# PIPBench: A Profile-Inclusive Framework for Personalized Image Generation Evaluation

Yuhang Wu[1], Shuxiang Zhang[2], WEE HIAN CHING[1], Chi Zhang[1,3], and Miao Liu[1,†]

[1] College of AI, Tsinghua University, Beijing, China
[2] Sun Yat-sen University, Guangzhou, China
[3] Shanghai Qi Zhi Institute, Shanghai, China

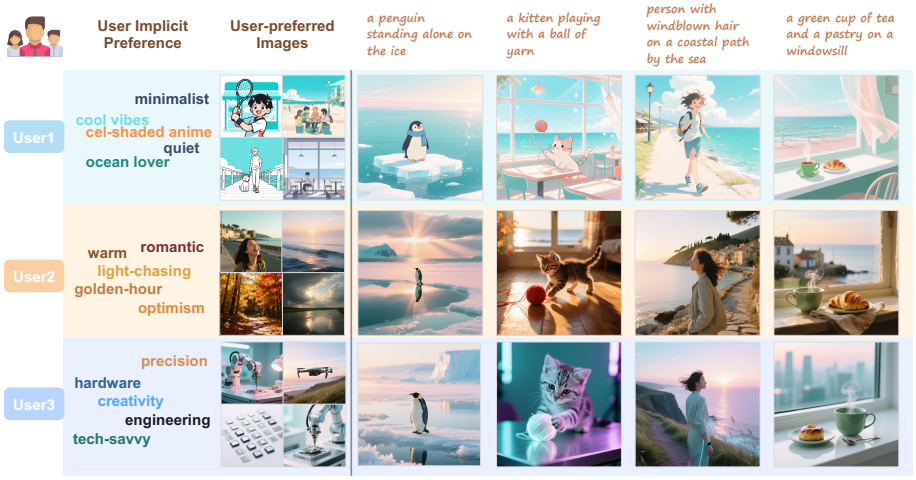


**Fig. 1: Problem setup of personalized image generation and examples from our PIPBench**. Our benchmark incorporates both user profiles and preferred images, enabling a comprehensive analysis of implicit and explicit visual aesthetic preferences.

**Abstract.** Recent text-to-image models such as DALL·E-3 excel at following diverse prompts yet remain blind to individual aesthetic preferences. We study personalized image generation, where models must align outputs with a user's implicit visual preferences based on a few historically preferred images and a short prompt. To this end, we introduce PIPBench, the first profile-inclusive benchmark for evaluating personalized image generation. We further propose a novel data construction pipeline that leverages psychological and demographic profiling dimensions for both real-user data collection and scalable agent-based data generation. Using PIPBench, we conduct a thorough evaluation of representative line of methods. Our experiments reveal key

[†]Corresponding author.

limitations in existing methods, suggesting new challenges and opportunities for personalized text-to-image synthesis. Project page: https://wuyuhang05.github.io/PIPBench/

**Keywords:** Personalized generation · Image generation · Benchmark

## 1 Introduction

Recent text-to-image generation models, such as DALL·E 3 [3], have attracted widespread attention. However, these systems [3, 10, 35] remain largely task-driven: they readily follow explicit user instructions, yet adopt a generic generative model without accounting for implicit individual preferences.

Consider the examples in Fig. 1, a typical user prompt conveys only the core concept. Existing systems first expand such prompts using large language models (LLMs), and then generate images based on enriched descriptions. This two-stage procedure introduces an enormous search space due to the stochasticity of both language enrichment and image synthesis. However, each user may resonate with only a small portion of this exploration space, resulting in a high interaction barrier for non-expert users, since personal preferences are often implicit and difficult to express through plain language.

An alternative paradigm for personalized text-to-image generation is illustrated in Fig. 1. Specifically, the generation model leverages historical user preference data to produce results that are better aligned with the user's aesthetic intent. This setting relates to prior work on style transfer and image editing, which align visual synthesis with visual or textual guidance. However, these methods do not consider user aesthetic preference modeling.

Notably, only a handful recent studies [16, 30, 38] have examined image generation conditioned on user preference data, largely due to the absence of a systematic benchmark with high-quality user data. Existing datasets [6, 7, 13] that include user interaction histories are largely domain-specific, focusing on applications such as sticker or poster preferences. In contrast, for general personalized image generation, individual implicit visual preferences are inherently difficult to articulate and quantify. For example, a user may consistently favor images depicting open natural landscapes over enclosed urban scenes, without being consciously aware of this tendency. Such inclinations are challenging to collect, as they are diverse and subjective, manifested through behavioral or profiling patterns [4, 19] rather than explicit statements. However, user profiling information is often overlooked in existing image generation works, leaving these implicit preference signals underrepresented.

To bridge this gap, we introduce **PIPBench**, the **P**rofile-**I**nclusive **P**ersonalized Image Generation **Bench**mark. Drawing on established psychological literature, we first formalize profile axes that implicitly shape visual preferences. These axes guide the design of a structured survey for real-user profile collection and enable the scalable construction of synthetic agents to diversify benchmark data distribution and facilitate large-scale training data generation. We then generate candidate image sets for preference selection by prompting Large Language Models

(LLMs) to produce profile-conditioned captions, followed by image generation. We evaluate representative prior methods on PIPBench, explore training-based methods for preference conditioning, and validate the benefits of our profile-inclusive framework. Our contributions are summarized as follows:

- We introduce the first personalized image generation benchmark, featuring real user profiles paired with their corresponding preferred images.
- We propose a novel agent-based pipeline for diverse benchmark data construction and scalable training data generation.
- We conduct comprehensive analyses of prevailing personalized image generation methods, providing insights into their strengths and limitations.

## 2 Related Work

**Personalized Modeling and Generation**. Our work is most closely related to existing studies on personalized generation. Prior research has formed two paradigms. The first is the test-time fine-tuning paradigm: optimizing model parameters specifically for each user or concept can achieve strong personalized effects, but it is accompanied by significant computational and storage overhead, making it difficult to scale in large-scale service scenarios [27]. The other, more relevant paradigm leverages conditional projection to extract conditioning information from reference images using foundation models [7, 16, 18, 21, 24, 30, 38]. Salehi *et al*. [30] proposed to mine core features of user styles from reference images via contrastive learning, converts them into conditional vectors, and injects them into the cross-attention modules of generative models to achieve personalized binding of style and content; Li *et al*. [18] employs multimodal large language models to mine global preference signals from reference images, enabling personalized generation without additional training. Kim *et al*. [16] incorporates user interaction history into condition modeling in the latent space through Transformer adapters, enabling adaptation to mainstream foundation models without additional fine-tuning. Chen *et al*. [7] leverages historical user prompts to personalized prompt rewriting via LLMs. Importantly, existing works have not addressed the implicit user preferences embedded in user profiles, primarily due to the limitations of existing benchmarks.

**Image Generation Benchmarks**. Existing evaluation systems for personalized generation are mainly constructed around fidelity, alignment, and semantic consistency, and have played an important role in object-level personalized assessment. DreamBench++ [25] introduces a personalized image-generation benchmark that uses a multimodal GPT to automatically align with human preferences, aiming to improve the agreement between automated evaluations and human subjective judgments. Meanwhile, some studies have begun to introduce subjective metrics such as user satisfaction to supplement objective evaluations [17, 20, 29, 36]. Xu *et al*. [36] constructs the ImageRewardDB dataset, which includes a substantial number of expert-annotated image comparison pairs and is used to train human preference reward models for text-to-image generation. Kirstain *et al*. [17] constructs a large-scale open dataset named Pick-a-Pic,

which collects real preference feedback from image generation enthusiasts via a dedicated web application and includes a vast number of image comparison pairs generated based on diverse text prompts. To specifically evaluate personalized generation capabilities, Chen *et al*. [7] propose a history-based benchmark that takes a pool of user historical prompts as input to assess how well models align with individual intentions. Salehi *et al*. [30] introduce the Viper data set, which uses user comments alongside reference images to systematically measure a model's ability to capture user-specific stylistic preferences. These previous benchmarks largely rely on synthetic data or isolated contextual cues—such as relying solely on a pool of prompts, scattered user comments, or specific reference images. Even the few datasets that incorporate real user feedback typically lack comprehensive demographic information and holistic user profiles, resulting in limited controllability over data quality and a failure to capture implicit, multi-dimensional user preferences. We provide a detailed comparison between our benchmark and existing ones in Tab. 5 of the Supplementary Material.

**Textual Inversion and Style Transfer**. Previous studies on text embedding adaptation and diffusion-based style transfer laid the foundation for personalized image generation [8, 11, 26, 27, 29, 40]. Text embedding methods achieve the encoding of target styles or objects by learning compact word vectors from a small number of reference images to represent specific concepts. Ramesh *et al*. [26] proposes a hierarchical text-conditional image generation method that captures the target style and semantics by leveraging CLIP's pre-trained representations to achieve image generation with high diversity and high photorealism ; diffusion-based style transfer accomplishes pixel-level style alignment through mechanisms such as cross-attention. Chung *et al*. [8] proposes a training-free diffusion-based style transfer method, achieving an accurate fusion of content and style by replacing self-attention key-value pairs. Zhang *et al*. [40] Combines pre-trained diffusion models with an implicit style prompt bank, realizing high-quality artistic style transfer while preserving content structure. However, these methods typically focus on a single concept or style, struggling to effectively integrate cross-image co-occurrence cues present in multiple reference images, thus having significant limitations in capturing implicit user preferences.

## 3 Benchmark Construction Pipeline

We adopt a synthetic–real hybrid design, where high-cost real-user data collection provides reliable evaluation resources, while a scalable agent-based engine both enriches benchmark diversity beyond the collected data and enables large-scale training data generation. A visual illustration of our benchmark construction process is provided in Fig. 2. Specifically, we design profiling axes that encode implicit visual preferences. Guided by these axes, we collect real-user profile data and construct synthetic agents. We then generate profile-conditioned candidate image sets. The final benchmark includes real-user selections as well as a synthetic component constructed via automatic ranking.

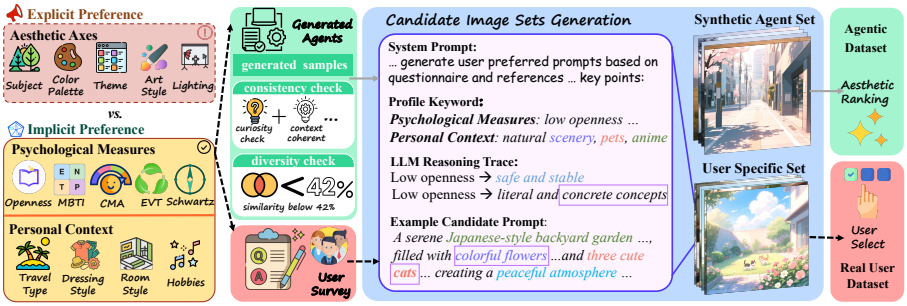


**Fig. 2: Overview of the PIPBench data collection pipeline**. We first define base user profiles using psychological measures and personal context. Real user profiles are collected through surveys, while synthetic agent profiles are generated by sampling from the profiling schema, followed by rigorous consistency and diversity checks. We then prompt Large Language Models (LLMs) to generate profile-aligned image prompts by reasoning over the relationship between latent user profiles and explicit aesthetic dimensions. Finally, given candidate image sets, preferred images selected by real users form a human-calibrated Real-User Dataset, while a synthetic Agentic Dataset is constructed automatically via aesthetic-score ranking within each set.

### 3.1 User Profile Definition

Unlike explicit aesthetic attributes, implicit visual preferences are inherently shaped by personal experience. Notably, such preferences are abstract and diverse, making them difficult to reliably articulate or systematically collect through direct surveys or discrete selection tasks. Nevertheless, established psychological studies suggest that such implicit inclinations are systematically correlated with individuals' psychological traits [19] and contextual backgrounds [4].

**Psychological Measures**. User visual preferences arise from a combination of personality traits and individual dispositions that jointly shape aesthetic judgment. Based on a comprehensive review of the psychological literature, we consider the following psychological measures:

- **Big Five Model** [5,33]. Defines personality along five dimensions: *Openness, Conscientiousness, Extraversion, Agreeableness,* and *Neuroticism*. Among these, **Openness** is most relevant to visual preference [5], reflecting imagination, aesthetic sensitivity, and receptivity to novelty. Individuals high in Openness favor abstract and unconventional imagery, whereas those low in Openness prefer realistic, structured compositions.
- **Ten-Item Personality Inventory (TIPI)** [12, 12]. A concise measure of the Big Five traits. We use it to estimate **Openness**, where higher scores indicate curiosity and preference for visual complexity and novelty.
- **Schwartz's Basic Human Values** [31]. Defines ten motivational values shaping human goals and artistic taste. High *Self-Direction* or *Stimulation* aligns with innovative, exploratory aesthetics, while high *Tradition* or *Security* corresponds to classical and balanced compositions.

- **Ecological Valence Theory (EVT)** [23]. Explains color preference through affective associations with real-world experiences. Warm palettes evoke comfort and nostalgia, while cool tones convey calmness and clarity.
- **Circumplex Model of Affect** [28]. Describes emotion along **Valence** (pleasant–unpleasant) and **Arousal** (activated–deactivated). Positive, high-arousal states promote preference for vivid, dynamic imagery, whereas calmer moods lead to softer, minimalist visuals.

**Personal Context**. In addition to psychological attributes, we collect a set of personal context variables, encompassing individual background and daily life factors that may shape aesthetic formation. We collect information on participants' academic background, which may influence cognitive and aesthetic styles across domains such as Arts, Humanities, STEM, and Business. We also include living styles to capture how lifestyle and identity relate to aesthetic tendencies—for example, artistic individuals often prefer creative or vintage visuals, whereas minimalist users favor clean and balanced compositions. To account for visual exposure and subcultural influence, we record participants' digital life and interests, including major platforms, content preferences, and hobbies. Finally, clothing and fashion preferences are collected, as dressing style often reflects aesthetic inclinations through preferred colors, textures, and compositions. Together, these personal context variables complement psychological factors to form a holistic profile of user visual preference. Details on the psychological measures, personal context variables, illustrative examples, and the survey design are provided in the Supplementary Material Sec. G $\sim$ I.

### 3.2 User Profile Collection and Agent Construction

**Real-user Profile Collection**. We collect demographic and psychological preference data through a 19-item online questionnaire derived from aforementioned profiling schema, including an additional item requesting users to upload their personally preferred images. After quality control and filtering, this stage yields 134 valid responses. The textual responses and uploaded images are then parsed and canonicalized to produce standardized user profiles.

**Synthetic-Agent Construction**. While the real-user dataset provides high-fidelity preference signals, collecting such data is inherently costly and difficult to scale. Notably, 68 out of 134 users who provided profiling data declined to participate in the subsequent stage requiring final preference selection. Moreover, we observe that the collected user profile distribution is highly skewed, likely due to the targeted outreach of the online survey. We refer readers to the Supplementary Material for additional details. To this end, we introduce a synthetic agent-based construction process that enhances benchmark diversity while enabling scalable training data generation for learning-based methods.

To generate scalable, self-consistent synthetic agents, we encode the user profile definitions into a unified schema and generate agent candidates by stochastically sampling from it. As outlined in Algorithm 1, each sampled candidate is validated against a consistency rule set $R$ containing hard and soft rules.

**Algorithm 1** Agent Dataset Construction

**Require:** Rule set $R = \{(r_i, \text{type}_i)\}_{i=1}^{M}$, where $\text{type}_i \in \{\text{hard}, \text{soft}\}$; thresholds $\tau_{\text{conf}}, \tau_{\text{jac}}$; target size $T$; maximum attempts $A_{\max}$;
**Ensure:** $D_{\text{agents}}$

Initialize $D_{\text{agents}} \leftarrow \emptyset$, $attempts \leftarrow 0$
**while** $|D_{\text{agents}}| < T$ **and** $attempts < A_{\max}$ **do**
  $attempts \leftarrow attempts + 1$; $x \leftarrow$ Sample(schema); $penalty \leftarrow 0$
  **for all** $(r_i, \text{type}_i) \in R$ **such that** $r_i(x) = \text{False}$ **do**
    **if** $\text{type}_i = \text{soft}$ **then** $penalty \leftarrow penalty + 1$ **else continue while**
  **end for**
  $sim\_max(x) \leftarrow \max_{s \in \text{Sig}(D_{\text{agents}})_{\text{last 200}}} \frac{|\text{Sig}(x) \cap s|}{|\text{Sig}(x) \cup s|}$
  **if** $(penalty \leq \tau_{\text{conf}})$ **and** $sim\_max(x) \leq \tau_{\text{jac}}$ **then**
    $D_{\text{agents}} \leftarrow D_{\text{agents}} \cup \{x\}$
  **end if**
**end while**
Assign IDs, export, and **return** $D_{\text{agents}}$

Hard rules correspond to logical contradictions that trigger immediate rejection, while soft rules penalize a confidence score. Specifically, the rules check: (A) ratings coherence: penalizing internally inconsistent personal context; (B) value antagonisms: detecting conflicting personal values based on curiosity, rigidity and excitement; (C) cross-domain coherence: penalizing aesthetic or behavioral mismatches across lifestyle and visual preferences; and (D) usage–behavior alignment: penalizing claims of low AI use paired with strong visual consumption patterns. Finally, candidates are filtered using a Jaccard-based diversity criterion against the signatures of previously accepted agents to ensure population heterogeneity. We refer readers to the Supplementary Material Sec. J∼L for detailed rule implementations, prompt templates, and examples.

### 3.3 Benchmark Construction

**Candidate Image Set Generation**. Since human implicit visual preferences are intrinsically diverse, highly individualized, and impossible to adequately encapsulate within fixed explicit axes, we design a chain-of-thought prompting strategy that leverages LLMs to reason from user profiles and construct a plausible, multi-dimensional sampling space for each user. Both real and synthetic profiles—spanning disciplinary background, cognitive style, openness, emotional state, and value orientation—are encoded in structured natural-language templates. Based on these profiles, our prompt guides the LLM through a three-stage deduction: (1) synthesizing demographic and lifestyle attributes to construct a holistic personal narritive; (2) utilizing psychological frameworks (*e.g.*, TIPI, Schwartz Values, EVT, CMA) as associative priors to build a high-dimensional space of implicit cognitive and emotional traits and (3) integrating these naunced signals to generate an unbounded, profile-conditioned sampling space for image generation prompts. We parse these outputs, remove redundancies, and compile

up to 20 enriched, stylistically coherent prompts per user, which are then used to generate a pool of candidate images via a diffusion model like Qwen-Image [35]. **Preference Image Selection**. For each real-user profile, we present the candidate image sets to participants, who select 6–8 images that best match their personal preferences. Participants may also optionally upload additional preferred images. After screening and quality control, this process yields a final set of 76 calibrated user records, each comprising the raw questionnaire responses, standardized profile, and human-selected preferred images. Additional details are provided in the Supplementary Material Sec. H. For synthetic agents, the candidate images are automatically ranked and selected based on aesthetic scores.

### 3.4 Benchmark Statistics

For every synthetic agent or real user, we construct standardized evaluation tuples. Given the agent/user's final preference pool, we sample $K \leq 5$ preferred images as the reference set $\{\mathcal{I}_k\}_{k=1}^{K}$. We then select another image from the same pool to serve as the ground-truth target $\mathcal{Y}_{\text{gt}}$, and employ a Vision-Language Model (VLM) to generate a short caption $\mathcal{X}$ focusing strictly on its primary visual content. An evaluated model receives $(\mathcal{X}, \{\mathcal{I}_k\}_{k=1}^{K})$ and must generate an image $\mathcal{Y}$ that aligns with both the prompt and the underlying preferences.

Following our pipeline, the consolidated benchmark comprises 1,369 testcases constructed from 1,876 images collected from 251 agents/users. The synthetic agent subset contributes 719 testcases (1,231 images, 175 agents), while the real-user subset contributes 650 testcases (645 images, 76 users).

## 4 Experiment

### 4.1 Problem Setting

Given a short image generation prompt $\mathcal{X}$ and a set of historical user-preferred images $\{\mathcal{I}_k\}_{k=1}^{K}$, we seek to generate an output image $\mathcal{Y}$ that is faithful to $\mathcal{X}$ while reflecting the user's visual preference. Formally, the target conditional distribution can be formulated as $p_\theta\big(\mathcal{Y} \mid \mathcal{X}, \{\mathcal{I}_k\}\big)$

**Insights on Problem Formulation**. Some existing works approach this problem by designing empirical proxies for the joint conditional distribution. In our experiments, we focus on four primary paradigms. First, we consider *test-time tuning* approach like DreamBooth [27], where a user-specific training strategy $f(\theta_u, \{\mathcal{I}_k\})$ is applied so that image sampling proceeds from $p_{f(\theta_u, \{\mathcal{I}_k\})}(\mathcal{I} \mid \mathcal{X})$. Second, we examine methods that directly generate images using a *joint condition* formulation $p_\theta\big(\mathcal{Y} \mid \mathcal{X}, \{\mathcal{I}_k\}\big)$. Third, we evaluate approaches that employ a *condition fusion* module to jointly summarize $\{\mathcal{I}_k\}$ and $\mathcal{X}$ into a unified conditioning representation $q_\xi(\mathcal{X}, \{\mathcal{I}_k\})$, and subsequently adopt the generation process using $p_\theta(\mathcal{Y} \mid q_\xi(\mathcal{X}, \{\mathcal{I}_k\}))$. Finally, we include *separate condition* approach that generate images based on decoupled pathways $p(\mathcal{Y} \mid \mathcal{X})$ and $p(\mathcal{Y} \mid \{\mathcal{I}_k\})$.

### 4.2 Metrics

To ensure reproducible evaluation of our benchmark, we adopt a hybrid protocol. Although automated metrics facilitate efficient model development, they often fall short in capturing the subjective nuances of personalized preferences. Consequently, as a complementary component, we introduce a persona-aware Elo rating system that employs LLMs as proxy judges to better assess the alignment between the generated images, user-specific preferences, and textual prompts.

**Automatic Metrics**. Following previous work [16, 38], we evaluate instruction fidelity via CLIP text-image similarity (**CLS-T**) [14], which quantify the semantic alignment between the generated image $\mathcal{Y}$ and the user's short prompt $\mathcal{X}$. To assess how well $\mathcal{Y}$ incorporates visual conditions from the user's reference image set $\{\mathcal{I}_k\}$, we report the average CLIP image-image similarity (**CLS-R**), DINO similarity (**DIS-R**) [22], and LPIPS (**LPIPS-R**) [39]. Note that we adopt DINO and LPIPS, as our personalized generation task extends beyond standard style transfer to capture preference-relevant background semantics, structural integrity, and latent aesthetic styles. Higher CLS/DIS and lower LPIPS scores denote superior alignment. For clarity, CLS and DIS are scaled by a factor of 100. We also calculate their grayscale variants in Supplementary Material Sec. C.

**Persona-aware Elo Rating**. Since real-world user preferences are inherently diverse and non-singleton, relying solely on image-based automated metrics is insufficient. Therefore, we adopt an arena-style [15] evaluation using **LLM-as-a-judge** to conduct pairwise comparisons on our Real-User dataset, ensuring reproducible assessment. The judge is provided with the complete **user profile** and explicitly instructed to act as that specific user. Operating under this assigned persona, the evaluator assesses two generated images and judges which one better aligns with the user's underlying aesthetic preferences.

To ensure fairness and mitigate position bias, we randomly swap the presentation order of the two evaluated images. Furthermore, to avoid self-enhancement bias, we employ multiple independent judge models, including GPT-5 [32], Gemini 2.5 Pro [9], and Qwen3-VL [1]. Following [37], we process each committee member's decision as an independent pairwise evaluation, expressly avoiding the information loss associated with majority voting. To validate the reliability of these persona-aware judging mechanism, we compare a sampled subset of the committee's results against human annotations. This comparison yields an agreement rate of $\sim$ 91%, demonstrating that our auto-judge serves as a highly consistent proxy for reproducible, human-aligned preference evaluation. Based on these fine-grained pairwise judgments, we compute an **Elo rating** for representative methods to quantify their capacity to capture individual tastes.

To further verify the reliability of the auto-ELO score, we conduct a user study. We select the top four models according to the ELO results in Table 1 and sample 100 cases for human evaluation. For each case, human annotators are given the user profile and reference image, and asked to select the best result among the four model outputs. The ranking is broadly aligned with our auto-ELO evaluation, suggesting that the persona-aware auto-judge provides

**Table 1:** Results of representative methods on our PIPBench. Best and second-best scores across conditioning methods are shown in **bold** and underlined, respectively.

| Model | Synthetic Agent | | | | Real-User | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | CLS-T ↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R ↑ | CLS-T ↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R↑ | Elo↑ |
| no-preference | 32.461 | 0.7559 | 62.784 | 10.195 | 32.677 | 0.7448 | 62.761 | 12.099 | 1427 |
| *Test-time Tuning* | | | | | | | | | |
| DreamBooth | **31.693** | 0.7374 | 63.444 | 10.705 | **31.756** | 0.7265 | 63.751 | 12.720 | 1452 |
| *Joint Conditioning Models* | | | | | | | | | |
| Qwen-Image-Edit (1-Ref) | 30.614 | 0.7326 | 64.615 | 15.461 | 31.048 | 0.7209 | 65.802 | 18.459 | 1521 |
| Qwen-Image-Edit (2-Ref) | 30.753 | 0.7652 | 62.319 | 12.259 | 30.559 | 0.7511 | 65.023 | 16.821 | 1354 |
| *VLM Conditioning Fusion* | | | | | | | | | |
| GPT-5 | 30.724 | **0.6977** | **68.119** | 17.085 | 30.550 | **0.6867** | **69.574** | **22.160** | **1765** |
| Gemini2.5-Pro | 30.621 | 0.7038 | 67.335 | 14.403 | 30.403 | 0.6910 | 69.174 | 20.090 | 1615 |
| QwenVL2.5-7B | 30.496 | 0.7434 | 64.639 | 13.313 | 30.510 | 0.7320 | 65.460 | 17.275 | 1445 |
| QwenVL2.5-32B | 30.937 | 0.7376 | 65.252 | 13.589 | 30.915 | 0.7229 | 66.368 | 17.322 | 1477 |
| QwenVL2.5-70B | 29.850 | 0.7183 | 65.894 | **17.336** | 29.797 | 0.7092 | 66.835 | 20.282 | 1531 |
| *Separate Conditioning* | | | | | | | | | |
| Fabric | 31.131 | 0.7277 | 64.531 | 13.367 | 31.213 | 0.7222 | 64.590 | 15.463 | 1412 |

a reasonable proxy for human preference while enabling efficient benchmark-scale evaluation. We provide implementation details in the Supplementary Material Sec. B.

The automatic metrics and the persona-aware Elo ratings in Tab. 1 indicate that automatic metrics closely align with their overall Elo rankings, with top-performing models like GPT-5 consistently achieving both the highest Elo ratings and the most best automatic scores. This strong correlation demonstrates that our proposed hybrid evaluation framework provides a highly accurate, reliable, and comprehensive assessment of personalized image generation quality.

### 4.3 Analysis of Preference Conditioning Methods

We denote the baseline Qwen-Image model without any preference conditions as **no-preference**. For *test-time tuning* approaches, which explicitly adapt the model parameters to a user's specific reference set prior to inference, we evaluate the widely adopted DreamBooth [27] framework on Qwen-Image [35]. For joint conditioning, we evaluate Qwen-Image-Edit [35] with randomly sampled references from $\mathcal{I}_k$. The (1-Ref) variant uses one reference image as guidance, while the (2-Ref) variant uses two references for joint conditioning.

For *condition fusion* methods, we examine different VLMs that enrich the short caption $\mathcal{X}$ based on the reference images $\{\mathcal{I}_k\}$. The expanded description is then fed to the same image generation model Qwen-Image [35]. Specifically, we consider the following VLMs for condition fusion, including proprietary models such as GPT-5 [32] and Gemini 2.5-Pro [9], as well as open-source Qwen models [2] of different scales. For fair comparison, all models adopt the same prompt, which is disclosed in the supplementary material Sec. D. Finally, for *separate condition* modeling, we only evaluate Fabric [34], as few studies have explored this setting for personalized image generation.

**Table 2:** Results on our benchmark using different generation backbones. We use GPT-5 as the conditioning fusion method.

| Model | Synthetic Agent | | | | Real-User | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | CLS-T ↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R ↑ | CLS-T ↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R↑ | Elo↑ |
| GPT-5 + Qwen-Image | 30.724 | 0.6977 | 68.119 | 17.085 | 30.550 | 0.6867 | 69.574 | 22.160 | 1603 |
| GPT-5 + Flux | 30.740 | 0.7189 | 67.520 | 15.880 | 30.287 | 0.7097 | 68.735 | 20.784 | 1503 |
| GPT-5 + SDXL | 30.510 | 0.7377 | 66.473 | 14.390 | 29.987 | 0.7331 | 67.951 | 18.690 | 1394 |

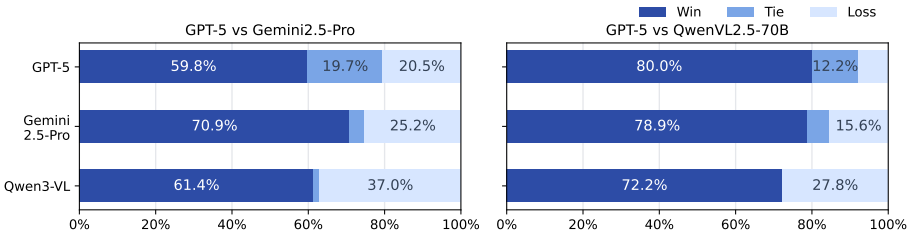


**Fig. 3:** Pairwise preference evaluation results. Win, tie, and loss rates of GPT-5 compared to Gemini 2.5 Pro (left) and QwenVL2.5-70B (right), as evaluated by different VLM judges (Qwen3-VL, Gemini 2.5 Pro, and GPT-5).

**The Effect of Preference Conditioning**. As shown in Tab. 1, most existing methods that incorporate reference images achieve better preference alignment performance compared with the no-preference baseline. Interestingly, results from joint conditional models further show that providing more reference images can actually degrade performance. This is primarily because even state-of-the-art image editing and generation models struggle to jointly interpret multiple reference images, highlighting the need to develop a new class of methods for personalized image generation. Furthermore, it is important to note that high **CLS-T** scores are not indicative of preference alignment. We therefore use CLIP-based instruction fidelity primarily as a sanity check and a trade-off diagnostic: the ultimate goal is to improve preference alignment while keeping the CLIP drop within a modest range of the no-preference upper bound, avoiding both prompt neglect and uncontrolled over-personalization.

For VLM condition fusion models, proprietary models demonstrate a stronger ability to convert multi-image inputs into coherent textual descriptions that capture users' visual preferences, thereby yielding superior preference alignment performance. In addition, model scaling shows notable benefits, as evidenced by the results on the Qwen-series VLMs, primarily due to the improved multi-image reasoning capability of larger models.

We note that the separate-condition method, Fabric, is built upon a much weaker image generation backbone. We include this method primarily for benchmarking completeness, and its lower performance should not be interpreted as evidence that separate-condition modeling is ineffective. However, it is important

to highlight that separate-condition designs introduce additional computational overhead, making them difficult to adapt to heavy generation backbones.

**Comparison Between Synthetic and Real Data**. A interesting finding is that aligning preferences for real users is easier than aligning those for synthetic agents. This is because participants in our user study reflect the prevailing preferences commonly seen on the internet, making their visual tastes easier for models to infer. In contrast, the synthetic agents cover a broader and more diverse spectrum of personalities, which makes preference extraction substantially more challenging. These findings highlight the benefits of our data collection strategy: the synthetic and real-user data complement each other under our carefully designed protocol. Moreover, the relative performance gaps among the evaluated methods remain consistent across real and synthetic data, supporting the validity of our synthetic pipeline. Our method offers a novel pathway for studying underrepresented social groups. We compare the demographic distribution of collected users and synthetic agents in the Supplementary Material Sec. I.

**Analysis of Persona-aware Judges**. Fig. 3 details the win, tie, and loss distributions produced by each judge during pairwise comparisons of the leading baselines. This analysis yields two key observations. First, self-enhancement bias remains minimal, as judges do not systematically favor models from their own families, benefiting from our carefully designed prompts. Second, we observe distinct evaluation behaviors under persona-conditioned assessment: GPT-5 tends to be more conservative, producing more tie verdicts, whereas Qwen3-VL is more decisive in issuing win–loss judgments. Using three independent judges therefore enhances the robustness of our evaluation. The full arena results are presented in the Supplementary Materials Sec. B.

**Results across Different Generation Backbones**. To assess how different image generation models affect preference alignment, we evaluate three widely used generation backbones, all paired with the GPT-5–based condition fusion method. The experimental results are presented in Tab. 2. We note that the reference images are generated using Qwen-Image, which gives this model an inherent advantage on automatic metrics. To mitigate this influence, we concurrently employ the profile-based Elo, providing a fairer comparison of the underlying backbone performances. As expected, Qwen-Image achieves the best alignment, followed by Flux. Their larger model capacities provide stronger expressiveness for capturing the nuanced textual descriptions of user preferences.

### 4.4 Optimize Preference Conditioned Generation

Our agent-based data construction process not only complements benchmark data construction, but also provides an effective mechanism for generating training data used to learn preference embeddings. To recap, each agent is associated with multiple preferred images, and we randomly sample from these images to compose the reference set $\{\mathcal{I}_k\}$ and target image $\mathcal{Y}_{\text{gt}}$. We denote the short caption for testing time input as $\mathcal{X}$. And the expanded caption that originally used to generate $\mathcal{Y}_{\text{gt}}$ is denoted as $\mathcal{E}$. We consider the following learnable approaches designed to optimize the encoding of image preferences:

**Table 3:** Results on our benchmark using different learnable methods to encode preference image embeddings. For each column, the best score (without no-preference baseline) is **bold**, and the second-best score is underlined.

| Model | Synthetic Agent | | | | Real-User | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | CLS-T ↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R ↑ | CLS-T ↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R↑ | Elo↑ |
| no-preference | 32.461 | 0.7559 | 62.784 | 10.195 | 32.677 | 0.7448 | 62.761 | 12.099 | 1485 |
| QwenVL2.5-7B | 30.496 | 0.7434 | 64.639 | 13.313 | 30.510 | 0.7320 | 65.460 | 17.275 | 1492 |
| QwenVL2.5-32B | 30.937 | 0.7376 | 65.252 | 13.589 | 30.915 | 0.7229 | 66.368 | 17.322 | 1504 |
| *VLM Instruction Tuning* | | | | | | | | | |
| Qwen2.5VL-7B-finetuned | 31.510 | **0.7002** | 65.283 | 13.590 | 31.588 | **0.6901** | 66.350 | 17.013 | **1597** |
| Qwen2.5VL-32B-finetuned | 31.322 | 0.7081 | 64.834 | **13.652** | 31.396 | 0.6950 | **66.814** | **17.444** | 1570 |
| *Preference Compression* | | | | | | | | | |
| Learnable Compressor | **31.716** | 0.7463 | **65.333** | 12.184 | **31.994** | 0.7270 | 65.264 | 14.528 | 1534 |
| Frozen Compressor | 29.409 | 0.7440 | 62.133 | 10.196 | 29.254 | 0.7248 | 63.612 | 11.883 | 1318 |

**VLM Instruction Tuning**. We fine-tune Qwen2.5-VL-7B and Qwen2.5-VL-32B to generate enriched captions conditioned on both the reference images and the short caption. Concretely, the VLM is optimized to model the joint distribution $p\big(\mathcal{E} \mid \mathcal{X}, \{\mathcal{I}_k\}\big)$. See the Supplementary Materials Sec. D for details.
**Preference Compression**. We first employ a VLM-based preference extractor to summarize a user's reference images into a compact set of preference tokens that encode their visual tastes. A lightweight cross-attention compressor further embed the dense multi-image features into these tokens, which are subsequently injected into the DiT blocks via in-block adapters to guide personalized generation. Since the VLM is already employed for visual reference comparison, we report model performance excluding VLM-based caption enrichment to ensure a fair comparison with respect to computational cost. We refer to the Supplementary Materials Sec. E for the exact compressor design and training details.

The experimental results are presented in Tab. 3. Visual instruction tuning improves overall performance by a small margin across all three levels of data, even though the training data comes solely from synthetic agents. Particularly, the improvement observed on real-user data further supports the validity of our agent construction pipeline. As for the preference compression methods, we find that freezing the VLM and training only the projection layer is suboptimal for encoding visual preferences. When we unfreeze the VLM, the model does learn to better attend to the reference images compared to the no-preference. However, this approach remains less effective compared to prompt re-writing method using VLMs. This observation highlights an exciting research direction focused on developing fully end-to-end models for visual preference learning.

### 4.5 Analysis on Importance of Profile-inclusive Framework

To demonstrate the necessity of our proposed profile-conditioned generation paradigm, we compare our synthetic agent pipeline against a *profile-free* baseline. While our framework constructs preference sets by conditioning LLM agents on rich, holistic user profiles, the baseline method generates data based solely on

**Table 4:** Comparison of VLM instruction-tuning performance on real-user data using Qwen2.5-VL-7B trained on data from the profile-free pipeline versus our pipeline.

| Method | CLS-T↑ | LPIPS-R ↓ | CLS-R ↑ | DIS-R ↑ | win rate ↑ |
|---|---|---|---|---|---|
| profile-free | **32.251** | 0.7192 | 64.596 | 13.574 | 47.6% |
| **profile-inclusive (Ours)** | 31.588 | **0.6901** | **66.350** | **17.013** | **52.4%** |

isolated, manually specified aesthetic tags. For a fair comparison, we generate both benchmark and training datasets using this profile-free pipeline at identical scale and comparable complexity to those constructed by our pipeline.

**Benchmark Data Quality**. By comparing the benchmark datasets generated by the profile-free pipeline and our profile-inclusive pipeline (denoted as Ours), we observe that Ours better reflects the practical complexities of human preferences. Specifically, Ours exhibits lower intra-user similarity (CLIP: **64.18** vs. 65.49; DINO: **12.35** vs. 23.70), indicating that profile-conditioned generation yields more diverse within-user preference sets, rather than aesthetic repetitions. Furthermore, at the inter-user level, Ours demonstrates lower mean Shannon entropy (**0.5811** vs. 0.9919) and higher Silhouette scores for $K \in [2, 50]$. These results indicate improved clusterability, suggesting that our method better captures distinct and diverse individual visual preferences as in real-world scenarios. These quantitative findings are corroborated by an user study assessing the diversity, coherence, and consistency of the reference-target pairs on a 1–5 scale, where Ours achieves a superior score of **3.61** compared to 3.28 for profile-free method. We refer to the Supplementary Material Sec. F for additional details.

**Training Data Quality**. We compare VLM instruction-tuning performance on real-user data using Qwen2.5-VL-7B as the base model, trained separately on the datasets generated by the two pipelines. As shown in Tab. 4, the profile-free pipeline lags behind ours by a notable margin across all visual preference alignment metrics (LPIPS-R, CLS-R, DIS-R) and persona-aware judge. This performance drop suggests that synthesizing training data using only explicit aesthetic tags fails to capture the multifaceted complexity and diversity of real users' visual choices. Note that the explicit model attains a slightly higher text alignment score (CLS-T), likely because it underutilizes visual reference signals and thus behaves closer to a no-preference baseline.

### 4.6 Visualizations

We provide visualizations of personalized image generation in Fig. 4, comparing the no-preference baseline, the Qwen-Image-Edit (1-Ref) *joint-condition* method, the GPT-5–based *preference fusion* method, and the Fabric *separate-condition method*. All preference-aware variants add user-style traits beyond the text input; in (c) and (d) these traits more clearly match the preferred images. However, Qwen-Edit (1-Ref), which relies on a single random reference, can be dominated by that image—for example, in (b) it almost copies the second preferred

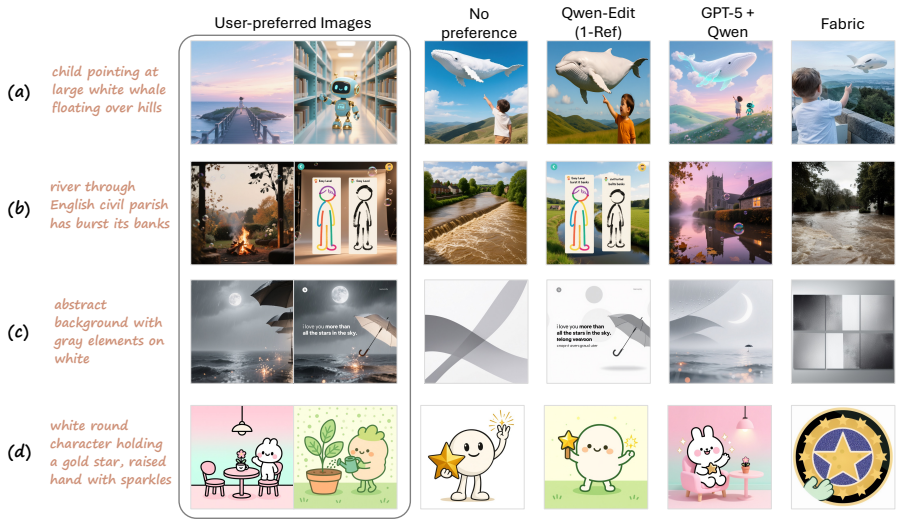


**Fig. 4:** Visualizations of personalized image generation by different methods

image—indicating that one picture is insufficient to represent user preference. GPT-5–based fusion may also over-emphasize preference: in (a), GPT-5 + Qwen adds a robot seen once in the reference set, contradicting the original prompt and illustrating a conflict between preferences following and instruction fidelity.

## 5 Conclusion

In this paper, we introduce PIPBench, the first profile-grounded benchmark designed to assess whether existing image generation models can effectively leverage historical user data for preference-aligned image generation. We propose a novel data collection pipeline that enables high-quality user profile collection and the construction of synthetic agents that emulate human personalities in visual preference. We conduct comprehensive experiments to demonstrate the limitations of existing methods in encoding user preferences and to validate the effectiveness of our proposed pipeline. We believe that PIPBench represents an important step toward user-driven image generation models and highlights promising research opportunities in developing end-to-end approaches for visual preference encoding. In addition, our synthetic data engine and user study protocol offer a meaningful attempt at constructing agents that can facilitate a deeper understanding of underrepresented user groups.

**Acknowledgments**. This work was supported in part by the Zhiyuan Scholar Program of the Beijing Municipal Science and Technology Commission (Z251100008125045), NSFC Grants, and a research grant from the ByteDance Seed Team.

## References


1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 9
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 10
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023) 2
4. Bourdieu, P.: Distinction: A social critique of the judgement of taste. In: Social Stratification, Class, Race, and Gender in Sociological Perspective, Second Edition, pp. 499–525. Routledge (2019) 2, 5
5. Chamorro-Premuzic, T., Reimers, S., Hsu, A., Ahmetoglu, G.: Who art thou? personality predictors of artistic preferences in a large uk sample: The importance of openness. British Journal of Psychology **100**(3), 501–516 (2009). https://doi.org/10.1348/000712608X366867 5
6. Chen, W., Huang, P., Xu, J., Guo, X., Guo, C., Sun, F., Li, C., Pfadler, A., Zhao, H., Zhao, B.: Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2662–2670 (2019) 2
7. Chen, Z., Zhang, L., Weng, F., Pan, L., Lan, Z.: Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7727–7736 (2024) 2, 3, 4
8. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8795–8805 (June 2024) 4
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 9, 10
10. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023) 2
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 4
12. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality **37**, 504–528 (04 2003). https://doi.org/10.1016/s0092-6566(03)00046-1 5
13. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. Acm transactions on interactive intelligent systems (tiis) **5**(4), 1–19 (2015) 2
14. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021) 9

15. Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., Chen, W.: Genai arena: An open evaluation platform for generative models. Advances in Neural Information Processing Systems **37**, 79889–79908 (2024) 9
16. Kim, H., Ahn, S., Seo, Y.D.: Draw your mind: Personalized generation via condition-level modeling in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17171–17180 (2025) 2, 3, 9
17. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 36652–36663. Curran Associates, Inc. (2023) 3
18. Li, Y., Yang, S., Han, X., Wang, W., Dong, J., Lyu, Y., Xue, Z.: Instant preference alignment for text-to-image diffusion models (2025) 3
19. McCrae, R.R.: Aesthetic chills as a universal marker of openness to experience. Motivation and Emotion **31**(1), 5–11 (2007) 2, 5
20. Mo, W., Ba, Y., Zhang, T., Bai, Y., Li, B.: Learning user preferences for image generation model (2025) 3
21. Mo, W., Zhang, T., Bai, Y., Han, L., Ba, Y., Metaxas, D.N.: Prefgen: Multimodal preference learning for preference-conditioned image generation. arXiv preprint arXiv:2512.06020 (2025) 3
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 9
23. Palmer, S.E., Schloss, K.B.: An ecological valence theory of human color preference. Proceedings of the National Academy of Sciences **107**(19), 8877–8882 (2010). https://doi.org/10.1073/pnas.0906172107 6
24. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2085–2094 (October 2021) 3
25. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., han, c., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Representation Learning. vol. 2025, pp. 46010–46032 (2025) 3
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022) 4
27. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023) 3, 4, 8, 10
28. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology **39**, 1161–1178 (12 1980). https://doi.org/10.1037/h0077714 6
29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494. Curran Associates, Inc. (2022) 3, 4

30. Salehi, S., Shafiei, M., Yeo, T., Bachmann, R., Zamir, A.: Viper: Visual personalization of generative models via individual preference learning. In: European Conference on Computer Vision. pp. 391–406. Springer (2024) 2, 3, 4
31. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online Readings in Psychology and Culture **2**(1) (12 2012) 5
32. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025) 9, 10
33. Soto, C.J., Jackson, J.J.: Five-factor model of personality. In: Oxford Bibliographies in Psychology. Oxford University Press (2020). https://doi.org/10.1093/obo/9780199828340-0120 5
34. Von Rütte, D., Fedele, E., Thomm, J., Wolf, L.: Fabric: Personalizing diffusion models with iterative feedback. In: European Conference on Computer Vision. pp. 385–400. Springer (2024) 10
35. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 2, 8, 10
36. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 15903–15935 (2023) 3
37. Xu, Y., Ruis, L., Rocktäschel, T., Kirk, R.: Investigating non-transitivity in llm-as-a-judge. arXiv preprint arXiv:2502.14074 (2025) 9
38. Xu, Y., Wang, W., Zhang, Y., Tang, B., Yan, P., Feng, F., He, X.: Personalized image generation with large multimodal models. In: Proceedings of the ACM on Web Conference 2025. pp. 264–274 (2025) 2, 3, 9
39. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 9
40. Zhang, Z., Zhang, Q., Xing, W., Li, G., Zhao, L., Sun, J., Lan, Z., Luan, J., Huang, Y., Lin, H.: Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. Proceedings of the AAAI Conference on Artificial Intelligence **38**(7), 7396–7404 (Mar 2024). https://doi.org/10.1609/aaai.v38i7.28570 4