

Personalized Reward Modeling for Text-to-Image Generation

Jeongeun Lee¹, Ryang Heo¹, and Dongha Lee^{1*}

¹Yonsei University

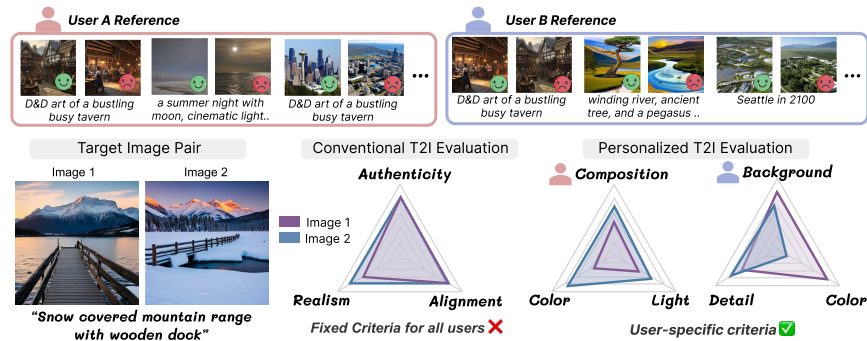


Fig. 1: Motivating the need for user-specific evaluation in personalized T2I assessment. From user histories, we observe that different users favor different visual attributes, suggesting that they prioritize different criteria when judging images. Consequently, applying a fixed, one-size-fits-all criterion space (left) can fail to capture these differences and predict which image a user would prefer, whereas inducing criteria from each user’s history (right) yields personalized evaluation spaces that more reliably predict user preferences. Notably, even when User A and User B prefer the same image, the dimensions they prioritize differ, highlighting that the same choice may be supported by different underlying factors.

Abstract. Recent text-to-image (T2I) models generate semantically coherent images from textual prompts, yet evaluating how well they align with individual user preferences remains an open challenge. Conventional evaluation methods, general reward functions or similarity-based metrics, fail to capture the diversity and complexity of personal visual tastes. We present PIGReward, a personalized reward modeling framework that dynamically generates user-conditioned evaluation dimensions and assesses images through CoT reasoning. To address the scarcity of user data, PIGReward adopts a self-bootstrapping strategy that reasons over limited reference data to construct rich user contexts, enabling personalization without user-specific training. Beyond evaluation, PIGReward provides personalized feedback that drives user-specific prompt optimization, improving alignment between generated images and individual intent. Extensive experiments demonstrate that PIGReward outperforms existing

* Corresponding author.

methods in both accuracy and interpretability, establishing a scalable, reasoning-based foundation for personalized T2I evaluation and optimization.

Keywords: Personalized reward modeling · Text-to-image generation · Visual chain-of-thought reasoning

1 Introduction

Recent text-to-image (T2I) generative models [28, 30, 33, 35] have shown remarkable performance in generating high-quality images from natural language prompts. Moving beyond generating generic, one-size-fits-all images, following works explores *personalization* in T2I generation to better align with individual preferences [8, 55, 62]. These methods typically adopt a history-driven approach [3, 4, 15], leveraging user-specific references (e.g., prior image preferences or interactions) to reflect unique visual tastes.

However, evaluating personalization remains a significant challenge under conventional T2I evaluation methods. Most widely-used approaches are based on reward functions fine-tuned on large-scale general human preference datasets [16, 56, 59]. They reduce human preferences to a single scalar score, failing to consider their inherent complexity. This has motivated multi-dimensional evaluation [21, 58, 69], which introduces a small set of pre-defined dimensions (e.g., Alignment, Detail). More recently, large vision-language models (LVLMs) have been adopted as generative reward models [51, 52, 76], offering interpretable explanations through their strong reasoning capabilities. Yet these approaches still rely on fixed criteria that are uniformly applied across users. In real world, different users may attend to distinct aspects of the same image pair, leading to divergent judgments. A limited set of predefined dimensions cannot fully cover such user-dependent evaluative perspective.

On the other hand, prior personalized T2I works [3, 15, 40, 50, 61] typically adopt three strategies to evaluate personalization. First, they employ statistical metrics such as CLIPScore [11], Fréchet Inception Distance (FID) [12], and Inception Score (IS) [37] to measure pixel- or embedding-level similarity between user reference and generated outputs. These approaches implicitly assume that user alignment can be approximated by visual recurrence. However, personalization is not just reproducing previously liked images, but generalizing underlying preference reflected in them [23, 60]. Second, some methods [15] naively prompt off-the-shelf LVLMs [13] to assess candidates given reference images in a few-shot manner. Lacking an explicit mechanism for user-specific preference reasoning, these approaches often result in heuristic judgments. Third, they conduct human evaluations, which are costly, time-consuming, and difficult to reproduce at scale.

In this work, we propose **PIGReward**, a LVLM-based personalized reward modeling framework for T2I generation. Unlike prior multi-dimensional approaches that rely on fixed dimensions, PIGReward adaptively derives user-tailored dimensions from limited user reference data, facilitating personalized assessment

without additional user-specific training. To model user preference from scarce user data, often limited to only pairwise choices, we introduce a self-bootstrapping preference reasoning process that converts implicit image-level preferences into explicit, language-based rationales, which result in a rich user context. Conditioned on this context, our reward model (1) generates user-conditioned evaluation dimensions and (2) performs chain-of-thought (CoT) reasoning to score along those dimensions with explicit rationales to produce a personalized reward. To support such structured reasoning, we construct a 5K CoT-formatted dataset and distill it into PIGReward via supervised fine-tuning. Thus, PIGReward supports interpretable, user-conditioned T2I evaluation under sparse user feedback.

Further, we demonstrate the potential of PIGReward beyond evaluation by leveraging its feedback for personalized prompt optimization. Unlike prior works [5, 20] that aim to optimize prompts toward general aesthetic quality, PIGReward incorporates individual context and provides user-conditioned signals that reflect individual visual taste. These signals identify preferred–rejected prompt variants, which are then used to fine-tune the prompt model via direct preference optimization (DPO) [31], enabling user-aligned generation.

Through extensive experiments, PIGReward demonstrates its effectiveness on user-level benchmarks, including Pick-a-Pic [16], PIP [3], and PASTA [60]. To further evaluate personalization under more realistic conditions, we introduce **PIGBench**, a benchmark constructed from open-ended, abstract prompts [65] that allow diverse visual interpretations across users. Across all datasets, PIGReward consistently outperforms single-scalar reward models, fixed multi-dimensional approaches, and recent LLM-based judges, achieving state-of-the-art performance in user-conditioned evaluation. These results highlight the importance of dynamically inducing user-specific evaluation dimensions rather than relying on predefined criteria. Moreover, PIGReward substantially surpasses similarity-based reference comparison methods, demonstrating that surface-level resemblance is insufficient for modeling personalized preferences and that structured, context-aware reasoning is essential. Beyond evaluation, we show that PIGReward effectively supports personalized prompt optimization, significantly improving user-aligned generation quality in both automatic and human evaluations.

2 Related Work

Visual Chain-of-Thought Reasoning. LLMs [1, 18, 49, 70] have demonstrated strong performance across diverse multimodal reasoning tasks [17, 57, 61]. As CoT reasoning has significantly improved the performance of LLMs through step-by-step reasoning [54, 63], recent studies [68, 71, 72] have extended CoT reasoning to the vision–language domain, showing benefits across diverse visual reasoning tasks, including mathematics [10] and robotic control [72]. However, a key challenge in applying CoT to vision–language models lies in the scarcity of high-quality, CoT-annotated datasets. To address this, recent works [6, 51, 68]

build data generation pipeline to collect task-specific CoT-formatted datasets and distill VLMs [1, 49] under supervision. Meanwhile, other works have explored LVLMs as generative reward models [19, 77], leveraging explicit rationales [39, 52, 57] for scalable and interpretable evaluation. Incorporating CoT has been shown to further enhance the reliability and robustness of reward modeling [51, 68].

Personalized Text-to-Image Generation. Unlike generic image generation, personalized T2I adapts outputs to user preferences by modeling intent from personal history (e.g., image interactions) [40, 50, 62]. PIP [3] proposes a personalized prompt rewriting method conditioned on historical inputs. ViPer [36] and PPD [4] extract user features, expressed either as natural-language profiles or as hidden-state embeddings from LVLMs [18], and directly inject these features into diffusion models [28]. Recent methods expand to multi-turn settings by giving iterative feedbacks [25, 46, 66] or with natural language instructions through conversation [22, 48].

Personalized Reward Modeling. Preference alignment has been studied for large language models (LLMs), particularly through reinforcement learning from human feedback [43]. Subsequent works use pairwise preference data to align models with human interests [7, 27, 74, 75]. In particular, DPO [31] has recently gained attention, also achieving great adaptation to diffusion models [9, 44, 47]. However, recent studies highlight that human preferences are diverse [2, 42], motivating personalized reward modeling [14, 32]. For instance, GPG [67] and SynthesizeMe [34] introduce natural-language user profiles that encode personal context for conditioning; GPO [73] targets group-level alignment by training an auxiliary predictor to steer models toward specific group preferences; VPL [29] formulates a latent-variable approach to capture multimodal, user-conditioned rewards; and PAL [2] models heterogeneous preferences with a sample-efficient personalized reward framework.

3 Personalized Reward Modeling for Text-to-Image Generation

Problem Formulation. The objective of personalized reward model $\mathcal{P}$ is to output a personalized reward r for a target pair $\mathbf{z} = (\mathbf{p}, \mathbf{I}_1, \mathbf{I}_2)$ given user reference data $\mathcal{D}_u = \{(\mathbf{p}_i, \mathbf{I}_i^+, \mathbf{I}_i^-)\}_{i=1}^N$, where $\mathbf{p}_i$ is the prompt, $\mathbf{I}_i^+$ is the user-preferred image, and $\mathbf{I}_i^-$ is the non-preferred one: $\mathbf{r} = \mathcal{P}(\mathbf{z}, \mathcal{D}_u)$. Here, $\mathbf{r} = +1$ if $\mathbf{I}_1$ is preferred, and $\mathbf{r} = -1$ otherwise.

3.1 PIGReward

Overview. We introduce the *preference reasoner* π , which processes a small set of user references $\mathcal{D}_u = \{(\mathbf{p}_i, \mathbf{I}_i^+, \mathbf{I}_i^-)\}_{i=1}^N$ to construct a user context $\mathcal{C}_u$:

$$\mathcal{C}_u = \pi(\mathcal{D}_u) = \{\mathbf{c}_i := \pi(\mathbf{p}_i, \mathbf{I}_i^+, \mathbf{I}_i^-)\}_{i=1}^N, \quad (1)$$

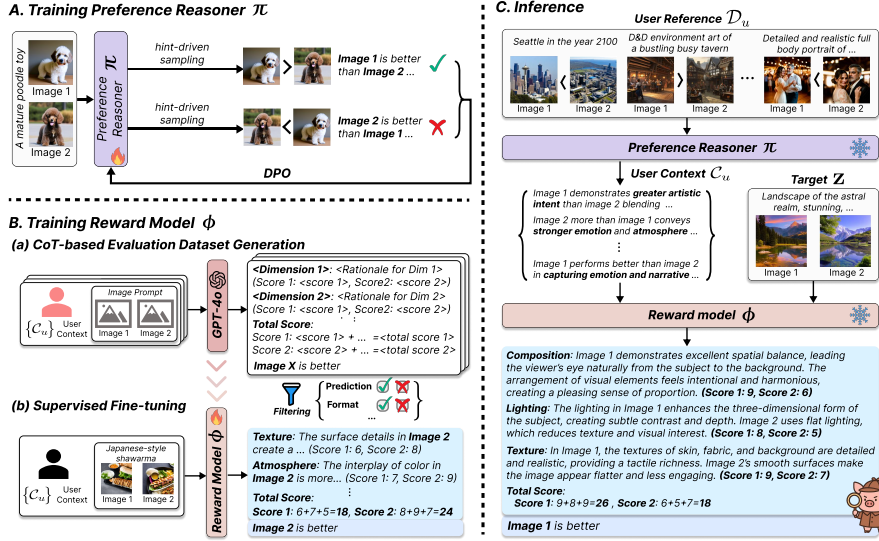


Fig. 2: Overview of PIGReward (A) **Training the preference reasoner π** : we construct contrastive rationale pairs via hint-driven sampling and fine-tune π with DPO, encouraging diverse rationales. (B) **Training the reward model ϕ** : (a) we use GPT-4o to generate CoT-formatted, dimension-wise evaluations and apply accept-reject filtering to remove incorrect/invalid outputs; (b) we then supervise fine-tune ϕ on the successful trajectories to distill structured, user-conditioned reasoning. (C) **Inference**: given user references $\mathcal{D}_u$, π builds a user context $\mathcal{C}_u$, and ϕ evaluates a target pair conditioned on $\mathcal{C}_u$ to output a personalized preference.

where each $\mathbf{c}_i$ is a language-based rationale explaining why $\mathbf{I}_i^+$ is preferred over $\mathbf{I}_i^-$. Conditioned on the resulting context $\mathcal{C}_u$, the *reward model* ϕ evaluates a target comparison $\mathbf{z} = (\mathbf{p}, \mathbf{I}^a, \mathbf{I}^b)$ and outputs a personalized reward $\mathbf{r}$:

$$\mathbf{r} = \phi(\mathbf{z}, \mathcal{C}_u). \quad (2)$$

While we leverage the inherent multimodal reasoning capability of LVLMS, we fine-tune both π and ϕ for task-specific objectives, as detailed below.

Training Setup. We adopt Qwen2-VL-7B [49] as the backbone LVLMS for both π and ϕ . The data used to train π , train ϕ , and evaluate are sampled from Pick-a-Pic [16], and each split is strictly disjoint.

Preference Reasoner. The objective of the preference reasoner π is to construct a rich user context $\mathcal{C}_u$ for user u from a small number of user reference data $\mathcal{D}_u$. Since each reference provides only a pairwise choice (i.e., which image was selected), it is too sparse to directly infer underlying preference factors. We therefore use π to convert implicit image-level preferences into explicit, language-based rationale that explain why the chosen image was preferred over the other, yielding more informative and interpretable user representations. Specifically, for each triplet $(\mathbf{p}_i, \mathbf{I}_i^+, \mathbf{I}_i^-) \in \mathcal{D}_u$, π generates a rationale and the resulting rationales

are aggregated into a single textual context $\mathcal{C}_u$. This user context enables the reward model ϕ to perform personalized evaluation without additional per-user fine-tuning [2, 25], which is more suitable in real-world scenarios where individual data is typically scarce and insufficient for reliable fine-tuning [34].

However, vanilla LVLMs often lack the granularity and diversity required to articulate subtle visual preferences between two images generated from the same text prompt [41, 45, 64]. This often leads to overly generic explanations across different users, thus failing to form distinctive user contexts. To enhance their explanatory richness and lexical diversity, we train π using contrastive judgments derived from a general preference dataset $\mathcal{D}_G = \{(\mathbf{p}_i, \mathbf{I}_i^+, \mathbf{I}_i^-)\}_{i=1}^M$. For each triplet, we apply hint-driven sampling [64], where we instruct π to generate a correct rationale j_i^+ (explaining why $\mathbf{I}_i^+$ is preferred over $\mathbf{I}_i^-$), and an incorrect rationale j_i^- (explaining why $\mathbf{I}_i^-$ is preferred over $\mathbf{I}_i^+$). We then optimize the preference reasoner π via the DPO loss, where $x_i := (\mathbf{p}_i, \mathbf{I}_i^+, \mathbf{I}_i^-)$, and π_{ref} is a frozen reference model.

$$\mathcal{L}_\pi = -\mathbb{E}_{x_i, j_i^+, j_i^-} \left[\log \sigma(\beta \Delta_i) \right], \text{ where } \Delta_i = \left(\log \frac{\pi(j_i^+ | x_i)}{\pi(j_i^- | x_i)} - \log \frac{\pi_{\text{ref}}(j_i^+ | x_i)}{\pi_{\text{ref}}(j_i^- | x_i)} \right). \quad (3)$$

Here, $\beta > 0$ is a temperature that controls the sharpness of the preference margin, and $\sigma(\cdot)$ denotes the logistic sigmoid that maps the scaled margin $\beta \Delta_i$ into a probability. This objective encourages π to assign higher relative likelihoods to correct rationales j_i^+ over incorrect ones j_i^- for the same image pair, thereby improving the fidelity and diversity of generated explanations during bootstrapping.

Personalized Reward Modeling. Given user context $\mathcal{C}_u$, the reward model ϕ performs personalized evaluation for a target image pair $\mathbf{z}$ and outputs a personalized reward $\mathbf{r} = \phi(\mathbf{z}, \mathcal{C}_u)$. As shown in Figure 2 (B), the evaluation follows a structured CoT format, involving two key steps: (1) inducing user-conditioned evaluation dimensions from $\mathcal{C}_u$ and (2) assessing images with generated dimensions into a final preference prediction. This CoT reasoning encourages more robust and consistent preference predictions, instead of shallow comparisons [51, 68].

To support such structured and personalized CoT reasoning with ϕ , we construct a dataset of 5K CoT-formatted samples. We first prepare pairs of user context $\mathcal{C}_u$ and target $\mathbf{z} = (\mathbf{p}, \mathbf{I}^+, \mathbf{I}^-)$. The user context is written in natural language, containing rationales explaining the user’s previous preference decisions. We then use GPT-4o [13] to infer evaluation dimensions from user context, then produce a detailed CoT reasoning that (i) scores each image along the inferred dimensions, (ii) aggregates per-dimension scores, and (iii) identifies the preferred image. We then apply a rejection sampling approach [51] to retain successful reasoning trajectories only. Specifically, we discard outputs that do not match the required structured format, and apply fully automated rule-based checks to filter out (i) inaccurate predictions whose final choice disagrees with the ground-truth human preference label, and (ii) invalid outputs, such as those with fewer than three dimensions or inconsistent score aggregation.

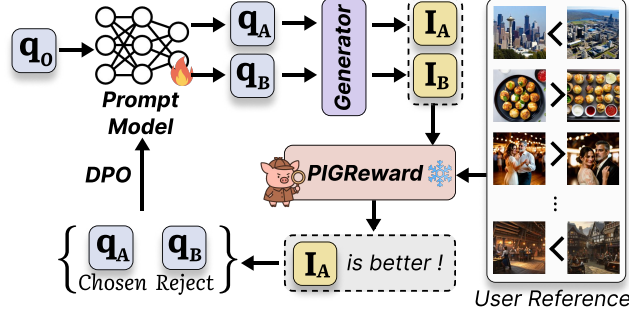


Fig. 3: Overview of personalized prompt optimization. PIGReward is instantiated as a user-conditioned reward function that compares images generated from candidate prompts and results in chosen–rejected prompt pairs, which are then used to optimize the prompt model.

Then we supervise fine-tune ϕ on the resulting dataset as:

$$\mathcal{L}_\phi = -\mathbb{E}_{(x_i, y_i)} \sum_{t=1}^{T_i} \log \phi(y_{i,t} \mid x_i, y_{i,<t}), \quad (4)$$

where $x_i = (\mathcal{C}_u, \mathbf{z})$ includes both the user context and the target image pair, and $y_i = (y_{i,1}, \dots, y_{i,T_i})$ is the corresponding GPT-4o-generated CoT output. Through this training, ϕ learns to produce structured, dimension-aligned evaluations grounded in contextual user information.

3.2 Personalized Prompt Optimization with PIGReward

Beyond evaluation, PIGReward can also serve as a personalized reward function for optimizing text prompts in T2I generation, guiding a prompt model toward generating images aligned with individual visual preferences. The overview is illustrated in Figure 3.

We train a prompt language model $\mathcal{M}$ [20] which refines an initial prompt $\mathbf{q}_0$ into expanded variants. For each user u , PIGReward is instantiated as a personalized reward function $\mathcal{P}_u$ conditioned on individual reference data $\mathcal{D}_u$. Given prompt $\mathbf{q}$, $\mathcal{M}$ generates two candidate expanded prompts, $\mathbf{q}_a$ and $\mathbf{q}_b$. With a T2I generator $\mathcal{G}$, each candidate prompt is rendered into an image, as $\mathbf{I}_a = \mathcal{G}(\mathbf{q}_a)$ and $\mathbf{I}_b = \mathcal{G}(\mathbf{q}_b)$. Then $\mathcal{P}_u$ assesses two images, then determines which image better aligns with the user preferences, considering corresponding prompts as chosen and rejected:

$$(\mathbf{q}_i^{\text{ch}}, \mathbf{q}_i^{\text{rej}}) = \begin{cases} (\mathbf{q}_a, \mathbf{q}_b), & \text{if } \mathcal{P}_u(\mathbf{q}, \mathbf{I}_a, \mathbf{I}_b) = +1, \\ (\mathbf{q}_b, \mathbf{q}_a), & \text{otherwise.} \end{cases} \quad (5)$$

The resulting chosen-rejected pairs form a user-conditioned preference signal for DPO training of $\mathcal{M}$. Let $x_i := (\mathbf{q}_{0,i}, u)$ denote the conditioning context, an

abstract prompt and user feature. We then optimize $\mathcal{M}$ using the DPO objective:

$$\mathcal{L}_{\mathcal{M}} = -\mathbb{E}_{(x_i, \mathbf{q}_i^{\text{ch}}, \mathbf{q}_i^{\text{rej}})} [\log \sigma(\beta \Delta_i)],$$

$$\text{where } \Delta_i = \log \frac{\mathcal{M}(\mathbf{q}_i^{\text{ch}} | x_i)}{\mathcal{M}(\mathbf{q}_i^{\text{rej}} | x_i)} - \log \frac{\mathcal{M}_{\text{ref}}(\mathbf{q}_i^{\text{ch}} | x_i)}{\mathcal{M}_{\text{ref}}(\mathbf{q}_i^{\text{rej}} | x_i)}. \quad (6)$$

This training encourages $\mathcal{M}$ to increase the likelihood of prompts that yield higher personalized rewards from PIGReward, producing a user-adaptive prompt generator that aligns more closely with individual visual preferences.

3.3 PIGBench: A Personalized T2I Dataset

Discussing Existing Datasets. Although several user-level preference datasets exist, they pose challenges for evaluating real-world personalization. Pick-a-Pic [16] is the most widely used pairwise preference dataset and includes user identifiers, allowing user-level analysis. However, as it was originally designed to capture general human preferences rather than individual characteristics, it lacks distinct user-specific variation [4]. PIP [3] is the first T2I dataset explicitly aimed at personalization, consisting of user-written prompts and generated images. Yet it lacks contrastive information—what users dislike—overlooking the importance of negative feedback in modeling preference [23]. ViPer [36] provides images with like/dislike comments simulated by LVLMs, but its data is limited to artistic domains, creating a gap from real-world imagery. Real-world personalization requires broad coverage of multifaceted visual interpretations (e.g., art style, domain) under sparse user feedback, where users provide only a few reference choices.

Constructing Dataset. We therefore collect user data to evaluate real-world personalization in T2I generation, named PIGBench. Unlike prior datasets, PIGBench focuses on how simple textual prompts can be interpreted in visually diverse ways according to different user preferences. PIGBench comprises user-level instances, where each instance includes a user-specific ranking of four images generated from a single, abstract prompt. To reflect realistic settings, we vary the amount of reference data per user, enabling evaluation under different context lengths.



Fig. 4: Examples from PIGBench. Each instance shows four image variants from the same prompt; checkmarks denote the selections of Users A/B. Users differ both across instances (above the dashed line) and within the same instance (below), highlighting divergent visual preferences.

We take the “*simple detail*” subset of PartiPrompts [65], which leaves substantial room for specification, and curate prompts spanning diverse 12 categories (e.g., world knowledge, outdoor scenes). For each prompt, we use GPT-4o [13] to expand it into four concrete variants that differ in visual style and semantic context, then generate the corresponding images with SDXL [30]. We then conduct a manual quality check to remove low-quality sets and filter out instances with insufficient personalization affordance, where the variants are unlikely to elicit divergent user preferences, resulting in 100 instances in total. Human annotators are asked to rank the four images for each expanded prompt according to their personal preferences. Each annotator is randomly assigned a varying number of image sets (between 5 and 15) to collect records with various lengths of user context, thereby resulting in 75 user records. Figure 4 shows an example from PIGBench, highlighting that different users often show distinct rankings for the same image instance.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on three conventional image preference datasets, Pick-a-Pic [16], PIP [3], and PASTA [25]—and PIGBench. Although these conventional datasets contain user identifiers, they are not intentionally designed for history-driven personalized evaluation and therefore we adapt them to a user-conditioned pairwise protocol. Also, for PIGBench, we convert each user ranking into a pairwise preference by comparing the top-ranked image against the bottom-ranked image among 4 candidates. Dataset statistics are in Table 1.

Table 1: Dataset statistics. Ref. size denotes the number of user reference.

Dataset	Ref. size	Users
Pick-a-Pic	8	10,000
PIP	5	1,002
PASTA	4	5,143
PIGBench	5–15	75

Baselines. We compare PIGReward against conventional and recent state-of-the-art T2I evaluation methods. (1) **w/o Reference:** Conventional evaluation methods assess image quality without user-specific context. We include CLIP/BLIP-based reward models such as CLIPScore [11], PickScore [16], HPS [56], Aesthetics [38], and ImageReward [59], all of which are fine-tuned on general single-dimensional preference datasets. We further consider multi-dimensional approaches such as MPS [69] and VisionReward [58], which include fixed criteria. Lastly, we include recent LVLM-based reward models: LLaVA-Reward [57] and UnifiedReward [52], as well as UnifiedReward-Think [51], which incorporates CoT reasoning process. (2) **w/ Reference:** Previous personalized T2I studies [4, 62] adopt the LVLM-as-a-judge paradigm, in which a large vision-language model (LVLM) evaluates image pairs based on given reference examples. We employ two open-source models—Qwen2-VL-7B and LLaVA-OneVision [18]—which support multi-image processing, and one closed-source model, GPT-4o [13].

Table 2: Quantitative comparison with conventional T2I evaluation methods. Baselines are evaluated without user references, and we compare against vanilla LVLMS as well given references. The best results are in bold, and the second-best results are underlined.

	Pick-a-Pic		PIP		PASTA		PIGBench	
	acc w/	tie acc w/o tie	acc w/	tie acc w/o tie	acc w/	tie acc w/o tie	acc w/	tie acc w/o tie
<i>w/o Reference</i>								
CLIPScore [11]	50.90	55.44	49.15	21.37	66.28	71.09	47.17	26.42
Aesthetics [38]	49.98	25.43	51.66	37.82	51.78	59.56	49.06	26.42
PickScore [16]	51.00	54.31	49.63	33.87	68.36	71.13	51.89	16.98
HPS (v2) [56]	50.14	6.65	49.68	12.61	69.03	<u>72.99</u>	43.40	28.30
ImageReward [59]	49.55	41.81	49.79	43.06	69.82	70.03	35.85	33.96
VisionReward [58]	49.81	30.40	49.73	31.73	65.54	71.18	43.40	34.78
MPS (overall) [69]	50.20	47.87	50.80	46.15	66.73	68.17	47.17	45.28
LLava-Reward [76]	41.57	48.95	25.60	25.48	68.17	50.57	64.15	64.15
UnifiedReward [52]	48.46	48.46	37.48	37.41	<u>69.86</u>	69.86	<u>67.92</u>	67.92
UnifiedReward-Think [51]	59.92	59.92	35.28	35.21	65.71	65.71	66.04	66.04
<i>w/ Reference</i>								
Qwen2.5-VL-7B [1]	<u>61.52</u>	<u>61.52</u>	49.95	49.95	50.34	50.34	54.72	54.72
LLaVA-OneVision [18]	55.04	55.04	50.75	50.76	50.12	50.12	56.60	56.60
GPT-4o [13]	46.90	46.70	<u>76.23</u>	<u>76.54</u>	68.80	68.80	64.15	<u>68.29</u>
PIGReward	63.76	62.14	77.84	78.59	75.43	74.43	84.91	85.85

4.2 Main Results

Table 2 presents the results of PIGReward and existing T2I evaluation methods under a personalized setting, where PIGReward consistently achieves the best performance. Both single-scalar and fixed multi-dimensional baselines underperform, indicating that a limited set of evaluation axes is insufficient to capture fine-grained human preference and user-dependent preferences, highlighting the need for adaptive assessment. Even LVLMS-based methods perform comparably to simple reward-function baselines, indicating that generic reasoning alone is insufficient to tailor individuals. Among baselines without reference data, UnifiedReward-Think shows strong performance, demonstrating that incorporating CoT reasoning significantly enhances evaluation capability. When provided with reference data, vanilla LVLMS show improved results but still fall short of PIGReward, suggesting that they struggle to infer complex and latent preferences embedded in user histories. Overall, these results confirm that context-conditioned evaluation dimensions and detailed CoT-based reasoning are both essential for effective personalized evaluation.

4.3 Comparisons with Similarity-based Metrics

Previous personalized T2I works [3, 15] typically evaluate personalization with similarity-based metrics that measure how closely generated outputs resemble reference images or prompts; they use maximum similarity as preference score. We include CLIPScore (Text-Align), SSIM, and DINOscore in our comparison: CLIPScore and Text-Align compute CLIP-based similarities [11], SSIM captures pixel-level perceptual similarity [53], and DINOscore measures visual similarity in the DINO feature space [26]. As shown in Figure 5, these metrics correlate

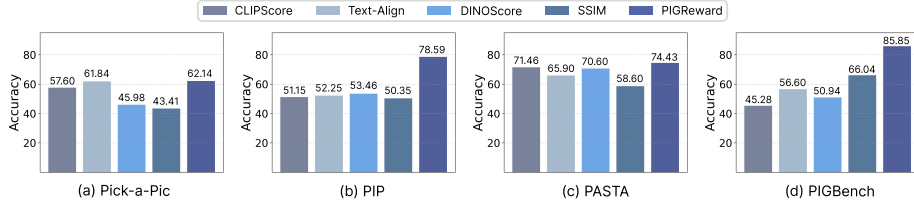


Fig. 5: Quantitative comparison with similarity-based metrics. CLIPScore, DINOscore, and SSIM compare generated images to reference images, while Text-Align compares to reference prompts. PIGReward leverages both images and prompts.

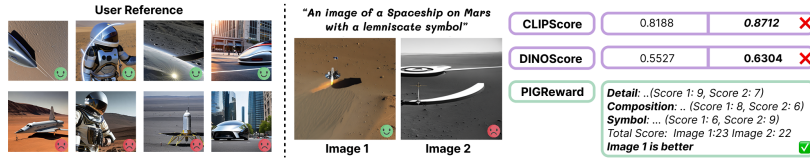


Fig. 6: A failure case of similarity-based metrics for personalized evaluation. From the user’s previously liked images (left), dominated by close-up metallic surfaces and sleek vehicle/space-themed visuals, CLIPScore ranks Image 2 higher by matching salient recurring patterns (a large, high-contrast, smooth structure), rather than the user’s underlying criteria. In contrast, PIGReward induces a user-specific evaluation space from the same history (e.g., detail, composition, symbolism) and correctly prefers Image 1. For brevity, we omit the full PIGReward output, including explicit rationales.

weakly with actual user preferences, as they are limited to capturing surface-level resemblance rather than contextual intent. Pixel- or embedding-level similarity often overlooks subjective visual preference aspects, insensitive to user-specific preferences, especially across diverse users or abstract prompts. In contrast, PIGReward reasons over user-specific context, enabling a more faithful understanding of individual preference than fixed similarity measures. In Figure 6, we provide an example showing why similarity-based metrics fail to capture fine-grained personalization.

4.4 More Discussions

Ablation on PIGReward components. To verify that the performance gain of PIGReward is not solely due to supervised fine-tuning on generated CoT data, we ablate the two core components of the reward model ϕ before SFT: (1) adaptive generation of user-conditioned evaluation dimensions and (2) dimension-wise CoT-based evaluation. Using the same vanilla Qwen2-7B backbone and evaluation protocol, we compare variants with no dimensions, fixed dimensions, and generated dimensions, both with and without CoT reasoning.¹ As shown in Table 3, incorporating generated dimensions with dimension-wise CoT evaluation achieves the best performance. This demonstrates that the improvement

¹ For fixed dimensions, we adopt the criteria from UnifiedReward-Think [51].

Table 3: Ablation on framework components before SFT. Evaluates whether the proposed framework itself improves personalized evaluation before GPT-4o-supervised fine-tuning.

Dimension	CoT	PIGBench	Pick-a-Pic
$\times$	$\times$	53.15	46.30
$\times$	$\checkmark$	54.72	51.30
Fixed	$\checkmark$	52.78	45.90
Generative	$\times$	50.00	50.10
Generative	$\checkmark$	67.83	53.40

Table 4: Ablation on training stages and CoT reasoning. Evaluates the impact of CoT reasoning, DPO-tuned preference reasoner π , and CoT-distilled reward model ϕ .

	DPO	π	SFT	ϕ	Pick-a-Pic	PIGBench
<i>w/o CoT</i>	$\checkmark$		$\times$		44.43	39.62
	$\times$		$\times$		48.94	56.94
	$\checkmark$		$\times$		45.63	70.83
<i>w/ CoT</i>	$\times$		$\checkmark$		49.98	58.49
	$\checkmark$		$\checkmark$		63.76	84.91

is not merely attributable to GPT-4o supervision or rejection filtering, but already emerges from the proposed reasoning framework itself. We then further strengthen these capabilities by applying SFT to ϕ .

Ablation of Each Training Stage. Table 4 reports the results of our ablations across training stages and the impact of CoT reasoning. Without CoT reasoning, PIGReward performs the worst, as it lacks a step-by-step, dimension-aware evaluation process. Introducing CoT reasoning improves performance by enabling structured, user-specific reasoning. Fine-tuning the preference reasoner π with DPO further boosts performance by producing more diverse and discriminative rationales. The best results are achieved when the reward model ϕ is further distilled on personalized CoT-formatted data, refining the ability of PIGReward to effectively perform personalized evaluation.

Model Instantiation for π and ϕ . Figure 7 (a) compares different instantiations of π and ϕ using either our task-specifically trained Qwen2-VL-7B or GPT-4o. Instantiating both π and ϕ with the finetuned Qwen2-VL (the most left) yields the best overall performance. In contrast, replacing either component with GPT-4o substantially degrades accuracy, and using GPT-4o for both π and ϕ results in the lowest performance. These results suggest that a strong general-purpose model cannot simply substitute our trained components and highlight the importance of task-specific finetuning for producing user-aligned rationales and reliable personalized evaluation.

Impact and Scalability of User Context Length. In Figure 7 (b), we study how the size of user reference (i.e., context length) affects personalization performance. Notably, PIGReward achieves strong performance even with very few (1-3) references, indicating effective preference inference under cold-start conditions. As the number of references increases, PIGReward steadily improves, suggesting that additional evidence helps identify user preferences. In contrast, GPT-4o degrades when the reference set becomes large, implying that naively injecting more references is not sufficient and that effective preference extraction and aggregation are critical. Unlike directly feeding all references into an LVLm, PIGReward performs preference reasoning for each reference individually and aggregates the resulting textual rationales into a compact user context,

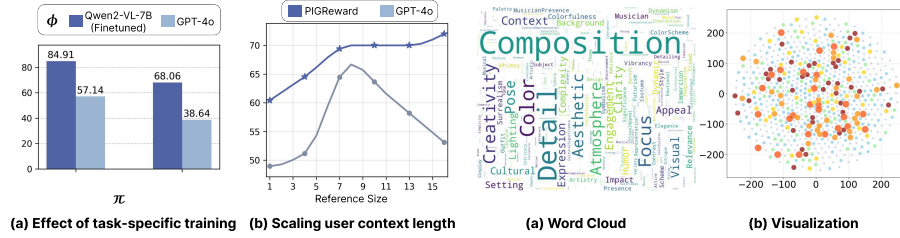


Fig. 7: Additional analyses. (a) compares model instantiations for π and ϕ and (b) evaluates scalability with respect to user reference size. **Fig. 8: Analysis of generated evaluation dimensions.** Shows that PIGReward generates semantically diverse dimensions beyond a fixed attribute space.

mitigating multi-image processing limits and better handling larger reference sets.

4.5 Analysis of Generated Evaluation Dimensions

The core idea of PIGReward is to dynamically induce user-conditioned evaluation dimensions from personalized contexts, unlike prior approaches that rely on predefined criteria. We analyze these generated dimensions from experiments.

Semantic diversity of generated dimensions. Figure 8 visualizes the generated dimensions during evaluation using (a) a word cloud and (b) a t-SNE embedding. The word cloud indicates that PIGReward produces a broad range of descriptive terms, while the t-SNE plot suggests that these dimensions span diverse evaluation perspectives rather than collapsing into a small set of fixed attributes. Quantitatively, PIGReward yields a much richer dimension space than prior approaches that rely on only 4–11 predefined dimensions: on Pick-a-Pic, it generates 671 distinct dimensions forming 2,609 unique dimension sets; on PIGBench, it generates 23 dimensions forming 42 sets. Overall, these results highlight that PIGReward can flexibly instantiate context-dependent criteria beyond a fixed attribute space.

Human-alignment of generated dimensions. To validate interpretability of PIGReward, we compare PIGReward-generated reasoning with justifications written by PIGBench users. Specifically, we measure *alignment*, the ratio of PIGReward-generated evaluation dimensions grounded in user-written explanations. PIGReward achieves a high alignment score of 83.34%, confirming that PIGReward produces user-aligned rationales that reflect actual human decision-making criteria.

4.6 Quantitative and Qualitative Study for Personalized Prompt Optimization

We compare images generated from (i) base prompts, (ii) prompts optimized by the general-purpose method Promptist [20], and (iii) prompts optimized by PIGReward. We use SDXL [30] as T2I generator. To initialize PIGReward, we use

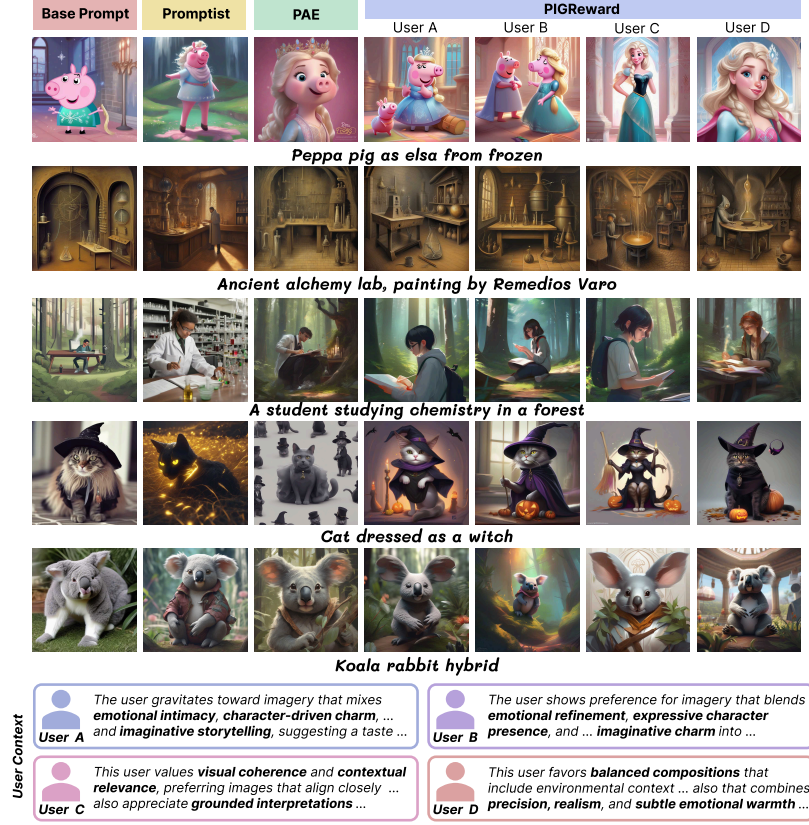


Fig. 9: Qualitative results for personalized prompt optimization. We additionally include PAE [24] to provide a broader view across representative prompt-optimization strategies. The depicted user contexts are derived from our preference reasoner and summarized with LLM. User A favors imaginative narrative elements, leading to more illustrative outputs, whereas User D prioritizes realism, resulting in highly detailed images with natural lighting.

user contexts constructed from the PIGBench and the same annotators who created PIGBench also conducted the quantitative evaluation. Also, we include an auxiliary automatic evaluation using PIGReward. As shown in Figure 10, ours achieves a substantially higher win rate, indicating that feedback from PIGReward effectively guides the generation process toward user-aligned outputs. Figure 9 presents qualitative examples, where the optimized

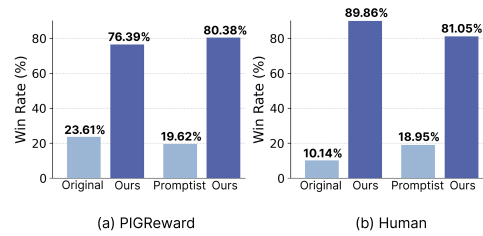


Fig. 10: Quantitative results for personalized prompt optimization. Win rates are computed using (a) automatic evaluation by PIGReward and (b) human evaluation.

prompts produce visually diverse results across users, suggesting that effective personalization requires tailoring the optimization direction to individual preferences rather than following a single globally optimal direction.

5 Conclusion and Future Directions

We introduced PIGReward, a personalized reward modeling framework for evaluating and optimizing T2I generation. PIGReward generates user-conditioned evaluation dimensions and performs structured CoT-based evaluation to produce interpretable preference predictions. We further show that PIGReward can guide personalized prompt optimization via user-conditioned rewards. To enable realistic evaluation, we present PIGBench, a user-study benchmark that elicits diverse rankings from abstract prompts. Experiments demonstrate that PIGReward consistently outperforms existing reward models and LVLMM-based judges in both accuracy and interpretability. Looking ahead, the reasoning traces produced by PIGReward can be leveraged as natural-language feedback to support more interactive and individually aligned T2I generation.

Acknowledgements

Please insert your acknowledgments here.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Chen, D., Chen, Y., Rege, A., Vinayak, R.K.: Pal: Pluralistic alignment framework for learning from heterogeneous preferences. arXiv preprint arXiv:2406.08469 (2024)
3. 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)
4. Dang, M., Singh, A., Zhou, L., Ermon, S., Song, J.: Personalized preference finetuning of diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8020–8030 (2025)
5. Datta, S., Ku, A., Ramachandran, D., Anderson, P.: Prompt expansion for adaptive text-to-image generation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3449–3476 (2024)
6. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9062–9072 (2025)
7. Dubois, Y., Li, C.X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P.S., Hashimoto, T.B.: AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems* **36**, 30039–30069 (2023)
8. Dunlop, C., Zheng, M., Venkatesh, K., Yanardag, P.: Personalized image editing in text-to-image diffusion models via collaborative direct preference optimization. arXiv preprint arXiv:2511.05616 (2025)
9. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems* **36**, 79858–79885 (2023)
10. Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al.: G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370 (2023)
11. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
13. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

14. Jang, J., Kim, S., Lin, B.Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y., Ammanabrolu, P.: Personalized soups: Personalized large language model alignment via post-hoc parameter merging. arXiv preprint arXiv:2310.11564 (2023)
15. Kim, H., Ahn, S., Seo, Y.D.: Draw your mind: Personalized generation via condition-level modeling in text-to-image diffusion models. arXiv preprint arXiv:2508.03481 (2025)
16. 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. *Advances in neural information processing systems* **36**, 36652–36663 (2023)
17. Lee, J., Yu, Y., Lee, D.: Hippo-video: Simulating watch histories with large language models for history-driven video highlighting. In: *Second Conference on Language Modeling*
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
19. Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., Kong, L., Liu, Q.: Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. arXiv preprint arXiv:2410.09421 (2024)
20. Li, W., Wang, J., Zhang, X.: Promptist: Automated prompt optimization for text-to-image synthesis. In: *CCF international conference on natural language processing and Chinese computing*. pp. 295–306. Springer (2024)
21. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19401–19411 (2024)
22. Liu, Y., He, M., Yao, F., Ji, Y., Tao, S., Du, J., Li, D., Gao, J., Zhang, L., Yang, H., et al.: What do you want? user-centric prompt generation for text-to-image synthesis via multi-turn guidance. arXiv preprint arXiv:2408.12910 (2024)
23. Mo, W., Ba, Y., Zhang, T., Bai, Y., Li, B.: Learning user preferences for image generation model. arXiv preprint arXiv:2508.08220 (2025)
24. Mo, W., Zhang, T., Bai, Y., Su, B., Wen, J.R., Yang, Q.: Dynamic prompt optimizing for text-to-image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26627–26636 (2024)
25. Nabati, O., Tennenholtz, G., Hsu, C., Ryu, M., Ramachandran, D., Chow, Y., Li, X., Boutilier, C.: Preference adaptive and sequential text-to-image generation. In: *Forty-second International Conference on Machine Learning* (2025), <https://openreview.net/forum?id=LCr6CIAEye>
26. 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)
27. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022)
28. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. arXiv preprint arXiv:2306.00637 (2023)
29. Poddar, S., Wan, Y., Ivison, H., Gupta, A., Jaques, N.: Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems* **37**, 52516–52544 (2024)

30. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
31. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* **36**, 53728–53741 (2023)
32. Rame, A., Couairon, G., Dancette, C., Gaya, J.B., Shukor, M., Soulier, L., Cord, M.: Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems* **36**, 71095–71134 (2023)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1(2), 3 (2022)
34. Ryan, M.J., Shaikh, O., Bhagirath, A., Frees, D., Held, W., Yang, D.: Synthesizeme! inducing persona-guided prompts for personalized reward models in llms. arXiv preprint arXiv:2506.05598 (2025)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **35**, 36479–36494 (2022)
36. 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)
37. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. *Advances in neural information processing systems* **29** (2016)
38. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* **35**, 25278–25294 (2022)
39. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
40. Shen, X., Zhang, R., Zhao, X., Zhu, J., Xiao, X.: Pmg: Personalized multimodal generation with large language models. In: *Proceedings of the ACM Web Conference 2024*. pp. 3833–3843 (2024)
41. Shypula, A., Li, S., Zhang, B., Padmakumar, V., Yin, K., Bastani, O.: Evaluating the diversity and quality of llm generated content. arXiv preprint arXiv:2504.12522 (2025)
42. Sorensen, T., Moore, J., Fisher, J., Gordon, M., Miresghallah, N., Rytting, C.M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al.: A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070 (2024)
43. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. *Advances in neural information processing systems* **33**, 3008–3021 (2020)
44. Su, H., Meng, L., Zhu, L., Lu, K., Li, J.: Ddpo: Direct dual propensity optimization for post-click conversion rate estimation. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 1179–1188 (2024)

45. Trivedi, P., Gulati, A., Molenschot, O., Rajeev, M.A., Ramamurthy, R., Stevens, K., Chaudhery, T.S., Jambholkar, J., Zou, J., Rajani, N.: Self-rationalization improves llm as a fine-grained judge. arXiv preprint arXiv:2410.05495 (2024)
46. 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)
47. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
48. Wang, C.: A diffusion-based method for multi-turn compositional image generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 381–391 (2024)
49. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
50. Wang, X., Wu, L., Yin, S., Li, Z., Chen, Y., Hufeng, H., Su, Y., Liu, Q.: I-am-g: Interest augmented multimodal generator for item personalization. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21303–21317 (2024)
51. Wang, Y., Li, Z., Zang, Y., Wang, C., Lu, Q., Jin, C., Wang, J.: Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), <https://openreview.net/forum?id=S43003uMGq>
52. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
53. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The thirty-seventh asilomar conference on signals, systems & computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
54. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)
55. Wei, Y., Zheng, Y., Zhang, Y., Liu, M., Ji, Z., Zhang, L., Zuo, W.: Personalized image generation with deep generative models: A decade survey. arXiv preprint arXiv:2502.13081 (2025)
56. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
57. Xiong, T., Wang, X., Guo, D., Ye, Q., Fan, H., Gu, Q., Huang, H., Li, C.: Llava-critic: Learning to evaluate multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13618–13628 (2025)
58. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
59. 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. *Advances in Neural Information Processing Systems* **36**, 15903–15935 (2023)
60. Xu, S., Zhou, X., Zhang, Y., Liu, M., Liang, T., Zhang, T., Bai, Y., Wu, Z., Zuo, W.: Preferthinker: Reasoning-based personalized image preference assessment. arXiv preprint arXiv:2511.00609 (2025)

61. 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)
62. Xu, Y., Zhang, J., Salemi, A., Hu, X., Wang, W., Feng, F., Zamani, H., He, X., Chua, T.S.: Personalized generation in large model era: A survey. *arXiv preprint arXiv:2503.02614* (2025)
63. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* **36**, 11809–11822 (2023)
64. Ye, Z., Li, X., Li, Q., Ai, Q., Zhou, Y., Shen, W., Yan, D., Liu, Y.: Learning llm-as-a-judge for preference alignment. In: *The Thirteenth International Conference on Learning Representations* (2025)
65. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* **2**(3), 5 (2022)
66. Zhang, H., Wu, T., Wei, Y.: Multi-view user preference modeling for personalized text-to-image generation. *IEEE Transactions on Multimedia* (2025)
67. Zhang, J.: Guided profile generation improves personalization with llms. *arXiv preprint arXiv:2409.13093* (2024)
68. Zhang, R., Zhang, B., Li, Y., Zhang, H., Sun, Z., Gan, Z., Yang, Y., Pang, R., Yang, Y.: Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198* (2024)
69. Zhang, S., Wang, B., Wu, J., Li, Y., Gao, T., Zhang, D., Wang, Z.: Learning multi-dimensional human preference for text-to-image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8018–8027 (2024)
70. Zhang, Y.S.Y.Q.M., Zheng, X.L.J.Y.X., Wu, K.L.X.S.Y., Fu, R.J.C., Chen, P.: Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394* **18** (2021)
71. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023)
72. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 1702–1713 (2025)
73. Zhao, S., Dang, J., Grover, A.: Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523* (2023)
74. Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., Liu, P.J.: Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425* (2023)
75. Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., et al.: Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964* (2023)
76. Zhou, S., Zhang, R., Zhu, H., Kveton, B., Zhou, Y., Gu, J., Chen, J., Chen, C.: Multimodal llms as customized reward models for text-to-image generation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19638–19648 (2025)
77. Zhou, Y., Fan, Z., Cheng, D., Yang, S., Chen, Z., Cui, C., Wang, X., Li, Y., Zhang, L., Yao, H.: Calibrated self-rewarding vision language models. *Advances in Neural Information Processing Systems* **37**, 51503–51531 (2024)