visual anchor.

11 OCT 2025 — Checking an ecology reference — Other session

User: The citation in my ecology report refers to a revised edition, but I have the older book. I will check the chapter in the newer edition before finishing.



MLLM: Check the edition in the bibliography, then compare chapter headings and page numbers in the revised volume. That should explain why the cited page differs.

25 OCT 2025 — Checking a chalky deposit — Other session

User: I found a chalky ring around my cactus pot and took a close-up for the horticulturist. I brought my watering dates and noted that I use tap water, so she can identify the deposits.



MLLM: She can compare the residue with your watering schedule and test the water's mineral content before suggesting a change in care.

10 NOV 2025 — Returning to the valley floor — Preference-relevant

User: I would revisit the valley floor, where I can study plants and drainage patterns at my feet. I keep looking for that grounded feeling on mountain trips.

MLLM: The valley floor shows how slope, water, and vegetation meet. It is a fitting route when you want to observe the mountain as a connected system.

13 NOV 2025 — Getting a clearer view of a demonstration — Other session

User: At a noodle-making demonstration, the worktop blocked the key hand movement. I will ask the instructor to move so everyone can see.

MLLM: Gathering in a shallow arc would help. The instructor could pause during the key motion and repeat it from the side, so seated viewers can follow.

23 NOV 2025 — Choosing bread for a rich soup — Preference-relevant

User: At dinner I served the soup with bread baked deeply enough to hold its structure while soaking up broth. I reached for that same bake level again; even after dipping, each slice kept a satisfying chew.

MLLM: That bake level suits a rich soup: the bread absorbs broth while still giving you something substantial to chew.

04 DEC 2025 — Noticing a familiar detail at the clinic — Preference-relevant

User: At the rescue clinic, I held a cat's carrier during its eye exam. After the all-clear, I studied its copper irises. I notice that detail whenever I meet a cat, but kept the examination first.

MLLM: You put the cat's health first, then noticed its eyes. That habit follows you from one cat encounter to the next.

Figure 13: An example of the long-term MLLM-human histories of persona Haruka.

K PRESENTATIONS OF PROMPTS

We present the core prompts used for VIMG-Bench construction, VIMGer, the baselines, and VIMG-Score evaluation. Text in braces (e.g., {query}) denotes run-time placeholders. Each construction prompt is paired with the required task inputs and a strict JSON output schema, which we omit for brevity. We use *category* and *subject* to refer to domain and sub-domain, respectively, and *scenario* to denote the everyday context grounding each interaction session.

K.1 CORE PROMPTS IN VIMG-BENCH CONSTRUCTION

Selecting Dialogue Topics and Story Beats for Interaction Sessions

Select persona-fitting scenarios **ONLY** from each session menu. Write concise English story beats grounded in the chosen scenario, with a credible everyday situation. Preserve subject and dimension scope: global evidence generalizes across subjects, category evidence across the category.

- For explicit/implicit units with several sessions select **DISTINCT** scenarios, i.e. distinct events supporting the same value; for evolving units connect the old inclination to a credible later update.
- For distractors pick the menu scenario that best fits the persona's everyday life, without eliciting a stable visual preference.
- A unit's `sibling_fixed` lists the persona's other fixed preferences on the same subject, each with the values that would contradict it.
- Never select an option whose `history_image_guidance` or scenario would put one of those competing values into the scene: the guidance is binding on the rendered image, so such a unit can never satisfy both scenario fidelity and the persona's other fixed preferences.
- Prefer an option that leaves those dimensions alone.
- Whether a photo is attached is decided after your selection, so do not make a story beat depend on one.
- Continuity explains cross-session causal consistency; use "Independent event" when no dependency exists.
- Do not write full dialogue, invent profile facts, invoke tools or delegate.
- **Return only the requested JSON.**

Authoring Interaction Sessions

Author natural ENGLISH daily chats between this persona and their assistant, using the exact scheduled sessions and turn counts.

SPEAKER CONTRACT: the field "`user`" is the HUMAN PERSONA speaking first. The field "`assistant`" is the AI responding to that human. The HUMAN USER owns every personal experience, memory, observation, preference and photo in the scenario.

Put the persona's first-person scenario narrative and preference evidence in "`user`", **NEVER** in "`assistant`". The assistant may help or comment but must not claim to have lived the user's life, taken their photos, or have their visual preferences. Even if JSON keys are presented alphabetically, the conversational order is USER then ASSISTANT. Each user/assistant pair is one turn.

- Do not mention benchmark, annotation, target value, IDs or any internal metadata.
- Follow each selected template `history_pattern` and scenario precisely, preserving subject, attribute, value, scope, and stable inclination.
- Use the story beat and continuity, do not invent a different scenario.
- Avoid persona biography recitations or repeated canned openings.
- **For image `carrier=image`**, the actual selected value must **NEVER** appear in either speaker's text, including synonyms or translations; the user still identifies the subject and attribute and naturally responds to the visible attribute.
- **For evidence `carrier=text` or none**, convey the selected value unambiguously in natural words, with user-grounded evidence.

Authoring Interaction Sessions

- **For every carrier other than none** return a detailed stand-alone English image prompt for one scene image attached to the first user message, following the session `image_guidance` ([X] is the user; [Y] is the selected value).
- **For `carrier=image`** the image prompt **MUST** include the exact selected value.
- **For `carrier=text`** the photo shows the real-world material or scene background in `image_guidance` while the text carries the value: it need not show the value, but wherever it shows the target attribute it must match the value, never another one.
- **For `carrier=context`** (a distractor photo) depict the scene material only, never any current target preference.

Use an ordinary photo unless the scenario itself concerns artwork. Do not depict the persona's face: use their first-person view or objects/other people needed by the scenario. Do not invent geographic locations for images; in particular, do not automatically place wildlife in the persona's country. The user may be viewing photos or livestreams from elsewhere when that fits the selected scenario. Do not put explanatory labels or preference statements inside images.

- Sessions of one explicit/implicit unit must be distinct events jointly supporting the same inclination; evolving sessions must clearly make the newer value supersede the earlier one through a credible life change.
- Write all linked sessions coherently, respecting their timestamps.
- Distractors should discuss the selected everyday situation neutrally: no assertion of any of the supplied current target preferences or conflicting inclinations.
- Other preferences are context for avoiding contradictions, not instructions to plant extra evidence.
- **Do not call tools or delegate.**
- If repair feedback is present, fix its actual defects while keeping all fixed plans and target values.
- **Return only the schema JSON.**

Auditing Interaction Sessions

- Independently audit these generated histories against their supplied persona, selected scenarios/templates, scopes and values.
- Return honest failures; a well-formed fail is a **VALID** audit result.
- **Do not rewrite dialogue or force a pass.**
- For every applicable check cite at least one **EXACT** case-sensitive quote with `session_id/message_id`.
- Entailment/stability must cite USER evidence; assistant suggestions are not user preferences.
- Entailment means the intended dimension/value and scope are clear; for image-carried values inspect the image prompt contract only (actual pixels are independently audited by the renderer and not supplied here).
- **Never claim to have seen pixels.**
- Stability requires generalizable inclination, not merely observing a one-off object.
- Check template expression fidelity, including no explicit preference announcement in implicit sessions.
- `Image_value_leakage` must catch synonyms/translations of the value anywhere in the corresponding IMAGE-carried session text, without rejecting a valid value in a separate TEXT-carried companion session.
- `Temporal_update` applies to evolving units: older A then newer B, with credible change and current B unambiguous.
- `Multi_session_evidence` applies to explicit/implicit units with more than one session: distinct events, jointly stable same value.
- `Distractor_neutrality` applies to distractors: no target preference/contradiction despite any shared subject words.
- `Target_conflicts` applies to all units, checking other fixed preferences and persona constraints; legitimate old value in an evolving unit is allowed.
- For distractors, `entailment/stability` are `not_applicable`.

Auditing Interaction Sessions

- `Image_value_leakage` is `not_applicable` if no image-carried sessions; `temporal_update/multi_session_evidence/distractor_neutrality` are `not_applicable` when their structure is absent.
- All remaining checks must pass or fail with exact citations.
- Explain repairable failures precisely.
- **Do not use tools or delegate.**

K.2 Prompts in VIMGer

Step 1 of VIMGer: Filtering Irrelevant Interaction Sessions (per history window)

Below is part {part} of {parts} of the long-term conversation history between a user and their personal AI assistant, in chronological order. Sessions are numbered.

{numbered sessions of this window}

New image-generation request from the user: {query}
Which of the numbered sessions above reveal this user's tastes, habits or circumstances that should shape how this image looks? List their numbers.

Target output: `Relevant sessions:` {session numbers} *(or* `Relevant sessions: none`*)*

Step 2 of VIMGer: Personalized Visual Imagination

Below are the past conversation sessions between a user and their personal AI assistant that reveal the user's preferences relevant to a new request, in chronological order.

{retained sessions with their shared images; if none: `(No earlier session is relevant to this request.)`}

New image-generation request from the user: {query}
Write one self-contained, detailed prompt for a text-to-image model that fulfils this request the way this user would want it: keep everything the request states, and settle every visual choice it leaves open from the user's preferences above. When a preference changed over time, follow the most recent one. Describe only the image.

Constructing the Oracle Specified Request P_u^* (`GPT-5.6-Sol`)

System prompt:

You write prompts for a text-to-image model. You receive one user's image-generation request and the complete list of visual choices this particular user wants for it, as '`dimension: value`'.

Write **ONE** self-contained prompt that fulfils the request and realises every listed value clearly and unambiguously, so that a viewer could check each value in the finished image:

- Keep the requested subject and everything the request itself states.
- Weave all values into one coherent scene in natural language. Use each value's own wording or an unmistakable paraphrase.
- Add no other specific visual choices beyond what a coherent image needs.
- Do not mention the user, preferences, dimensions, or these instructions.

40 to 90 words. Output only the prompt.

User message:

Request: {query}
Visual choices:

- {dimension}: {value}

...

Reward Judge for R_{imagine} (`GPT-5.6-Sol`)

System prompt:

You check whether an image-generation prompt specifies a list of required visual values. For each numbered value, answer "`yes`" if the prompt explicitly asks for that value or an unmistakable equivalent of it, and "`no`" if the prompt omits it, leaves it vague, or asks for something different.

Return only a JSON object: {"`answers`": ["`yes`" | "`no`", ...]} with one entry per value, in order.

User message:

Required values:
1. {dimension}: {value}
...

Prompt to check:
{specified request}

Reward Judge for R_{realize} (`GPT-5.6-Sol`)

System prompt:

You are a strict visual judge. You see one image and a numbered list of required visual values ('`dimension: value`'). For each value, answer "`yes`" only if the image clearly shows it, and "`no`" if it is absent, ambiguous, or contradicted.

Return only a JSON object: {"`answers`": ["`yes`" | "`no`", ...]} with one entry per value, in order.

User message:

{rollout image}

Required values:
1. {dimension}: {value}
...

K.3 Prompts in Baselines and Evaluation

Personalized Prompt Writing in the Two-Stage Baselines

System prompt:

You are a personal multimodal AI assistant with a long-term memory of your conversations with one user. You will receive your full interaction history with this user (their messages and the photos they shared, in chronological order), followed by a new image-generation request from them.

Write **ONE** self-contained prompt for a text-to-image model that fulfils the request for this particular user:

- Keep the requested subject and everything the user explicitly specifies in the request.
- The request leaves most visual choices open. Resolve each of them the way this user would most likely want, based on what their history reveals about their tastes, habits and circumstances — stated outright, implied, or visible in their photos. When a preference changed over time, follow the most recent one.
- Do not mention the user, the history, or your reasoning; describe only the image to be generated.

Output only the image prompt.

User message:

{entire interaction history with its shared images}

New image-generation request from the user:
{query}

Context of the BAGEL History and RAG Baselines

History:

Below is the full interaction history between a user and their personal AI assistant, in chronological order.

{full interaction history with its shared images}

The user now asks the assistant to generate an image for them. {query}

RAG:

Below are past conversation sessions between a user and their personal AI assistant that may reveal the user's preferences, in chronological order.

{top-8 retrieved sessions with their shared images}

The user now asks the assistant to generate an image for them. {query}

Retrieval instruction for requests:

Given a user's image-generation request, retrieve past user-assistant conversation sessions that reveal the user's visual preferences relevant to it.

Retrieval instruction for sessions:

Represent this user-assistant conversation session for retrieval.

VQA Judge of VIMG-Score

System prompt:

You are a strict visual question answering judge. Look at the image and answer the question about it.

User message:

{generated image}

{question}

Answer with exactly one word: Yes or No.