

# Rethinking Direct Preference Optimization in Diffusion Models

Junyong Kang<sup>\*1</sup>, Seohyun Lim<sup>\*1</sup>, Kyungjune Baek<sup>2</sup>, Hyunjung Shim<sup>†1</sup>

<sup>1</sup>Korea Advanced Institute of Science and Technology

<sup>2</sup>Sejong University

{jykang,seohyunlim,kateshim}@kaist.ac.kr

{kyungjune.baek}@sejong.ac.kr

## Abstract

Aligning text-to-image (T2I) diffusion models with human preferences has emerged as a critical research challenge. While Direct Preference Optimization (DPO) has established a foundation for preference learning in large language models (LLMs), its extension to diffusion models remains limited in alignment performance. In this work, we propose an enhanced version of Diffusion-DPO by introducing a stable reference model update strategy. This strategy facilitates the exploration of better alignment solutions while maintaining training stability. Moreover, we design a timestep-aware optimization strategy that further boosts performance by addressing preference learning imbalance across timesteps. Through the synergistic combination of our exploration and timestep-aware optimization, our method significantly improves the alignment performance of Diffusion-DPO on human preference evaluation benchmarks, achieving state-of-the-art results.

## 1 Introduction

Diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019; Song et al. 2021) have emerged as a powerful generative framework, achieving remarkable success in text-to-image (T2I) generation (Podell et al. 2023; Saharia et al. 2022). By leveraging large-scale image-text pairs during training, these models can synthesize high-fidelity images conditioned on natural language prompts. However, due to the uncurated and noisy nature of web-scale datasets, their outputs often misalign with human aesthetic and semantic preferences.

To address these challenges, the field of aligning with human feedback has emerged as a crucial research direction. Inspired by advances in aligning language models with human feedback (Ouyang et al. 2022; Rafailov et al. 2023), recent efforts have extended alignment techniques to the vision domain. These methods can be broadly categorized into two prominent approaches: reward model-based fine-tuning (Black et al. 2024; Fan et al. 2023; Xu et al. 2024; Clark et al. 2024; Prabhudesai et al. 2023) and Direct Preference Optimization (DPO) (Wallace et al. 2024; Li et al. 2024; Yang, Chen, and Zhou 2024; Zhu, Xiao, and Honavar 2025).

Reward model-based approaches typically rely on large vision-language models, such as PickScore (Kirstain et al. 2023) and ImageReward (Xu et al. 2024). They are known to suffer from unstable training and reward overoptimization problems (Hu et al. 2025; Kim, Kim, and Park 2025). In contrast, DPO (Rafailov et al. 2023) offers a more stable alternative by directly optimizing the human preference data without the use of an explicit reward model. Extensions of DPO to diffusion models, such as Diffusion-DPO (Wallace et al. 2024) and D3PO (Yang et al. 2024), have shown early promise in the image generation domain. However, their alignment performance remain suboptimal compared to recent state-of-the-art methods (Ethayarajh et al. 2024; Zhu, Xiao, and Honavar 2025), as shown in Figure 1(a).

In this work, we identify a key limitation in current DPO adaptations in diffusion as constrained model exploration. Naïve Diffusion-DPO has relatively small divergence from the pre-trained model, suggesting limited traversal in model space (Figure 1(b)). This motivates our key hypothesis: encouraging greater exploration can help the model discover improved alignment solutions.

To this end, we adopt a reference update framework to promote exploratory behavior. We find that updating the reference model forces the model to quickly change its prediction, leading to more exploration. However, unrestricted reference updates lead to a model divergence problem, where the model loses its generative prior and degrades image quality.

Based on the observation that model error grows as the reference model diverges from the pre-trained model, we introduce a regularization algorithm to constrain the deviation of the reference model. This adaptive strategy restricts excessive updates to the reference model when the deviation becomes large, preserving the generative prior while enabling meaningful exploration. Despite its simplicity, this method offers important insights into the training stability of DPO for diffusion models and significantly improves alignment performance.

In addition, we observe that the impact of preference optimization with our exploration is imbalanced across diffusion timesteps, showing the need for emphasizing the learning of early timesteps. As several prior works (Balaji et al. 2022; Wang and Vastola 2024) demonstrated that diffusion models synthesize semantic structures during early timesteps, we aim to prioritize preference learning in early timesteps. To accom-

<sup>\*</sup>These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

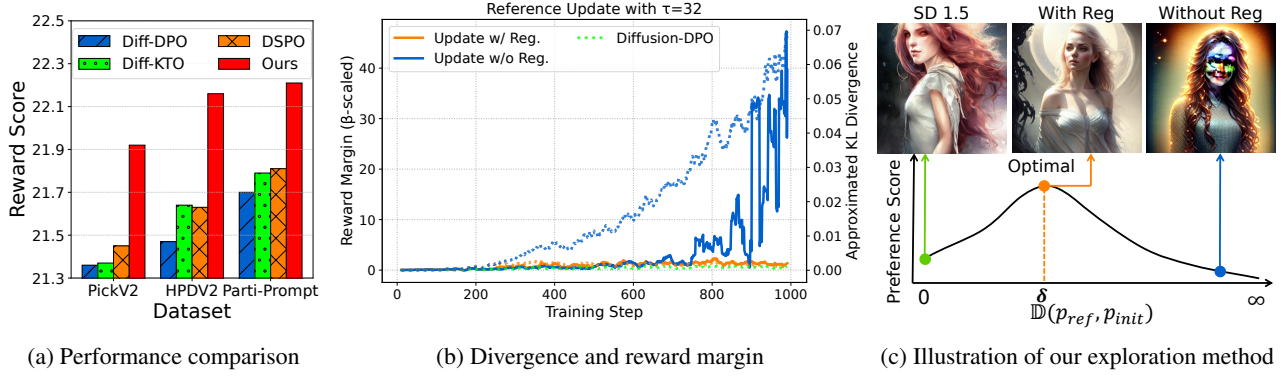


Figure 1: (a) Alignment performance of Diffusion-DPO, baselines, and our proposed method on SD1.5 with PickScore reward. Our method significantly improves the alignment performance over Diffusion-DPO. (b) (solid lines) Implicit reward margin under the reference update strategy, with and without our regularization. (dotted lines) Approximated KL divergence between the training model and the pre-trained model (Diffusion-DPO), and between the reference model and the pre-trained model (ours). (c) Relationship between the divergence from the pre-trained model and the preference score. The illustration shows that controlled divergence enables effective exploration while excessive deviation results in a decline in preference score.

plish this, we propose a timestep-aware optimization strategy for our exploration method. Specifically, we oversample early timesteps during loss computation and apply a decreasing reward scale schedule to balance reward magnitudes across timesteps.

The contributions of this paper are summarized as follows:

1. We propose a novel recipe to improve direct preference optimization for T2I preference alignment, by introducing a stable reference model update method combined with a timestep-aware optimization strategy.
2. Our analysis provides new insights into reference model relaxation and timestep-dependent behavior of preference optimization in diffusion models, distinguishing from existing methods.
3. By combining the reference model update strategy with a timestep-aware optimization strategy, our method significantly enhances Diffusion-DPO’s alignment performance and achieves state-of-the-art performance. This success highlights that effective exploration is key to maximizing the performance of DPO for diffusion models.

## 2 Preliminaries

**Diffusion Models.** Diffusion models are a class of generative models that learn to reverse a gradual noising process applied to data. Following the DDPM (Ho, Jain, and Abbeel 2020) formulation, the forward process is defined as a Markov chain with a noise schedule  $\alpha_t$ , resulting in a sequence of latent variables  $\mathbf{x}_{1:T}$ :

$$q(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (1)$$

where  $q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I})$ .

The goal of the diffusion model is to learn a reverse process parameterized by a neural network  $p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  to obtain generated samples

$p_\theta(\mathbf{x}_0)$ . Given  $\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\bar{\alpha}_t \mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$ , where  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ , the model estimates the noise  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  via  $\epsilon_\theta(\mathbf{x}_t, t)$ . The training objective is derived from the evidence lower bound (ELBO) on the data likelihood:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} [\lambda(t) \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2], \quad (2)$$

where  $t \sim \mathcal{U}(0, T)$  and  $\lambda(t)$  denotes timestep-wise weighting function. Recent works (Choi et al. 2022; Hang et al. 2023; Yu et al. 2024) suggest advanced weighting schedules for  $\lambda(t)$  to improve sample quality and convergence.

**Preference Optimization in Diffusion.** To align the conditional distribution  $p_\theta(\mathbf{x}_0 | \mathbf{c})$  with human preferences, where  $\mathbf{c} \sim \mathcal{D}_c$  denotes the prompt condition, RLHF methods (Ouyang et al. 2022; Xu et al. 2024; Black et al. 2024) utilize a reward model  $r(\mathbf{c}, \mathbf{x}_0)$ . These methods aim to maximize the reward of the generated sample  $\mathbf{x}_0$  while keeping the distribution close to a reference distribution  $p_{\text{ref}}$  in terms of KL-divergence regularization:

$$\max_{p_\theta} \mathbb{E}_{\mathbf{c} \sim \mathcal{D}_c, \mathbf{x}_0 \sim p_\theta(\mathbf{x}_0 | \mathbf{c})} [r(\mathbf{c}, \mathbf{x}_0)] - \beta \mathbb{D}_{\text{KL}} [p_\theta(\mathbf{x}_0 | \mathbf{c}) \| p_{\text{ref}}(\mathbf{x}_0 | \mathbf{c})]. \quad (3)$$

The reward model is typically learned from preference-annotated datasets under the Bradley-Terry model, where each data entry consists of a triplet  $(\mathbf{c}, \mathbf{x}_0^w, \mathbf{x}_0^l)$ , representing a prompt, a preferred image, and a dispreferred image, respectively. Rather than training a reward model, Direct Preference Optimization (DPO) (Rafailov et al. 2023) parametrizes the *implicit reward* using the current and the reference model:

$$r(\mathbf{c}, \mathbf{x}_0) = \beta \log \frac{p_\theta(\mathbf{x}_0 | \mathbf{c})}{p_{\text{ref}}(\mathbf{x}_0 | \mathbf{c})}, \quad (4)$$

where we omit the partition function  $Z(\mathbf{c}) = \sum_{\mathbf{x}_0} p_{\text{ref}}(\mathbf{x}_0 | \mathbf{c}) \exp(r(\mathbf{c}, \mathbf{x}_0)/\beta)$  as it does not contribute to the loss formulation. Diffusion-DPO (Wallace et al. 2024) expands the RLHF objective (Eq. 3) into the diffusion

trajectory  $p_\theta(\mathbf{x}_{0:T})$ , and then plugs the implicit reward into the Bradley-Terry model to obtain the following loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}} \log \sigma \left( \beta \mathbb{E}_{\mathbf{x}_{1:T}^w \sim p_\theta(\mathbf{x}_{1:T}^w | \mathbf{x}_0^w), \mathbf{x}_{1:T}^l \sim p_\theta(\mathbf{x}_{1:T}^l | \mathbf{x}_0^l)} \left[ \log \frac{p_\theta(\mathbf{x}_{0:T}^w)}{p_{\text{ref}}(\mathbf{x}_{0:T}^w)} - \log \frac{p_\theta(\mathbf{x}_{0:T}^l)}{p_{\text{ref}}(\mathbf{x}_{0:T}^l)} \right] \right). \quad (5)$$

This is intractable as the loss requires sampling from  $p_\theta(\mathbf{x}_{0:T})$ . Note that we omit the prompt  $c$  for simplicity. Utilizing Jensen’s inequality and approximating the reverse process  $p_\theta$  with the forward process  $q$ , Diffusion-DPO derives the final tractable loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0, T), \mathbf{x}_t^w \sim q(\mathbf{x}_t^w | \mathbf{x}_0^w), \mathbf{x}_t^l \sim q(\mathbf{x}_t^l | \mathbf{x}_0^l)} \left[ \log \sigma(\beta T \lambda(t) (r_t(\mathbf{x}_0^w) - r_t(\mathbf{x}_0^l))) \right], \quad (6)$$

where we denote  $r_t(\mathbf{x}_t) = -(\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 - \|\epsilon - \epsilon_{\text{ref}}(\mathbf{x}_t, t)\|_2^2)$ . From this approximation, we interpret  $r_t(\mathbf{x}_t)$  as a timestep-wise implicit reward. Thus, the above loss can be regarded as forcing the model to maximize the margin between  $r_t(\mathbf{x}_t^w)$  and  $r_t(\mathbf{x}_t^l)$ .

### 3 Method

Our goal is to improve the preference alignment of Diffusion-DPO by addressing two limitations: insufficient exploration in the model space and imbalance in timestep-level learning. To encourage exploration of the model, we begin by replacing the fixed reference model with the training model. We find that naively updating the reference model leads to error scaling behavior, which can result in model divergence.

To mitigate this issue, we constrain the divergence of the reference model from the pre-trained model, which allows the model to explore new solutions while preserving its generative quality. However, we observe that our exploration method learns the preference signal unevenly across timesteps. To facilitate the preference learning in early steps, we introduce a timestep-aware training strategy to address the imbalance problem. By integrating this strategy with our exploration method, we further improve the performance of Diffusion-DPO.

#### Reference Model Update with Regularization

In standard DPO, the reference model remains fixed to the initial pre-trained model  $p_{\text{init}}$ . While this design maintains training stability, it limits the model’s capacity to explore diverse alignment solutions. Recent works (Pang et al. 2024; Zhang et al. 2025) have challenged this constraint by proposing multiple training stages using reward models, where the reference model is updated at each stage to improve preference alignment. In language model alignment, TR-DPO (Gorbatovski et al. 2024) demonstrates that updating the reference model during training can mitigate overoptimization and improve performance. Motivated by these findings, we extend this reference update strategy to the diffusion setting by periodically replacing the reference model with the current training model.

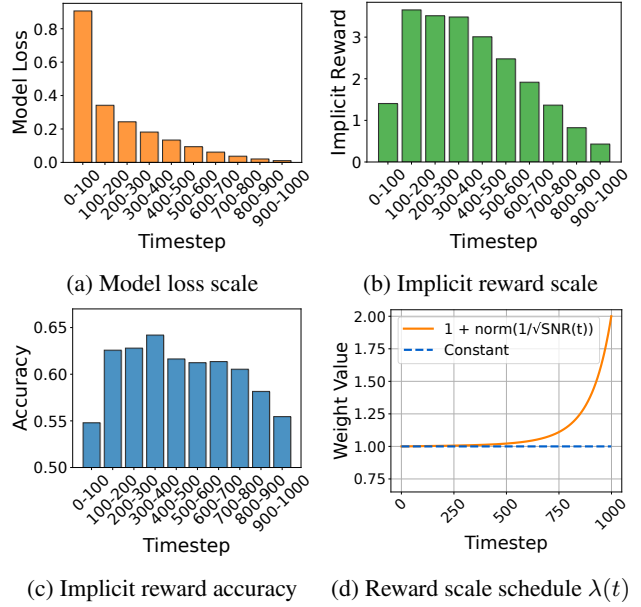


Figure 2: Imbalance problem in our reference update method. We present the scale of (a) model losses and (b) implicit rewards, (c) the preference accuracy of implicit rewards, (d) and our proposed reward scale schedule  $\lambda(t)$ .

We observe that a naïve reference update strategy in diffusion models leads to a critical model divergence problem. To analyze this, we examine the training dynamics with a reference update period of  $\tau = 32$  (see Appendix C for other values). Figure 1(b) shows the growing divergence of the reference model from the pre-trained model, along with the implicit reward margin  $r_t(\mathbf{x}_t^w) - r_t(\mathbf{x}_t^l)$  (Update w/o Reg.).

As training progresses, both divergence and reward margin increase, indicating that the model is actively optimizing toward the DPO objective through exploration. This also implies growing prediction error, quantified as  $\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2$ . Smaller  $\tau$  values amplify this effect by reducing the gap between training and reference models, forcing the model to scale its prediction error more aggressively. Although moderate exploration may help the model discover better solutions, the uncontrolled error explosion ultimately causes the model to diverge. This behavior contrasts with observations in TR-DPO (Gorbatovski et al. 2024) for language models, where updating the reference model tends to reset the reward margin toward zero during training. In the diffusion setting, however, excessive reference drift degrades image quality due to error scaling.

To balance exploration and training stability, we propose to *regularize the reference model* by constraining its divergence from the pre-trained model (Figure 1(c)). Our key insight is that excessive divergence leads to increased prediction error. By limiting this divergence, we can suppress error scaling while enabling controlled exploration.

We define a divergence metric  $\mathbb{D}(p_{\text{ref}}, p_{\text{init}})$ , to quantify the deviation of the current reference model  $p_{\text{ref}}$  from the initial model  $p_{\text{init}}$ . Estimating this divergence requires computing

the expectation under the joint distribution of  $p_{\text{ref}}$  or  $p_{\text{init}}$  across timesteps, which is intractable. Instead, we approximate the divergence using the forward process  $q$ . Specifically, the (reverse) KL divergence can be approximated as follows:

$$\mathbb{D}_{\text{KL}}(p_{\text{ref}}, p_{\text{init}}) \approx \mathbb{E}_{\mathbf{x}_0 \sim \mathcal{D}, \mathbf{x}_{1:T} \sim q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[ \log \frac{p_{\text{ref}}(\mathbf{x}_{0:T})}{p_{\text{init}}(\mathbf{x}_{0:T})} \right]. \quad (7)$$

Using a similar derivation to equation in (Wallace et al. 2024), we obtain the following:

$$\begin{aligned} \mathbb{D}_{\text{KL}}(p_{\text{ref}}, p_{\text{init}}) &\approx T \mathbb{E}_{\mathbf{x}_0 \sim \mathcal{D}, t, \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \\ &\left[ \mathbb{D}_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_{0,t}) \parallel p_{\text{ref}}(\mathbf{x}_{t-1} | \mathbf{x}_t)) \right. \\ &\quad \left. - \mathbb{D}_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_{0,t}) \parallel p_{\text{init}}(\mathbf{x}_{t-1} | \mathbf{x}_t)) \right]. \quad (8) \end{aligned}$$

We empirically find that this approximation proves sufficient for divergence monitoring. To reduce computational overhead, we evaluate the divergence on a small subset of preferred images  $\mathbf{x}_0^w$  from the training batch. After establishing the divergence metric, the next step is to choose a reference model that ensures training stability. When the divergence of the current reference model  $p_{\text{ref}}$  exceeds a threshold  $\delta$ , we freeze the reference model near the  $\delta$ -boundary to prevent further updates, as shown in Figure 1(b) (Update w/ Reg.). We also explore a re-initialization option in Section 4, which resets the reference model to the pre-trained model.

### Boosting with Timestep-Aware Optimization

Despite the benefits of our reference update method, we find that the learned preference signal is unevenly distributed across diffusion timesteps. This leads to suboptimal alignment, particularly in early steps where semantic structures are formed (Balaji et al. 2022; Wang and Vastola 2024). In fact, several diffusion training studies (Yu et al. 2024; Choi et al. 2022) discovered that optimization is more difficult in early timesteps and emphasizing these steps improves the output quality (Figure 2(a)).

In the preference optimization setting, we observe a similar trend during our exploration. To investigate this, we analyze the implicit reward  $r_t(\mathbf{x}_t)$  using a model trained with our reference update strategy. We randomly sample 5,000 image pairs from the Pick-a-Pic v2 validation set (Kirstain et al. 2023) and compute both the average scale of implicit reward and preference accuracy (the number of cases where  $r_t(\mathbf{x}_t^w)$  is greater than  $r_t(\mathbf{x}_t^l)$ ) across 10 evenly partitioned intervals  $[0, T]$ .

As shown in Figure 2(b) and (c), both the scale and accuracy of  $r_t(\mathbf{x}_t)$  are marked lower at early timesteps. This finding indicates that the reward signal is weaker in early steps, leading to imbalanced preference learning difficulty. Motivated by this observation, we aim to develop a timestep-aware preference optimization strategy that accounts for this imbalance.

To encourage preference learning in early steps, we apply an oversampling approach inspired by (Yang, Chen, and Zhou 2024), drawing a single timestep  $t$  instead of multi-sample expectations. In this method, timesteps are drawn from a skewed categorical distribution  $\text{Cat}(\gamma^t)$  towards early

steps, with probability vector  $\gamma^t / \sum_{t'} \gamma^{t'}$ , where  $\gamma \in [0, 1]$ . Moreover, we introduce a timestep-dependent reward scaling schedule  $\lambda(t)$  to directly mitigate imbalance. Although Eq. 6 already presents the weighting schedule, it has been ignored in previous works and treated as a constant in practice (Wallace et al. 2024; Zhu, Xiao, and Honavar 2025). Instead, we design  $\lambda(t)$  to decrease over timesteps, assigning larger values than the constant schedule during early steps. As an example, we define  $\lambda(t) = 1 + \text{norm}(1/\sqrt{\text{SNR}(t)})$ , where  $\text{SNR}(t)$  denotes the signal-to-noise ratio,  $\text{norm}(\cdot)$  indicates the normalization operator over time (Figure 2(d)). As  $\lambda(t)$  controls the implicit regularization via  $\beta$ , we also interpret this schedule as a means to reduce the risk of overfitting at early timesteps. We explore other choices in Appendix C, verifying the advantage of the proposed schedule.

We note that the timestep-aware strategy alone may not yield performance gains in isolation (Section 4). Our key contribution lies in its synergy with our exploration method, which unlocks the potential of DPO for diffusion models.

## 4 Experiment

### Experimental Setup

**Dataset.** Following prior works (Wallace et al. 2024; Li et al. 2024), we use Pick-a-Pic v2 train dataset (Kirstain et al. 2023) for training. For evaluation, we employ test set prompts from the Pick-a-Pic v2 dataset (500 entries), HPDv2 benchmark (Wu et al. 2023) (3,200 entries), and the PartiPrompts dataset (Yu et al. 2022) (1,632 entries). As Pick-a-Pic v2 has a small number of prompts, we generate a total of 2,500 images using five different seeds.

**Evaluation Protocol.** To quantitatively evaluate the proposed method, we adopt five reward models as evaluation metrics: PickScore (Kirstain et al. 2023), HPSv2 (Wu et al. 2023), CLIP (Radford et al. 2021), Aesthetics Score (Schuhmann 2022), and ImageReward (Xu et al. 2024). For each reward model, we compare the win rates of our method against the baseline approaches. The win rate is the proportion of images with higher reward scores than those generated by the baseline model, under the same seed.

**Baseline Methods.** We evaluate our method against baseline preference optimization algorithms, Diffusion-DPO (Wallace et al. 2024), Diffusion-KTO (Li et al. 2024), and DSPO (Zhu, Xiao, and Honavar 2025). We reproduce Diffusion-DPO and DSPO, and use a public checkpoint for Diffusion-KTO. When reproducing the baseline methods, we maintain consistency by employing the same hyperparameters reported in the original paper. We also include supervised fine-tuning (SFT) as a baseline, but we exclusively use the preferred images.

**Implementation Details.** In this paper, we conduct experiments on Stable Diffusion v1.5 (SD1.5) (Rombach et al. 2022) and SDXL (Podell et al. 2023). We tune the reference model update period,  $\tau$ , by searching over  $\{16, 32, 64\}$  steps and select the optimal value for each model. The monitoring divergence threshold  $\delta$  is empirically determined as 0.005 for SD1.5 and 0.002 for SDXL. For the timestep-aware training strategy, we set the discount factor  $\gamma$  for the timestep sampling to 0.9 as the default. Other details and hyperparameters are provided in Appendix A.

Dataset	Model	PickScore	HPSv2	CLIP	Aesthetic	ImageReward	Average
PickV2	vs. SD1.5*	<b>89.96</b>	<b>83.84</b>	<b>64.56</b>	<b>78.04</b>	<b>77.76</b>	<b>78.83</b>
	vs. Diff-KTO*	<b>74.52</b>	<b>52.16</b>	<b>56.16</b>	<b>56.00</b>	<b>51.80</b>	<b>58.13</b>
	vs. SFT	<b>71.72</b>	<b>50.40</b>	<b>55.00</b>	49.04	<b>53.08</b>	<b>55.85</b>
	vs. Diff-DPO	<b>75.20</b>	<b>70.80</b>	<b>53.64</b>	<b>69.16</b>	<b>66.36</b>	<b>67.03</b>
	vs. DSPO	<b>71.36</b>	<b>51.76</b>	<b>53.72</b>	<b>51.32</b>	<b>51.16</b>	<b>55.86</b>
PartiPrompts	vs. SD1.5*	<b>84.25</b>	<b>84.31</b>	<b>60.66</b>	<b>81.00</b>	<b>80.82</b>	<b>78.21</b>
	vs. Diff-KTO*	<b>71.57</b>	<b>56.80</b>	<b>53.80</b>	<b>65.32</b>	<b>62.56</b>	<b>62.01</b>
	vs. SFT	<b>71.38</b>	<b>56.43</b>	<b>55.76</b>	<b>59.07</b>	<b>64.46</b>	<b>61.42</b>
	vs. Diff-DPO	<b>72.18</b>	<b>75.80</b>	<b>53.19</b>	<b>75.37</b>	<b>73.47</b>	<b>70.00</b>
	vs. DSPO	<b>69.73</b>	<b>56.56</b>	<b>53.74</b>	<b>60.48</b>	<b>62.68</b>	<b>60.64</b>
HPDv2	vs. SD1.5*	<b>91.44</b>	<b>89.34</b>	<b>63.62</b>	<b>82.66</b>	<b>84.22</b>	<b>82.26</b>
	vs. Diff-KTO*	<b>73.12</b>	<b>53.69</b>	<b>52.75</b>	<b>56.59</b>	<b>55.31</b>	<b>58.29</b>
	vs. SFT	<b>73.88</b>	<b>57.88</b>	<b>54.94</b>	<b>53.75</b>	<b>58.38</b>	<b>59.77</b>
	vs. Diff-DPO	<b>77.22</b>	<b>77.81</b>	<b>53.87</b>	<b>69.62</b>	<b>73.50</b>	<b>70.40</b>
	vs. DSPO	<b>72.28</b>	<b>57.75</b>	<b>53.97</b>	<b>53.09</b>	<b>57.44</b>	<b>58.91</b>

Table 1: Win rates of our method against baseline preference optimization methods using SD1.5 as the base model. \* indicates model checkpoints released by the original authors. Higher win rates indicate better alignment performance and win rates exceeding 50% are marked in bold.

Model	PickScore	HPSv2	CLIP	Aesthetic	ImageReward	Average
vs. SDXL*	<b>81.24</b>	<b>81.76</b>	<b>57.64</b>	<b>59.28</b>	<b>70.96</b>	<b>70.18</b>
vs. MaPO*	<b>81.16</b>	<b>74.88</b>	<b>58.16</b>	45.12	<b>65.92</b>	<b>65.05</b>
vs. InPO*	<b>64.80</b>	<b>56.56</b>	<b>54.76</b>	<b>55.00</b>	<b>56.76</b>	<b>57.58</b>
vs. Diff-DPO	<b>68.40</b>	<b>73.76</b>	<b>50.28</b>	<b>57.52</b>	<b>54.40</b>	<b>60.87</b>
vs. DSPO	<b>60.88</b>	<b>64.68</b>	<b>51.44</b>	<b>55.52</b>	49.28	<b>56.36</b>

Table 2: Win rates of our method using SDXL as the base model, evaluated on the Pick-a-Pic v2 test set.

## Experiment Results

**Quantitative Results.** To verify the effectiveness of the proposed method, we compare our method with the original Diffusion-DPO and baseline preference optimization algorithms. Table 1 presents the experimental results, measured in win rates from five reward metrics and their average. Notably, when comparing our method to Diffusion-DPO, the average win rate ranges from 67% to 70% across datasets, indicating significant improvement of alignment. These findings underscore that model exploration plus the timestep-aware training strategy can unlock the potential of Diffusion-DPO. We further report our results on SDXL in Table 2, including public checkpoints of MAPO (Hong et al. 2024) and InPO (Lu et al. 2025) as baselines. Due to space constraints, we report results for the remaining test prompt sets and raw reward scores in Appendix B.

**Qualitative Results.** Figure 3 presents images generated by baselines and by our method. We find that Diffusion-DPO tends to show only subtle changes compared to the original model, due to limited exploration. Diffusion-KTO and DSPO also struggle to produce images faithful to the text prompt. For example, they fail to generate *burgers* in the first row, and miss compositional objects such as *cyberpunk + cat* (DSPO) or *pixel + bulldog* (Diffusion-KTO). Overall, our method correctly identifies objects and compositional relationships

Model	PickScore	HPSv2	CLIP	Aesthetic	IR
Diff-DPO	21.36	27.19	33.84	5.53	0.32
LR=5e-8	19.75	24.95	29.81	4.79	-0.39
$\beta=1000$	21.26	27.16	33.55	5.55	0.23
Re-Init	21.50	27.16	34.17	5.57	0.38
Ours	<b>21.93</b>	<b>27.84</b>	<b>34.42</b>	<b>5.75</b>	<b>0.65</b>

Table 3: Ablation study of alternative exploration strategies. Raw scores for each reward metric are reported. The highest value for each metric is displayed in bold.

described in the text prompts and generates aesthetically appealing images compared to other models. We display more qualitative results in Appendix D.

## Ablation Study

**Comparison with Alternative Exploration Strategies.** Table 3 compares our method (SD1.5) on the Pick-a-Pic v2 test set with alternative exploration strategies: (1) increasing learning rates, (2) reducing the implicit regularization coefficient  $\beta$ , and (3) re-initializing the reference model in the update strategy, when its divergence exceeds the threshold. (1) Increasing the learning rate from 1e-8 to 5e-8 leads to

3D digital illustration, Burger with wheels speeding on the race track, supercharged, detailed, hyperrealistic, 4K



Cyberpunk cat



<pixel art> gray French bulldog



a pair of headphones on a pumpkin



(a) SD1.5 (Rombach et al. 2022)

(b) Diff-DPO (Wallace et al. 2024)

(c) Diff-KTO (Li et al. 2024)

(d) DSPO (Zhu, Xiao, and Honavar 2025)

(e) Ours

Figure 3: Qualitative comparison. We compare the generated outputs from various preference optimization algorithms based on SD1.5, including our method.

model collapse and a substantial drop in all metrics. (2) Reducing  $\beta$  from 5,000 to 1,000 does not improve performance. (3) Re-initializing scheme yields a minor improvement, since the strong constraint of the initial model restricts exploration.

**Effect of Reference Update Period.** Figure 4 illustrates that, without our reference regularization, frequent model update ( $\tau$  decreases) causes model divergence, leading to a sharp performance drop. By constraining the update boundary with the divergence monitoring, our method consistently outperforms Diffusion-DPO, reducing the sensitivity to the update period  $\tau$ .

**Effect of Timestep-Aware Strategy.** Table 4 shows that

combining timestep-aware optimization with exploration improves performance, while using it alone may degrade Diffusion-DPO. This suggests that exploration is critical for enabling effective preference learning at early timesteps, highlighting the synergistic effect between the two components. We also find that reward scale scheduling further enhances oversampling. Figure 5 presents the relative increase in model divergence induced by our timestep-aware strategy, compared to using only the reference update. The scheduled method exhibits a lower divergence budget in early timesteps, indicating a regularization effect that helps prevent overfitting and leads to better performance.

## 5 Related Work

### RLHF in Diffusion Models

Reinforcement Learning from Human Feedback (RLHF) has proven highly effective in aligning human preference in the large language model domain (Ouyang et al. 2022; OpenAI 2024). Recently, similar approaches have been explored in the T2I diffusion domain, leveraging human feedback and various quality metrics as reward signals. Previous works in RLHF to diffusion models have re-formulated the diffusion process as a Markov Decision Process (MDP). DDPO (Black et al. 2024) and DPOK (Fan et al. 2023) compute rewards at the final timestep and apply the policy gradient method to fine-tune the model. Alternatively, methods such as ReFL (Xu et al. 2024), DRaFT (Clark et al. 2024), and AlignProp (Prabhudesai et al. 2023) propose differentiable reward frameworks, enabling direct policy updates through backpropagation.

### Direct Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al. 2023) has emerged as a promising alternative to RLHF, because it obviates the need to train a separate reward model. Building on the success of DPO, numerous variants have recently been explored in the language domain (Azar et al. 2024; Gorbatsovski et al. 2024; Meng, Xia, and Chen 2024; Wu et al. 2024; Hong, Lee, and Thorne 2024; Zhao et al. 2025). DPO has also been extended to diffusion models to enhance alignment between generated images and human preferences. Notably, Diffusion-DPO (Wallace et al. 2024) and D3PO (Yang et al. 2024) adapt the DPO loss to diffusion models. Diffusion-KTO (Li et al. 2024) substitutes the standard DPO loss with Kahneman-Tversky Optimization (KTO), training with single-instance data without requiring pairwise comparisons. Meanwhile, some recent works consider the innate structure of diffusion models instead of naively applying the language model losses. Yang et al., (Yang, Chen, and Zhou 2024) modify the uniform timestep sampling in Diffusion-DPO, deriving the loss from the densely defined rewards across timesteps. InPO (Lu et al. 2025) introduces DDIM inversion in Diffusion-DPO instead of random noise injection for training efficiency, and DSPO (Zhu, Xiao, and Honavar 2025) fine-tunes diffusion models by aligning with human preferences using score matching principles.

## 6 Conclusion

We present a novel training framework for enhancing DPO in diffusion models. Our method enables the stable model exploration by updating the reference model under a divergence constraint and addressing reward scale imbalance across denoising steps to further improve exploration. Experiments show that our strategy significantly improves the alignment performance of Diffusion-DPO across multiple benchmarks, achieving new state-of-the-art results. We believe our work opens for future research on the training dynamics of preference optimization and motivates further development of DPO-based methods in diffusion models.

Model	Pick	HPSv2	CLIP	Aesthetic	IR
Diff-DPO	21.36	27.19	33.84	5.53	0.32
Time. only	21.10	26.57	33.56	5.54	0.16
Exploration only	21.88	27.63	34.40	5.70	0.58
$\gamma = 0.8$	21.66	27.49	34.25	5.66	0.52
$\gamma = 0.8 + \text{Scale}$	21.90	27.65	34.42	5.69	0.61
$\gamma = 0.9$	21.82	27.71	34.39	5.71	0.64
$\gamma = 0.9 + \text{Scale (Ours)}$	<b>21.93</b>	<b>27.84</b>	<b>34.42</b>	<b>5.75</b>	<b>0.65</b>

Table 4: Ablation study of the timestep-aware optimization strategy. (Top) Our timestep-aware strategy shows a synergistic effect when combined with exploration. (Bottom) The reward scale schedule further enhances performance. Raw scores for each reward metric are reported.

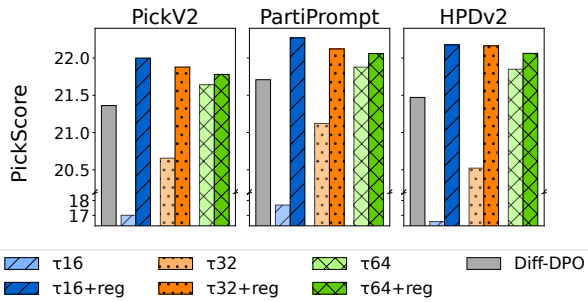


Figure 4: Results of reference model regularization with  $\tau \in \{16, 32, 64\}$ , evaluated using the PickScore reward.

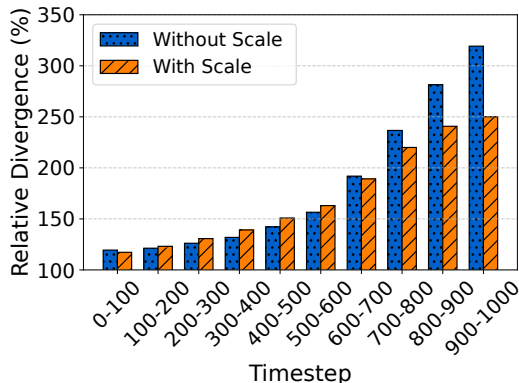


Figure 5: Relative increase in divergence with and without reward scale scheduling. In each interval, 100% represents the divergence of our reference update method.

## Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the MSIP (RS-2025-00520207, RS-2023-00219019), IITP grant funded by the Korean government (MSIT) (RS-2024-00457882, RS-2025-02217259), KEIT grant funded by the Korean government (MOTIE) (No. 2022-0-00680, No. 2022-0-01045), and Artificial Intelligence Graduate School Program (KAIST) (RS-2019-II190075).

## References

- Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.
- Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.
- Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2024. Training Diffusion Models with Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.
- Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception Prioritized Training of Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11462–11471.
- Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2024. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *The Twelfth International Conference on Learning Representations*.
- Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 79858–79885.
- Gorbatovski, A.; Shaposhnikov, B.; Malakhov, A.; Surnachev, N.; Aksenov, Y.; Maksimov, I.; Balagansky, N.; and Gavrilov, D. 2024. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*.
- Hang, T.; Gu, S.; Li, C.; Bao, J.; Chen, D.; Hu, H.; Geng, X.; and Guo, B. 2023. Efficient Diffusion Training via Min-SNR Weighting Strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7441–7451.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.
- Hong, J.; Lee, N.; and Thorne, J. 2024. ORPO: Monolithic Preference Optimization without Reference Model. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11170–11189. Miami, Florida, USA: Association for Computational Linguistics.
- Hong, J.; Paul, S.; Lee, N.; Rasul, K.; Thorne, J.; and Jeong, J. 2024. Margin-aware Preference Optimization for Aligning Diffusion Models without Reference. *arXiv:2406.06424*.
- Hu, Z.; Zhang, F.; Chen, L.; Kuang, K.; Li, J.; Gao, K.; Xiao, J.; Wang, X.; and Zhu, W. 2025. Towards Better Alignment: Training Diffusion Models with Reinforcement Learning Against Sparse Rewards. *arXiv:2503.11240*.
- Kim, S.; Kim, M.; and Park, D. 2025. Test-time Alignment of Diffusion Models without Reward Over-optimization. In *The Thirteenth International Conference on Learning Representations*.
- Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.
- Li, S.; Kallidromitis, K.; Gokul, A.; Kato, Y.; and Kozuka, K. 2024. Aligning Diffusion Models by Optimizing Human Utility. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Lu, Y.; Wang, Q.; Cao, H.; Wang, X.; Xu, X.; and Zhang, M. 2025. InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28629–28639.
- Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.
- OpenAI. 2024. GPT-4 Technical Report. *arXiv:2303.08774*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.
- Pang, R. Y.; Yuan, W.; He, H.; Cho, K.; Sukhbaatar, S.; and Weston, J. E. 2024. Iterative Reasoning Preference Optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv:2307.01952*.
- Prabhudesai, M.; Goyal, A.; Pathak, D.; and Fragkiadaki, K. 2023. Aligning Text-to-Image Diffusion Models with Reward Backpropagation. *arXiv:2310.03739*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Saharia, C.; Chan, W.; Saxena, S.; Lit, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Gontijo-Lopes, R.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.
- Schuhmann, C. 2022. LAION-AESTHETICS. <https://laion.ai/blog/laion-aesthetics/>. Accessed: 2023 - 11 - 10.
- Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.
- Wang, B.; and Vastola, J. J. 2024. Diffusion Models Generate Images Like Painters: an Analytical Theory of Outline First, Details Later. arXiv:2303.02490.
- Wu, J.; Xie, Y.; Yang, Z.; Wu, J.; Gao, J.; Ding, B.; Wang, X.; and He, X. 2024. beta-DPO: Direct Preference Optimization with Dynamic beta. *Advances in Neural Information Processing Systems*, 37: 129944–129966.
- Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *CoRR*.
- Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.
- Yang, K.; Tao, J.; Lyu, J.; Ge, C.; Chen, J.; Shen, W.; Zhu, X.; and Li, X. 2024. 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*, 8941–8951.
- Yang, S.; Chen, T.; and Zhou, M. 2024. A dense reward view on aligning text-to-image diffusion with preference. In *Proceedings of the 41st International Conference on Machine Learning*, 55998–56032.
- Yu, H.; Shen, L.; Huang, J.; Li, H.; and Zhao, F. 2024. Unmasking Bias in Diffusion Model Training. In *The 18th European Conference on Computer Vision ECCV 2024*. Springer.
- Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.*
- Zhang, X.; Yang, L.; Li, G.; Cai, Y.; xie jiake; Tang, Y.; Yang, Y.; Wang, M.; and CUI, B. 2025. IterComp: Iterative Composition-Aware Feedback Learning from Model Gallery for Text-to-Image Generation. In *The Thirteenth International Conference on Learning Representations*.
- Zhao, H.; Winata, G. I.; Das, A.; Zhang, S.-X.; Yao, D.; Tang, W.; and Sahu, S. 2025. RainbowPO: A Unified Framework for Combining Improvements in Preference Optimization. In *The Thirteenth International Conference on Learning Representations*.
- Zhu, H.; Xiao, T.; and Honavar, V. G. 2025. DSPO: Direct Score Preference Optimization for Diffusion Model Alignment. In *The Thirteenth International Conference on Learning Representations*.