electronics

MDPI

Article

Diffusion Preference Alignment via Attenuated Kullback–Leibler Regularization

Xinjian Zhang [1] **and Wei Xiang** [2,3,*]

[1] School of Automation, Beijing Information Science and Technology University, Beijing 100192, China; 2022020459@bistu.edu.cn
[2] School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC 3086, Australia
[3] College of Science and Engineering, James Cook University, Cairns, QLD 4878, Australia
* Correspondence: w.xiang@latrobe.edu.au

Abstract

Direct preference optimization (DPO) has been successfully applied to align large language models (LLMs) with human preferences. In recent years, DPO has also been used to improve the generation quality of text-to-image diffusion models. However, existing techniques often rely on a single type of reward model. They are also prone to overfitting to inaccurate reward signals. As a result, model quality cannot be continuously improved. To address these limitations, we propose xDPO. This method introduces a novel regularization approach that implicitly defines reward functions for both preferred and non-preferred samples. This design greatly enhances the flexibility of reward modeling. The experimental results show that, after fine-tuning Stable Diffusion v1.5, xDPO achieves significant improvements in human preference evaluations compared to previous DPO methods. It also improves training efficiency by approximately 1.5 times. Meanwhile, xDPO maintains image–text alignment performance that is comparable to the original model.

Keywords: machine learning; text-to-image diffusion model; preference alignment; direct preference optimization (DPO)

Academic Editors: Junchao Zhang and Chuanli Wang

Received: 24 June 2025
Revised: 17 July 2025
Accepted: 21 July 2025
Published: 23 July 2025

Citation: Zhang, X.; Xiang, W. Diffusion Preference Alignment via Attenuated Kullback–Leibler Regularization. *Electronics* **2025**, *14*, 2939. https://doi.org/10.3390/electronics14152939

1. Introduction

In recent years, diffusion models have emerged as a powerful generative framework capable of efficiently sampling from and learning complex data distributions [1–3]. However, effectively aligning the outputs of such models with human aesthetics and preferences—particularly in the text-to-image (T2I) domain—remains a significant challenge. The prevailing approach is to first pretrain on large-scale unlabeled datasets to acquire general generative capabilities, followed by fine-tuning on specific downstream tasks or directly leveraging human feedback. This fine-tuning aims to make the model outputs better conform to human preferences while preserving the performance of the pretrained model [4]. Fine-tuning is typically formulated as a distribution optimization problem, where a regularization term is added to encourage the optimized distribution to remain close to the pretrained distribution. Representative methods include reinforcement learning from human feedback (RLHF) [5] and direct preference optimization (DPO) [6,7].

However, empirical studies have shown that DPO suffers from over-optimization issues [8], leading to a degraded quality of generated samples. To address this problem, we propose a novel joint regularization technique that transforms the implicit reward

function from a fixed form into a more flexible and designable one, and apply this method to diffusion models.

Our main contributions include:

- We introduce a higher-order divergence regularization method, incorporating a tunable coefficient η and higher-order divergence terms into the conventional KL regularization. This enables a more refined implicit reward function design, surpassing previous approaches that rely solely on simple KL penalties.
- For diffusion models, we redefine the data likelihood and construct the xDPO framework, from which we derive a concise and efficient loss function to achieve direct preference optimization.
- Through comprehensive experiments, xDPO demonstrates approximately 1.6 times the training efficiency of Diffusion-DPO significantly improves the quality of generated images (see Figure 1), and receives consistent approval from both human evaluators and preference evaluation models .

Figure 1. xDPO is a novel framework that implicitly constructs a reward function to effectively distinguish between preferred and non-preferred samples, thereby aligning text-to-image diffusion models with human preferences. As shown in the figure, the model fine-tuned with xDPO generates images that better reflect human preferences in terms of both visual quality and text alignment. The presented results are based on fine-tuning Stable Diffusion v1-5, with image generation conditioned on prompts sampled from the Pick-a-Pic [9] and PartiPrompts [10] datasets. The corresponding prompts are provided in Appendix B , and all image samples are randomly selected.

2. Related Work

2.1. Text-to-Image Generative Models

Diffusion models convert Gaussian noise into images via an iterative denoising process [2,11,12]. Text-to-image (T2I) generative models are designed to produce high-quality, high-fidelity images conditioned on input text prompts. Moreover, these models have driven advances in related areas such as image editing [13], video generation [14], and 3D modeling [15]. However, state-of-the-art diffusion models are typically trained on large-scale, noisy datasets and often exhibit systematic biases relative to human preferences, indicating that there remains substantial room for improvement.

2.2. Diffusion Models Alignment

The alignment of text-to-image (T2I) diffusion models with human preferences has garnered widespread attention. In this domain, supervised fine-tuning (SFT) remains the most commonly used approach for preference alignment. Inspired by the success of Reinforcement Learning from Human Feedback (RLHF) in large language models [4,16,17], researchers have begun developing image-preference reward models. Representative examples include an aesthetic predictor leveraging learned visual embeddings [18], the ImageReward framework [19], the PickScore metric for aesthetic comparison [20], and the HPSv2 human preference scorer [9].

Building upon these human-preference reward models, [21] proposed Diffusion Policy Optimization for Knowledge (DPOK), while [22] introduced Denoising Diffusion Policy Optimization (DDPO). Both methods formalize the denoising (or sampling) process of diffusion models as a Markov decision process and employ policy gradient techniques for fine-tuning. Some approaches improve preference alignment by adjusting the training data distribution to favor samples with strong visual appeal and highly consistent textual descriptions [23–25]. Additionally, regenerating captions for pre-collected web images has been used to enhance the accuracy and richness of textual annotations [26,27], further advancing preference alignment in text-to-image diffusion models.

Additionally, Diffusion-DPO [7] and D3PO [28] leverage the duality between reward learning and policy optimization inherent in direct preference optimization (DPO), fine-tuning solely on offline human preference pairs at each denoising step; DenseReward [29] further assigns higher weights to early denoising steps within this framework. Diffusion-KTO [30] replaces DPO with Kahneman–Tversky Optimization (KTO) [31], relying exclusively on binary feedback per image for fine-tuning. However, DPO depends on KL regularization, resulting in overly simplistic implicit reward functions. DPO techniques often suffer from the learned policy p_θ gradually diverging from the reference policy p_{ref} during fine-tuning, resulting in degraded model performance. To mitigate this over-optimization issue, the χPO framework [32] introduces a χ^2-divergence term alongside the KL divergence to constrain the distance between the reference distribution p_{ref} and the target distribution p_θ. Inspired by this, we incorporate a tunable coefficient η into the original KL regularization and integrate a higher-order divergence regularizer to more precisely control the discrepancy between p_θ and p_{ref}. Experimental results demonstrate that our method achieves significant performance improvements compared to Diffusion-DPO.

3. Preliminaries

3.1. Diffusion Models

This process comprises a forward (diffusion) phase and a reverse (denoising) phase [1–3,33]. In Denoising Diffusion Probabilistic Models (DDPMs) [2], image generation is cast as a Markov chain. In the forward pass, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is gradually added to a clean sample $x_0 \sim p_{\text{data}}(x_0)$ over T timesteps under a noise schedule governed

by α_t and σ_t. Concretely, for each $t \in [0, T]$, one has $x_t = \alpha_t x_0 + \sigma_t \epsilon$. In the reverse (denoising) phase, generation proceeds by starting from pure Gaussian noise and iteratively removing noise via a learned network ϵ_θ, ultimately recovering a high-quality image. During training, the model parameters θ are optimized by minimizing the evidence lower bound (ELBO). A common reparameterization of this objective is

$$L_{\text{DDPM}} = \mathbb{E}_{x_0, t, \epsilon}\Big[\lambda(t)\,\big\|\epsilon - \epsilon_\theta(x_t, c, t)\big\|^2\Big], \tag{1}$$

where $t \sim \mathcal{U}(0, T)$,$x_t \sim q(x_t \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0, (1 - \bar{\alpha}_t)\, I\big)$, $\lambda(t)$ is a weighting function dependent on the timestep, and c denotes any conditioning information.

3.2. Human Preference Optimization

3.2.1. RLHF

RLHF employs maximum likelihood estimation to fit the Bradley–Terry (BT) model on the dataset $\mathcal{D}$, thereby obtaining an explicit parameterized reward model r. The BT loss is defined as

$$L_{BT}(\phi) := -\,\mathbb{E}_{(x_0^w, x_0^l, c)\sim\mathcal{D}}\Big[\log\sigma\big(r_\phi(x_0^w, c) - r_\phi(x_0^l, c)\big)\Big]. \tag{2}$$

Reinforcement Learning from Human Feedback (RLHF) proceeds by first learning a reward model $r(x, c)$, then optimizing the conditional generation distribution $p_\theta(x_{0:T} \mid c)$ to maximize expected reward while regularizing its divergence from a pretrained reference distribution $p_{\text{ref}}(x_0 \mid c)$. Concretely, the training objective can be written as

$$L_{\text{RLHF}} = \mathbb{E}_{c\sim D}\,\mathbb{E}_{x_0\sim p_\theta(x_0|c)}\big[r(x_0, c)\big] - \beta\, D_{\text{KL}}\big[p_\theta(x_0 \mid c)\,\big\|\,p_{\text{ref}}(x_0 \mid c)\big]. \tag{3}$$

where $D_{\text{KL}}[p_\theta \| p_{\text{ref}}]$ measures the KL divergence between the learned policy $p_\theta(y \mid x)$ and the reference policy $p_{\text{ref}}(y \mid x)$, preventing p_θ from collapsing and helping to maintain output diversity. The hyperparameter $\beta > 0$ balances the expected reward against the KL penalty.

3.2.2. Direct Preference Optimization

Direct preference optimization (DPO) is a fine-tuning paradigm that leverages the correspondence between the optimal policy $p^*(y \mid x)$ and the reward model $r(x, y)$ via $r(x, c) = \beta \log \frac{p_\theta(x|c)}{p_{\text{ref}}(x|c)} + \beta \log Z(c)$, and then optimizes the model parameters using human-preference triplets (c, x_0^+, x_0^-). The DPO training objective can be written as

$$L_{\text{DPO}}(\theta) = -\,\mathbb{E}_{c, x_0^+, x_0^-}\Big[\log\sigma\big(\beta \log \tfrac{p_\theta(x_0^+|c)}{p_{\text{ref}}(x_0^+|c)} - \beta \log \tfrac{p_\theta(x_0^-|c)}{p_{\text{ref}}(x_0^-|c)}\big)\Big] \tag{4}$$

where β controls the extent to which $p_\theta(x \mid c)$ may deviate from the reference distribution $p_{\text{ref}}(x \mid c)$.

Diffusion-DPO adapts the DPO algorithm from Equation (4), which is used to optimize p_θ, for application to diffusion models by casting the generation process as a Markov decision process (MDP). The corresponding objective is defined as follows:

$$L_{\text{Diffusion-DPO}} = -\mathbb{E}\Big[\log\sigma\Big(\beta \log\frac{p_\theta(x_t^w \mid x_{t+1}^w, c)}{p_{\text{ref}}(x_t^w \mid x_{t+1}^w, c)} - \beta \log\frac{p_\theta(x_t^l \mid x_{t+1}^l, c)}{p_{\text{ref}}(x_t^l \mid x_{t+1}^l, c)}\Big)\Big] \tag{5}$$

where x_0^w, x_0^l, $c \sim \mathcal{D}$, $\quad t \sim \mathcal{U}(0, T)$, $\quad x_{t,t+1}^w \sim p(x_{t,t+1}^w \mid x_0^w)$, $\quad x_{t-1,t}^l \sim p(x_{t-1,t}^l \mid x_0^l)$. The DPO-based approach can directly optimize the target distribution p_θ without relying on reinforcement learning algorithms.

4. Method

Consider a dataset annotated with human preferences, denoted as $\mathcal{D}_{\text{pref}} = (c, x_0^w, x_0^l)$, where each entry comprises a text prompt c and two associated images: x_0^w (the "preferred" sample) and x_0^l (the "non-preferred" sample). We wish to train a generation policy π_θ that, when sampling $x_0 \sim p_\theta(\cdot \mid c)$, preferentially outputs higher-quality, human-favored images. To improve the optimization of the generation policy π_θ, we implicitly design the reward function. In particular, xDPO introduces a coefficient η to adjust the weight of the KL regularizer and further incorporates a higher-order divergence term defined by a generator function: $g(z) = \eta z + \frac{1}{n+1}\gamma z^{n+1}$, thereby imposing a stronger penalty on deviations from the reference policy p_{ref}. Accordingly, the resulting reinforcement learning objective becomes

$$\begin{aligned} \max_{p_\theta} \; & \mathbb{E}_{c\sim\mathcal{D}_c}\, \mathbb{E}_{x_0\sim p_\theta(x_0|c)}\big[r(x_0,c)\big] \\ & - \beta\, \mathbb{E}_{c\sim\mathcal{D}_c}\Big[\eta D_{\text{KL}}\big(p_\theta(\cdot \mid c) \,\|\, \pi_{\text{ref}}(\cdot \mid c)\big) + D_g\big(p_\theta(\cdot \mid c) \,\|\, p_{\text{ref}}(\cdot \mid c)\big)\Big] \end{aligned} \tag{6}$$

where the hyperparameter $\beta > 0$ controls the overall strength of the combined regularization. Treating the diffusion process as a multi-step Markov decision process (MDP) and following [7,21,28,30], we optimize a reward model at each time step to obtain the following objective:

$$\begin{aligned} L_{\text{RLHF}} = \; & \mathbb{E}_{c\sim D}\, \mathbb{E}_{x_{0:T}\sim p_\theta(\cdot|c)}\Big[\sum_{t=0}^{T-1} r(x_t, c)\Big] \\ & - \beta\Big(\eta\, D_{\text{KL}}\big[p_\theta(x_{0:T} \mid c) \,\|\, p_{\text{ref}}(x_{0:T} \mid c)\big] + D_g\big[p_\theta(x_{0:T} \mid c) \,\|\, p_{\text{ref}}(x_{0:T} \mid c)\big]\Big), \end{aligned} \tag{7}$$

where $p_{\text{ref}}(x_{0:T} \mid c)$ is the reference distribution given by the pretrained diffusion model, and $\beta > 0$ weights the combined regularization terms.

Here, we define the composite divergence $D_f(\pi \| \pi_{\text{ref}}) = \mathbb{E}_{c\sim D}\Big[\eta D_{\text{kl}}\big[\pi(\cdot \mid c) \| \pi_{\text{ref}}(\cdot \mid c)\big] + D_{\text{g}}\big[\pi(\cdot \mid c) \| \pi_{\text{ref}}(\cdot \mid c)\big]\Big]$ with generator $f(z) := \eta(z \log z - z) + \frac{1}{n+1}\gamma z^{n+1}$, and use β to control the overall strength of this regularization. We present the divergences and corresponding derivatives of xDPO and Diffusion-DPO in Table 1. The derivative of the combined KL + $g(z)$ regularizer is $f'_\chi(z) = \eta \log z + \gamma z^n$, which satisfies $0 \notin \text{dom}\big(f'_\chi(z)\big)$. Consequently, in the text-to-image diffusion setting (see Figure 2), one can reparameterize the objective as in Equation (7) of [34].

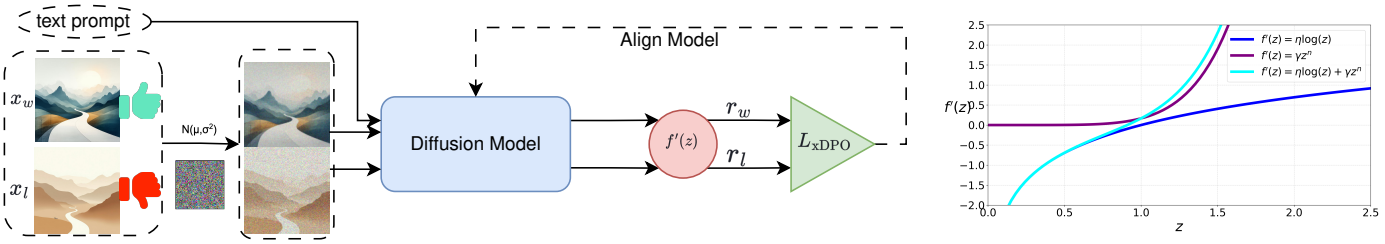

Figure 2. We propose xDPO, a novel method based on the KL regularization framework of direct preference optimization (DPO). Specifically, xDPO introduces a tunable coefficient η into the KL regularizer and incorporates higher-order divergence terms to implicitly construct the reward function. This design enables the reward mechanism to decouple the treatment of positive and negative samples, thereby improving the model's alignment with human preferences. Moreover, we extend the optimization objective to every step of the diffusion process, thus avoiding the high computational cost of back-propagating reward signals through the entire sampling trajectory.We present the xDPO training objective in Appendix C.

Table 1. The divergences and corresponding derivatives of xDPO and Diffusion-DPO.

	$f(z)$	$f'(z)$
Diffusion-DPO (Reverse KL)	$z \log z$	$\log z$
xDPO	$\eta(z \log z - z) + \frac{1}{n+1}\gamma z^{n+1}$	$\eta \log z + \gamma z^n$

$$\sum_{t=0}^{T-1} r(x_t, c) = \beta\Big(\eta \ln \frac{p_\theta(x_{0:T} \mid c)}{p_{\text{ref}}(x_{0:T} \mid c)} + \gamma\Big(\frac{p_\theta(x_{0:T} \mid c)}{p_{\text{ref}}(x_{0:T} \mid c)}\Big)^n\Big). \tag{8}$$

Accordingly, we define the per-step reward as

$$r(x_t, c) = \beta\Big(\eta \ln \frac{p_\theta(x_t \mid x_{t+1}, c)}{p_{\text{ref}}(x_t \mid x_{t+1}, c)} + \gamma\Big(\frac{p_\theta(x_t \mid x_{t+1}, c)}{p_{\text{ref}}(x_t \mid x_{t+1}, c)}\Big)^n\Big). \tag{9}$$

Invoking the reverse process of the T2I diffusion model $p_\theta(x_t \mid x_{t+1}, c)$ (see Equation (1), this reward can be written in closed form:

$$\begin{aligned} r(x_t, c) = -\beta\omega\Big(& \\ & \eta \left\|\epsilon_\theta(x_{t+1}, t+1) + \epsilon_{t+1}\right\|_2^2 + \gamma \exp\big(n \left\|\epsilon_\theta(x_{t+1}, t+1) - \epsilon_{t+1}\right\|_2^2\big) \\ & - \eta \left\|\epsilon_{\text{ref}}(x_{t+1}, t+1) - \epsilon_{t+1}\right\|_2^2 - \gamma \exp\big(n \left\|\epsilon_{\text{ref}}(x_{t+1}, t+1) - \epsilon_{t+1}\right\|_2^2\big)\Big). \end{aligned} \tag{10}$$

where $\omega = \frac{\beta_t \alpha_{t-1}}{2(1-\bar{\alpha}_{t-1})\alpha_t}$ is typically set as a constant in practice [2,35], $\epsilon_{t+t} \sim \mathcal{N}(0, I)$ represents the noises added in the forward (diffusion) process for "preferred" (w) and "less preferred" (l) samples, and ϵ_{ref} denotes the denoising network of the reference model. This loss optimizes the difference in mean-squared error between the learned denoiser ϵ_θ and the reference denoiser ϵ_{ref} on preferred versus non-preferred noisy images.

We generalize the optimization of p_θ in the Bradley–Terry preference model (Equation (2)) to the diffusion setting. By substituting Equation (10) into Equation (2) and rearranging terms, we obtain the following overall training objective:

$$L_{\text{xDPO}} = \mathbb{E}_{\substack{(x_0^w, x_0^l, c)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), \\ x_t^w \sim p(x_t \mid x_0^w, c), x_t^l \sim p(x_t \mid x_0^l, c)}} \Big[\log \sigma\big(r(x_t^w, c) - r(x_t^l, c)\big)\Big], \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function, c the text prompt, and $(x_0^w, x_0^l, c) \sim \mathcal{D}$. Here, $t \sim \mathcal{U}(0, T)$, and x_t^w, x_t^l are the intermediate noisy images at step t for the "winning" and "losing" samples, respectively, drawn via $(x_t^w, x_{t+1}^w) \sim p(x_{t+1} \mid x_t)$and $(x_t^l, x_{t+1}^l) \sim p(x_{t+1} \mid x_t)$.

5. Experiments

5.1. Model and Dataset

In this section, we demonstrate the effectiveness of xDPO for personalized preference alignment through a series of experiments. Following the Diffusion-DPO [7] approach, we fine-tune the open-source Stable Diffusion 1.5 (SD1.5) base model [12] on the human-preference image pairs in the Pick-a-Picv2 dataset [20], using the xDPO objective function Equation (11). The Pick-a-Picv2 dataset comprises 0.8 million preference triplets of the form (preferred image, non-preferred image, input prompt).

5.2. Hyperparameters

For all SD1.5 experiments, we employ the AdamW optimizer [36] . Training is performed on two NVIDIA RTX 4090 GPUs (NVIDIA, Santa Clara, CA, USA), with a per-GPU micro-batch size of one preference pair and gradient accumulation over 128 steps. The

learning rate is set to 1×10^{-8}. Empirically, we observed stable and strong performance for β values in the range [1000, 2000], and therefore report results using the configuration $\beta = 1000$.

5.3. Evaluation

We evaluated the performance of the xDPO model in aligning with human preferences using both automated metrics and a user study. In the automated image quality evaluation, we employed four metrics: the first three were PickScore [20], HPSV2 [9], and ImageReward [19], each trained on human-preference datasets to predict which image a user is more likely to prefer under the same prompt. These three metrics not only assess preference prediction but also incorporate aesthetic judgment to some extent. In addition, we used CLIP [37] to measure prompt responsiveness by computing the image–text similarity score.

We generated images with xDPO and four baseline models—pretrained SD1.5, Diffusion-DPO [7], SePPO [38], and Diffusion-KTO [30]—on two benchmarks: PartiPrompt [10] (1632 prompts) and HPSv2 [9] (3200 prompts). We then applied the aforementioned automated metrics to all outputs and reported both average scores and win-rate statistics. For a fair comparison, Diffusion-DPO [7], Diffusion-KTO [30], and SePPO [38] were evaluated using their officially released checkpoints, and all models (including xDPO) employed the default hyperparameters from [7]: a guidance scale of 7.5 and 50 denoising steps. For the user study, we presented human evaluators with paired images—one from xDPO and one from Diffusion-DPO—generated under the same prompt, and asked two questions: 1. General Preference (Q1): Given the same prompt, which image do you personally prefer? 2. Prompt Alignment (Q2): Which image better matches the textual description? Each comparison was rated by five annotators, and the majority vote ($\geq$3 out of 5) determined the group decision. We aggregated their choices into win-rate percentages over two sets of prompts: 100 randomly sampled from PartiPrompts [10] and 200 randomly sampled from the HPSv2 benchmark [9].

5.4. Quantitative Results

Table 2 reports the absolute reward scores of each reward model across all datasets. Compared to SD1.5 and Diffusion-DPO, xDPO achieves substantial improvements on every reward metric. The bottom of Figure 3 presents a side-by-side comparison of xDPO, Diffusion-DPO, Diffusion-KTO, and the original SD1.5 on the HPSv2 dataset. On HPSv2, xDPO achieves the highest PickScore, ImageReward, and CLIP scores. It is also worth noting that xDPO delivers a significant improvement in training efficiency, training 1.6 times faster than Diffusion-DPO (as shown in Figure 4), while generating higher-quality images. Furthermore, in human evaluations (see the top of Figure 3), annotators demonstrated a clear preference for our xDPO model, which received higher support rates than both baseline models, DPO-SD1.5 and Diffusion-DPO. Specifically, for the "Overall Preference" (Q1) task on the HPSv2 dataset, xDPO outperformed DPO-SD1.5 with a 76.4% win rate and surpassed Diffusion-DPO with a 73.4% win rate. These results clearly demonstrate that our approach provides a more efficient solution for fine-tuning diffusion models based on human preferences.

Table 2. The results of the model-feedback evaluation on the HPSv2 and Parti-Prompt datasets are presented below, with the highest scores highlighted in bold.

Dataset	Method	HPSV2 ↑	PickScore↑	Image Reward ↑	CLIP↑	Aesthetic ↑
HPSV2	SD v1-5 [12]	26.97	20.690	0.125	0.349	5.46
	Diffusion-DPO [7]	27.28	21.124	0.315	0.354	5.56
	Diffusion-KTO [30]	**27.99**	21.332	0.696	0.352	5.69
	SePPO [38]	27.88	21.496	0.616	0.354	**5.76**
	xDPO (ours)	27.98	**21.694**	**0.737**	**0.354**	5.66
PartiPrompts	SD v1-5 [12]	26.956	21.240	0.244	0.336	5.260
	Diffusion-DPO [7]	27.193	21.489	0.396	0.341	5.339
	Diffusion-KTO [30]	**27.740**	21.544	0.633	0.339	5.473
	SePPO [38]	27.611	21.667	0.572	0.338	**5.504**
	xDPO (ours)	27.677	**21.823**	**0.725**	**0.344**	5.439

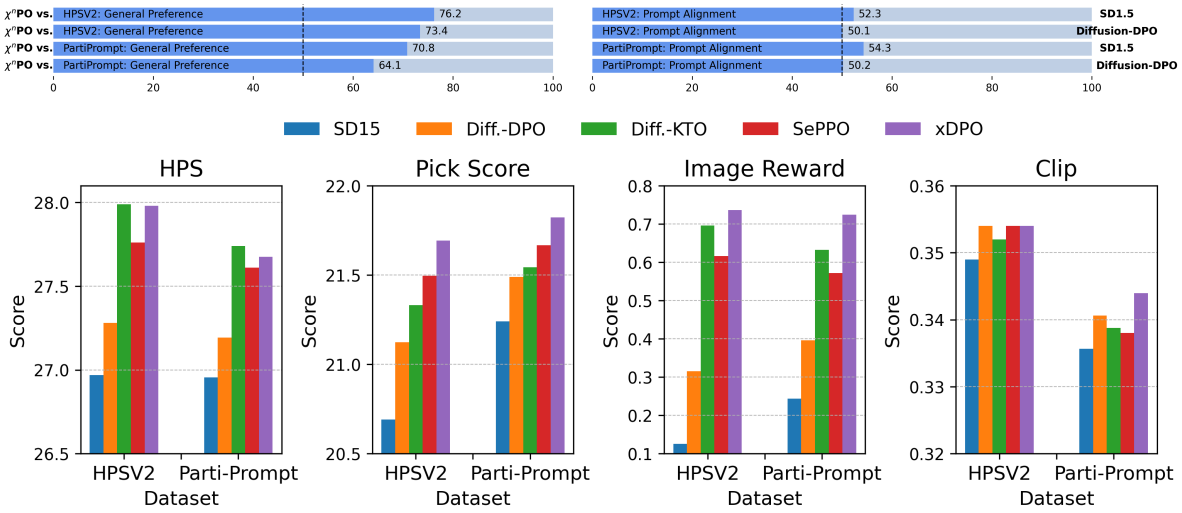

Figure 3. (**Top**) User study win rates (%) comparing xDPO (SD v1-5) with the original SD v1-5 and Diffusion-DPO (SD v1-5) indicate that xDPO significantly enhances the alignment performance of the base SD v1-5 model on both evaluation tasks of the HPSv2 and PartiPrompt benchmarks. (**Bottom**) We compare reward scores across all datasets for each baseline under the different reward models.

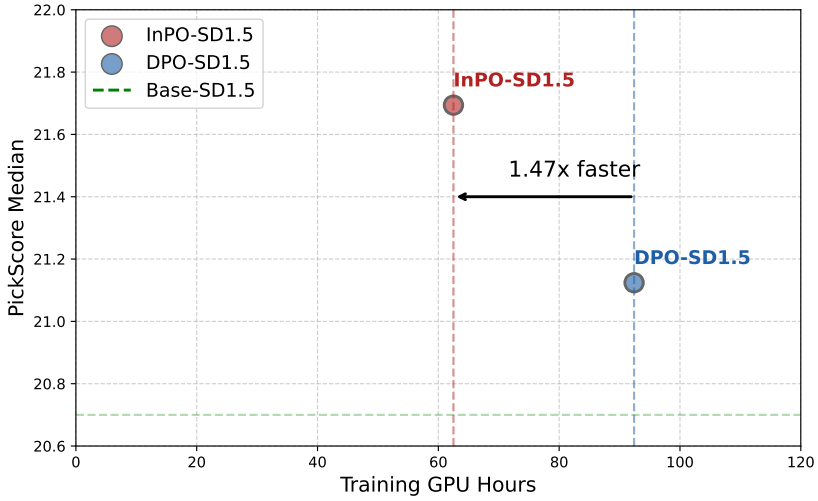

Figure 4. On the HPDv2 test set, we compare the trade-off between image generation quality and training efficiency after human preference optimization of SD1.5. Our method, xDPO, not only achieves higher image quality, but also attains a training speed that is 1.5 times faster than Diffusion-DPO [7], demonstrating superior overall performance.

5.5. Qualitative Results

Figure 1 illustrates images generated by xDPO, which demonstrate outstanding appeal, vivid color palettes, dramatic lighting, strong compositional balance, and lifelike human and animal anatomical structure. Figure 5 presents a direct visual comparison of xDPO with Diffusion-DPO and other baseline methods. In the first row, although all models faithfully reproduce the "pixel art" aesthetic, Diffusion-DPO fails to capture the critical "on the moon" detail; by contrast, xDPO accurately renders this element, achieving superior fine-grained fidelity and overall image quality. In the second row, our xDPO aligned model not only depicts the vibrant hues and mist-enshrouded rock formations with remarkable precision but also outperforms Diffusion-DPO in rendering the glass sphere—faithfully reproducing perspective distortion, specular highlights, reflections, and the subtle stretching and compression at the periphery. In the bottom row, when conditioned on prompts containing detailed descriptions of hand anatomy, xDPO generates anatomically more accurate hand structures than those produced by the Diffusion-DPO method.

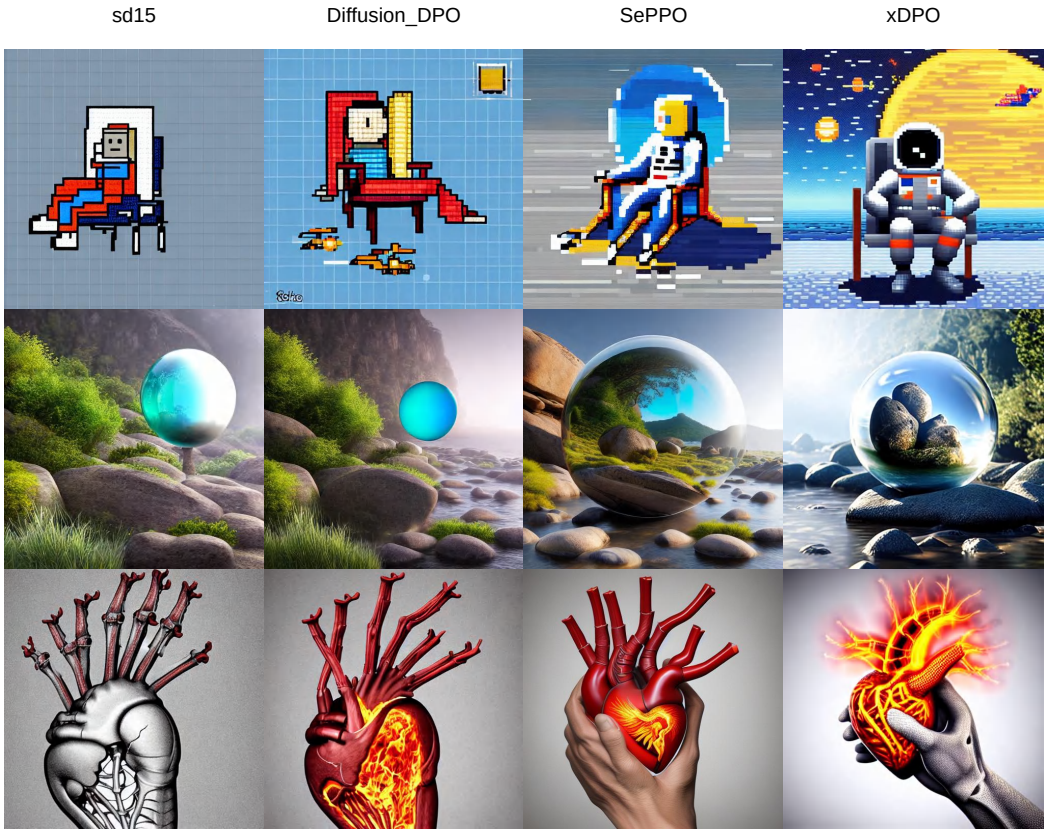

Figure 5. We illustrate images generated by different models from the same textual prompts. From top to bottom, the prompts are as follows: "A space man sat on a beach chair on the moon, pixel art.", "A photorealistic image of a giant floating glass sphere in a rocky landscape surrounded by a gentle mist.", "A 3D-rendered, anatomical flaming heart is held by skeletal metal hands in this expressive image." More images can be found in Appendix A.

5.6. Ablations

As the alignment process progresses, the model increases the conditional probability of the preferred sample, $p_\theta(x_t^w \mid x_{t+1}^w, c)$, while decreasing that of the non-preferred sample, $p_\theta(x_t^l \mid x_{t+1}^l, c)$. Consequently, the ratio for the preferred sample $z_w = \frac{p_\theta(x_t^w|x_{t+1}^w,c)}{p_{\text{ref}}(x_t^w|x_{t+1}^w,c)}$ tends to become greater than 1, whereas the ratio for the non-preferred sample $z_l = \frac{p_\theta(x_t^l|x_{t+1}^l,c)}{p_{\text{ref}}(x_t^l|x_{t+1}^l,c)}$ tends to fall below 1. Based on this observation, we introduce the x^n divergence penalty to further amplify the function values in the region where $z > 1$, thereby strengthening

the gradient signals associated with preferred samples during training. As illustrated in Figure 2 (right), this design enhances the model's responsiveness to human preferences. In our preliminary exploration, we directly introduced the x^n divergence penalty and found that it significantly constrained the model's flexibility in adjustment, slowed the convergence of the training loss, and had a limited effect on improving the reward score. To mitigate this issue, we introduced a coefficient γ and applied $1/n$ weighting scheme to reduce the impact of the x^n term. The experimental results confirm the effectiveness of this strategy. To investigate the impact of different values of n on model performance, we conducted comparative experiments with $n = 1, 3, 6$. Rows 3 to 5 of Table 3 show the performance of xDPO under different n values. The results indicate that the model performs best when $n = 6$. In addition, we explored balancing the optimization strength between positive and negative samples by reducing the value of η. Rows 6 to 8 of Table 3 present the results of experiments where only η was adjusted. The findings show that setting η to 0.05 or 0.1 enables the model to achieve a more balanced generation quality across various reward functions.

Table 3. Ablation and comparison results on SD1.5 using the HPDv2 test set,with the highest scores highlighted in bold.

	Median (HPDv2)			
	HPSV2	**PickScore**	**Image Reward**	**CLIP**
Base-SD1.5	26.97	20.690	0.125	0.349
DPO-SD1.5	27.28	21.124	0.315	0.354
$\eta = 1, n = 1, \gamma = 1/1$	27.83	21.53	0.643	0.357
$\eta = 1, n = 3, \gamma = 1/3$	27.87	21.62	0.644	0.355
$\eta = 1, n = 6, \gamma = 1/6$	27.88	21.57	0.653	0.355
$\eta = 0.5, n = 6, \gamma = 1/6$	27.98	21.70	0.702	0.355
$\eta = 0.1, n = 6, \gamma = 1/6$	27.97	**21.709**	0.731	**0.357**
$\eta = 0.05, n = 6, \gamma = 1/6$	27.98	21.694	**0.737**	0.354

6. Conclusions and Limitations

6.1. Conclusions

This paper proposes xDPO, an efficient diffusion model alignment method based on human preferences. xDPO directly fine-tunes diffusion models by implicitly designing reward functions. Experiments on fine-tuning T2I diffusion models with the Pick-a-Pic v2 dataset demonstrate that xDPO significantly improves training efficiency compared to Diffusion-DPO. Moreover, xDPO achieves robust alignment across various reward models, fully showcasing its potential to enhance the performance of text-to-image diffusion models.

6.2. Limitations

Due to its reliance on offline fine-tuning, the performance of xDPO is highly sensitive to the quality of the training dataset. If the dataset contains harmful, violent, or pornographic content, the model may generate images with inappropriate elements. Therefore, we believe that adopting an online training paradigm holds greater promise and stability for future developments.

Author Contributions: Methodology, X.Z.; Software, X.Z.; Resources, W.X.; Writing—original draft, X.Z.; Writing—review & editing, X.Z.; Supervision, W.X.; Project administration, W.X.; Funding acquisition, W.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All original contributions presented in this paper are contained herein, and the source code is available at https://github.com/StarshipZhang/xDPO (accessed on 15 July 2025). If you have any questions, please contact the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Further Qualitative Results

We sample 12 images from the "paintings" subset of the HPS dataset at intervals of 50, starting from the 0th image. The results are shown in Figures A1 and A2.

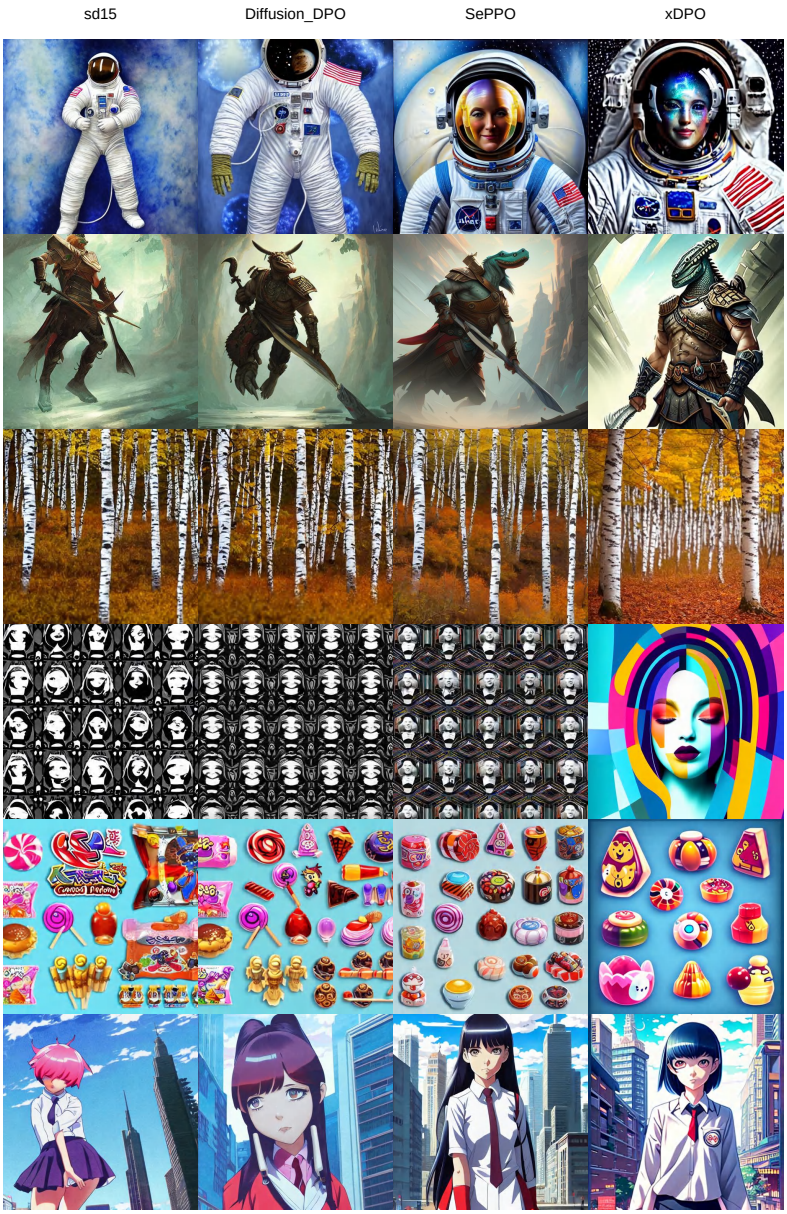

Figure A1. We illustrate images generated by different models from the same textual prompts. From top to bottom, the prompts are as follows: 'Realistic portrait painting of an astronaut suit with a 3D fractal lace design and iridescent bubble texture.', 'A digital painting of a warrior with a crocodile face in a heroic pose, viewed from the side, by Ross Tran.', 'A birch forest in autumn with falling leaves that resemble flying butterflies and dancing elves.', 'An abstract wallpaper with a portrait design.', 'The image showcases a collection of stylized candy designs and RPG assets created by Takeshi Murakami.', 'A gouache of a giantess in school uniform standing in a city, with an anime style and created by various artists including Ilya Kuvshinov and Magali Villeneuve'.

Figure A2. We illustrate images generated by different models from the same textual prompts. From top to bottom, the prompts are as follows: 'Portrait of Archduke Franz Ferdinand by Charlotte Grimm, depicting his detailed face.', 'The image features a surreal fox and skulls in highly detailed, liquid oilpaint style.', 'A portrait painting of Priscilla from Claymore with intricate details and an eerie, realistic style, created by Artgerm, Greg Rutkowski, and Alphonse Mucha.', 'An art print by Barry Moser.', 'A portrait art of a necromancer, referencing DND and War craft.', 'Bust portrait of a gothic goddess crying at dawn by various artists'.

Appendix B. The Appendix is an Optional Section

We summarize the detailed text prompts used in Figure 1 in the following Table A1.

Table A1. Detailed prompts used for generated images in Figure 1.

Dataset	Method
Figure 1, Row 1, Col1	Pippi is tethered to the international space station in her space suit amidst stars and galaxies.
Figure 1, Row 1, Col2	Two girls holding hands while watching the world burn in the style of various artists.
Figure 1, Row 1, Col3	A galaxy-colored DnD dice is shown against a sunset over a sea, in artwork by Greg Rutkowski and Thomas Kinkade that is trending on Artstation.
Figure 1, Row 1, Col4	Image of Earth reflected in a human eye, rendered with Octane, in high resolution.
Figure 1, Row 1, Col5	A full body portrait of a sorceress with a long glowing hooded cloak, by Maciej Kuciara and Jason Chan.
Figure 1, Row 1, Col6	Portrait of a young goth girl in warhammer armor, art by Kuvshinov Ilya, Wayne Barlowe, Gustav Klimt, Artgerm, and Wlop.
Figure 1, Row 2, Col1	A metal bat bird with a red heart head, golden body, joints, and wings as if it is taking off.
Figure 1, Row 2, Col2	A stylized image of fish resembling mythical fantasy creatures in the style of Moebius.
Figure 1, Row 2, Col3	Goro Fujita's illustration depicts a big city on the left and a forest on the right, separated by a highway filled with cars leaving the city.
Figure 1, Row 2, Col4	The image is a concept art character design sheet featuring anime-style women in tek gear, French maid, pinup, cyberpunk, sci-fi, and fantasy styles by various artists.
Figure 1, Row 2, Col5	Lionel Messi portrayed as a sitcom character.
Figure 1, Row 2, Col6	A forest scene depicted in the morning light by Rumiko Takahashi.
Figure 1, Row 3, Col1	A sailboat emoji with a rainbow-colored sail.
Figure 1, Row 3, Col2	A girl in a school uniform playing an electric guitar.
Figure 1, Row 3, Col3	A cute plush griffon with a lion body and seagull head.
Figure 1, Row 3, Col4	The image depicts a person playing Warhammer.
Figure 1, Row 3, Col5	A comical magazine poster of an ancient golden palace.
Figure 1, Row 3, Col6	Anime character holding a axolotl with a black mouth mask.
Figure 1, Row 4, Col1	A cat inside a rocket on a planet with cactuses.
Figure 1, Row 4, Col2	A yellow striped monster panics while using a laptop.
Figure 1, Row 4, Col3	A head-on centered symmetrical portrait of Elisha Cuthbert as a holy paladin, wearing steel armour and with blonde hair, depicted in a highly detailed digital painting with dramatic lighting, in the style of Artgerm and Anna Podedworna.
Figure 1, Row 4, Col4	Dwayne Johnson depicted as a philosopher king in an academic painting by Greg Rutkowski.
Figure 1, Row 4, Col5	Illustration of a cottage designed by Salvador Dali in a blooming forest during spring with a nearby stream, created by Goro Fujita.
Figure 1, Row 4, Col6	An oil painting of Audrey Hepburn portraying Cersei Lannister from Game of Thrones.
Figure 1, Row 5, Col1	A sunset panorama showing a graveyard of souls, with backlight and painted by Frazetta.
Figure 1, Row 5, Col2	The image is of Roman ruins featuring silver and gold artifacts, depicted in hyper-detailed art style by artists Greg Rutkowski and Gustave Dore, and has been shared on various online platforms including Artstation, Worth1000.com, CGSociety, and DeviantArt.
Figure 1, Row 5, Col3	A detailed digital painting of an ancient overgrown statue in a clearing, with vibrant colors and mystical lighting.
Figure 1, Row 5, Col4	A bald general with an angry expression in an intricately detailed and elegant digital painting.
Figure 1, Row 5, Col5	A beautiful Arabian angel wearing a niqab and adorned with jewelry by various artists.
Figure 1, Row 5, Col6	Psytrance artwork by HR Giger.

Appendix C. The xDPO Training Objective

```
import torch
def xDPO_loss(model, ref_model, x_w, x_l, c, beta):
    ''''''
    This is an example psuedo-code snippet for calculating the xDPO loss on
        a single image pair with corresponding caption
    model: Diffusion model that accepts prompt conditioning c
        and time conditioning t
    ref_model: Frozen initialization of model
    x_w: Preferred Image (latents in this work)
```

```
x_l: Non-Preferred Image (latents in this work)
c: Conditioning (text in this work)
beta: Regularization Parameter
returns: xDPO loss value
''''''
timestep = torch.randint(0, 1000)
noise = torch.randn_like(x_w)
noisy_x_w = add_noise(x_w, noise, t)
noisy_x_l = add_noise(x_l, noise, t)
model_w_pred = model(noisy_x_w, c, t)
model_l_pred = model(noisy_x_l, c, t)
ref_w_pred = ref(noisy_x_w, c, t)
ref_l_pred = ref(noisy_x_l, c, t)
model_w_err = (model_w_pred - noise).norm().pow(2)
model_l_err = (model_l_pred - noise).norm().pow(2)
ref_w_err = (ref_w_pred - noise).norm().pow(2)
ref_l_err = (ref_l_pred - noise).norm().pow(2)
exp_w =  torch.exp(6*(model_w_err - ref_w_err)) /6
                        + 0.05 * (model_w_err - ref_w_err)
exp_l = torch.exp(6*(model_l_err - ref_l_err)) /6
                        + 0.05 * (model_l_err - ref_l_err)
inside_term = -0.5 * beta * (exp_w - exp_l)
implicit_acc = (inside_term > 0).sum().float() / inside_term.size(0)
loss = -1 * F.logsigmoid(inside_term)
```

References

1. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2256–2265.
2. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
3. Song, Y.; Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11918–11930
4. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
5. Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-tuning language models from human preferences. *arXiv* **2019**, arXiv:1909.08593.
6. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 53728–53741.
7. Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; Naik, N. Diffusion Model Alignment Using Direct Preference Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 8228–8238.
8. Song, Y.; Swamy, G.; Singh, A.; Bagnell, J.A.; Sun, W. Understanding Preference Fine-Tuning Through the Lens of Coverage. *arXiv* **2024**, arXiv:2406.01462. [CrossRef]
9. Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv* **2023**, arXiv:2306.09341.
10. Yu, J.; Xu, Y.; Koh, J.Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B.K.; et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv* **2022**, arXiv:2206.10789.
11. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.

12. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
13. Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Cao, L.; Chen, S. Diffusion model-based image editing: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4409–4437. [CrossRef] [PubMed]
14. Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv* **2023**, arXiv:2310.19512.
15. Poole, B.; Jain, A.; Barron, J.T.; Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv* **2022**, arXiv:2209.14988.
16. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* **2022**, arXiv:2204.05862. [CrossRef]
17. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
18. Schuhmann, C. LAION-AESTHETICS. 2022. Available online: https://laion.ai/blog/laion-aesthetics/ (accessed on 10 November 2023).
19. Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 15903–15935.
20. Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 36652–36663.
21. Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; Lee, K. Reinforcement learning for fine-tuning text-to-image diffusion models. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS) 2023, New Orleans, LA, USA, 10–16 December 2023.
22. Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; Levine, S. Training diffusion models with reinforcement learning. *arXiv* **2023**, arXiv:2305.13301.
23. Dai, X.; Hou, J.; Ma, C.Y.; Tsai, S.; Wang, J.; Wang, R.; Zhang, P.; Vandenhende, S.; Wang, X.; Dubey, A.; et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv* **2023**, arXiv:2309.15807. [CrossRef]
24. Lee, K.; Liu, H.; Ryu, M.; Watkins, O.; Du, Y.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Gu, S.S. Aligning text-to-image models using human feedback. *arXiv* **2023**, arXiv:2302.12192.
25. Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; Li, H. Better aligning text-to-image models with human preference. *arXiv* **2023**, arXiv:2303.14420.
26. Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. Improving image generation with better captions. *Comput. Sci.* **2023**, *2*, 8.
27. Segalis, E.; Valevski, D.; Lumen, D.; Matias, Y.; Leviathan, Y. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv* **2023**, arXiv:2310.16656. [CrossRef]
28. Yang, K.; Tao, J.; Lyu, J.; Ge, C.; Chen, J.; Shen, W.; Zhu, X.; Li, X. 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, Seattle, WA, USA, 16–22 June 2024; pp. 8941–8951.
29. Yang, S.; Chen, T.; Zhou, M. A dense reward view on aligning text-to-image diffusion with preference. *arXiv* **2024**, arXiv:2402.08265.
30. Li, S.; Kallidromitis, K.; Gokul, A.; Kato, Y.; Kozuka, K. Aligning diffusion models by optimizing human utility. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 24897–24925.
31. Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv* **2024**, arXiv:2402.01306. [CrossRef]
32. Huang, A.; Zhan, W.; Xie, T.; Lee, J.D.; Sun, W.; Krishnamurthy, A.; Foster, D.J. Correcting the mythos of kl-regularization: Direct alignment without overparameterization via chi-squared preference optimization. *arXiv* **2024**, arXiv:2407.13399.
33. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
34. Wang, C.; Jiang, Y.; Yang, C.; Liu, H.; Chen, Y. Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
35. Kingma, D.; Salimans, T.; Poole, B.; Ho, J. Variational diffusion models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21696–21707.
36. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.

37. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual , 18–24 July 2021.
38. Zhang, D.; Lan, G.; Han, D.J.; Yao, W.; Pan, X.; Zhang, H.; Li, M.; Chen, P.; Dong, Y.; Brinton, C.; et al. SePPO: Semi-Policy Preference Optimization for Diffusion Alignment. *arXiv* **2024**, arXiv:2410.05255.