PERSONALIZE AT TEST TIME: LEARNING USER PREFERENCES FOR IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models can generate high-quality images, yet aligning their outputs with individual user preferences remains challenging. A key bottleneck is accurately modeling diverse user preferences from limited feedback. Existing approaches often rely on labor-intensive manual preference annotations or vision-language models(VLM) to extract preference information from user interaction histories, introducing substantial annotation or computational costs that limit scalability. We propose an approach that learns personalized reward models directly from users' historical image preference pairs. First, we use an autoencoder to compress hundreds of visual attributes into 50 attribute-anchored preference dimensions and train an evaluator to score images along these dimensions. We then represent each user's preferences as a linear combination of the shared dimension scores, estimating the user-specific weights by maximizing the likelihood of their observed pairwise preferences under the Bradley–Terry model. This formulation reduces per-user adaptation to optimizing a low-dimensional weight vector, simplifying optimization and enabling data-efficient personalization from sparse feedback. The learned personalized rewards guide image generation at inference time while keeping the diffusion model frozen. Experiments on real-user preference data show that our approach achieves approximately 77% held-out pairwise preference prediction accuracy and improves the alignment of generated images with individual user preferences.

1 INTRODUCTION

Text-to-image diffusion models have made remarkable progress in visual fidelity, prompt adherence, and compositional generation, powering increasingly capable systems such as Stable Diffusion (Podell et al., 2024; Rombach et al., 2022), DALL-E (Ramesh et al., 2021), and Seedream (Seedream et al., 2025). To adapt these general-purpose models to downstream requirements, prior work has explored supervised or reinforcement-learning-based fine-tuning, the injection of additional conditioning signals, and reward-guided inference (Wallace et al., 2024; Prabhudesai et al., 2023). These approaches have proven effective for applications including subject-driven generation, style customization, and general preference alignment.

Despite their effectiveness in many applications, personalizing diffusion models to individual users remains challenging, as it requires accurately modeling each user's preferences from limited feedback. Existing approaches generally obtain user preferences in two ways: using vision-language models to summarize a user's historical preference records into a profile or embedding (Dang et al., 2025; Kim et al., 2025a), or collecting explicit human annotations of users' likes and dislikes and converting them into structured visual preferences to guide generation (Salehi et al., 2024). These approaches have two important limitations. First, incorporating preference representations requires additional fine-tuning of the diffusion model or training of a specialized adapter, introducing substantial computational cost and limiting scalability across different generative models. Second, acquiring reliable preference representations remains difficult: profiles inferred by vision-language models may be inaccurate when user histories are short, long, or heterogeneous, while collecting detailed human annotations imposes considerable burden on users.

In this work, we propose PAT(**P**ersonalized **a**t **T**est Time for diffusion models), which instantiates a personalized reward model at test time directly from a user's historical preference feedback. For

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

each pair of images generated for a prompt, a user only needs to indicate which image they prefer. Our approach assumes that user preferences admit a low-dimensional latent representation, with individual differences captured by the weights each user assigns to shared preference dimensions. We therefore formulate personalization as a two-stage problem: first, identifying a compact, shared visual attribute space that captures diverse preference-relevant attributes; and second, quantifying each user's preferences along these dimensions. To identify the shared preference dimensions, we draw on the collection of over 500 visual attributes compiled by Salehi et al. (2024). While these attributes provide broad coverage of visual preferences, estimating a separate weight for every attribute would require fitting hundreds of user-specific parameters from only a small number of pairwise observations, making preference estimation prone to overfitting. We therefore seek a compact shared representation that preserves meaningful variation across visual attributes while reducing the number of parameters to be estimated for each user. To this end, we use an autoencoder to derive a low-dimensional set of shared preference dimensions grounded in these attributes. Unlike covariance-based linear methods such as PCA (Pearson, 1901; Hotelling, 1933) and factor analysis(Spearman, 1961; Thurstone, 1931), the autoencoder can capture a nonlinear low-dimensional representation that better preserves interactions among visual attributes. The advantage is supported by our empirical results in Table 6. An evaluator then maps a prompt–image pair (c, x) to its scores $s_\theta(c, x) \in \mathbb{R}^d$ along these dimensions, providing a shared basis for evaluating the visual characteristics of images.

To capture each user's individual preferences, we construct a personalized reward model that can be efficiently learned at test time. We quantify the reward of user u as a linear combination $r_u(c, x) = \lambda_u^\top s_\theta(c, x)$, where λ_u is the user-specific preference weight. Given the user's preference history represented as a collection of n prompt–image preference pairs $(c_k, x_k^+, x_k^-)_1^n$, we estimate the user-specific weights using the Bradley–Terry objective,

$$\lambda_u^* = \arg\max_{\lambda_u} \sum_{k=1}^{m_u} \log \sigma \left(\lambda_u^\top \left[s_\theta(c_k, x_k^+) - s_\theta(c_k, x_k^-) \right] \right). \tag{1}$$

This factorization reduces test-time personalization to estimating a low-dimensional weight vector, keeping per-user optimization compact and enabling data-efficient learning from sparse feedback. The resulting reward is used to guide generation at inference time while the underlying diffusion model remains frozen. To improve generalization for users with very few preference pairs, we further introduce a mixed-effects parameterization $\lambda_u = G + b_u$, where G captures population-level preferences shared across users and b_u captures user-specific deviations. By pooling preference information through G, this formulation further enables more reliable personalization in few-shot settings. We evaluate our approach on real-user preference data, achieving approximately 78% held-out pairwise preference prediction accuracy. When transferred to personalized image generation, our method achieves a 63.0% win rate against the non-personalized generator. Our main contributions are as follows:

- We introduce the first mathematical formulation of user profiling for personalized image generation based directly on offline pairwise preferences. Our formulation represents each user by a low-dimensional preference vector over shared visual dimensions, turning personalization into a tractable parameter-estimation problem.
- We propose PAT, an end-to-end pipeline that fits a compact personalized reward and uses it to guide a frozen diffusion model through test-time noise optimization, requiring neither per-user model fine-tuning nor separate personalized generators.
- We empirically demonstrate that our approach accurately models real users' preferences under limited feedback and that the learned personalized rewards transfer effectively from preference prediction to personalized image generation.

2 Related Work

Diffusion Model Alignment. A growing body of work aligns pretrained diffusion models with downstream objectives such as human preferences, aesthetic quality, and safety-awareness. One class of approaches fine-tunes the diffusion model using reward supervision, thereby incorporating the desired objective into the model parameters (Dai et al., 2023; Betker et al., 2023; Clark et al.,

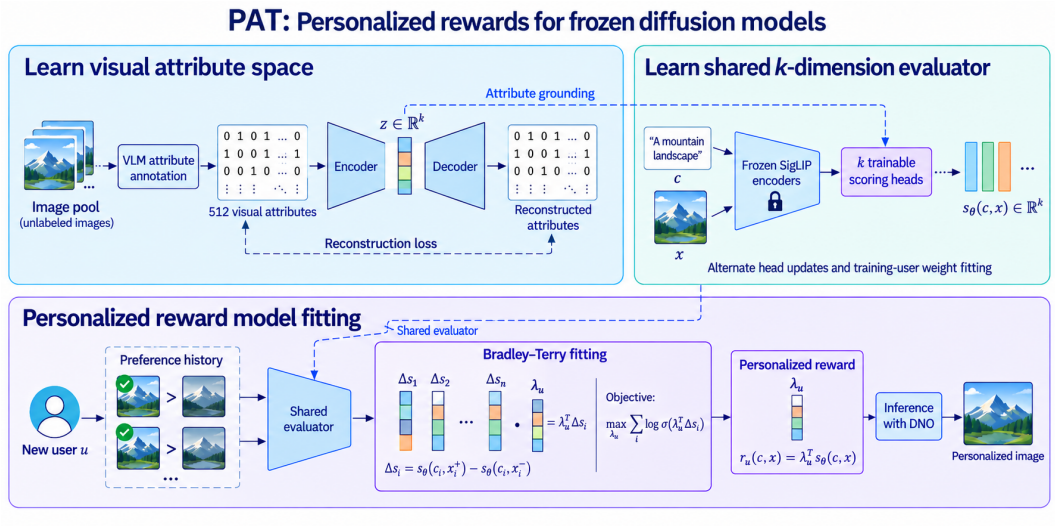


Figure 1: **Overview of PAT.** We learn a shared low-dimensional visual preference space and evaluator, fit a personalized reward from each user's pairwise preference history, and use the reward to guide a frozen diffusion model at inference time.

2023; Prabhudesai et al., 2023; Fan et al., 2023; Black et al., 2023). Another class performs alignment at inference time, steering the denoising trajectory or directly optimizing the latent variables while keeping the pretrained diffusion model frozen (Jiang et al., 2026; Goyal et al., 2026; Tang et al., 2024; Yeh et al., 2024; Kim et al.; Zhai et al., 2025). More recently, reinforcement learning has been introduced to diffusion model alignment by treating denoising as a sequential decision-making process. Existing methods employ policy-optimization algorithms such as GRPO (Zheng et al., 2026; Liu et al., 2026; Xue et al., 2025; Li et al., 2025a;b; 2026c), as well as RLHF and direct preference optimization objectives such as DPO (Wallace et al., 2024; Jiang et al., 2026; Liang et al., 2024; Wang et al., 2025; Lee et al., 2025b; Gu et al., 2024; Yang et al., 2024a; Zhu et al., 2025), to improve generation according to human feedback or learned rewards. While most of these methods perform alignment by updating the diffusion model during training, a small number of concurrent works, besides ours, have also explored alignment at inference time. Jin et al. (2025) and Xie et al. (2026) investigate training-free inference-time reward transfer from RL aligned-base difference between tuned model and base model. Chen et al. (2026) treats intermediate flow latents as implicit policies and optimizes them at test time using PRDP and GRPO objectives.

Personalization for Generative Models. Personalization has been extensively studied for large language models through approaches such as user-profile conditioning, memory augmentation, parameter-efficient adaptation, and personalized reward modeling (Yang et al., 2024b; Li et al., 2026a; Poddar et al., 2024; Wang et al., 2024; Song et al., 2026; Yu et al., 2026; Wang et al., 2026a; Wu et al., 2026; Tan et al., 2026; Barreto et al., 2025; Cai et al., 2026; Iyer et al.; Chen et al., 2025; Shenfeld et al., 2025; Kim et al., 2025b). Personalizing diffusion models poses additional challenges. An LLM can jointly interpret heterogeneous user histories, extract preference-relevant information, and incorporate the resulting profile into generation through natural-language conditioning. In contrast, a diffusion model primarily serves as a generator and does not natively provide a comparable interface for reasoning over user histories or constructing explicit user models. Existing approaches therefore commonly rely on external vision-language models to summarize a user's prior choices or encode them into preference representations, while others use human-defined attributes or manual annotations to characterize the relevant dimensions of variation (Dang et al., 2025; Lee et al., 2025a; Salehi et al., 2024; Kim et al., 2025a; Wang et al., 2026b; Li et al., 2026b). Our approach makes diffusion personalization more scalable by directly learning shared preference dimensions and user-specific rewards from pairwise feedback. It requires neither external VLM-based nor manual analysis of each user's history, and personalizes generation at inference time without fine-tuning the underlying diffusion model.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

For each user u, we observe a small set of offline preference pairs, $\mathcal{D}_u = \{(c_i, x_i^+, x_i^-)\}_{i=1}^{m_u}$, where in i-th image pair, c_i is a text prompt and x_i^+ is preferred to x_i^-. We assume that user preferences lie in a shared k-dimensional space, represented by an evaluator $s(c, x) \in \mathbb{R}^k$. The personalized reward is a linear combination

$$r_u(c, x) = \lambda_u^\top s(c, x), \qquad \lambda_u \in \mathbb{R}^k, \tag{2}$$

where λ_u is the only user-specific parameter. At inference time, we aim to generate images maximizing the personal reward $r_u(c, x)$.

3.2 LEARNING AN AUTOENCODER PREFERENCE EVALUATOR

Preference Dataset. We construct the data for learning the shared preference evaluator from the training split of Pick-a-Pic v1 (Kirstain et al., 2023), where each comparison consists of a text prompt c, two generated images (x^+, x^-), and the image preferred by the user. We sample a dimension-learning pool of 50,000 images from light raters. To obtain structured supervision describing the visual properties of each image, we annotate the pool with Qwen2.5-VL-72B (Bai et al., 2025) using the ViPer attribute taxonomy (Salehi et al., 2024). The taxonomy contains approximately 512 attributes organized into 21 visual categories, including art style, color palette, composition, and mood. For each image, Qwen2.5-VL-72B assigns a binary label indicating whether an attribute is present in the image. Aggregating these annotations produces a binary attribute matrix with size $50{,}000 \times 512$, which is subsequently used to train the autoencoder.

Autoencoder for dimensionality reduction. Assigning a separate preference weight to each of the 512 attributes would require estimating a high-dimensional user vector from limited feedback, making reliable estimation difficult. We therefore learn a shared representation with $k \ll 512$ dimensions, reducing the number of user-specific parameters to estimate. The bottleneck dimension k controls the trade-off between representing diverse preferences and learning them from sparse feedback. To effectively reduce dimension, we use an autoencoder to capture potentially nonlinear attribute dependencies. Specifically, let $a_i \in \{0, 1\}^{512}$ denote the attribute vector of image x_i in the annotated pool of $N = 50{,}000$ images. An encoder E_η produces a latent representation, and a decoder D_ψ reconstructs the attribute probabilities:

$$z_i = E_\eta(a_i) \in \mathbb{R}^k, \qquad \hat{a}_i = D_\psi(z_i) \in (0, 1)^{512}. \tag{3}$$

We learn the encoder and decoder parameters (η, ψ) by minimizing

$$\mathcal{L}_{\mathrm{AE}} = \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathrm{rec}}(a_i, \hat{a}_i)\,, \tag{4}$$

where ℓ_{rec} is the binary cross-entropy loss across attributes. After training, we discard the decoder and use z_i as shared supervision targets for the evaluator $s_\theta(c_i, x_i)$. These compact scores encode visual attribute structure rather than user preferences; each user's preferences are subsequently learned through the weights λ_u over the evaluator's k outputs.

Evaluator for each dimension. We train a differentiable evaluator $s_\theta(c, x) \in \mathbb{R}^k$ that combines general human preference learning with dimension-specific visual attribute supervision. A frozen SigLIP model encodes the caption and image, whose representations are concatenated and passed to MLP scoring heads with output dimension of k. Following the pairwise preference-learning principle of PickScore (Kirstain et al., 2023), each head learns to assign higher scores to human-preferred images. We additionally supervise each head using the corresponding autoencoder latent dimension: for each dimension, we supervise the top 33% of pairs by absolute latent difference using the sign of that difference as the ordering target, then apply an attribute-ranking loss only to image pairs with different binary labels in that dimension. This auxiliary objective encourages the heads to retain distinct visual attribute signals while learning general preferences. We train the scoring heads on around 36k image pairs while keeping the SigLIP encoders and autoencoder fixed. The resulting scores form a shared basis for personalized reward fitting. Details are provided in Appendix A.

3.3 PERSONALIZED REWARD LEARNING

Given the shared evaluator and user u's offline preference pairs $\mathcal{D}_u$, we infer the weight vector λ_u that best explains the user's choices. Under the Bradley–Terry model, the probability of preferring x_i^+ to x_i^- is

$$P_u(x_i^+ \succ x_i^- \mid c_i) = \sigma\big(r_u(c_i, x_i^+) - r_u(c_i, x_i^-)\big) = \sigma\big(\lambda_u^\top \Delta s_i\big)\,, \tag{5}$$

where $\sigma(a) = (1 + e^{-a})^{-1}$ and $\Delta s_i = s(c_i, x_i^+) - s(c_i, x_i^-)$. The basic regularized maximum-likelihood estimator is

$$\hat{\lambda}_u = \arg\max_{\lambda \in \mathbb{R}^k} \left\{ \sum_{i=1}^{m_u} \log \sigma(\lambda^\top \Delta s_i) - \gamma \|\lambda\|_2^2 \right\}, \tag{6}$$

where γ controls regularization and m_u is the number of total preference pairs for user u. This formulation reduces personalization to estimating a low-dimensional vector, but its effectiveness still depends on the shared evaluator providing meaningful preference coordinates.

Alternating evaluator and weight optimization. The collected image preference pairs can further train the shared evaluator to better reflect preferences across users. We alternate between fitting each user's λ_u with the evaluator fixed and updating the evaluator heads with the user weights fixed. Repeating these two steps allows the shared preference representation and user-specific weights to adapt together, while the evaluator's image and text encoders and the diffusion model remain frozen.

Mixed-effects preference modeling. We further extend Equation 6 to a mixed-effect model, parameterizing each user's weights as $\lambda_u = G + b_u$. Here, $G \in \mathbb{R}^k$ is a shared fixed effect capturing population-level preferences, and $b_u \in \mathbb{R}^k$ is a user-specific random effect capturing deviations from this shared component. With an isotropic Gaussian prior $b_u \sim \mathcal{N}(0, I)$, we jointly estimate the shared component and user deviations by maximizing the regularized log-likelihood:

$$(\hat{G}, \{\hat{b}_u\}) = \arg\max_{G, \{b_u\}} \sum_{u \in \mathcal{U}_{\text{train}}} \sum_{i=1}^{m_u} \log \sigma\big((G + b_u)^\top \Delta s_{u,i}\big) - \gamma_b \sum_{u \in \mathcal{U}_{\text{train}}} \|b_u\|_2^2, \tag{7}$$

where $\mathcal{U}_{\text{train}}$ denotes the set of training users, and γ_b controls the allowed user-level variation, and m_u is the number of total preference pairs for user u. This mixed-effect model is particularly useful when each user provides only a few preference pairs. The shared component G is learned across users, allowing data-sparse users to benefit from population-level evidence, while the penalty on b_u limits unsupported individual deviations. This can reduce the variance of user-specific estimates and improve preference prediction under limited feedback.

3.4 INFERENCE-TIME GENERATION WITH DIRECT NOISE OPTIMIZATION

After learning the personalized reward r_u, we use Direct Noise Optimization (DNO) (Tang et al., 2024) to guide generation without updating the diffusion model. Let $M_\phi(c, x_T)$ denote the image produced by the frozen diffusion model ϕ from prompt c and initial noise $x_T \sim \mathcal{N}(0, I)$. Given a user's personal reward model $r_u(c, x)$, we aim to find the locally optimal noise $x_T^\star$ such that

$$x_T^\star = \arg\max_{x_T} r_u(c, M_\phi(c, x_T)).$$

The optimal noise is derived from gradient ascent. At the k-th optimization step, the $k+1$ noise x_T^{k+1} is updated by,

$$x_T^{k+1} := x_T^k + \ell \nabla_{x_T} r_u(c, M_\phi(c, x_T^k)). \tag{8}$$

Note that while Tang et al. (2024) mainly uses an SDE sampler and optimizes over both initial noise and step-wise injected noises, we use ODE sampler thus only the initial noise is optimized. Additional implementation details are provided in Appendix D.

4 EXPERIMENTS

Our experiments evaluate two central questions: (1) whether the learned personalized reward accurately represents a user's preferences, and (2) whether this reward can be used to generate images that match those preferences. We first study a controlled toy setting with manually specified synthetic preference profiles, and then extend the evaluation to real user preferences from Pick-a-Pic.

4.1 TOY EXAMPLE SETTING

Goal, data, and evaluation. We begin with a controlled toy experiment involving five synthetic users, each represented by a manually specified ground-truth preference vector $\lambda^\star$ (Table 5). Their preferences are defined over five candidate reward dimensions: `clip` (Radford et al., 2021), `aes` (Schuhmann, 2022), `color` (Hasler & Suesstrunk, 2003), `hps` (Wu et al., 2023), and `pickscore` (Kirstain et al., 2023). These scores capture complementary aspects of visual appearance and text–image alignment; details are provided in Appendix B.1. We repeat the experiment three times, sampling 200 captioned image pairs from Pick-a-Pic v1 in each repetition and splitting them into 100 training and 100 test pairs. After standardizing each reward dimension over the sampled images, we construct synthetic preference labels by selecting the image with the higher ground-truth weighted reward, $(\lambda^\star)^\top r(c, x)$. We then estimate λ from the training preferences by maximizing the Bradley–Terry likelihood with regularization 0.1. Evaluation considers the preference prediction accuracy on the held-out pairs. We compare the learned multi-dimensional reward with the ground-truth reward and a best one-hot baseline, which uses the highest-performing individual reward dimension available in each configuration. This comparison tests whether learning a combination of reward signals better captures a user's preferences than relying on a single evaluator.

Table 1: Held-out pairwise preference prediction accuracy(%) for the synthetic configurations. Best one-hot uses the strongest individual reward dimension. The learned combination achieves higher accuracy.

Config	PAT	Best one-hot
Uniform	**90.0** $\pm$ 4.3	77.0 $\pm$ 1.6
Normal-1	**90.7** $\pm$ 1.2	80.3 $\pm$ 3.3
Normal-2	**88.7** $\pm$ 3.6	79.0 $\pm$ 3.7
Heavy-1	**91.0** $\pm$ 5.7	81.7 $\pm$ 4.8
Heavy-2	**90.7** $\pm$ 4.4	81.0 $\pm$ 5.4

Table 2: Held-out pairwise preference accuracy (%) on 631 Pick-a-Pic users, grouped by the number of training pairs. Each user has 50 test pairs. Bold indicates the best accuracy in each group.

Training pairs	11–20	21–30	31–60	61+
PAT	**73.1** $\pm$ 1.3	**74.6** $\pm$ 0.9	**77.1** $\pm$ 0.3	**78.2** $\pm$ 0.1
w/o Mixed Effect	**73.1** $\pm$ 1.4	73.5 $\pm$ 0.9	76.4 $\pm$ 0.4	77.9 $\pm$ 0.1
w/o Alternating Training	71.2 $\pm$ 1.1	72.3 $\pm$ 0.9	75.2 $\pm$ 0.7	75.4 $\pm$ 0.2
w/o Alternating Training & Mixed Effect	71.0 $\pm$ 1.0	71.6 $\pm$ 0.7	74.7 $\pm$ 0.7	74.7 $\pm$ 0.3
Universal λ	67.6 $\pm$ 0.4	66.9 $\pm$ 0.6	69.3 $\pm$ 0.2	70.3 $\pm$ 0.3
w/ Frozen Evaluator	66.9 $\pm$ 0.4	65.3 $\pm$ 0.6	65.1 $\pm$ 0.8	66.0 $\pm$ 0.4

Results. As shown in Table 1, our learned reward combination outperforms the best single reward dimension across all five synthetic preference configurations, achieving around 90% held-out accuracy. These results demonstrate that fitting user-specific weights from limited pairwise feedback better captures diverse preference profiles than relying on any individual metric. Figure 2 qualitatively illustrates how different preference weights affect generation under the same caption. Increasing the weight on colorfulness produces more vivid, saturated images, while other weight configurations yield more muted palettes and different architectural compositions. These examples show that varying the preference weights enables different visual styles without changing the text prompt.

4.2 REAL USER: PERSONALIZED WEIGHT MODELING

We next extend our evaluation from synthetic preferences to real user choices using the training split of Pick-a-Pic (Kirstain et al., 2023). We evaluate 631 users, holding out 50 preference pairs per user for testing and using their remaining pairs for training. To examine performance under different amounts of personal feedback, we groups users by their training-set sizes, with 66, 55, 120, and 390 users in the four groups, respectively. We measure held-out pairwise preference accuracy to assess how well each method predicts a user's unseen choices. We repeat the train-test split and model user preference for 3 times and report the average held-out pair test accuracy. We evaluate variants of PAT to examine the contribution of each pipeline component. The full method combines alternating evaluator and user-weight optimization, hierarchical preference modeling, and user-specific personalization. *w/o mixed effect* replaces the joint parameterization $\lambda_u = G + b_u$ with separate per-user weight fitting regularized toward a fitted global preference vector, pushing back to an objective as



Figure 2: Qualitative examples of preference-controlled generation under a fixed prompt. Each image is generated using a different weight vector λ over (`clip`, `aes`, `color`, `hps`), producing distinct visual styles. Corners emphasize individual reward dimensions, while intermediate positions combine multiple dimensions. The corresponding weights and metric scores are shown beside each image.

Equation 6. *w/o Alternating Training* freezes the shared evaluator during user-weight fitting. We also remove both components to evaluate their combined contribution. *Universal* λ replaces user-specific weights with a shared population-level vector. Finally, *Universal* λ *w/ Frozen Evaluator* uses a fixed evaluator and a shared weight vector, removing all three components.

Table 2 shows that PAT achieves strong preference prediction performance across different amounts of user feedback. Personalization provides substantial gains: compared with using shared population-level weights, user-specific modeling improves accuracy by 5.5%–7.9% across all four groups, highlighting the importance of capturing individual preferences. Mixed-effects modeling is particularly useful when feedback is limited: it improves accuracy by 1.1% and 0.7% percentage points in the 20–30 and 30–60 groups, respectively, while matching independent fitting in the smallest group. Alternating training further improves accuracy across all groups, supporting joint adaptation of the evaluator and user-specific weights.

4.3 Ablation Studies

Autoencoder vs. other dimension reduction methods. We compare the autoencoder with two linear dimension reduction methods:(a) a correlation-based baseline that groups correlated visual attributes and averages their binary labels to construct shared preference dimensions and (b) Principle Component Analysis(PCA) that uses SVD to factorize the binary label matrix and use the top k principle dimension as shared preference dimensions. As detailed in Appendix B.5, the autoencoder achieves highest held-out pairwise accuracy at every tested dimensionality. At $k = 50$, accuracy improves from 75.39% to 77.14%. These results support Autoencoder approach constructs a more effective basis for personalization than other linear dimension reduction methods.

Preference-space dimensionality. Table 3 examines the effect of the autoencoder dimension k. Increasing k from 20 to 50 improves PAT from 75.28% to 77.14% accuracy, suggesting that additional dimensions help capture heterogeneous user preferences. Beyond $k = 50$, accuracy plateaus,

Table 3: Effect of autoencoder dimension k on held-out pairwise preference accuracy (%) for 631 users. Variants follow the component ablations in Table 2. Bold indicates the best accuracy at each dimensionality.

Variant	k=20	k=40	k=50	k=80	k=100
PAT	75.28	76.88	**77.14**	77.17	77.10
w/o Mixed Effect	75.14	76.93	76.73	76.85	76.74
w/o Alternating Training	72.99	74.80	74.65	74.60	74.51
w/o Alternating Training & Mixed Effect	72.60	74.24	74.04	74.23	74.36
Universal λ	69.58	70.36	69.53	69.45	69.53
w/ Frozen Evaluator	64.72	65.88	65.86	65.51	65.47

Table 4: General-purpose image-quality and prompt-alignment metrics for PAT and vanilla SD3.5. Win rate is the fraction of matched pairs in which our image receives a higher score.

Metric	Baseline	PAT	PAT win rate
PickScore	0.2207	0.2203	47.6%
HPS	0.2817	0.2817	52.0%
CLIP	0.2825	0.2813	49.1%
AES	5.913	5.856	45.4%

reaching 77.17% and 77.10% at $k = 80$ and $k = 100$, respectively. These results suggest that 50 dimensions are sufficient to capture the preference variation needed for accurate prediction in this setting, with further increases providing no additional benefit for the full method.

Regularization strength. We vary the regularization coefficient γ to examine its effect on preference estimation, with detailed results provided in Appendix B.4. For the sake of notation brevity, we use γ to represent the regularization strength including mixed-in effect. Held-out pairwise accuracy peaks with $\gamma = 0.03$. This trend suggests that moderate regularization helps generalization, whereas excessive regularization restricts the model's ability to capture user-specific preferences.

4.4 Real User: Personalized Image Generation

Experimental setup. We evaluate whether the learned personalized rewards improve image generation beyond ranking existing images. We include all 631 users and generate 10 images per user for each method. We use Stable Diffusion 3.5 (SD3.5) as the base generator, with 30 sampling steps and a classifier-free guidance scale of 4.5. For PAT, we optimize the initial noise using Direct Noise Optimization (DNO) with the learned personalized reward, while keeping the diffusion model frozen. DNO runs for 30 optimization steps with a learning rate of 0.005. We use GPT-5.5 as the judge under a blind, profile-aware evaluation protocol, comparing each method's outputs against vanilla SD3.5 for alignment with the user's preferences. latex We first provide GPT-5.5 with 10 image pairs and the user's recorded preferences, one pair for each test prompt, to construct a user preference profile. We then use this profile to guide GPT-5.5 in judging which generated image better aligns with the user's preferences. The prompt templates are provided in Appendix C.

Baselines. We compare against our reproduction of Personalized Preference Fine-tuning of Diffusion Models (PPD). For each user, PPD randomly selects four historical prompt–image pairs and uses a 72B Qwen vision-language model to extract the user's preferences. The resulting VLM embeddings serve as additional conditioning signals for diffusion model fine-tuning, and the texts summarized from VLM are treated as user profile. At generation time, we set PPD's user-signal guidance scale to 1. We also include a *Profile-as-Prompt* baseline, which appends the textual user profile produced by PPD's VLM to the original generation prompt, directly conditioning SD3.5 on the preference description.

Generation results. Images generated by PAT are preferred over vanilla SD3.5 in **63.0%** of comparisons, compared with 58.5% for PPD and 55.7% for Profile-as-Prompt. Our method therefore improves the win rate against the shared baseline by 4.5% and 7.3%, respectively. These results demonstrate that personalized rewards learned from offline preference pairs transfer effectively to generation-time optimization, achieving the highest win rate among the evaluated methods while keeping the diffusion model frozen.

Image-quality control. To assess whether the personalized generation gains reflect a general improvement in image quality, we compare PAT with vanilla SD3.5 over the image pairs generated

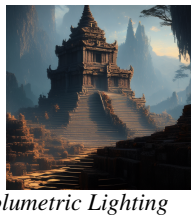
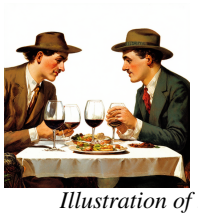
User Profile	SD 3.5	SD 3.5+Profile	PPD	PAT
Visually, they favor clean, centered compositions with strong subject isolation, soft natural textures, warm light, lush greens, creamy whites, and controlled saturated accents.				
	a mature poodle toy			
Visually, they prefer earthy textures, architectural intricacy, and compositions ... consistently reject overly glossy, cartoonish, exaggerated, or poster-like imagery; extreme close-ups, simplified fantasy shapes...				
	a Majestic Ancient Temple, Moody, Fantasy Landscape, Volumetric Lighting Digital painting, Marc Allante, Studio Ghibli, dramatic, Intricate detail			
They like refined palettes, stylized realism, clean graphic shapes, purposeful costume/armor design, and compositions that place the subject in a meaningful environment.				
	Illustration of two men sitting at a table eating food and drinking wine			
Aesthetically, they favor clean, centered compositions with a strong iconic subject, ... reject cluttered scenes, busy narrative cartoons, muddy or overly literal landscapes.				
	Cow made of Cactus oil painting			

Figure 3: **Qualitative comparison of personalization methods.** For each user, we compare images generated by vanilla SD3.5, Profile-as-Prompt, PPD, and PAT under the same prompt. We highlight the user preferences in green and dispreferences in red. More qualitative results are in Appendix E.

from the same prompts. We evaluate each pair using four general-purpose quality and alignment metrics: PickScore, HPS, CLIP, and AES. As shown in Table 4, average scores remain similar across methods, while metric-based win rates range from 45.4% to 52.0%, showing no consistent advantage for PAT on these metrics. Together with the 63.0% win rate under profile-aware evaluation, these results suggest that our gains reflect better alignment with individual preferences rather than a broad increase in generic image quality.

5 Conclusion

We introduced PAT, a framework for personalizing diffusion models at test time from historical pairwise preferences. By combining shared, attribute-grounded reward dimensions with user-specific weights, PAT enables preference estimation from limited feedback and guides generation while keeping the diffusion model frozen. Experiments demonstrate improved preference prediction and better personalized generation win rate over vanilla SD3.5, without consistent gains in generic quality metrics. These results support compact personalized rewards as an effective interface between individual preferences and inference-time image generation.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, organizing experimental results, preparing figures, providing feedback on research methodology and experiments, and summarizing existing literature. We also used LLMs to assist with implementing our method and employed a vision-language model as a judge in part of our experimental evaluation. Generative AI tools were not used to originate research ideas, formulate the methodology, design experiments, or search for related literature. The authors reviewed all AI-assisted work and verified the correctness of LLM-generated code. We take responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses existing public preference data from Pick-a-Pic and involves no new participant recruitment or manual preference annotation. Automated visual attribute annotations and model-based generation evaluations may reflect biases in the underlying models and data. Personalized generation may reinforce biased preferences or produce harmful content, while user histories and learned preference vectors may reveal sensitive information. Practical deployment should therefore protect preference data and retain appropriate content-safety safeguards.

REPRODUCIBILITY STATEMENT

Appendix A details the evaluator architecture and training objectives, Appendix A provides dataset descriptions, reward metrics, and ablation details. We will publicly release our code and experimental configurations to support reproduction and further research.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

André Barreto, Vincent Dumoulin, Yiran Mao, Mark Rowland, Nicolas Perez-Nieves, Bobak Shahriari, Yann Dauphin, Doina Precup, and Hugo Larochelle. Capturing individual human preferences with reward features. *Advances in Neural Information Processing Systems*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Hongru Cai, Yongqi Li, Tiezheng Yu, Fengbin Zhu, Wenjie Wang, Fuli Feng, and Wenjie Li. One adapts to any: Meta reward modeling for personalized llm alignment. *arXiv preprint arXiv:2601.18731*, 2026.

Jili Chen, Changqin Huang, Qionghao Huang, Yaxin Tu, Zhonglong Zheng, and Xiaodi Huang. Test-time reinforcement learning for flow matching. In *Forty-third International Conference on Machine Learning*, 2026.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. Pad: Personalized alignment of llms at decoding-time. In *International Conference on Learning Representations*, volume 2025, pp. 9014–9039, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Meihua Dang, Anikait Singh, Linqi Zhou, Stefano Ermon, and Jiaming Song. Personalized preference fine-tuning of diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8020–8030. IEEE, 2025.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Maurya Goyal, Anuj Singh, and Hadi Jamali-Rad. Unified control for inference-time guidance of denoising diffusion models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5437–5446. IEEE, 2026.

Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024.

David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pp. 87–95. SPIE, 2003.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Vivek Iyer, Mozhdeh Gheini, Christian Federmann, Liling Tan, and Alexandra Birch. Personalized reward modelling from implicit user preferences.

Zhou Jiang, Yandong Wen, and Zhen Liu. Rethinking preference alignment for diffusion models with classifier-free guidance. *arXiv preprint arXiv:2602.18799*, 2026.

Luozhijie Jin, Zijie Qiu, Jie Liu, Zijie Diao, Lifeng Qiao, Ning Ding, Alex Lamb, and Xipeng Qiu. Inference-time alignment control for diffusion models with reinforcement learning guidance. *arXiv preprint arXiv:2508.21016*, 2025.

Hyungjin Kim, Seokho Ahn, and Young-Duk Seo. Draw your mind: Personalized generation via condition-level modeling in text-to-image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17171–17180. IEEE, 2025a.

Minbeom Kim, Kang-il Lee, Seongho Joo, Hwaran Lee, Thibaut Thonet, and Kyomin Jung. Drift: Decoding-time personalized alignments with implicit user preferences. In *EMNLP (Findings)*, pp. 6107–6126, 2025b.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Jeongeun Lee, Ryang Heo, and Dongha Lee. Personalized reward modeling for text-to-image generation. *arXiv preprint arXiv:2511.19458*, 2025a.

Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18465–18475, 2025b.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, et al. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Xinyu Li, Ruoming Jin, Jianfeng Zhu, Ruixin Guo, and Zhi Liu. Aplaud: Adaptive personalized low-rank decomposition for user-specific llm. *arXiv preprint arXiv:2609.04738*, 2026a.

Yang Li, Songlin Yang, Xiaoxuan Han, Wei Wang, Jing Dong, Yueming Lyu, and Ziyu Xue. Instant preference alignment for text-to-image diffusion models. *IEEE Transactions on Visualization and Computer Graphics*, 2026b.

Yuming Li, Qingyu Li, Chengyu Bai, Xiangyang Luo, Zeyue Xue, Wenyu Qin, Meng Wang, Yikai Wang, and Shanghang Zhang. Aegpo: Adaptive entropy-guided policy optimization for diffusion models. *arXiv preprint arXiv:2602.06825*, 2026c.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025b.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(5):7, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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

Christoph Schuhmann. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*, 2025.

Yutong Song, Jiang Wu, Weijia Zhang, Chengze Shen, Shaofan Yuan, Weitao Lu, Jian Wang, Yu Wang, Nikil Dutt, and Amir M Rahmani. Card: Cluster-level adaptation with reward-guided decoding for personalized text generation. *arXiv preprint arXiv:2601.06352*, 2026.

Charles Spearman. " general intelligence" objectively determined and measured. 1961.

Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin, et al. Instant personalized large language model adaptation via hypernetwork. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23557–23580, 2026.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Tuning-free alignment of diffusion models with direct noise optimization. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024.

Louis Leon Thurstone. Multiple factor analysis. *Psychological review*, 38(5):406, 1931.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. *arXiv preprint arXiv:2505.11245*, 2025.

Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Kumar Avinava Dubey, et al. Conditional language policy: A general framework for steerable multi-objective finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2153–2186, 2024.

Xuefei Wang, Jun Han, Zixuan Wang, Qingkai Zeng, Xiao Wang, Ruijie Wang, and Jianxin Li. Learning to adapt cross-domain preferences via meta-lora for llm personalization. *arXiv preprint arXiv:2608.12389*, 2026a.

Zihao Wang, Yuxiang Wei, Xinpeng Zhou, Tianyu Zhang, Tao Liang, Yalong Bai, Hongzhi Zhang, and Wangmeng Zuo. Premier: Personalized preference modulation with learnable user embedding in text-to-image generation. *arXiv preprint arXiv:2603.20725*, 2026b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Yue Wu, Chengbing Wang, Yimeng Bai, Xiaoyan Zhao, Yang Zhang, and Fuli Feng. Prefreward: Learning user preference matrix for personalized text generation. *arXiv preprint arXiv:2607.21067*, 2026.

Xin Xie, Fan Zhang, and Dong Gong. Test-time weak-to-strong alignment: Transferring implicit rewards from weak to strong flow models. *arXiv preprint arXiv:2609.05968*, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024a.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024b.

Po-Hung Yeh, Kuang-Huei Lee, and Jun-Cheng Chen. Training-free diffusion model alignment with sampling demons. *arXiv preprint arXiv:2410.05760*, 2024.

Xin Yu, Hanwen Xing, and Lingzhou Xue. Exact: Explicit attribute-guided decoding-time personalization. *arXiv preprint arXiv:2602.17695*, 2026.

Kevin Zhai, Utsav Singh, Anirudh Thatipelli, Souradip Chakraborty, Anit Kumar Sahu, Furong Huang, Amrit Singh Bedi, and Mubarak Shah. Mira: Towards mitigating reward hacking in inference-time alignment of t2i diffusion models. *arXiv preprint arXiv:2510.01549*, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. Dspo: Direct score preference optimization for diffusion model alignment. International Conference on Learning Representations (ICLR 2025), 2025.

A EVALUATOR TRAINING

Architecture and supervision. Our evaluator produces a vector of scores $s_\theta(c, x) = [s_{\theta,1}(c, x), \ldots, s_{\theta,k}(c, x)]^\top$ for a caption–image pair (c, x). We encode the caption and image using frozen SigLIP encoders and concatenate their representations into a 2,304-dimensional feature vector(1152 dimension for both image and caption embeddings). This vector is passed through an MLP consisting of a linear layer with 512 hidden units, a GELU activation, and a linear output layer producing k scalar scores. The output dimensions share the hidden representation but have separate output weights. We train the MLP on approximately 36,000 image pairs using two sources of supervision: human pairwise preferences and dimension-specific targets derived from the pretrained autoencoder. The former provides a common preference-learning objective across dimensions, while the latter encourages distinct visual attribute responses.

Human preference supervision. Let (c_i, x_i^+, x_i^-) denote a training pair, where x_i^+ is preferred to x_i^- by the user. For dimension j, we define the predicted score difference as

$$d_{ij} = s_{\theta,j}(c_i, x_i^+) - s_{\theta,j}(c_i, x_i^-). \tag{9}$$

Following the pairwise preference-learning principle of PickScore (Kirstain et al., 2023), we encourage each output dimension to assign a higher score to the preferred image. We use the binary cross-entropy loss

$$\ell_{\mathrm{BCE}}(d, y) = -y \log \sigma(d) - (1 - y) \log\big(1 - \sigma(d)\big), \tag{10}$$

where $d \in \mathbb{R}$ is a predicted score difference, $y \in \{0, 1\}$ indicates whether the first image should receive a higher score than the second, and σ is the sigmoid function. Since each pair is ordered by human preference, the target for this objective is always 1. For a minibatch $\mathcal{B}$, the preference loss is

$$L_{\mathrm{prefer}} = \frac{1}{|\mathcal{B}|k} \sum_{i \in \mathcal{B}} \sum_{j=1}^{k} \ell_{\mathrm{BCE}}(d_{ij}, 1). \tag{11}$$

Visual attribute supervision. To construct dimension-specific supervision, we use the fixed encoder of the pretrained autoencoder:

$$z_i^+ = E_\eta(a(x_i^+)), \qquad z_i^- = E_\eta(a(x_i^-)), \tag{12}$$

where $a(x) \in \{0, 1\}^{512}$ is the binary visual attribute annotation of image x, E_η is the autoencoder encoder with fixed parameters η, and $z_i^+, z_i^- \in \mathbb{R}^k$ are the corresponding continuous latent representations.

We first compute the latent difference for each pair and dimension:

$$\delta_{ij} = z_{ij}^+ - z_{ij}^-. \tag{13}$$

For each dimension j, we rank the training pairs by $|\delta_{ij}|$ and retain the top 33% for attribute supervision. Let $\mathcal{I}_j$ denote the indices of the retained pairs. We define the supervision mask and binary ordering target as

$$m_{ij} = \mathbf{1}[i \in \mathcal{I}_j], \qquad y_{ij}^{\mathrm{attr}} = \mathbf{1}[\delta_{ij} > 0]. \tag{14}$$

Thus, pair selection is based on the magnitude of the continuous latent difference, while the binary target records its direction. We do not binarize the individual images' latent values. This procedure focuses attribute supervision on pairs with relatively large differences within each dimension.

For a minibatch $\mathcal{B}$, the masked attribute loss is

$$L_{\mathrm{attr}} = \frac{\sum_{i \in \mathcal{B}} \sum_{j=1}^{k} m_{ij}\, \ell_{\mathrm{BCE}}(d_{ij}, y_{ij}^{\mathrm{attr}})}{\max\left(1, \sum_{i \in \mathcal{B}} \sum_{j=1}^{k} m_{ij}\right)}. \tag{15}$$

Pairs outside $\mathcal{I}_j$ contribute no attribute loss for dimension j, but still contribute to the human preference loss. We optimize the sum of the two losses:

$$L_{\mathrm{eval}} = L_{\mathrm{prefer}} + L_{\mathrm{attr}}. \tag{16}$$

Together, these objectives encourage the evaluator to balance general preference prediction with dimension-specific visual attribute supervision. Only the MLP parameters are updated; the SigLIP encoders and autoencoder remain fixed during this training stage. The autoencoder outputs define supervision masks and targets, whereas both losses operate on the evaluator's predicted score differences. The evaluator therefore learns continuous, attribute-grounded preference scores rather than directly reconstructing the autoencoder latent values. These scores are subsequently combined using user-specific weights to construct personalized rewards.

B More Details about Experiments

B.1 Reward Metrics

Our controlled experiments define synthetic user preferences over five reward dimensions: `clip`, `aes`, `color`, `hps`, and `pick`. These scores measure text–image alignment, visual aesthetics, colorfulness, and learned human preferences. Although the dimensions may be correlated, they provide distinct signals that can be combined to construct different preference profiles.

CLIP similarity (`clip`). CLIP (Radford et al., 2021) learns a shared image–text embedding space through contrastive training on image–text pairs. We use the similarity between the embeddings of a generated image and its corresponding prompt to measure semantic alignment. Higher scores indicate stronger correspondence between the image content and the text description. This dimension captures prompt relevance, although it does not directly measure visual quality or individual user preferences.

Aesthetic score (`aes`). The aesthetic predictor (Schuhmann, 2022) estimates an image's aesthetic appeal using a prediction head applied to its CLIP image embedding. It is trained to predict human aesthetic ratings and provides an image-only quality signal, without conditioning on the prompt. This dimension captures visual appeal that may not be reflected by text–image similarity alone.

Colorfulness (`color`). The colorfulness metric of Hasler & Suesstrunk (2003) quantifies image colorfulness using statistics of red–green and yellow–blue opponent channels. It combines the variability and mean magnitude of these color signals to produce an image-only score. Unlike the learned reward models, this metric measures a specific visual property without preference supervision. Higher values indicate more colorful images, which need not be preferred by every user; its contribution to the synthetic reward is determined by the user's weight vector.

Human Preference Score (`hps`). Human Preference Score v2 (Wu et al., 2023) is a CLIP-based scoring model fine-tuned on the Human Preference Dataset v2 (HPD v2). Given a prompt and an image, it predicts a score intended to reflect human preferences over generated images. Compared with the original CLIP similarity, this score incorporates explicit human preference supervision and provides a broader signal of perceived image quality and prompt compatibility.

PickScore (`pickscore`). PickScore (Kirstain et al., 2023) is a CLIP-based preference model trained on pairwise human feedback from Pick-a-Pic. It assigns a scalar score to a prompt–image pair using the scaled similarity between its learned text and image embeddings. Training encourages the relative scores of competing images to match observed human choices. Like HPS, PickScore provides a population-level preference signal; the two models are trained on different preference datasets and may therefore rank the same images differently.

Combining reward dimensions. Because the five metrics have different numerical scales, we standardize each dimension over the sampled images before constructing synthetic preferences, as described in Section 4.1. Let $\widetilde{r}_j(c, x)$ denote the standardized score of dimension j. For a synthetic user with ground-truth weights $\lambda^\star$, the reward is

$$r_{\lambda^\star}(c, x) = \sum_{j=1}^{5} \lambda_j^\star \widetilde{r}_j(c, x).$$

For each image pair associated with the same prompt, we assign the synthetic preference label to the image with the higher weighted reward. This construction allows us to evaluate preference learning against a known reward function.

B.2 PICK-A-PIC DATASET

Pick-a-Pic (Kirstain et al., 2023) is an open dataset of real user preferences for text-to-image generation, collected through a web application in which users submit prompts and compare generated images. Each comparison associates a text prompt with two candidate images and the user's preference feedback. The dataset also permits ties and includes user identifiers, enabling preference observations to be grouped into individual user histories. Its prompts and judgments arise from users interacting with image generation systems, providing a setting for studying heterogeneous visual preferences.

We use Pick-a-Pic v1 in two experimental settings. In the controlled toy experiments, it supplies captioned image pairs, while preference labels are constructed from the synthetic weighted reward rather than the original human choices. Each repetition samples 200 pairs, with 100 used for fitting the preference weights and 100 for evaluation. In the real-user experiments, we use the recorded human choices to fit and evaluate personalized rewards. Following Section 4.2, we select the 631 users with the largest numbers of preference pairs from the training split, hold out 50 pairs per user for evaluation, and use the remaining pairs for training. These experiments assess, respectively, recovery of controlled synthetic preferences and prediction of unseen choices from real users.

B.3 TOY EXAMPLE λ WEIGHTS

Table 5 reports the ground-truth and fitted preference weights for the five synthetic users introduced in Section 4.1, with fitted weights averaged over three independently sampled datasets. The configurations cover both broadly distributed preferences (uniform and half-normal) and more concentrated preferences (Pareto). Weight recovery is strongest for the heavy-1 configuration, with a reported cosine similarity of 0.990, and the fitted vector preserves its emphasis on the third and fourth reward dimensions. Recovery is less accurate for the uniform and half-normal configurations, where the fitted weights place greater emphasis on the fifth dimension. These results highlight that recovering the exact preference weights remains challenging with limited pairwise feedback, even in this controlled setting.

Table 5: Ground-truth and fitted preference weights for five synthetic users with uniform, half-normal, or Pareto weight configurations. Each row reports the ground-truth vector $\lambda^\star$, the fitted vector $\hat{\lambda}$ averaged over three datasets, and the reported cosine similarity between fitted and ground-truth weights (higher is better).

Config	Family	Ground-truth $\lambda^\star$	Fitted $\hat{\lambda}$ (mean over 3 datasets)	$\cos(\hat{\lambda}, \lambda^\star)$
uniform	5-dim uniform	$(0.20, 0.20, 0.20, 0.20, 0.20)$	$(0.12, 0.09, 0.16, 0.22, 0.41)$	0.869
normal-1	half-normal	$(0.22, 0.17, 0.19, 0.18, 0.24)$	$(0.15, 0.06, 0.13, 0.19, 0.47)$	0.870
normal-2	half-normal	$(0.24, 0.11, 0.18, 0.25, 0.22)$	$(0.17, 0.00, 0.12, 0.27, 0.44)$	0.883
heavy-1	Pareto	$(0.06, 0.03, 0.44, 0.40, 0.07)$	$(0.00, 0.00, 0.51, 0.39, 0.10)$	0.990
heavy-2	Pareto	$(0.38, 0.06, 0.10, 0.26, 0.20)$	$(0.34, 0.00, 0.04, 0.25, 0.37)$	0.939

B.4 ABLATION ON REGULARIZATION

Figure 4 examines the effect of the regularization coefficient γ on held-out pairwise preference accuracy. Among the tested values, accuracy peaks at approximately 77.3% with $\gamma = 0.03$. Increasing γ beyond this value progressively reduces accuracy, reaching approximately 74.7% at $\gamma = 1$. This pattern suggests that moderate regularization provides a useful balance between fitting observed preferences and generalizing to unseen comparisons, whereas excessive regularization limits the model's ability to capture user-specific preferences.

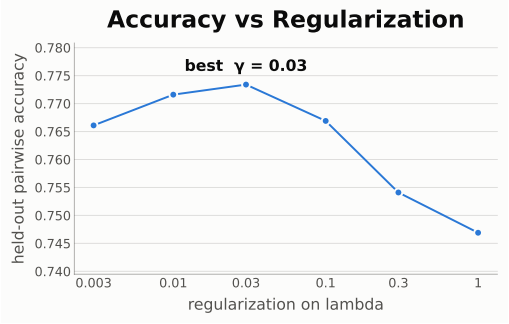


Figure 4: Ablation on the regularization coefficient γ.

Table 6: Comparison of correlation-based dimension reduction, PCA based dimension reduction, and an autoencoder across four preference-space dimensions k. Values are reported as percentages.

k	Correlation	PCA	Autoencoder
20	74.52%	74.41%	75.28%
30	74.81%	75.61%	76.29%
40	74.87%	76.05%	76.88%
50	75.39%	76.63%	77.14%

B.5 Dimension Reduction with Correlation

Correlation-based dimension construction. Let $a(x) \in 0, 1^{512}$ denote the attribute annotation vector for image x. We first compute pairwise correlations between attribute dimensions across the annotated images, then group highly correlated attributes into k clusters, denoted by $\mathcal{C}_1, \ldots, \mathcal{C}_k$. Each cluster defines one shared preference dimension. For an image x, we compute its score along dimension j by averaging the binary labels of the attributes belonging to that cluster:

$$z_j^{\text{corr}}(x) = \frac{1}{|\mathcal{C}_j|} \sum_{\ell \in \mathcal{C}_j} a_\ell(x), \qquad j = 1, \ldots, k.$$

For example, if a cluster contains ten attributes and four are present in an image, its score is 0.4. This aggregation converts the original binary attribute vector into a compact representation $z^{\text{corr}}(x) \in [0, 1]^k$, where each coordinate measures the fraction of attributes present within a correlated group.

Principal component analysis. We also consider PCA as a linear alternative to the autoencoder for constructing the shared attribute representation. We use the same annotated image pool, with each image represented by its binary attribute vector $a(x) \in \{0, 1\}^{512}$. Let $\bar{a}$ denote the mean attribute vector over this pool. We center the attribute vectors and compute the k principal directions with the largest eigenvalues of the empirical covariance matrix. Collecting these directions as the columns of $W_k \in \mathbb{R}^{512 \times k}$, we represent each image as

$$z^{\text{PCA}}(x) = W_k^\top \big(a(x) - \bar{a}\big) \in \mathbb{R}^k. \tag{17}$$

The mean vector and projection matrix are estimated from the representation-learning pool and subsequently held fixed. Unlike correlation-based grouping, which averages attributes within each cluster, PCA constructs orthogonal linear combinations that retain the greatest attribute variance at the chosen dimensionality. In contrast to the autoencoder, its representation is restricted to a linear projection.

Evaluator and personalized reward. For both alternatives, we replace the autoencoder latent targets with the corresponding k-dimensional representation while retaining the evaluator architecture and training objectives described in Appendix A. Specifically, we compute the representation difference between the two images in each training pair. For each dimension, we select the top 33% of pairs by absolute difference and use the sign of the difference as the binary attribute-ordering target. The evaluator is trained using this dimension-specific supervision together with human preference supervision. We then fit user-specific weights under the same personalized reward-learning procedure:

$$r_u(c, x) = \lambda_u^\top s_\theta(c, x). \tag{18}$$

This comparison varies the construction of the shared attribute representation while keeping the downstream evaluator and personalization procedure unchanged.

Table 7: **Prompt templates for user profiling and GPT-5.5 evaluation.** The first prompt constructs a taste profile from preferred–dispreferred image pairs. The second compares two generated images using that profile. Here, {k} is the number of preference pairs and {n} is the number of attached images ($n = 2k$); {profile} and {caption} denote the resulting profile and evaluation caption.

Stage	Prompt template
1. User profiling	You are profiling one user's personal visual taste from {k} preference examples. {n} images are attached in **THIS EXACT ORDER**, as {k} consecutive pairs: image 1 = **PREFERRED**, image 2 = **DISPREFERRED** (same caption) image 3 = **PREFERRED**, image 4 = **DISPREFERRED** (same caption) ... and so on for all {k} pairs. Within each pair the two images answer the same prompt, so what separates them is taste, not subject matter. Study what the preferred image has that the dispreferred one lacks, and what the dispreferred one has that the user rejected. Write a concise profile of this user's taste, in two short paragraphs: 1. **TOPICS/subjects** they are drawn to (what the images are of). 2. Their **VISUAL/AESTHETIC preferences**, described using these aspects where relevant: Art Style, Brushstrokes, Color Palette, Composition, Skill, Detail, Hues, Saturation, Line, Shape, Value, Pattern, Texture, Artistic Medium, Mood, Perspective, Depth, Movement, Form, Juxtaposition, Iconography. State what they favour **AND** what they consistently reject. **IMPORTANT:** describe only **CONTENT** topics and **STYLISTIC** preferences. Do **NOT** mention image quality, resolution, sharpness, realism-as-quality, rendering fidelity, artifacts, or which images are 'better made' — those are not part of the user's taste. Output only the profile text.
2. Pairwise judging	Here is a profile of a user's visual taste (topics + stylistic preferences): {profile} Two new AI-generated images are attached: img_P (first) and img_Q (second), both for the caption "{caption}". Based **ONLY** on how well each image matches this user's **TOPICS** and **STYLISTIC** preferences above — **NOT** on raw image quality or which looks better made — which image would this user more likely prefer? Answer with "P" or "Q".

C Prompts for Profile-Aware Evaluation

Table 7 presents the two-stage prompting protocol used for profile-aware evaluation. First, GPT-5.5 constructs a user preference profile from 10 image pairs, one for each test prompt. Within each pair, the preferred image is followed by the dispreferred image, allowing the model to summarize both favored and rejected visual characteristics. The profile covers subject matter and stylistic preferences while explicitly excluding judgments of image quality or rendering fidelity. Second, GPT-5.5 receives this profile, an evaluation caption, and two generated images, and selects the image that better matches the user's preferences. In the templates, {k} and {n} are set to 10 and 20, respectively; {profile} and {caption} are replaced with the generated profile and evaluation caption.

D Image Generation with Direct Noise Optimization

Deterministic sampling process. We apply Direct Noise Optimization (DNO) (Tang et al., 2024) to maximize the learned personalized reward at inference time while keeping the image generator frozen. Stable Diffusion 3.5 uses a flow-matching model with a deterministic ODE-based sampler in our implementation. Once the prompt, initial latent, and sampling configuration are fixed, no additional randomness is introduced during denoising. Consequently, the complete generation process can be represented as a deterministic mapping

$$x = M_\phi(c, x_T), \tag{19}$$

where c is the text prompt, x is the clean image, $x_T \in \mathbb{R}^d$ is the initial latent noise, and ϕ denotes the frozen diffusion generator, including its denoising sampler and VAE decoder. The initial latent is sampled from the Gaussian prior, $x_T^{(0)} \sim \mathcal{N}(0, I)$. This setting differs from the main stochastic formulation studied by Tang et al. (2024), which jointly optimizes the initial noise and the additional noises injected at individual sampling steps. Because our deterministic sampler introduces no step-wise noise, its output is fully determined by x_T, and the initial latent is the only optimization variable.

Noise optimization. For a user u with personalized reward function $r_u(c, x)$, the unregularized DNO objective is

$$\max_{x_T} \quad r_u(c, M_\phi(c, x_T)) \,. \tag{20}$$

At each optimization step, we run the complete differentiable sampling trajectory from the current initial latent and evaluate the personalized reward on the resulting image. Gradients are then backpropagated through the frozen generator to x_T. Although gradients pass through M_ϕ and r_u, all parameters of the generator and personalized reward model remain fixed; only x_T is updated. Starting from $x_T^{(0)} \sim \mathcal{N}(0, I)$, a gradient-ascent update can be written as

$$x_T^{(k+1)} = x_T^{(k)} + \ell \, \nabla_{x_T} r_u(c, M_\phi(c, x_T))|_{x_T = x_T^{(k)}} \,, \tag{21}$$

where ℓ denotes the optimization step size. In practice, we perform this update using Adam for 30 optimization steps with learning rate 0.005.

E More Qualitative Results

We add more qualitative examples in Figure 5. In general, PAT generates images that are closer to user profile.

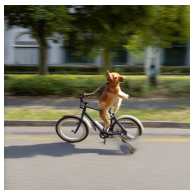
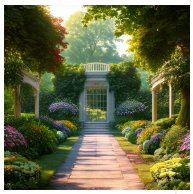
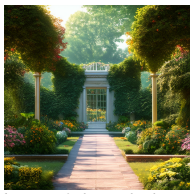
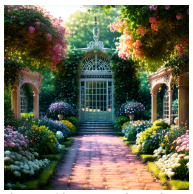
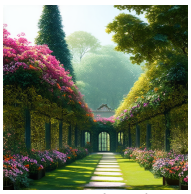
User Profile	SD 3.5	SD 3.5+Profile	PPD	PAT
They favor cooler or earthy palettes with selective strong accents, textured landscapes... They consistently reject overly polished fantasy spectacle, high-saturation blue/green casts				
	photo of a dirt road in a forest, pine trees			
Visually, they favor clean, polished, high-saturation images with bright eyes, direct engagement, centered compositions, warm or vivid palettes, smooth surfaces, and clear character silhouettes.				
	a dog riding bicycle			
Aesthetically, they prefer vivid color, warm natural light, ... They favor recognizable forms with emotional immediacy, saturated blues / oranges, and compositions centered on a single living figure.				
	A cactus playing a guitar on a beach			
Visually, they favor ornate, composed images with strong framing, clear central forms, full-scene context, rich pattern, decorative linework, and a sense of depth.				
	concept art of a beautiful english garden, 4k, vivid colors, intricate details very realistiv, cinematic lighting, volumetric lighting			
Aesthetically, they favor vivid, strong value contrast, glowing bokeh or star patterns, and theatrical central compositions... reject latter, quieter, greyer, more literal scenes; sparse compositions..				
	sorrow			
They favor grounded compositions with environmental context, gentle depth, tactile textures, and stylized anime/painterly looks when the subject is fantasy.				
	Egirl with red hair, gorgeous, high-quality, beautiful			
Aesthetically, they favor saturated color, strong lighting, warm highlights, rich contrast, ... They reject flatter, emptier, more muted or gray compositions				
	concept art, drawing of fantasy ballroom at night with faceless dancers			

Figure 5: More qualitative comparison of personalization methods.