This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Rethinking DPO-style Diffusion Aligning Frameworks

Xun Wu[1] Shaohan Huang[1*] Lingjie Jiang[2] Furu Wei[1]
[1] Microsoft Research [2] Peking University
{xunwu, shaohanh, fuwei}@microsoft.com lingjiejiang@stu.pku.edu.cn

Abstract

Direct preference optimization (DPO) has shown success in aligning diffusion models with human preference. However, We identify two potential risks for existing DPO algorithms: First, current DPO methods for estimating the rewards of step-wise intermediate samples are biased, leading to inaccurate preference ordering for step-wise optimization. Second, existing DPO methods may inadvertently increase the sampling probabilities of dispreferred samples, potentially introducing application risks. To address these issues, we propose Revised Direct Preference Optimization (RDPO), a simple but effective step-wise DPO-based text-to-image diffusion model alignment method. By designing a more theoretically grounded and efficient intermediate-step reward estimation and introducing an additional regularization terms to constrain the sampling probability of dispreferred samples, RDPO can achieve more effective and stable text-to-image alignment performance. Our experiments on two datasets, with base models including Stable Diffusion v1.5 and SDXL, demonstrate that RDPO can effectively learn and construct reward signals for each step of the model, improving alignment performance while ensuring better generalization. Code is available at https://github.com/yushuiwx/RDPO.git

1. Introduction

Text-to-image generative models [28, 33, 35, 38] have seen significant advancements in recent years. Notably, large-scale text-to-image diffusion models such as Imagen [34] and DALL·E 2 [28] have demonstrated remarkable capabilities in generating high-quality and creative images based on textual prompts. However, despite these advancements, current generative models still suffer from misalignment with human preferences, such as discrepancies with the provided text prompts or the generation of incorrect content [25].

Direct preference optimization (DPO), which fine-tunes the model on paired data to align the model generations with human preferences, has demonstrated its success in large language models (LLMs) [27]. Recently, researchers generalized this method to diffusion models for text-to-image generation [2, 20, 29, 42, 48]. Given a pair of images generated from the same prompt and a ranking of human preferences for them, DPO aims to increase the likelihood of generating the preferred sample while reducing the likelihood of generating the less preferred sample. This process enables the model to produce more visually appealing and aesthetically aligned images that better reflect human preferences.

However, after carefully revisiting existing DPO methods for text-to-image diffusion alignment, we identify two potential risk: (1) **First**, when estimating the rewards of intermediate step-wise samples, existing DPO algorithm either relies on strong assumptions that are difficult to satisfy [41, 47] or introduces bias [20, 29], leading to an inaccurate determination of the preference order of step-wise samples. (2) **Second**, we find that existing DPO methods can increase the sampling probabilities of dispreferred samples, as long as the relative probability between the preferred and dispreferred classes increases, which may introduce application risks and potential issues with generalization performance. We conduct detailed theoretical analysis of both two issues in § 3.2 and § 3.3, respectively.

Therefore, in this paper, we propose Revised Direct Preference Optimization (RDPO) to address abovementioned two risk. First, our theoretical analysis reveals that the intermediate reward used in existing DPO algorithms is a Q-value, representing the expected final reward of the entire diffusion trajectory up to step t. However, a true step-wise reward at step t should be the difference between the Q-values (i.e., ΔQ) at steps t and $t-1$, capturing the impact of the denoising process at that step. Thus, unlike previous methods that approximate step-wise rewards using the full trajectory's expectation, RDPO leverages this expectation difference ΔQ, enabling more accurate intermediate-step reward estimation. Additionally, to address risk 2—ensuring that DPO follows the principle of increasing the generation probability of preferred samples while decreasing that of dispreferred samples—RDPO introduce an additional penalty term. This term is jointly optimized with the loss function to constrain the sampling probability of dispreferred samples.

*Corresponding author.

In our experiments, we fine-tune Stable Diffusion v1.5 and SDXL on two datasets using RDPO, improving their ability to generate images with higher aesthetic scores and better alignment with human preferences. Compared to strong baselines, including both RLHF and DPO-style methods, RDPO achieves superior preference alignment and better generalization. Our main contributions are:

- We demonstrate that the reward estimation for intermediate-step samples in existing DPO methods is biased and propose a more accurate estimation approach.
- We introduce an additional penalty term to mitigate the risks of DPO methods unintentionally increasing the sampling probability of dispreferred samples.
- Based on the two contributions above, we propose RDPO and demonstrate through extensive experiments that RDPO achieves more effective, stable, and generalized alignment.

2. Related Works

2.1. Diffusion Generative Models

The generative potential of diffusion models [9, 37, 39] has been extensively explored across various domains, enabling the synthesis of high-fidelity data from Gaussian noise. This includes the generation of images [6, 11, 23, 32], audio signals [21], video sequences [10, 36], three-dimensional shapes [7, 24], and robotic motion trajectories [3, 13], all achieved through iterative denoising processes. Furthermore, diffusion models have been shown to outperform traditional generative models, such as GANs [14], in certain applications due to their stable training dynamics and capacity for more diverse outputs. Their flexible architecture allows them to be adapted to various types of generative tasks, including conditional generation, where the model is guided by additional information, such as labels or other inputs, enabling targeted synthesis.

2.2. Reinforcement Learning from Human Feedback

Diffusion models have demonstrated high-quality generation capabilities [28, 33–35, 38] by training on large-scale datasets, but the mixed quality of these datasets often leads to visually unappealing and misaligned outputs. An effective solution is to align generative models with human preference by using Reinforcement Learning with Human Feedback (RLHF) [16, 30, 46], whose effectiveness has been widely validated [1, 5, 15, 25, 40]. Proximal Policy Optimization (PPO) based methods [18, 25, 49, 50] first train a reward model [15, 43, 45] using labeled pair-wise preference data, and subsequently utilize it to provide preference signal for fine-tuning diffusion models. However, this approach requires hosting an additional reward model for training, which incurs high computational costs and can lead to instability during the optimization process.

Inspired by the success of Direct Preference Optimization (DPO) in the NLP domain [17, 26], some works introduce DPO into the alignment process of diffusion models [20, 22, 41, 47]. Given a pair of images or videos generated from the same prompt and a ranking of human preference for them, DPO aims to increase the probability of generating the preferred sample while decreasing the probability of generating another sample, which enables the model to generate more visually appealing and aesthetically pleasing images that better align with human preferences.

3. Background

3.1. Preliminaries

Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) generate high-quality samples by progressively refining random noise into complex data distributions through an iterative denoising process.

In the forward process, given an input x_0 sampled from the real distribution p_{data}, diffusion models gradually add Gaussian noises to x_0 at each step $t \in [1, T]$, as follows:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (1)$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ denotes the Gaussian noise at step t. $\alpha_{1:T}$ denotes the variance schedule and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

In the reverse denoising process, the diffusion model is trained to learn $p(x_{t-1}|x_t)$ at each step t. Specifically, following [39], the denoising step at step t is formulated as

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\hat{x}_0(x_t),\ \text{predicted } x_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t)}_{\text{direction pointing to } x_t} + \underbrace{\sigma_t \epsilon'_t}_{\text{random noise}} \quad (2)$$

where $\epsilon_\theta(\cdot)$ is a noise prediction network with trainable parameters θ, which aims to use $\epsilon_\theta(x_t, t)$ to predict the noise ϵ in Eq. (1) at each step t. $\epsilon'_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is sampled from the standard Gaussian distribution. In fact, x_{t-1} is sampled from the estimated distribution $\mathcal{N}(\mu_\theta(x_t), \sigma_t^2 \boldsymbol{I})$. According to the reverse process, $\hat{x}_0(x_t) = (x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t))/\sqrt{\bar{\alpha}_t}$ represents the predicted x_0 at step x.

Direct Preference Optimization (DPO). The DPO method [41, 47] was originally proposed to fine-tune large language models to align with human preferences based on paired datasets. Given a prompt x, two responses y_0 and y_1 are sampling from the generative model π_θ, *i.e.,* $(y_0,\ y_1) \sim \pi_\theta(y|x)$. Then, y_0 and y_1 are ranked based on human preferences. Let y_w denote the preferred response in $(y_0,\ y_1)$ and y_l denote the dis-preferred response. DPO

optimizes parameters θ in π_θ by minimizing the following loss function.

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \quad (3)$$

where σ is the sigmoid function, and β is a hyper-parameter. π_{ref} represents the reference model, usually set as the pre-trained models before fine-tuning.

DPO for Diffusion Models. We take D3PO [47] as an example for explanation. For a text-to-image diffusion model π_θ parameterized by θ, given a text prompt c, D3PO first samples a pair of generation trajectories $[x_T^0, \ldots, x_0^0]$ and $[x_T^1, \ldots, x_0^1]$. Then, they compare the reward scores $r(c, x_0^0)$ and $r(c, x_0^1)$ of generated images, using the reward model $r(\cdot)$, and rank their preference order. The preferred image is denoted by x_0^w and the dis-preferred image is denoted by x_0^l. D3PO assumed that the preference order of final images (x_0^0, x_0^1) represents the preference order of (x_t^0, x_t^1) at all intermediate steps t. Subsequently, the diffusion model is fine-tuned by minimizing the following DPO-like loss function for $\Phi = (c, x_t^w, x_t^l, x_{t-1}^w, x_{t-1}^l)$ at the step level:

$$\mathcal{L}_{\text{D3PO}}(\theta) = -\mathbb{E}_\Phi\left[\log\sigma\left(\beta\log\frac{\pi_\theta(x_{t-1}^w|x_t^w,c)}{\pi_{\text{ref}}(x_{t-1}^w|x_t^w,c)} - \beta\log\frac{\pi_\theta(x_{t-1}^l|x_t^l,c)}{\pi_{\text{ref}}(x_{t-1}^l|x_t^l,c)}\right)\right]. \quad (4)$$

Without loss of generality, we take D3PO as the baseline in the following sections (§ 3.2 and § 3.3) to analyze the potential risk of existing DPO algorithms applied to diffusion models.

3.2. Potential Risk 1 of DPO in Diffusion

In the context of text to-image diffusion models, the denoising process is typically conceptualized as a multi-step Markov Decision Process (MDP). Therefore, when optimizing with the DPO loss, determining the perferred and disperferred samples at each step is crucial. Since the intermediate states consist of noise and partially generated images, it is challenging for humans to accurately judge which segment is better. Existing works can be categorized based on two different assumptions for assigning rewards to intermediate steps:

Assumption 1 *The preference order between the final generations (x_0^0, x_0^1) can consistently represent the preference order between corresponding noisy samples (x_t^0, x_t^1) at all intermediate steps.*

Both Diffusion-DPO [41] and D3PO [47] introduce this strong assumption to label the preference order of intermediate steps directly based on the preference order of final generations. However, prior work [20, 29] has demonstrated that this assumption does not always hold. To better align

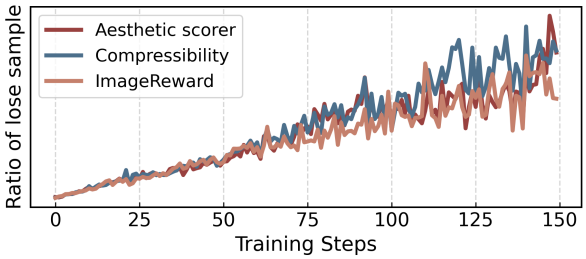


Figure 1. The change curve of the ratio of dispreferred sample $\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$ during the D3PO [47] fine-tuning process with different reward signal (aesthetic scorer, compressibility and ImageReward [45]). We find that the ratio keep increasing while using different reward signals.

the preference labels with the denoising performance at each step, both SPO [20] and TailorPO [29] biuld their methods upon below assumption 2:

Assumption 2 *the step-wise reward of x_t can be estimated by the corresponding reward of the predicted x_0 (shown in Eq.2) at t step.*

Here, we carefully revisit the assumption 2, as the assumption 1 has already been proven to be invalid. Previous studies [4, 8, 29] have proven that $\mathbb{E}[x_0|c, x_t] = \hat{x}_0(x_t)$ and the expectation of image rewards $\mathbb{E}[r(c, x_0)|c, x_t]$ can be approximated by the reward of the expected image $r(c, \mathbb{E}[x_0|c, x_t])$, then they take the reward of $\hat{x}_0(x_t)$ as the reward for t step to determine which sample is preferred.

$$r_t(c, x_t) \triangleq \mathbb{E}[r(c, x_0)|c, x_t] \approx r(c, \hat{x}_0(x_t)) \quad (5)$$

We indicate that $r(c, \hat{x}_0(x_t))$ is actually a Q-value in the MDP process, i.e., $Q(c, x_{T:t}) = r(c, \hat{x}_0(x_t))$, representing the expected reward of the final image generated by the entire diffusion chain from step T (the initial step) to step t. This reflects the cumulative effect of the entire chain. However, the true unbiased step-wise reward at step t should be the difference between the Q-values at steps t and $t-1$., i,e, the expected change in expectation reward:

$$r_t^*(c, x_t) \triangleq \Delta Q \approx \left(r\left(c, \hat{x}_0\left(x_t\right)\right) - r\left(c, \hat{x}_0\left(x_{t-1}\right)\right)\right) \quad (6)$$

This difference bwtween Q-value represents the expected reward change before and after the denoising process at step t, which more accurately reflects the quality of the sampling at step t.

3.3. Potential Risk 2 of DPO in Diffusion

In this section, we take a step back and examine the DPO loss shown in Eq. 4. The gradient of $\mathcal{L}_{\text{D3PO}}$ with respect to parameters θ is given by (proven by Ren et al. [29]):

$$\nabla_\theta\mathcal{L}_{\text{D3PO}}(\theta) \propto -\mathbb{E}\left[\left(f_t/\sigma_t^2\right)\cdot\widehat{A}\right], \quad (7)$$

while

$$f_t \triangleq \beta(1-\sigma(\beta\log\frac{\pi_\theta(x_{t-1}^w|x_t^w,c)}{\pi_{\text{ref}}(x_{t-1}^w|x_t^w,c)}-\beta\log\frac{\pi_\theta(x_{t-1}^l|x_t^l,c)}{\pi_{\text{ref}}(x_{t-1}^l|x_t^l,c)})). \quad (8)$$

Here $\mu_\theta(\cdot)$ and σ_t represents the mean and variance of the conditional distribution. Eq. 8 means that standard DPO loss can lead to a increasing of the model's likelihood of the dispreferred completions, as long as the relative probability between the preferred and dispreferred classes increases. We further validate this phenomenon by visualizing the change of the ratio $\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$ during preference aligning process in Figure 1. We find that the ratio of disperferred samples keep increasing, which further evidence of our findings.

Why is this an issue? First, fine-tuning a diffusion model using DPO increases the likelihood of generating dispreferred samples, which may contain undesirable content such as violence, distortions, or other human dispreferred artifacts. This poses potential risks when deploying the fine-tuned model in real-world applications. Additionally, a large body of research focuses on distilling multi-step diffusion models into few-step or parameter-efficient variants to accelerate the generation process. However, the tendency of DPO to increase negative sample probability introduces potential risks in the distilled models, making them more prone to generating undesired content. Moreover, fine-tuning with DPO on a specific dataset increases the model's probability mass on that dataset, reducing generalization ability and making the training process unstable, often leading to mode collapse.

4. Revised Direct Preference Optimization

To address the aforementioned two potential risk, we propose the Revised Direct Preference Optimization (RDPO) for better aligning diffusion models with human preferences. Specifically, given a text prompt c and the time step t, we have two noisy samples x_{t-1}^+ and x_{t-1}^- both sampled from x_t.

To determine the preference order between x_{t-1}^+ and x_{t-1}^-, we compute the step-wise reward of $r_t^*(c,x_{t-1}^+)$ and $r_t^*(c,x_{t-1}^-)$ by using Eq. 6. The sample with the larger reward is assigned x_{t-1}^w, while the other, the dispreferred sample, is assigned x_{t-1}^l. To address the potential risk 2 described in Sec 3.3, we add a penalty term $\Psi = \min\left(0, \log\frac{\pi_{\text{ref}}(x_{t-1}^l|x_t,c)}{\pi_\theta(x_{t-1}^l|x_t,c)}\right)$ to the DPO loss to incentivise maintaining a low log-likelihood of the dispreferred sample. This penalty term is 0 when $\pi_{\text{ratio}}(x_{t-1}^l|x_t,c)\leq 1$ and decreases as the ratio goes below 1.

We present the full algorithm details of RDPO in Algorithm 1 and the final optimization objective of RDPO is given by:

$$\mathcal{L}_{\text{RDPO}}(\theta) = -\mathbb{E}_{(c,x_t,x_{t-1}^w,x_{t-1}^l)}$$

Algorithm 1 RDPO: Revised Direct Preference Optimization

Input: Diffusion model $\pi_\theta(\cdot)$, reference model $\pi_{\text{ref}}(\cdot)$, reward model $r(\cdot)$, text prompt c

Initialize $x_T\sim\mathcal{N}(0,\boldsymbol{I})$ **for** $t=T,\dots,1$ **do**
  Sample x_{t-1}^+, x_{t-1}^- from $\pi_\theta(\cdot|x_t,c)$ Rank x_{t-1}^+ and x_{t-1}^- based on their step-wise rewards r^* to obtain x_{t-1}^w and x_{t-1}^l
  // Fix the potential risk 1 in Sec. 3.2
  if $r_t^*(c,x_{t-1}^+) > r_t^*(c,x_{t-1}^-)$ **then**
    $x_{t-1}^w \leftarrow x_{t-1}^+$ $x_{t-1}^l \leftarrow x_{t-1}^-$
  end
  else
    $x_{t-1}^l \leftarrow x_{t-1}^+$ $x_{t-1}^w \leftarrow x_{t-1}^-$
  end
  Optimize $\pi_\theta(\cdot)$ using Eq. (9) $x_{t-1}\leftarrow x_{t-1}^w$
  // Fix the potential risk 2 in Sec. 3.3
end

Output: The fine-tuned diffusion model $\pi_\theta(\cdot)$.

$$\left[\log\sigma\left(\delta-\lambda\cdot\underbrace{\min\left(0,\log\frac{\pi_{\text{ref}}(x_{t-1}^l|x_t,c)}{\pi_\theta(x_{t-1}^l|x_t,c)}\right)}_{\text{penalty term }\Psi}\right)\right], \quad (9)$$

while

$$\delta = \beta\left(\log\frac{\pi_\theta(x_{t-1}^w|x_t,c)}{\pi_{\text{ref}}(x_{t-1}^w|x_t,c)}-\log\frac{\pi_\theta(x_{t-1}^l|x_t,c)}{\pi_{\text{ref}}(x_{t-1}^l|x_t,c)}\right). \quad (10)$$

Here $\lambda>0$ is a hyperparameter. By adding the term Ψ, the model can no longer minimise the loss by increasing the log-likelihood of the both preferred and dispreferred examples while keeping the relative probability of picking the preferred completion over the dispreferred. It must also ensure that the log-likelihood of the dispreferred examples remains low relative to the log-likelihood under the reference model.

5. Experiments

5.1. Experimental Setting

Compared Baselines. We compare our RDPO with all strong relevant baselines: Diffusion-DPO [41], D3PO [47], DDPO [2], SPO [20] and TailorPO [29]. We implement our RDPO based on the our implementation of TailorPO-G and set two versions of our RDPO for better validating each component in our methods. Here we use RDPO* denotes using better intermediate step reward mentioned in § 3.2, while RDPO indicates using both two components we proposed for two failure modes.

Table 1. Reward values of images generated by diffusion models fine-tuned using different methods. The prompts are related to common animals [47]. Experiments were conducted for three runs and we report the average results for fair comparison. – denotes the corresponding reward can not be applied to this method.

	Aesthetic scorer ↑	ImageReward ↑	HPSv2 ↑	PickScore ↑	Compressibility ↑	VP-Score ↑
Stable Diffusion v1.5	5.79	0.65	27.51	20.20	-105.51	25.61
DDPO [2]	6.57	0.99	28.00	20.24	-37.37	25.77
D3PO [48]	6.46	0.95	27.80	20.40	-29.31	26.09
SPO [19]	5.89	0.95	27.88	20.38	–	26.15
TailorPO [29]	6.66	1.20	28.37	20.34	-6.71	26.34
TailorPO-G [29]	6.96	1.26	28.03	20.68	–	26.27
RDPO* (Ours)	7.28	1.39	**28.57**	20.83	-4.41	26.59
RDPO (Ours)	**7.80**	**1.48**	28.42	**21.10**	**-2.37**	**26.93**

Table 2. Reward values of images generated by diffusion models fine-tuned using different methods. The prompts are randomly selected from Pick-a-pic [15]. Experiments were conducted for three runs and we report the average results for fair comparison. – denotes the corresponding reward can not be applied to this method.

	Aesthetic scorer ↑	ImageReward ↑	HPSv2 ↑	PickScore ↑	Compressibility ↑	VP-Score ↑
Stable Diffusion v1.5	5.28	0.14	23.36	19.60	-143.82	22.96
DDPO [2]	5.41	0.16	24.07	19.79	-93.44	23.43
D3PO [48]	5.47	0.16	24.11	19.67	-97.57	23.30
SPO [19]	5.30	0.17	24.08	19.91	–	23.70
TailorPO [29]	5.44	0.16	24.93	20.02	-82.81	23.68
TailorPO-G [29]	5.58	0.25	24.67	19.86	–	23.87
RDPO* (Ours)	5.71	0.29	25.11	20.23	-76.60	24.07
RDPO (Ours)	**5.90**	**0.37**	**25.71**	**20.27**	**-69.63**	**24.32**

For SPO, we used the officially released implementation and adopted the same hyperparameters as specified in the original paper. For all other methods, we followed the hyperparameter settings outlined in [48], except for reducing the batch size across all methods. Specifically, within all our frameworks, image generation was performed using $T = 20$, and $T_{\text{fine-tune}} = 5$ time steps were uniformly sampled for fine-tuning. That is, we fine-tuned the model only at specific steps $t = 20, 16, 12, 8, 4$.

Training Details. Following the experimental settings described in D3PO [48], and TailorPO [29], we use the prompts of animals released by Yang et al. [47] and prompts in the Pick-a-pic [15], respectively. For training rewards, i.e., the preference scorers, we follow existing works by selecting commonly used ones, Aesthetic scorer, ImageReward [45], HPSv2 [43], PickScore [15], Compressibility and VP-Score [44].

We utilized the DDIM scheduler [39] with $\eta = 1.0$ and employed $T = 20$ inference steps. The generated images were produced at a resolution of 512×512. To fine-tune the UNet parameters, we leveraged LoRA [12] with a dataset comprising a total of 10,000 samples and a batch size of 2. The baseline model was initialized as the pre-trained Stable Diffusion v1.5[1]. We run all experiments with 8 × H100 GPUs (each with 80GB). We set $\lambda = 1,00$ by default and provide the corresponding ablation study results in § 5.3.

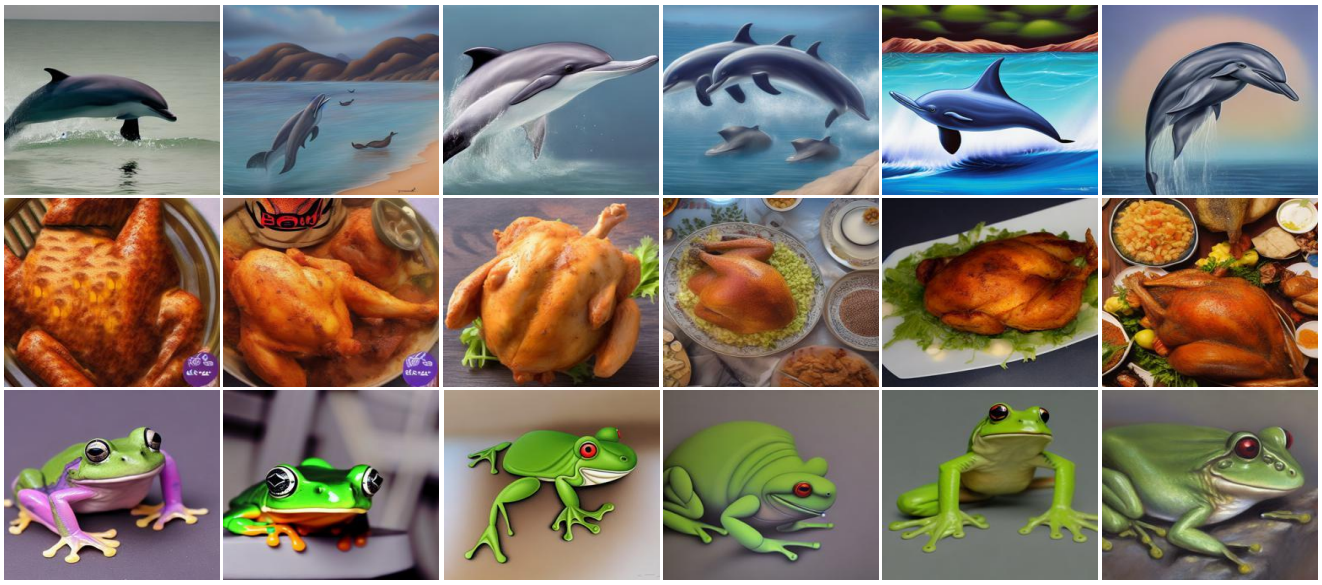
[1] https : / / huggingface . co / runwayml / stable - diffusion-v1-5

5.2. Main Results

In this section, we compare our RDPO with all strong relevant baselines from both quantitative and qualitative perspectives.

Quantitative results. Following Ren et al. [29] , we randomly sampled five images for each prompt and computed the average reward value of all the images for quantitative evaluation. We present the quantitative results on the animal prompts and pick-a-pic prompts on Tab. 1 and Tab. 2, respectively. We find that: (1) Since our RDPO* builds upon TailorPO-G by incorporating the better revised intermediate step reward mentioned in § 3.2, we compare RDPO* and TailorPO-G and find that RDPO* consistently achieves superior performance across different settings where various rewards are used as training signals, which demonstrates the effectiveness of our proposed the better revised intermediate step reward. (2) By comparing RDPO* and RDPO, while the only difference between them is using penalty term Ψ or not, we find that RDPO achieves overall improvements over RDPO*, further validating the effectiveness of our proposed penalty term Ψ.

Qualitative results. We visualize the corresponding generation results in Fig. 2 and Fig 4, and these visual results show that the images generated by RDPO align better with the prompts, are more aesthetically pleasing, contain richer



SD v1.5 [31] D3PO [47] DDPO [2] SPO [19] TailorPO-G [29] RDPO (Ours)

Figure 2. Visualization of images generated by diffusion models fine-tuned using different methods on animal prompts [47]. Among these results, we can find that the images generated by RDPO are more aligned to text and with higher fidelity and more visually pleasing.

Table 3. Ablation study results for penalty term Ψ in Eq. 9. Prompts are related to common animals provided by Yang et al. [47]. Δ means performance gains. All results are reported as the mean over three independent runs.

	Aesthetic scorer	ImageReward
Stable Diffusion v1.5	5.79	0.65
D3PO [48]	6.46	0.95
D3PO + Penalty TermΨ	**7.33**$_{\Delta=+0.87}$	**1.14**$_{\Delta=+0.19}$
SPO [19]	5.89	0.95
SPO + Penalty TermΨ	**6.54**$_{\Delta=+0.65}$	**1.21**$_{\Delta=+0.26}$
TailorPO [29]	6.66	1.20
TailorPO + Penalty TermΨ	**7.51**$_{\Delta=+0.85}$	**1.44**$_{\Delta=+0.24}$

details, and exhibit fewer distortions and deformations.

5.3. Ablation Studies

Hyperparameter λ. We conduct an ablation study over the value of λ in Eq. 9 to determine the sensitivity of the model's performance to this parameter. We test $\lambda \in \{10, 100, 10,000\}$ on animal prompts with aesthetic scorer and show the results in Fig 3 (a). We find that adding penalty term Ψ always leads to better performance compared to baseline, but a too-large value of λ may lead to model collapse (see the $\lambda = 10,000$ at step 175).

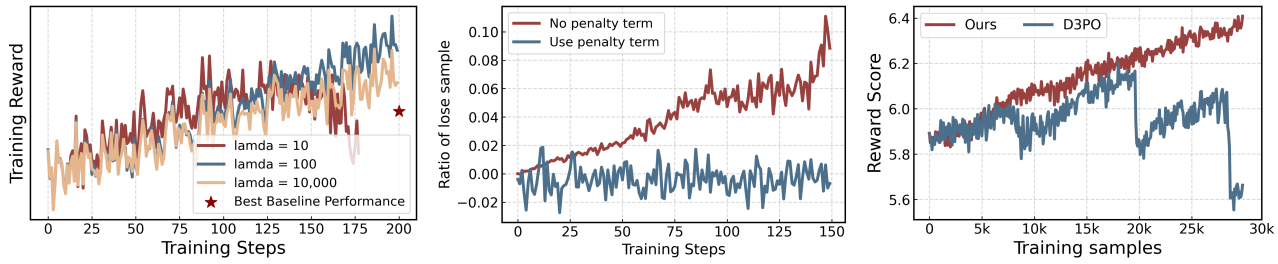
Effectiveness of Penalty Term Ψ in Eq. 9. First, we visualize the change trend of ratio of RDPO* and RDPO in Eq 9 at Fig 3 (b), while the only difference RDPO uses the penalty term Ψ. We find that by adding this penalty term, we effectively limit the model's tendency to increase the sampling probability of dispreferred samples. This ensures that the model does not generate potentially harmful content and prevents overfitting, which could occur if both preferred and dispreferred sample probabilities were increased without the penalty term, leading to reduced generalization.

To further validate the effectiveness and generalization of the penalty term Ψ, we add this penalty term to baselines, and train these baselines and their variants with this term (denoted as $_{\Psi}$) on the animal prompts dataset, using the aesthetic scorer and imagereward as the reward signal. The results, shown in Table 3, demonstrate that this penalty term not only improves the performance and stability of our own model but also enhances the performance and stability of other baselines.

5.4. Further Analysis

In this section, we present further analysis of RDPO to validate the effectiveness.

r^* in Eq. 6 better represents the reward at each step. To validate that r^* is more effective, we adopt an intermediate-step best-of-N sample testing approach. Specifically, at each step from x_t to x_{t-1}, we sample N candidates for x_{t-1} and score them based on the corresponding intermediate-step rewards, selecting the best x_{t-1} for the next sampling step. To evaluate the accuracy of the intermediate rewards, we measure the quality of the final generated image. We set $N = 8, 32$ and use SD v1.5 as the generative model on animal prompts for the experiments. We compare our $r*$ with r in Eq. 5 [29] and step-preference reward model proposed by Liang et al. [20]. The results are shown in Table 4. We find that our r^* achieves the best intermediate-



(a) (b) (c)

Figure 3. (a) Visualization of reward model values change with different λ used in Eq. 9. (b) Using penalty tern Ψ can limit the model's tendency to increase the ratio of dispreferred sample $\log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$ during the aligning process. (c) RDPO is more stable than baseline DPO methods.

Figure 4. Visualization of images generated by diffusion models fine-tuned using different methods on Pick-a-pic prompts [15]. Among these results, we can find that the images generated by RDPO are more aligned to text and with higher fidelity and more visually pleasing.

Table 4. Intermediate-step best-of-N results. Prompts are common animal prompts [47]. We take Stable Diffusion v1.5 as the base model and use aesthetic scorer as the final generation evaluator.

	N = 8	N = 32
Stable Diffusion v1.5	5.79	5.79
+ step-preference reward model [20]	5.93	6.31
+ r in Eq. 5 [29]	6.02	6.17
+ r^* in Eq. 6 (Ours)	**6.24**	**6.40**

step Best-of-N performance, indicating that r^* is better at selecting the optimal sample from multiple intermediate steps. In other words, r^* is a more effective estimator of the intermediate-step reward.

Training process of RDPO is more stable. A potential concern is that the DPO algorithm is sensitive to out-of-distribution preference data, as it assumes that π_{ref} can adequately capture and represent the distribution of preference data. This could lead to instability during DPO training. To address this, we increased the number of samples (i.e., training steps) from 10,000 to 30,000 using Aesthetic scorer on animal prompts. We show the changes in training rewards for our model and the baseline D3PO in Figure 3 (c). We observe that both D3PO and SPO exhibit collapse, while our model continuously and stably optimizes, maintaining an increase in reward.

Generalize to Out-of-Distribution Data. As discussed in § 3.3, we are concerned that the standard DPO algorithm, by increasing the ratio of win and lose samples, may lead to overfitting on the training data, resulting in poor generalization. We hypothesize that our penalty term Ψ (see in § 3.3) can mitigate this by constraining the ratio of dispreferred samples, allowing the model to effectively focus on win samples and maintain better generalization. here we conduct experiments to test this hypothesis using out-of-distribution (OOD) prompt data. Specifically, we optimize our model on common animal prompts [47] and test it on prompts sampled from the Pick-a-Pic dataset [15]. The results are summarized

Table 5. Out-of-Distribution evaluation. Models are fine-tuned using common animals prompts provided by Yang et al. [47] and evaluated on prompts sampled from Pick-a-pic [15]. All results are reported as the mean over three independent runs.

	Aesthetic scorer	ImageReward
Stable Diffusion v1.5	5.28	0.14
D3PO [48]	5.20	0.13
SPO [19]	5.17	0.14
TailorPO-G [29]	5.32	0.15
RDPO	**5.77**	**0.25**

Table 6. Generalize experiments on other diffusion scheduler. We use DDPM here and models are fine-tuned using prompts provided by Pick-a-pic [15]. All results are reported as the mean over three independent runs.

	Aesthetic scorer	ImageReward
Stable Diffusion v1.5	5.37	0.21
D3PO [48]	5.44	0.23
SPO [19]	5.40	0.23
TailorPO-G [29]	5.48	0.28
RDPO	**5.60**	**0.31**

in Table 5. We find that (1) some dpo methods achieve lower performance than baseline SD v1.5 (e.g., D3PO and SPO), indicating that DPO may harm the generalization ability of diffusion models by increasing the sampling probability of both preferred and dispreferred samples. (2) our RDPO achieves the best OOD performance, which validating that he penalty term Ψ enhances the model's generalization by constraining the sampling probability of dispreferred samples.

Generalize to Other Diffusion Scheduler. In the main experiments of § 5, we primarily conducted experiments on DDIM [39]. To further validate the effectiveness of our model, we also performed related experiments on DDPM [9]. The experimental settings followed those outlined in § 5.1, where we used animal prompts as the training set and SD v1.5 as the baseline. The results are presented in Table 6. We observe that RDPO achieves comparable performance on DDPM, indicating that RDPO is robust to different schedulers.

Generalize to Other Strong Diffusion Models. While our main experiments utilize SD v1.5 as the base model for fine-tuning, to further validate the effectiveness of our RDPO, we extend our study by adopting SDXL[2] as the new foundation model and follow the experimental setup described in § 5.1. Specifically, we use animal-related prompts provied by Yang et al. [47] as the training prompt set. The results are summarized in Table 7, wihch demonstrate that our RDPO continues to achieve superior alignment performance, and further validates that our RDPO can generalizes well to stronger diffusion models.

[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

Table 7. Generalize experiments on other strong diffusion baseline model. Prompts are provided by Pick-a-pic [15]. All results are reported as the mean over three independent runs.

	Aesthetic scorer	ImageReward
Stable Diffusion XL (SDXL)	6.33	1.79
D3PO [48]	6.52	1.90
SPO [19]	6.54	1.83
TailorPO-G [29]	6.62	1.80
RDPO	**6.77**	**1.85**

Table 8. Reward generalization: the model fine-tuned towards a reward model also exhibited higher reward values on other different but related reward models.

Train \ Evaluate	Aesthetic scorer	ImageReward	HPSv2	PickScore
SD v1.5	5.61	0.69	27.77	20.51
Aesthetic scorer	**7.33**	<u>1.12</u>	27.99	<u>20.57</u>
ImageReward	6.14	**1.46**	<u>28.30</u>	20.81
HPSv2	5.34	0.99	**28.29**	20.65
PickScore	<u>6.31</u>	1.03	27.93	**21.02**

Reward generalization. We follow Ren et al. [29] to conduct reward generalization expirements. We selected one reward model from the aesthetic scorer, ImageReward, HPSv2, and Pickscore for fine-tuning, and used the other three reward models for evaluation. Table 8 shows that after being fine-tuned towards the aesthetic scorer, ImageReward, and PickScore, the model usually exhibited higher performance on all these four reward models. In other words, our method boosted the overall ability of the model to generate high-quality images.

6. Conclusion

In this work, we presented findings on two potential risk of DPO in text to-image diffusion models. First we indicates that the reward estimation for intermediate-step samples in existing DPO methods is biased and propose a more accurate estimation approach. Then, we find that the standard DPO loss can lead to an increased likelihood of the model generating dispreferred completions, which may pose potential risks in real-world applications. In order to mitigate above two issues, we introduce a new diffusion alignment method named RDPO, which we show overcomes the two potential risk of DPO and is competitive with strong relevant baselines, both quantitatively and qualitatively.

Potential societal impacts. Our method focuses on aligning the text to-image diffusion model with human preferences. While our approach aims to improve this alignment process, a potential risk lies in the possibility of inherent biases in the human-provided preference signals (e.g., gender, race, etc.). These biases may pose risks during the alignment process of the diffusion model and debiasing may become an important future direction.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*. OpenReview.net, 2024. 1, 4, 5, 6

[3] Chang Chen, Fei Deng, Kenji Kawaguchi, Caglar Gulcehre, and Sungjin Ahn. Simple hierarchical planning with diffusion. In *ICLR*. OpenReview.net, 2024. 2

[4] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*. OpenReview.net, 2023. 3

[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] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 2

[7] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M. Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, pages 11808–11826. PMLR, 2023. 2

[8] Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *CoRR*, abs/2404.14743, 2024. 3

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 8

[10] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 2

[11] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2022. 2

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022. 5

[13] Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *ICML*, pages 9902–9915. PMLR, 2022. 2

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[15] 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. 2, 5, 7, 8

[16] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024. 2

[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] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *CoRR*, abs/2406.04314, 2024. 5, 6, 7, 8

[20] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2 (5):7, 2024. 1, 2, 3, 4, 6, 7

[21] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *ICML*, pages 21450–21474. PMLR, 2023. 2

[22] Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024. 2

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804. PMLR, 2022. 2

[24] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*. OpenReview.net, 2023. 2

[25] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 1, 2

[26] 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. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2

[27] 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 *NeurIPS*, 2023. 1

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 1, 2

[29] Jie Ren, Yuhang Zhang, Dongrui Liu, Xiaopeng Zhang, and Qi Tian. Refining alignment framework for diffusion models with intermediate-step preference ranking. *arXiv preprint arXiv:2502.01667*, 2025. 1, 3, 4, 5, 6, 7, 8

[30] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023. 2

[31] 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 (CVPR)*, pages 10684–10695, 2022. 6

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022. 2

[33] 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, 2

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*. OpenReview.net, 2023. 2

[37] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. JMLR.org, 2015. 2

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 2, 5, 8

[40] Zhiwei Tang, Dmitry Rybin, and Tsung-Hui Chang. Zeroth-order optimization meets human feedback: Provable learning via ranking oracles. *arXiv preprint arXiv:2303.03751*, 2023. 2

[41] 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. *arXiv preprint arXiv:2311.12908*, 2023. 1, 2, 3, 4

[42] 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 (CVPR)*, pages 8228–8238, 2024. 1

[43] 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

[44] Xun Wu, Shaohan Huang, and Furu Wei. Multimodal large language model is a human-aligned annotator for text-to-image generation. *arXiv preprint arXiv:2404.15100*, 2024. 5

[45] 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. *arXiv preprint arXiv:2304.05977*, 2023. 2, 3, 5

[46] Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. Rlcd: Reinforcement learning from contrast distillation for language model alignment, 2023. 2

[47] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[48] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8941–8951, 2024. 1, 5, 6, 8

[49] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024. 2

[50] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. *arXiv preprint arXiv:2312.12490*, 2023. 2