

Personalized Preference Fine-tuning of Diffusion Models

Meihua Dang^{1*} Anikait Singh^{1*} Linqi Zhou^{1,2} Stefano Ermon¹ Jiaming Song²
¹Stanford University, ²Luma AI

{mhdang, anikait, ermon}@cs.stanford.edu, {alexzhou, jiaming}@lumalabs.ai

Abstract

*RLHF techniques like DPO can significantly improve the generation quality of text-to-image diffusion models. However, these methods optimize for a **single reward** that aligns model generation with population-level preferences, neglecting the nuances of individual users' beliefs or values. This lack of personalization limits the efficacy of these models. To bridge this gap, we introduce PPD, a **multi-reward optimization objective** that aligns diffusion models with **personalized preferences**. With PPD, a diffusion model learns the individual preferences of a population of users in a few-shot way, enabling generalization to unseen users. Specifically, our approach (1) leverages a vision-language model (VLM) to extract personal preference embeddings from a small set of pairwise preference examples, and then (2) incorporates the embeddings into diffusion models through cross attention. Conditioning on user embeddings, the text-to-image models are fine-tuned with the DPO objective, simultaneously optimizing for alignment with the preferences of multiple users. Empirical results demonstrate that our method effectively optimizes for multiple reward functions and can interpolate between them during inference. In real-world user scenarios, with as few as four preference examples from a new user, our approach achieves an average win rate of 76% over Stable Cascade, generating images that more accurately reflect specific user preferences.*

1. Introduction

Diffusion models [12, 40, 42, 43] have achieved great success in text-to-images generation tasks, as demonstrated by models such as DALL-E 2 [33], Imagen [35], Stable Diffusion [29, 34], and Stable Cascade [27]. These large-scale models, often trained through supervised fine-tuning on text-image datasets containing high-quality aesthetic images such as LAION-5B [38], demonstrate impressive image quality and a remarkable visual understanding of text prompts. Recent advances focus on fine-tuning diffusion

models to better align with human preferences. Methods such as DDPO [2] and DPOK [9] employ reinforcement learning from human feedback (RLHF) techniques to optimize human-alignment reward functions. Alternatively, Diffusion-DPO [45] utilize direct preference optimization (DPO) [32] to directly optimizes for human preferences using an offline ranking dataset. SD3 [7] shows that fine-tuning diffusion models with human feedback techniques can significantly enhance overall image quality.

Despite these promising results, Diffusion-DPO aligns a model with a single reward that represents population-level human preferences. These fine-tuning techniques are thus insufficient for achieving personalized preference alignment. As studied in domains such as language [10, 36, 44], individual users have unique preferences—for instance, some may favor images with brighter colors, while others might prefer images with centralized foregrounds. Fine-tuning a separate model for each person based on their preference data is impractical and does not scale well, limiting the feasibility of personalized alignment. Additionally, this may limit cross-user generalization.

A complementary research direction introduces additional control mechanisms into diffusion models to enable customization and personalization. Approaches such as IP-Adapter [51] and CtrlNet [54] allow for controlled image generation by using an additional conditioning image that encodes personalization features. However, these methods are constrained by their reliance on single-image inputs and do not directly learn or adapt to human preferences.

To address these limitations, we propose PPD, a general framework that incorporates personalization into a diffusion model preference alignment through multi-reward optimization. Our approach learns a diverse set of human preferences within a single model and generalizes effectively to new users. Personalization in our approach requires minimal additional data, as we label samples in the preference dataset with a unique identifier for each user. This annotation is inexpensive to obtain and commonly included in datasets such as Pick-a-Pic [16]. We leverage vision-language models (VLMs) such as LLaVA-OneVision [20] to extract user preference embeddings from

*Equal contributions.

Work done during MD's internship at Luma AI.

few-shot pairwise preference examples for each user. We show that these few-shot text embeddings from preference pairs serve as effective features to represent reward functions under the Bradley-Terry model [3] and thus are suitable for conditioning the diffusion model. Next, adopting Stable Cascade [27] as our base model, we condition the diffusion model on these user embeddings through additional cross-attention layers and fine-tune it with a variant of the Diffusion-DPO [45] objective for preference alignment.

Our empirical evaluations demonstrate that PPD can effectively optimize multiple rewards including CLIP [13, 31], Aesthetic [37], and HPS [47], significantly outperforming baselines. It also interpolates smoothly between different rewards during inference. Additionally, in real-world user scenarios, PPD aligns more closely with individual user preferences on the dataset Pick-a-Pic [16]. Notably, with only a few-shot ranking examples for a new user, PPD generate images that align closely with that user’s specific preference and outperforms prior approaches such as Stable Cascade, achieving an average win rate of 76%, as evaluated by GPT4o [26].

2. Related Work

Aligning Diffusion Models Aligning text-to-image diffusion models has been extensively studied. Some methods optimize with online RL using a Bradley Terry reward function [2, 9, 55]. Others fine-tune diffusion models to increase the reward of generated images by gradient back-propagation on the reward function [5, 30, 49]. Recently proposed approaches [45, 50] leverage the duality of the reward learning and policy optimization from DPO [32] to eliminate the need of on-policy samples and only utilize offline human preference data. Alternatively, DiffusionKTO [22] replaces DPO with KTO [8], eliminating the need for paired data. Additionally, some methods modify the training data distribution on visually appealing and textually cohered data [6, 18, 48], while others improve text accuracy by re-captioning pre-collected web images [11, 39].

Controllable Generation for Diffusion Models To address the lack of fine-grained control in image synthesis, various methods add conditional controls to text-to-image diffusion models. CtrlNet [54] and T2I-Adapter [25] introduce additional UNet modules that encode spatial information, such as edges and human poses, derived from conditioning images. IP-Adapter [51] introduces additional cross-attention layers to incorporate high-level semantic information from reference images.

LLM Alignment The alignment of large language models (LLMs) using RLHF objectives has been extensively explored [1, 32, 56], demonstrating effective instruction-following and alignment to population values. However,

a key challenge remains in representing the diversity of human preferences [36], as these models often capture an averaged population preference rather than accounting for individual differences, leading to biases and reduced alignment with underrepresented communities. *Pluralistic Alignment* [10, 44] addresses this issue by modeling a distribution of individuals, each with distinct values and perspectives, with methods like distributional and overton alignment enhancing calibration across diverse user groups. Several works have investigated LLM personalization through the re-parameterization of reward models [23, 28, 52], enabling a more tailored user experience. However, these works leverage synthetic users and study a limited number of users, particularly in text and robotic control domains.

3. Preliminary

We introduce the preliminaries behind diffusion models, DPO and its adaption to diffusion models.

Diffusion Models Diffusion models are generative models consisting two processes: a forward (diffusion) process and a backward (denoising) process [12, 40, 42, 43]. In the forward process, Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is progressively added to a data distribution $P_{\text{data}}(\mathbf{x}_0)$ over T steps, guided by noise scheduling functions α_t and σ_t . At timestep $t \in [0, T]$, noisy images \mathbf{x}_t are generated via $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$. In the backward process, samples are generated from Gaussian noise with a learnable model, which can also be conditioned on additional inputs, such as text in the case of text-to-image diffusion models. Typically, diffusion models – denoted as ϵ_θ – are trained to predict the Gaussian noise ϵ from noisy images \mathbf{x}_t at timestep t with conditioning \mathbf{c} , using a denoising objective:

$$L_{\text{DM}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\omega(\lambda_t) \|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)\|_2^2], \quad (1)$$

where $\lambda_t = \log [\alpha_t^2 / \sigma_t^2]$ is a signal-to-noise ratio, $\omega(\lambda_t)$ is a pre-specified weighting function [14]. Once the model ϵ_θ is trained, images can be generated from random noise using accelerated sampling methods such as DDIM [41].

Direct Preference Optimization (DPO) DPO [32] is a fine-tuning paradigm that aligns autoregressive large language models (LLMs) with human preferences without relying on traditional reinforcement learning techniques. Unlike Reinforcement Learning from Human Feedback (RLHF), which involves training a reward model and then fine-tuning the LLM using reinforcement learning, DPO directly optimizes the model based on human preference data. Given a dataset containing ranking pairs $(\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-)$ where \mathbf{c} being the prompt and $\mathbf{x}_0^+, \mathbf{x}_0^-$ being the preferred and dis-preferred responses labeled by human, the training objective optimizes a simple loss under which we can obtain an

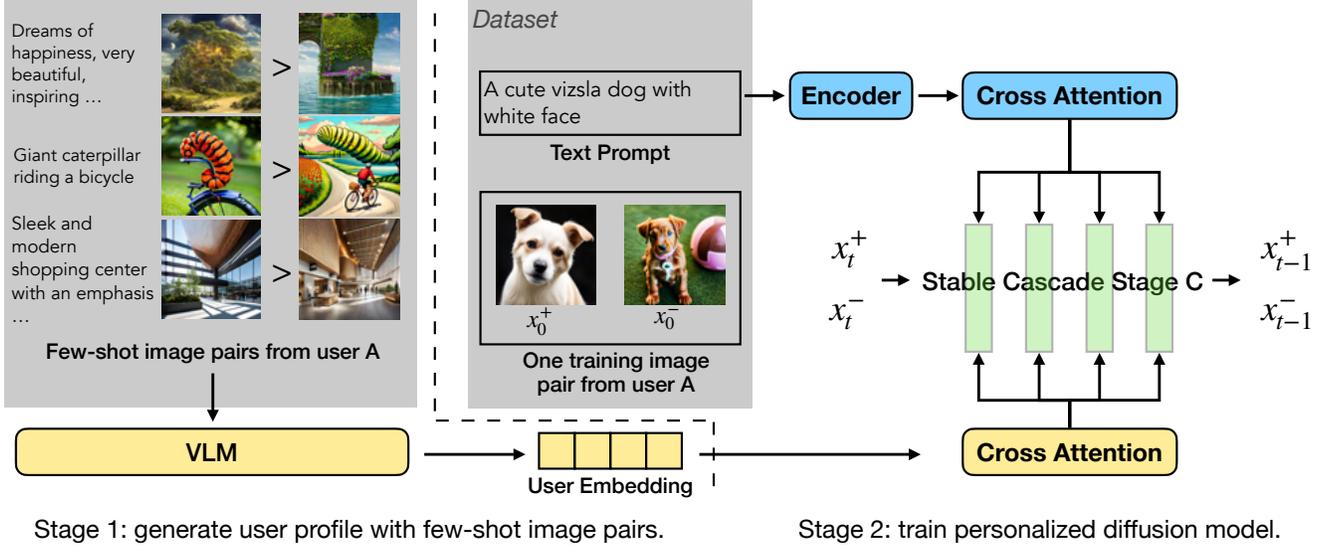


Figure 1. **The overall architecture of PPD.** In Stage 1, user embedding are generated with few-shot preference examples using a VLM. In Stage 2, we fine-tune diffusion models on the preference datasets with the user embedding as conditioning added to cross-attention.

optimal policy:

$$L_{\text{DPO}}(\theta) = -\mathbb{E}_{\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-} [\log \sigma(\beta \Delta)], \text{ where} \quad (2)$$

$$\Delta = \left(\log \frac{p_{\theta}(\mathbf{x}_0^+ | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_0^+ | \mathbf{c})} - \log \frac{p_{\theta}(\mathbf{x}_0^- | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_0^- | \mathbf{c})} \right),$$

where σ is the sigmoid function and β is a hyperparameter determining how much $p_{\theta}(\mathbf{x}_0 | \mathbf{c})$ may deviate from a reference distribution $p_{\text{ref}}(\mathbf{x}_0 | \mathbf{c})$. Intuitively, the objective aims to increase the likelihoods of preferred responses while decreasing that of dispreferred ones.

Diffusion DPO The DPO objective in Eq. (2) requires optimization of log-likelihoods $\log p_{\theta}(\mathbf{x}_0^+ | \mathbf{c})$, which is analytically intractable for diffusion models. Diffusion-DPO [45] leverages the fact that the (appropriately weighted) training objective of diffusion models as in Eq. (1) can be viewed as the evidence lower bound (ELBO) of their log-likelihoods. It derives an upper bound to DPO for diffusion models to align them with human preferences:

$$L_{\text{Diff-DPO}}(\theta) = -\mathbb{E}_{\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-} [\log \sigma(-\beta T \omega(\lambda_t) \Delta)], \text{ where} \quad (3)$$

$$\Delta = \|\epsilon^+ - \epsilon_{\theta}(\mathbf{x}_t^+, \mathbf{c}, t)\|_2^2 - \|\epsilon^+ - \epsilon_{\text{ref}}(\mathbf{x}_t^+, \mathbf{c}, t)\|_2^2$$

$$- (\|\epsilon^- - \epsilon_{\theta}(\mathbf{x}_t^-, \mathbf{c}, t)\|_2^2 - \|\epsilon^- - \epsilon_{\text{ref}}(\mathbf{x}_t^-, \mathbf{c}, t)\|_2^2),$$

where \mathbf{c} is the caption; \mathbf{x}_0^+ and \mathbf{x}_0^- are the preferred and dispreferred images; $\mathbf{x}_t^+ = \alpha_t \mathbf{x}_0^+ + \sigma_t \epsilon^+$ and $\mathbf{x}_t^- = \alpha_t \mathbf{x}_0^- + \sigma_t \epsilon^-$ are noisy images obtained from forward process with Gaussian noise ϵ^+ and ϵ^- at timestep t . This objective computes 4 terms as in Eq. (1): the denoising loss of training and preference distributions ϵ_{θ} and ϵ_{ref} on preferred and dispreferred

noisy images \mathbf{x}_t^+ and \mathbf{x}_t^- respectively. Intuitively, it aims to decrease the loss of preferred images while increasing that of dispreferred ones.

4. Method

Despite impressive image quality and alignment control achieved in prior works, a key limitation among works that use DPO for diffusion models is that it assumes a uniform preference among all users for generated images. In practice, this assumption often fails, as individuals may have diverse preferences related to style, color, lighting, and other visual attributes. While such preferences can be easily expressed by human through pairwise comparisons, it is challenging to summarize them accurately through text or single-image prompts, due to the inaccuracy of natural language and the limitations inherent in single-prompt formats.

Motivated by this problem, we propose PPD, a method that aligns diffusion models with individual-level human preferences. We first formulate the problem of personalization and define a new DPO variant for this task (Sec. 4.1), and then show that individual-level preference learning can be achieved using few-shot comparison examples using a VLM (Sec. 4.2), finally we introduce a parameterization of denoiser that can take advantage of the VLM embeddings (Sec. 4.3). The overall architecture is in Fig. 1.

4.1. Problem Setup

We consider a dataset containing $(\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-, \mathbf{u})$ where \mathbf{c} is the text prompt, \mathbf{x}_0^+ , \mathbf{x}_0^- are preferred and dispreferred generated images and \mathbf{u} represents the user labeling the preference. User features \mathbf{u} can be high-dimensional including

information like age or gender, but often \mathbf{u} consists only of scalar user IDs $\mathbf{u} = [u]$. We define $r(\mathbf{c}, \mathbf{x}_0, \mathbf{u})$ as the reward on image \mathbf{x}_0 for user \mathbf{u} given prompt \mathbf{c} . We would like to fine-tune a text-to-image models $p_{\theta_i}(\mathbf{x}_0|\mathbf{c})$ for individual user \mathbf{u}_i such that reward is maximized for this user while keeping close to a reference model $p_{\text{ref}}(\mathbf{x}_0|\mathbf{c})$ in terms of KL-divergence as regularization :

$$\begin{aligned} \max_{p_{\theta_i}} \mathbb{E}_{\mathbf{c}, \mathbf{x}_0} [r(\mathbf{c}, \mathbf{x}_0, \mathbf{u}_i)] \\ - \beta \mathbb{D}_{\text{KL}} [p_{\theta_i}(\mathbf{x}_0|\mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_0|\mathbf{c})], \end{aligned} \quad (4)$$

where β is a parameter controlling how much $p_{\theta_i}(\mathbf{x}_0|\mathbf{c})$ deviates from $p_{\text{ref}}(\mathbf{x}_0|\mathbf{c})$. Following [32], the optimal p_{θ_i} for user u_i takes the form

$$p_{\theta_i}(\mathbf{x}_0|\mathbf{c}) \propto p_{\text{ref}}(\mathbf{x}_0|\mathbf{c}) \exp(r(\mathbf{c}, \mathbf{x}_0, \mathbf{u}_i)/\beta). \quad (5)$$

Suppose model p_{θ} has sufficient capacity, we can represent $p_{\theta_i}(\mathbf{x}_0|\mathbf{c})$ as $p_{\theta}(\mathbf{x}_0|\mathbf{c}, \mathbf{u}_i)$, where \mathbf{u}_i serves as a conditioning. This allows us to jointly optimize a single model $p_{\theta}(\mathbf{x}_0|\mathbf{c}, \mathbf{u})$ for all users. Following [45], we have the following objective to learn a personalized diffusion models:

$$\begin{aligned} L_{\text{PPD}}(\theta) = -\mathbb{E}_{\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-, \mathbf{u}} [\log \sigma(-\beta T \omega(\lambda_t) \Delta)], \text{ where} \\ \Delta = \|\epsilon^+ - \epsilon_{\theta}(\mathbf{x}_t^+, \mathbf{c}, \mathbf{u}, t)\|_2^2 - \|\epsilon^+ - \epsilon_{\text{ref}}(\mathbf{x}_t^+, \mathbf{c}, t)\|_2^2 \\ - (\|\epsilon^- - \epsilon_{\theta}(\mathbf{x}_t^-, \mathbf{c}, \mathbf{u}, t)\|_2^2 - \|\epsilon^- - \epsilon_{\text{ref}}(\mathbf{x}_t^-, \mathbf{c}, t)\|_2^2). \end{aligned} \quad (6)$$

This objective is similar to Eq. (3) but with the addition of \mathbf{u} , representing extra features conditioned on for each user. But what user features should be conditioned on?

4.2. Generating User Features from a VLM with In-Context Examples

Discussion on User Features in Preference Datasets
One fundamental challenge in personalized image generation lies in the effective parameterization of the user representation, denoted as \mathbf{u} , for each individual. This parameterization is critical because it directly influences the model’s capacity to generate outputs tailored to specific user preferences. Existing image preference datasets, such as Pick-a-Pic [16], typically lack comprehensive user information beyond basic identifiers, often focusing on generic preference data rather than capturing the nuanced individual preferences necessary for effective personalization.

Even in preference datasets that include user information (e.g. PRISM [15]), this information is frequently obtained through self-reported surveys, which can lead to incomplete or biased user profiles. Discrepancies often arise between stated preferences and actual behavior; users may not accurately articulate their likes and dislikes, or they may respond in a manner that they perceive as socially desirable rather than truthful. Such inconsistencies present significant challenges when attempting to model and predict user preferences accurately.

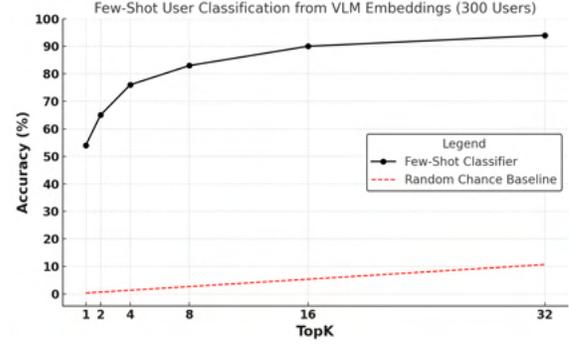


Figure 2. **Top-K accuracy of the User Classification.** We fine-tune a user-classifier from the frozen embeddings from the VLM on few-shot preference examples for 300 users. This classifier significantly outperforms a random chance baseline.

To address these issues, we adopt a simplified approach by assuming access solely to a unique user ID for each individual in the dataset. This user ID serves as a key to differentiate users within the labeled preference dataset, represented as $\mathcal{D}_{\text{pref}} = \{(\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-, \mathbf{s})^{(i)}\}$, where \mathbf{s} denotes a unique user identifier.

Conditioning on a Unique User Identifier The primary challenge, then, is to leverage this user ID i to construct a meaningful user embedding that can effectively guide the model’s predictions. A naïve approach might involve representing the embedding \mathbf{u} using a one-hot encoding of the user IDs, which can be formally defined by a function $f : \mathbf{S} \mapsto \mathbf{U}$, establishing a bijective mapping from the set of user IDs $\mathbf{s} \in \{1, 2, \dots, |D|\}$ to the space of one-hot encoded vectors $\mathbf{u}_i \in \{0, 1\}^{|D|}$, where $|D|$ represents the total number of unique users in the dataset. While this method uniquely encodes each user, it suffers from a critical limitation: the inability to generalize to new, unseen users, as the embedding space is inherently tied to the users present in the training data.

To enhance generalization and adaptability, we reconsider the parameterization of u_i in the context of learning the distribution of reward functions, denoted as $r(\mathbf{c}, \mathbf{x}_0, \mathbf{u}_i)$. Here, a more flexible and robust parameterization leverages few-shot examples that reflect an individual user’s preferences. These examples can be extracted from the offline preference dataset for each user, providing concrete instances of preferred and dispreferred content. Conditioning on these examples allows for a more accurate estimation of the reward function $r(\mathbf{c}, \mathbf{x}_0, \mathbf{u}_i)$ for each user.

In this work, we utilize the features of a pre-trained vision-language Model (VLM), specifically LLaVA-OneVision [21], an open-source multimodal model with multi-image processing capabilities. For each user, a set of $N = 4$ few-shot examples is selected, comprising a caption describing the context, a preferred image, and a

dispreferred image. These examples are processed by the VLM, from which we extract an intermediate hidden state.

To validate that the user embedding is effective, we condition on the frozen embedding to classify which user provided the preference annotations. As seen in Figure 2, the classifier significantly outperforms chance with a top-16 accuracy of 90%, indicating that the embedding can disentangle different users in the preference dataset effectively. By incorporating this user-specific embedding into the conditioning of a text-to-image diffusion model, we enable user-aware reward modeling and generation.

4.3. Personalization as Conditioning

In order to add the user embedding as conditioning to the diffusion model in image generation, we utilize additional cross-attention layers akin to IP-Adapter [51]. Here we learn to condition on the user information using the VLM embeddings and then integrate the image features into pre-trained UNet by adapted modules with decoupled cross-attention. There are many other ways to add conditioning information to diffusion models; we chose this instantiation due to its training and parameter efficiency.

We utilize Stable Cascade [27] as our base model which comprises of three stages: Stages A-C. Stage A and B are image compressors, similar to the VAE in Stable Diffusion [34] but achieve a much higher compression rate; and Stage C is the generation module conditioned on text. In Stage C, the text features \mathbf{c} from pretrained CLIP text encoder are plugged into the model by feeding into the cross-attention layers. Given the query features \mathbf{Z} and text features \mathbf{c}_t , the output is $\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}$, where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \mathbf{c}_t\mathbf{W}_k$, $\mathbf{V} = \mathbf{c}_t\mathbf{W}_v$ are query, key and values, and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are trainable layers.

Following IP-Adapter [51], we utilize a decoupled cross-attention mechanism to condition on the user embedding. First, we add a new cross-attention layer to each cross-attention layer in the Stage C model, using the original query \mathbf{Q} and introducing new keys \mathbf{K}' and values \mathbf{V}' for user features. Then, we simply add the output of user cross-attention to the output of text cross-attention. Hence the final formulation of cross-attention is

$$\mathbf{Z}' = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} + \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^T}{\sqrt{d}})\mathbf{V}', \quad (7)$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \mathbf{c}_t\mathbf{W}_k$, $\mathbf{V} = \mathbf{c}_t\mathbf{W}_v$, $\mathbf{K}' = \mathbf{u}_t\mathbf{W}'_k$, $\mathbf{V}' = \mathbf{u}_t\mathbf{W}'_v$ and $\mathbf{W}'_k, \mathbf{W}'_v$ are new introduced trainable parameters for \mathbf{u} . During training, we optimize only the added cross-attention layers while keeping the parameters of pre-trained diffusion model frozen, using the training objective as in Eq. (6). We also randomly drop user features (zero out the user embedding) in the training stage as regularization.

	CLIP	Aesthetic	HPS
Stable Cascade	31.97	5.33	23.87
Diffusion-DPO	32.48	5.46	25.96
SFT	32.26	5.56	25.78
ours	32.66	5.92	27.51
Diffusion-DPO (CLIP)	32.96	-	-
Diffusion-DPO (Aesthetic)	-	6.42	-
Diffusion-DPO (HPS)	-	-	28.61

Table 1. **Averaged scores on three reward function.** We condition on each reward to generate images both in SFT and our methods. Diffusion-DPO (*) are upper-bounds.

This allows us to also prompt the model unconditionally of the user, by passing a zero embedding for the user.

5. Experiments

Models and Dataset In this section, we demonstrate the performance of PPD across a range of experiments designed to assess its capabilities in personalized preference alignment. We use the objective from Eq. (6) to fine-tune the open-source diffusion model, Stable Cascade [27] – specifically Stage C, which is responsible for text-conditioned generation. Following Wallace et al. [45], we fine-tune our model on the Pick-a-Pic v2 dataset [16], a preference dataset containing 58K text prompts and 0.8M image pairs labeled by 5K users.

Training Details During training of PPD, we only optimize the added cross-attention layers while keeping the pre-trained diffusion model fixed, as described in Sec. 4.3. We use AdamW [24] with an effective batch size of 768 pairs and a learning rate of 1×10^{-5} for all experiments, training for one epoch. We find $\beta \in [0.1, 2]$ for hyperparameter tuning. Details are in Appendix B.

User Information Conditioning We begin by demonstrating the model’s ability to condition on synthetic user information, optimizing three distinct reward functions, we further show that generalize by linearly interpolating the weights of different reward functions (Sec. 5.1). We then show that the model can also condition on real user preferences, generating personalized images based on few-shot examples from new users (Sec. 5.2).

5.1. Synthetic User Alignment

Synthetic User We utilize three reward functions, each representing an aspect of image-to-text generation quality, to simulate three distinct synthetic users: (i) CLIP (OpenCLIP ViT-H/14) [13, 31] for prompt image alignment, (ii) Aesthetic Predictor [37] for image visual appeal, and (iii) HPSv2 (Human Preference Score v2) [47] for human preference estimate. We relabel pairs $(\mathbf{c}, \mathbf{x}_0^+, \mathbf{x}_0^-, \mathbf{u})$ based

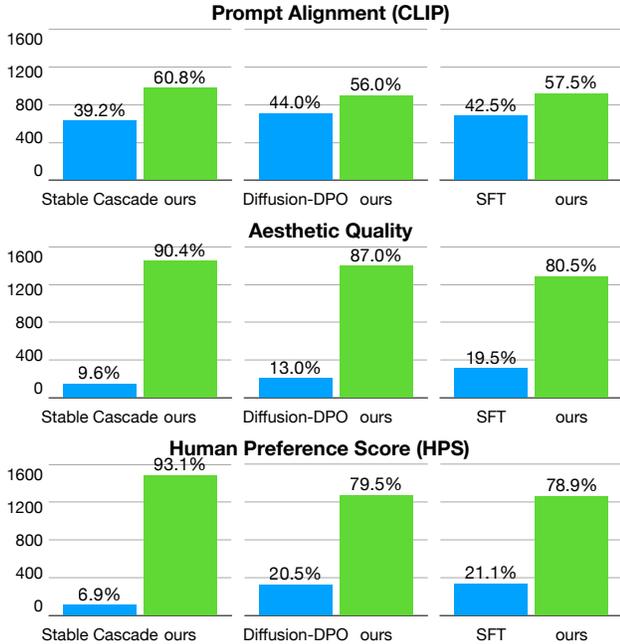


Figure 3. **Automatic win rate evaluation with reward functions.** We compare against Stable Cascade, Diffusion-DPO, and SFT.

on the above reward functions, where \mathbf{u} is a one-hot vector indicating the reward function, taking one of 3 values: $\mathbf{u} \in \{[1, 0, 0], [0, 1, 0], [0, 0, 1]\}$.

Baseline We compare our method with: (1) Stable Cascade [27] pretrained model, which is also the reference model during training; (2) Diffusion-DPO, fine-tuned on Pick-a-Pic label without user conditioning [45]; (3) SFT, supervised fine-tuning on relabeled data with user conditioning. We also compare with (4) Diffusion-DPO (CLIP/Aesthetic/HPS), each annotation indicating a separately fine-tuned model with corresponding reward, this serves as the upper-bound of our method. For a fair comparison, all baselines are fine-tuned using Stable Cascade with the same architecture as our model, optimizing only the added cross-attention layers.

Text-to-Image Generation For each method, we generate images with 30 inference steps using captions from PartiPrompts [53], a benchmark containing 1.6K examples across various categories and challenge aspects. For SFT and our method, we generate three images per prompt, each conditioned on one reward function. For Diffusion-DPO (CLIP/Aesthetic/HPS), images for each annotation are generated from its corresponding model. We report averaged reward scores in Tab. 1 and win rate comparisons in Fig. 3, we can see that our method significantly outperforms Stable Cascade in different reward functions, achieving win rates of 60%, 90%, and 93% in terms of CLIP, Aesthetic, and HPS respectively. Our method effectively optimizes each

reward by conditioning on individual preference, outperforming supervised fine-tuning by a large margin. Unlike Diffusion-DPO, our method can jointly optimize multiple user rewards, achieving win rates comparable to Diffusion-DPO models fine-tuned separately on each user’s data.

Generalization via Interpolation During training we fine-tune the model using one-hot encoding of each objective. Specifically, the user embedding vector \mathbf{u} being $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$ corresponding to optimization for CLIP, Aesthetic, and HPS. The user embedding can also be seen as weighting mechanism for each objective. However, during inference, we are also interested in whether we can trade-off among different objectives by interpolating their respective weights. In Fig. 4, we generate images conditioned on various objective weightings, namely interpolating the given user embedding, we can see that as the weight of an objective increases, the corresponding score also rises, and vice versa. Therefore, our method can generalize to different reward combinations, even only trained on single objective data. This can not be achieved by optimizing separate models and each with an objective.

This shows that by linearly interpolating the user embedding, we can adjust the rewards of the generated images accordingly. This finding aligns with the real user experiments presented next, as real user embeddings exhibit a linear correlation in embedding space.

5.2. Personalized Real User Alignment

We now consider a realistic setting with preferences from the Pick-a-Pic dataset for each user. In this setting, we utilize the user features which are generated from a VLM on in-context preference examples as outlined in Section 4.2. We scale from the 3 synthetic “users” in Section 5.1 to a filtered subset of the 4800 unique users in Pick-a-Pic.

Experiment setting We repartition the Pick-a-Pic dataset into a set of 4 splits: (1) Train, (2) Validation, (3) Test with Seen Users + Held-out Captions and (4) Test with Unseen Users + Held-out Captions. This decomposition allows us to study independently the generalization of our methods to new prompts and unseen users. Additionally, for each user, we take sets of $N = 4$ few-shot preference examples for each user and use these groupings for the conditioning of the VLM on few-shot preference pairs. We thus have to filter out 824 users in Pick-a-Pic which have less than 4 examples per user, as no user embedding for this user can be formed. We generate the user embedding through the procedure outlined in Sec. 4.2.

Generation Results We perform qualitative analysis on the set of images generated for each user. We perform this evaluation by generating a user profile per user. We use a set of few-shot preference pairs per user which we condition a

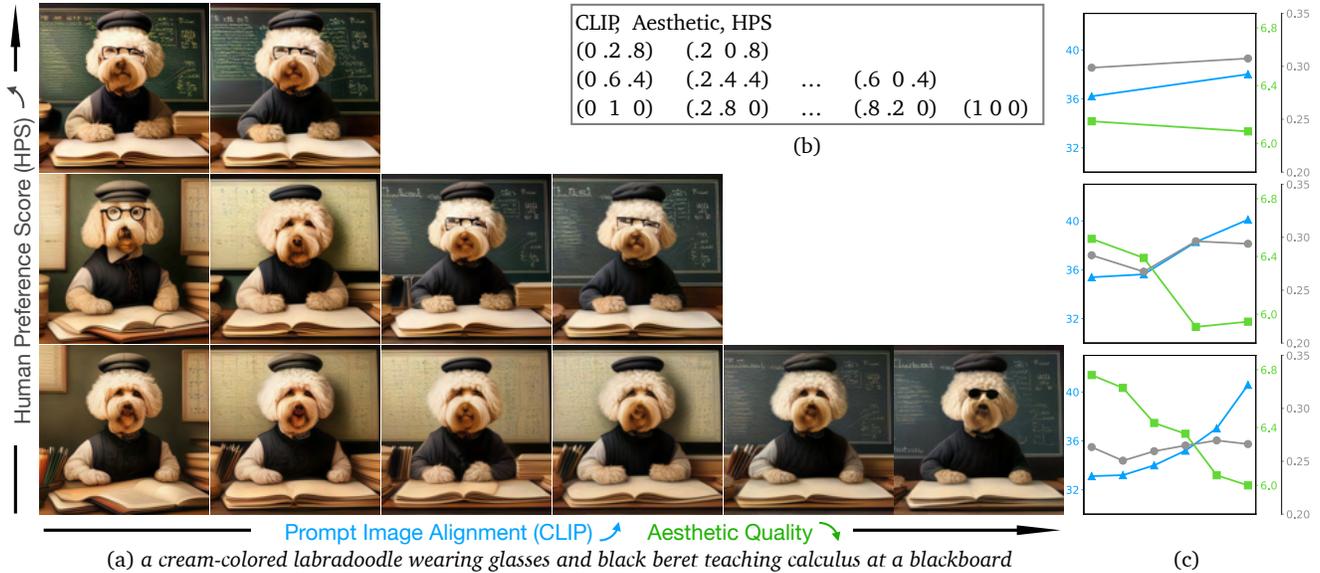


Figure 4. **PPD is able to interpolate among three distinct objectives during inference.** (a) generated images conditioned on (b) various weights, with three axes representing CLIP, Aesthetic, and HPS; (c) reward scores for each image. The score for each objective increases as its respective weight increases, and decreasing otherwise. For each row from left to right, the **CLIP** score increases alongside its weight, leading to a decrease in the **Aesthetic** score as its weight decreases. From bottom to top, the **HPS** score increases with its weight.

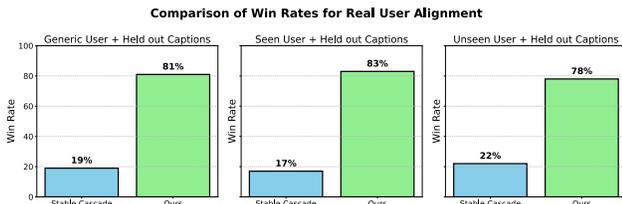


Figure 5. **Automatic Win Rate Evaluation with GPT-4o** We evaluate on seen users, held out users, and aggregate to see win-rates for generic users, comparing Stable Cascade and PPD.

VLM on along with Chain-of-Thought prompting (COT) to construct a user-profile consistent with the preference pairs. The COT prompt asks the model to summarize each image in the preference pair and describe the differences between them, which is then aggregated to form a user profile. Sample user profiles and the full Chain-of-Thought prompt is found in Appendix B. As seen in Fig. 6, our method has closer alignment to the user profile with respect to Stable Cascade and Diffusion-DPO, being able to pick up details about the user such as the highlighted part. While Stable Cascade and Diffusion-DPO with augmented captions can capture certain aspects of the user description, they often overfit to the user profile, resulting in generated images that overlook the intended image caption.

Personalized Preference Alignment Evaluation We evaluate whether the generated model aligns with specific user preference using a VLM as a judge, leveraging strong

VLMs such as GPT4o-vision [26], similar to recipes in prior work [4, 19]. In particular, we condition the VLM on few-shot examples from the user as well as a pair of images for a given test time caption and ask the model to judge which of the two images would be preferred by the user. To ensure consistency of the response the model selects, we judge each pair of images twice using the VLM, flipping the ordering of the images, filtering out pairs where the VLM is inconsistent. With this paradigm and 4 shots per user, we find that the VLM is able to match the ground truth personalized human preference with 83% accuracy.

We decompose our evaluation into two settings: (1) evaluating users that were seen during training but on captions that were held out and (2) evaluating held-out users and captions. We aggregate over these settings to also compute metrics for the general user in the dataset. For the 4341 users in Pick-a-Pic we select 192 held out and 277 in-distribution users. For each of these users, we choose 4 + 1 examples, where 4 examples are used for the in-context conditioning for the user and 1 example is used as the caption for evaluation. We compute an automatic win-rate using pairs of models, generating one image for each caption and evaluating with the VLM-as-Judge pipeline outlined above.

We report win rate comparisons in Fig. 5. Here, PPD significantly outperforms Stable Cascade in all three evaluation settings, achieving a win rate of 83% on seen users, and 78% on unseen users, aggregated to 81% on all users.

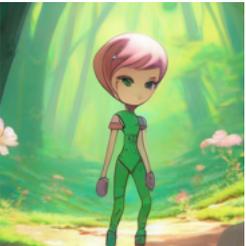
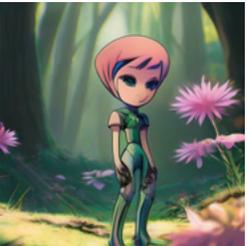
User Profile	Stable Cascade	Diffusion-DPO	Diffusion-DPO	Ours
<p>The preferred images share common themes of wood grain, classic or futuristic design elements, and a focus on aesthetics. They often feature unique front grille patterns, large windows, and a mix of traditional and modern design features. ...</p>	<p>Append user profile to caption as augmentation</p> 			
<i>A needle-felted palm tree</i>				
<p>This user prefers images that have a personal and intimate feel with a focus on the subject's appearance and expression. They also appreciate images with soft focus backgrounds and warm lighting that highlight texture and detail. ...</p>				
<i>An anime girl, masterpiece, good line art, trending in pixiv</i>				
<p>The preferred images ... such as a softer color palette, more dynamic lighting, and a balanced composition that enhances the visual appeal of the scene. The characters are often positioned off-center, which adds to the dynamism of the composition. The backgrounds are rich in detail. ...</p>				
<i>Aelita from Code Lyoko in the forest, gathering flowers</i>				
<p>The user is drawn to images that blend elements of reality with fantastical elements, creating a sense of wonder and curiosity. ... The user may be interested in fantasy art, animation, and storytelling through visual media.</p>				
<i>A cat on a propaganda poster</i>				

Figure 6. **Qualitative Analysis of Images Generated by PPD and Baselines.** Compared to Diffusion-DPO, PPD achieves closer alignment with the generated user profile, highlighted in green. The caption-augmented method captures user profile details; however, it often leads to unintended image alterations, causing the image to disregard the caption itself, as indicated in red.

6. Discussion, Conclusion, and Limitations

We propose a framework for personalizing text-to-image diffusion models using user feature embeddings from a VLM. By conditioning on these embeddings via decoupled cross-attention and fine-tuning with a DPO objective, our method enables accurate reward modeling and high-quality personalized image generation. It generalizes to unseen users with only a few preference examples at test time.

Several open questions and limitations remain. Our current approach that conditioning on few-shot examples is caption-agnostic, selecting random user preferences; future work could explore active preference selection to improve personalization. Extending this method to video generation and studying the trade-off between implicit vs. stated preferences are also promising directions.

Acknowledgment

This research is done during MD’s internship at Luma AI and supported in part by ARO (W911NF-21-1-0125), ONR (N00014-23-1-2159), and the CZ Biohub. AS gratefully acknowledges the support of the NSF Graduate Research Fellowship Program.

References

- [1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. [1](#)
- [2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. [1](#), [2](#)
- [3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. [2](#)
- [4] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? *arXiv e-prints*, art. arXiv:2407.04842, 2024. [7](#)
- [5] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. [2](#)
- [6] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiao-fang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. [2](#)
- [7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. [1](#)
- [8] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. [2](#)
- [9] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2024. [1](#), [2](#)
- [10] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. *arXiv e-prints*, art. arXiv:2406.15951, 2024. [1](#), [2](#)
- [11] Gabriel Goh, James Betker, Li Jing, Aditya Ramesh, Tim Brooks, Jianfeng Wang, Lindsey Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Prafulla Dhariwal, Casey Chu, Joy Jiao, Jong Wook Kim, Alex Nichol, Yang Song, Lijuan Wang, and Tao Xu. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. [2](#)
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, pages 6840–6851, 2020. [1](#), [2](#)
- [13] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. [2](#), [5](#)
- [14] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems 34*, 2021. [2](#)
- [15] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv e-prints*, art. arXiv:2404.16019, 2024. [4](#)
- [16] 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. *arXiv preprint arXiv:2305.01569*, 2023. [1](#), [2](#), [4](#), [5](#), [3](#)
- [17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. [3](#)
- [18] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. [2](#)
- [19] Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. *arXiv e-prints*, art. arXiv:2401.06591, 2024. [7](#)
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. [1](#), [3](#)
- [21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv e-prints*, art. arXiv:2408.03326, 2024. [4](#)

- [22] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024. 2
- [23] Xinyu Li, Zachary C Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024. 2
- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [25] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2
- [26] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2, 7
- [27] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 5, 6
- [28] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024. 2
- [29] 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 *The Twelfth International Conference on Learning Representations*, 2024. 1
- [30] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 2
- [31] 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*, pages 8748–8763. PMLR, 2021. 2, 5
- [32] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2024. 1, 2, 4
- [33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 5
- [35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494, 2022. 1
- [36] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? *arXiv e-prints*, art. arXiv:2303.17548, 2023. 1, 2
- [37] Christoph Schuhmann. CLIP MLP aesthetic score predictor. <https://github.com/christophschuhmann/improved-aesthetic-predictor>, 2022. Accessed: 2024-11-06. 2, 5
- [38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294, 2022. 1
- [39] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023. 2
- [40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015. 1, 2
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2
- [42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1, 2
- [43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2
- [44] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A Roadmap to Pluralistic Alignment. *arXiv e-prints*, art. arXiv:2402.05070, 2024. 1, 2
- [45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 1, 2, 3, 4, 5, 6
- [46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 3
- [47] 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. 2, 5

- [48] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. [2](#)
- [49] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2024. [2](#)
- [50] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiabin Chen, Weihao 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*, pages 8941–8951, 2024. [2](#)
- [51] Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. [1](#), [2](#), [5](#)
- [52] Chao Yu, Hong Lu, Jiayuan Gao, Qixin Tan, Xinting Yang, Yu Wang, Yi Wu, and Eugene Vinitzky. Few-shot in-context preference learning using large language models. *arXiv preprint arXiv:2410.17233*, 2024. [2](#)
- [53] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gungjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification. [6](#)
- [54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. [1](#), [2](#)
- [55] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 4, 2024. [2](#)
- [56] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Secrets of rlhf in large language models part i: Ppo, 2023. [2](#)