Rethinking Direct Preference Optimization in Diffusion Models

Junyong Kang[1*]   **Seohyun Lim**[1*]   **Kyungjune Baek**[2]   **Hyunjung Shim**[1†]

[1] Korea Advanced Institute of Science and Technology
[2] 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 [10, 29, 30] have emerged as a powerful generative framework, achieving remarkable success in text-to-image (T2I) generation [22, 27]. 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 [20, 25], 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 [3, 7, 35, 5, 23] and Direct Preference Optimization (DPO) [31, 16, 37, 42].

Reward model-based approaches typically rely on large vision-language models, such as PickScore [15] and ImageReward [35]. They are known to suffer from unstable training and reward overoptimization problems [13, 14]. In contrast, DPO [25] 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 [31] and D3PO [36], have shown early promise in the image generation domain. However, their alignment performance remain suboptimal compared to recent state-of-the-art methods [6, 42], as shown in Figure 1(a).

---

[*]Equal contribution.

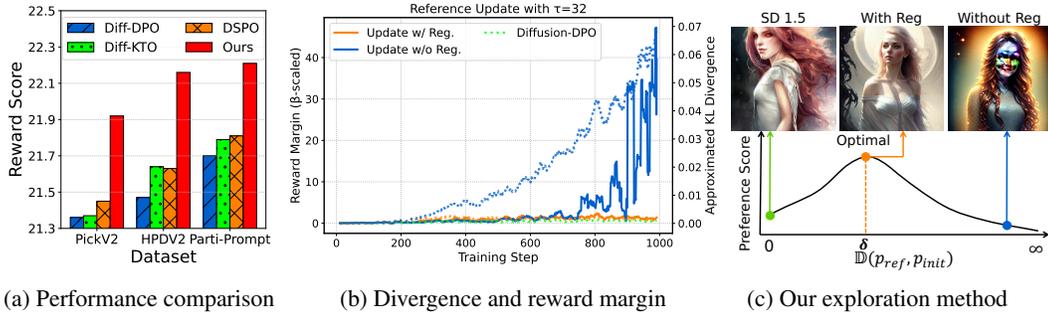| (a) Performance comparison | (b) Divergence and reward margin | (c) Our exploration method |

Figure 1: (a) Alignment performance of Diffusion-DPO, baselines, and our proposed method on SD1.5 with PickScore reward. Our method significantly improves alignment over Diffusion-DPO. (b) (solid lines) Implicit reward margin under the reference update strategy, with and without our regularization. (dotted lines) KL divergence between the training model and the pre-trained model (Diffusion-DPO), and between the reference and the pre-trained model (ours). (c) Illustration of our reference-regularized exploration method.

In this work, we identify a key limitation in current DPO adaptations in diffusion as constrained model exploration. Naíve 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 [2, 32] demonstrated that diffusion models synthesize semantic structures during early timesteps, we aim to prioritize preference learning in early timesteps. To accomplish 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 [10] formulation, the forward process is

defined as a Markov chain with a noise schedule α_t, resulting in a sequence of latent variables $\boldsymbol{x}_{1:T}$:

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}), \text{where} \quad q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\boldsymbol{I}). \quad (1)$$

The goal of the diffusion model is to learn a reverse process parameterized by a neural network $p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ to obtain generated samples $p_\theta(\boldsymbol{x}_0)$. Given $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\bar{\alpha}_t\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I})$, where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the model estimates the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ via $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$. The training objective is derived from the evidence lower bound (ELBO) on the data likelihood:

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

where $t \sim \mathcal{U}(0, T)$ and $\lambda(t)$ denotes timestep-wise weighting function. Recent works [4, 9, 38] suggest advanced weighting schedules for $\lambda(t)$ to improve sample quality and convergence.

Preference Optimization in Diffusion. To align the conditional distribution $p_\theta(\boldsymbol{x}_0|\boldsymbol{c})$ with human preferences, where $c \sim \mathcal{D}_c$ denotes the prompt condition, RLHF methods [20, 35, 3] utilize a reward model $r(\boldsymbol{c}, \boldsymbol{x}_0)$. These methods aim to maximize the reward of the generated sample $\boldsymbol{x}_0$ while keeping the distribution close to a reference distribution p_{ref} in terms of KL-divergence regularization:

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

The reward model is typically learned from preference-annotated datasets under the Bradley-Terry model, where each data entry consists of a triplet $(c, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)$, representing a prompt, a preferred image, and a dispreferred image, respectively. Rather than training a reward model, Direct Preference Optimization (DPO) [25] parametrizes the *implicit reward* using the current and the reference model:

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

where we omit the partition function $Z(\boldsymbol{c}) = \sum_{\boldsymbol{x}_0} p_{\text{ref}}(\boldsymbol{x}_0|\boldsymbol{c}) \exp(r(\boldsymbol{c}, \boldsymbol{x}_0)/\beta)$ as it does not contribute to the loss formulation. Diffusion-DPO [31] expands the RLHF objective (Eq. 3) into the diffusion trajectory $p_\theta(\boldsymbol{x}_{0:T})$, and then plugs the implicit reward into the Bradley-Terry model to obtain the following loss:

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

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

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

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

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 naïvely 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.
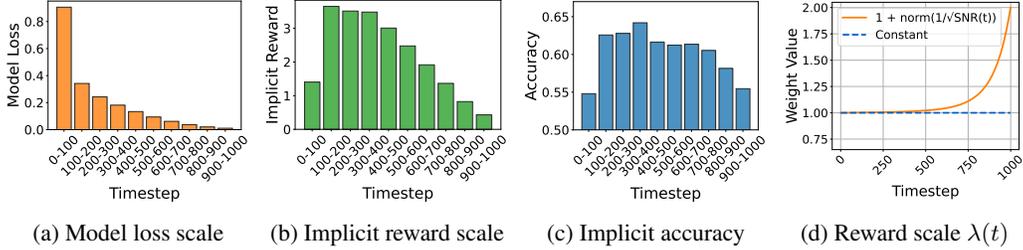
3

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 $\lambda(t)$.

3.1 Reference Model Update with Regularization

In standard DPO, the reference model remains fixed to the initial pre-trained model p_{init}. While this design maintains training stability, it limits the model's capacity to explore diverse alignment solutions. Recent works [21, 40] 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 [8] 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.

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 $r_t(\boldsymbol{x}_t^w) - r_t(\boldsymbol{x}_t^l)$ (Update w/o Reg.).

As training progresses, both the divergence and the reward margin increase, indicating that the model is actively optimizing toward the DPO objective through exploration. This also implies growing prediction error, quantified as $||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||_2^2$. Smaller τ 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 [8] 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_{ref} from the initial model p_{init}. Estimating this divergence requires computing the expectation under the joint distribution of p_{ref} or p_{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:

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

Using a similar derivation to equation in [31], we obtain the following:

$$\mathbb{D}_{\text{KL}}(p_{\text{ref}}, p_{\text{init}}) \approx T\, \mathbb{E}_{\boldsymbol{x}_0 \sim \mathcal{D},\, t,\, \boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \Big[\mathbb{D}_{\text{KL}}\big(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_{0,t}) \,\big\|\, p_{\text{ref}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \big)$$
$$- \mathbb{D}_{\text{KL}}\big(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_{0,t}) \,\big\|\, p_{\text{init}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \big) \Big] \tag{8}$$

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 $\boldsymbol{x}_0^w$ 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_{ref} exceeds a threshold δ, we freeze the reference model near the δ-boundary to prevent further updates, as shown in Figure 1(b) (Update w/ Reg.). We also explore a re-initialization option in Section 4.3, which resets the reference model to the pre-trained model.

3.2 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 [2, 32]. In fact, several diffusion training studies [38, 4] 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 $r_t(\boldsymbol{x}_t)$ using a model trained with our reference update strategy. We randomly sample 5,000 image pairs from the Pick-a-Pic v2 validation set [15] and compute both the average scale of implicit reward and preference accuracy (the number of cases where $r_t(\boldsymbol{x}_t^w)$ is greater than $r_t(\boldsymbol{x}_t^l)$) across 10 evenly partitioned intervals $[0, T]$.

As shown in Figure 2(b) and (c), both the scale and accuracy of $r_t(\boldsymbol{x}_t)$ 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 [37], drawing a single timestep t instead of multisample 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 [31, 42]. Instead, we design $\lambda(t)$ to decrease over timesteps, assigning larger values than the constant schedule during early steps. As an example, we define $\lambda(t) = 1 + \text{norm}(1/\sqrt{SNR(t)})$, where $SNR(t)$ 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 β, 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 4.2). Our key contribution lies in its synergy with our exploration method, which unlocks the potential of DPO for diffusion models.

4 Experiment

4.1 Experimental Setup

Dataset. Following prior works [31, 16], we use Pick-a-Pic v2 train dataset [15] for training. For evaluation, we employ test set prompts from the Pick-a-Pic v2 dataset (500 entries), HPDv2 benchmark [34] (3,200 entries), and the PartiPrompts dataset [39] (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 [15], HPSv2 [34], CLIP [24], Aesthetics Score [28], and ImageReward [35]. 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 [31], Diffusion-KTO [16], and DSPO [42]. We reproduce Diffusion-DPO and DSPO, and use a public checkpoint for Diffusion-KTO. When reproducing the baseline methods, we maintain

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.

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

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

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

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) [26] and SDXL [22]. We tune the reference model update period, τ, by searching over $\{16, 32, 64\}$ steps and select the optimal value for each model. The monitoring divergence threshold δ 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 γ for the timestep sampling to 0.9 as the default. Other details and hyperparameters are provided in Appendix A.

4.2 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 [12] and InPO [17] 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 described in the text prompts and generates aesthetically appealing images compared to other models. We display more qualitative results in Appendix D.

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 [26] | (b) Diff-DPO [31] | (c) Diff-KTO [16] | (d) DSPO [42] | (e) Ours |

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

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.

Model	PS	HPSv2	CLIP	Aes.	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
β=1000	21.26	27.16	33.55	5.55	0.23
Re-Init	21.50	27.16	34.17	5.57	0.38
Ours	**21.93**	**27.84**	**34.42**	**5.75**	**0.65**

Table 4: Ablation of timestep-aware optimization. (Top) Combining with exploration yields synergy. (Bottom) Scale scheduling boosts performance.

Model	PS	HPSv2	CLIP	Aes.	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
$\gamma = 0.8$ + Scale	21.90	27.65	34.42	5.69	0.61
$\gamma = 0.9$	21.82	27.71	34.39	5.71	0.64
$\gamma = 0.9$ + Scale (Ours)	**21.93**	**27.84**	**34.42**	**5.75**	**0.65**

4.3 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 β, 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 model collapse and a substantial drop in all metrics. (2) Reducing β 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.
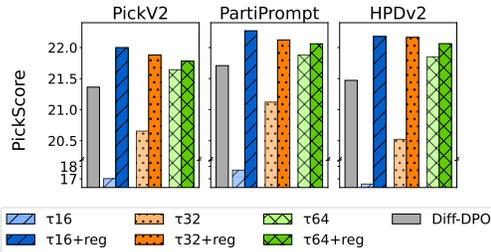
Figure 4: Reference model regularization with $\tau \in \{16, 32, 64\}$ evaluated using PickScore.
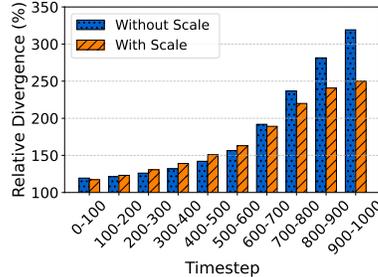

Figure 5: Effect of reward scale scheduling, where 100% corresponds to the divergence of our reference update.

Effect of Reference Update Period. Figure 4 illustrates that, without our reference regularization, frequent model update (τ 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 τ.

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

5.1 RLHF in Diffusion Models

Reinforcement Learning from Human Feedback (RLHF) has proven highly effective in aligning human preference in the large language model domain [20, 19]. 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 [3] and DPOK [7] compute rewards at the final timestep and apply the policy gradient method to fine-tune the model. Alternatively, methods such as ReFL [35], DRaFT [5], and AlignProp [23] propose differentiable reward frameworks, enabling direct policy updates through backpropagation.

5.2 Direct Preference Optimization

Direct Preference Optimization (DPO) [25] 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 [1, 8, 18, 33, 11, 41]. DPO has also been extended to diffusion models to enhance alignment between generated images and human preferences. Notably, Diffusion-DPO [31] and D3PO [36] adapt the DPO loss to diffusion models. Diffusion-KTO [16] 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., [37] modify the uniform timestep sampling in Diffusion-DPO, deriving the loss from the densely defined rewards across timesteps. InPO [17] introduces DDIM inversion in Diffusion-DPO instead of random noise injection for training efficiency, and DSPO [42] 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.

Acknowledgements

References

[1] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11462–11471, 2022.

[5] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

[8] Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

[9] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7441–7451, October 2023.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[11] Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference, 2024.

[13] Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, and Wenwu Zhu. Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards, 2025.

[14] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

[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. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[16] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[17] Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. Inpo: Inversion preference optimization with reparametrized ddim for efficient diffusion model alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28629–28639, 2025.

[18] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

[19] OpenAI. Gpt-4 technical report, 2024.

[20] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 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, 2022. Curran Associates Inc.

[21] Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

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

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

[25] 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.

[26] 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.

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 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, 2022. Curran Associates Inc.

[28] Christoph Schuhmann. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022. Accessed: 2023 - 11- 10.

[29] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 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, 2021.

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

[32] Binxu Wang and John J. Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later, 2024.

[33] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. beta-dpo: Direct preference optimization with dynamic beta. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024.

[34] 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. *CoRR*, 2023.

[35] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[36] 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*, pages 8941–8951, 2024.

[37] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55998–56032, 2024.

[38] Hu Yu, Li Shen, Jie Huang, Hongsheng Li, and Feng Zhao. Unmasking bias in diffusion model training. In *The 18th European Conference on Computer Vision ECCV 2024*. Springer, 2024.

[39] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022.

[40] Xinchen Zhang, Ling Yang, Guohao Li, YaQi Cai, xie jiake, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin CUI. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[41] Hanyang Zhao, Genta Indra Winata, Anirban Das, Shi-Xiong Zhang, David Yao, Wenpin Tang, and Sambit Sahu. RainbowPO: A unified framework for combining improvements in preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Supplementary Materials

Dataset	Model	PickScore	HPSv2	CLIP	Aesthetic	IR	Average
PickV2	vs. SDXL*	**81.24**	**81.76**	**57.64**	**59.28**	**70.96**	**70.18**
	vs. MaPO*	**81.16**	**74.88**	**58.16**	45.12	**65.92**	**65.05**
	vs. InPO*	**64.80**	**56.56**	**54.76**	**55.00**	**56.76**	**57.58**
	vs. Diff-DPO	**68.40**	**73.76**	**50.28**	**57.52**	**54.40**	**60.87**
	vs. DSPO	**60.88**	**64.68**	**51.44**	**55.52**	49.28	**56.36**
PartiPrompts	vs. SDXL*	**71.45**	**79.84**	**53.00**	**64.71**	**74.20**	**68.64**
	vs. MaPO*	**73.77**	**79.47**	**58.82**	47.30	**69.12**	**65.70**
	vs. InPO*	**53.86**	**57.90**	**52.39**	**56.50**	**56.92**	**55.51**
	vs. Diff-DPO	**59.68**	**73.04**	45.34	**58.21**	**54.66**	**58.19**
	vs. DSPO	**59.13**	**67.10**	48.35	**54.04**	**52.14**	**56.15**
HPDv2	vs. SDXL*	**78.91**	**84.38**	**51.88**	**58.38**	**73.22**	**69.35**
	vs. MaPO*	**77.69**	**73.19**	**52.56**	49.03	**67.41**	**63.97**
	vs. InPO*	**58.06**	**53.25**	49.97	**54.59**	**56.34**	**54.44**
	vs. Diff-DPO	**62.59**	**73.72**	47.09	**57.41**	**55.34**	**59.23**
	vs. DSPO	**58.84**	**56.00**	42.84	**64.34**	**51.41**	**54.69**

Table S1: Win rates of our method against baseline preference optimization methods using SDXL 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.

Dataset	Model	PickScore	HPSv2	CLIP	Aesthetic	IR
PickV2	SD1.5*	20.66±0.03	26.52±0.04	32.59±0.12	5.39±0.01	-0.07±0.02
	Diff-KTO*	21.37±0.03	27.77±0.04	33.83±0.12	5.69±0.01	0.59±0.02
	SFT	21.42±0.03	27.77±0.04	33.79±0.12	**5.76**±0.01	0.57±0.02
	Diff-DPO	21.36±0.03	27.19±0.04	33.84±0.12	5.53±0.01	0.32±0.02
	DSPO	21.46±0.03	27.78±0.04	34.00±0.12	5.74±0.01	0.61±0.02
	Ours	**21.93**±0.03	**27.84**±0.04	**34.42**±0.11	5.75±0.01	**0.65**±0.02
PartiPrompts	SD1.5*	21.31±0.03	26.96±0.04	32.70±0.14	5.28±0.01	-0.08±0.03
	Diff-KTO*	21.79±0.03	28.10±0.04	33.79±0.14	5.54±0.01	0.49±0.03
	SFT	21.78±0.03	28.09±0.04	33.50±0.14	5.59±0.01	0.43±0.03
	Diff-DPO	21.71±0.03	27.46±0.04	33.57±0.14	5.38±0.01	0.22±0.03
	DSPO	21.81±0.03	28.11±0.04	33.74±0.14	5.59±0.01	0.49±0.03
	Ours	**22.21**±0.03	**28.31**±0.04	**34.29**±0.14	**5.66**±0.01	**0.73**±0.02
HPDv2	SD1.5*	20.73±0.02	26.63±0.03	33.95±0.09	5.38±0.01	-0.25±0.02
	Diff-KTO*	21.64±0.02	28.26±0.03	35.55±0.09	5.76±0.01	0.57±0.02
	SFT	21.58±0.02	28.16±0.03	35.23±0.09	5.79±0.01	0.51±0.02
	Diff-DPO	21.47±0.02	27.49±0.03	35.31±0.09	5.60±0.01	0.21±0.02
	DSPO	21.63±0.02	28.18±0.03	35.36±0.10	5.80±0.01	0.54±0.02
	Ours	**22.16**±0.02	**28.38**±0.03	**35.91**±0.09	**5.83**±0.01	**0.68**±0.02

Table S2: Average reward scores for each method on SD1.5, with 1-sigma error bars. The highest score in each column is shown in bold, and the second highest is underlined.

A   Further Implementation Details

We train our models on the Pick-a-Pic v2 dataset [15]. The Pick-a-Pic v2 training set comprises about 900K image pairs with approximately 58k distinct prompts. Images were ranked by human evaluators, consisting of a preferred and a non-preferred image for each given input prompt.

During training, we follow the hyperparameter configurations of prior works [31, 42, 16]. We use the AdamW optimizer with a learning rate of 2.048×10^{-8}. Training is performed with a batch size of 4 per GPU, 128 gradient accumulation steps, and 4 NVIDIA A6000 GPUs, resulting in an effective batch size of 2048. We train the models for 1000 iterations on SD1.5 and 600 iterations on SDXL.

Dataset	Model	PickScore	HPSv2	CLIP	Aesthetic	IR
PickV2	SDXL	22.16±0.07	27.98±0.09	36.09±0.29	6.01±0.03	0.57±0.05
	MaPO	22.25±0.07	28.32±0.09	36.23±0.28	**6.15±0.02**	0.70±0.04
	InPO	<u>22.68±0.03</u>	<u>28.88±0.04</u>	36.89±0.12	6.09±0.01	<u>0.98±0.02</u>
	Diff-DPO	22.65±0.07	28.46±0.08	37.23±0.26	6.02±0.03	0.89±0.04
	DSPO	22.66±0.07	28.81±0.08	**37.58±0.26**	5.96±0.02	0.95±0.04
	Ours	**22.94±0.07**	**29.06±0.08**	<u>37.28±0.26</u>	6.09±0.02	**1.01±0.04**
PartiPrompts	SDXL	21.31±0.03	26.96±0.04	32.70±0.14	5.28±0.01	-0.08±0.03
	MaPO	22.62±0.03	28.58±0.05	35.35±0.14	**5.91±0.01**	0.79±0.02
	InPO	<u>23.01±0.03</u>	<u>29.14±0.05</u>	35.89±0.15	5.86±0.01	1.01±0.02
	Diff-DPO	22.94±0.03	28.80±0.04	<u>36.36±0.14</u>	5.85±0.01	1.08±0.02
	DSPO	22.95±0.03	29.06±0.04	**36.60±0.14**	5.84±0.01	<u>1.16±0.02</u>
	Ours	**23.09±0.03**	**29.39±0.05**	36.22±0.14	<u>5.90±0.01</u>	**1.17±0.02**
HPDv2	SDXL	22.78±0.02	28.63±0.03	38.16±0.09	6.13±0.01	0.78±0.01
	MaPO	22.84±0.02	29.01±0.03	38.14±0.09	**6.22±0.01**	0.88±0.01
	InPO	<u>23.27±0.02</u>	<u>29.55±0.03</u>	38.46±0.09	6.18±0.01	1.04±0.01
	Diff-DPO	23.20±0.02	29.08±0.03	38.59±0.09	6.17±0.01	1.06±0.01
	DSPO	23.24±0.02	29.48±0.03	**38.94±0.08**	6.11±0.01	<u>1.12±0.01</u>
	Ours	**23.41±0.02**	**29.63±0.03**	<u>38.67±0.08</u>	6.21±0.01	**1.13±0.01**

Table S3: Average reward scores for each method in SDXL with 1-sigma error bar. The highest score in each column is shown in bold, and the second highest is underlined.

Dataset	PickScore	HPSv2	CLIP	Aesthetic	ImageReward	Average
PickV2	**70.80**	**67.80**	**55.00**	**69.20**	**64.60**	**65.48**
PartiPrompts	**58.21**	**62.01**	49.51	**52.02**	**55.82**	**55.51**
HPDv2	**70.69**	**68.72**	**51.47**	**59.84**	**62.81**	**62.71**

Table S4: Win rates of our method against Diffusion-DPO using SD3 as the base model. Higher win rates indicate better alignment performance and win rates exceeding 50% are marked in bold.

We set the base regularization coefficient (or signal scale) $\beta = 5000$ for Diffusion-DPO and DSPO (SDXL), and $\beta = 0.001$ for DSPO (SD1.5), consistent with their original settings.

For evaluation, we use 50 inference timesteps and set the classifier-free guidance scale to 7.5 (5.0 for SDXL). To evaluate our method with a sufficient amount of images, we evaluate with 5 different random seeds on the Pick-a-Pic v2, generating a total of 2,500 images. As the number of prompts in PartiPrompts and HPDv2 test dataset is large (1,632 and 3,200 prompts, respectively), the evaluation is conducted using a single seed.

B Quantitative Results

In this section, we provide more detailed quantitative results. We include the win rate results for the SDXL model in the PartiPrompts and HPDv2 test prompts in Table S1. Our method consistently achieves average win rates above 50% against Diffusion-DPO and baseline methods, with particularly strong performance on human preference metrics such as PickScore, HPSv2, and ImageReward.

Additionally, we present the raw reward scores from each method with 1-sigma error bars in Table S2 (SD1.5) and S3 (SDXL). In the SD1.5 results, our method significantly improves the reward scores of Diffusion-DPO, achieving the highest scores on most metrics. For example, on the Pick-a-Pic v2 test set, PickScore and ImageReward increase by 0.57 and 0.33, respectively. In SDXL results, while there are exceptions in the CLIP score (which was not trained on human preference prediction tasks) and the Aesthetic Score metric (which does not consider the text prompt), our method records the best performance in all other metrics.

To demonstrate the generalizability of our method, we also present results on Stable Diffusion 3, which modernizes diffusion models by introducing flow matching and a multimodal transformer. Due to limited computational resources, we reduce the batch size from 2048 to 128 and train the model for

200 iterations with a learning rate of 3e-7. The reference update period τ is set to 32, the monitoring threshold to 0.03, and all other hyperparameters remain the same as those used for SD1.5 and SDXL.

Table S4 compares Diffusion-DPO with our method under the same training configuration, showing that our method again outperforms the standard Diffusion-DPO. This result demonstrates that the effectiveness of our method is independent of the structural components of diffusion models.

C   Further ablation study

C.1   Analysis on Training Dynamics of τ
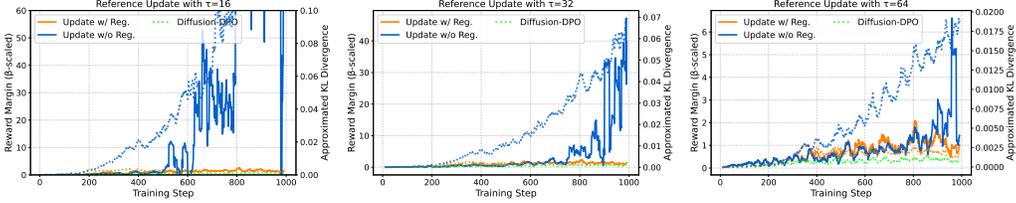
Figure S1: Training dynamics of reference update method, with $\tau = \{16, 32, 64\}$ (SD1.5). (solid lines) Implicit reward margin under the reference update strategy, with and without our regularization. (dotted lines) Approximated KL divergence between the training model and the pre-trained model (Diffusion-DPO), and between the reference model and the pre-trained model (ours).

Figure S1 presents the training dynamics for the reference update period $\tau \in \{16, 32, 64\}$. As discussed in Section 3, we observe a consistent increase in both the reference model's divergence and the implicit reward margin. For smaller values of τ, the reference model stays closer to the training model, resulting in more aggressive scaling of the prediction error. Increasing τ can improve training stability by slowing reference updates, but it also reduces alignment performance due to limited exploration, as shown in Figure 4. In contrast, our reference regularization method effectively prevents divergence and error explosion, ensuring controlled exploration for the training model.

To gain deeper theoretical insight into the model divergence issue, we compare the DPO gradient behavior between diffusion models and autoregressive language models. Suppose we sample a pairwise data $\boldsymbol{x}_0^w, \boldsymbol{x}_0^l$ and timestep t for the Diffusion-DPO loss defined in Eq. 6. Then, the gradient of the loss is computed as:

$$\nabla_\theta \mathcal{L}(\theta) = 2\beta\sigma(r_t(\boldsymbol{x}_0^l) - r_t(\boldsymbol{x}_0^w)) \cdot [(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t) - \boldsymbol{\epsilon})\nabla_\theta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t) - (\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t) - \boldsymbol{\epsilon})\nabla_\theta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t)], \tag{S1}$$

where β absorbs constant terms for simplicity. Empirically, we find that frequent updates to the reference model cause the gradient magnitude to become dominated by the model error term $||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||_2^2$.

In contrast, autoregressive language models are more robust to such error scaling, as reported in TR-DPO [8]. We hypothesize that this robustness stems from the fundamental modeling differences between language models and diffusion models. Specifically, under the Lipschitz condition, the gradient of the DPO loss in language models is bounded.

Theorem 1. *Suppose $f = f_\theta(x) \in \mathbb{R}^V$ denote the output logits for a vocabulary of size V. Also, assume that f_θ is K-Lipschitz with respect to θ. Let y^w (preferred) and y^l (dispreferred) be two responses for x, with the same length T. Then, $\|\nabla_\theta\mathcal{L}_{DPO}\| \le 2\sqrt{2}\beta \cdot TK$.*

Proof. We firstly show the upper bound of the logit. Let $y \in \{1, \ldots, V\}$ be a token. The softmax distribution for y is:

$$\pi_\theta(y \mid x) = \frac{\exp(f_\theta(y))}{\sum_{j=1}^V \exp(f_\theta(j))}, \tag{S2}$$

and the log-likelihood is computed as:

$$\log \pi_\theta(y \mid x) = f_\theta(y) - \log\left(\sum_{j=1}^V \exp(f_\theta(j))\right). \tag{S3}$$

15

The gradient with respect to the logits is:

$$\nabla_f \log \pi_\theta(y \mid x) = \mathbf{e}_y - \pi_\theta(\cdot \mid x), \tag{S4}$$

where $\mathbf{e}_y$ is the one-hot vector.

Then,

$$\|\nabla_f \log \pi_\theta(y \mid x)\|_2^2 = \sum_{i=1}^{V} \left(\mathbf{e}_y(i) - \pi_\theta(i \mid x) \right)^2$$

$$= (1 - 2\pi_\theta(y \mid x)) + \sum_{i=1}^{V} \pi_\theta(i \mid x)^2 \leq 1 + \sum_{i=1}^{V} \pi_\theta(i \mid x) = 2, \tag{S5}$$

where we use $\pi(i) \in [0, 1]$. Hence, we have

$$\|\nabla_f \log \pi_\theta(y \mid x)\|_2^2 \leq 2. \tag{S6}$$

Now consider the log probability for two responses y^w and y^l:

$$\log \pi_\theta(y^i \mid x) = \sum_{t=1}^{T} \log \pi_\theta(y_t^i \mid x, y_{<t}^i), i \in \{w, l\} \tag{S7}$$

Since Equation S6 holds for all y_t, and f_θ is K-Lipschitz with respect to θ, it follows that:

$$\|\nabla_\theta \log \pi_\theta(y^i \mid x)\| \leq \sqrt{2}TK. \tag{S8}$$

Now, consider the gradient of the DPO loss:

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta \cdot \sigma(-\beta z) \cdot \left(\nabla_\theta \log \pi_\theta(y^w \mid x) - \nabla_\theta \log \pi_\theta(y^l \mid x) \right), \tag{S9}$$

where $z := \log \pi_\theta(y^w \mid x) - \log \pi_\theta(y^l \mid x) - (\log \pi_{\text{ref}}(y^w \mid x) - \log \pi_{\text{ref}}(y^l \mid x))$.

From the Equation S9, the DPO gradient is bounded in norm by:

$$\|\nabla_\theta \mathcal{L}_{\text{DPO}}\| \leq 2\beta \cdot \sigma(-\beta z) \cdot \sqrt{2}TK.$$

Since $\sigma(\cdot) \leq 1$, we arrive at the final upper bound. $\qquad\square$

In Diffusion-DPO, even under the Lipschitz assumption for the model, the gradient can diverge due to the unbounded nature of the noise prediction error. This again highlights the need for our reference model regularization, which controls model divergence while still allowing effective exploration.

C.2 Analysis on Monitoring Threshold

We conduct experiments across varying the monitoring threshold δ values, which defines the safe region for the reference model update strategy. Table S5 presents results for three different δ, based on Diffusion-DPO on SD1.5, evaluated on the Pick-a-Pic v2 test prompts. The results show that $\delta = 0.005$ achieves the best performance, while either an excessively small or large δ leads to performance degradation. A small δ does not allow the model to explore enough, while a large δ fails to provide adequate regularization for the reference model. Although we experimentally choose δ based on this trade-off, the optimal value of δ may vary across models. In future work, we hope to explore methods for the optimal selection of δ.

C.3 Analysis on Timestep-aware Training Strategy

To further investigate the impact of timestep weighting strategies, we compare our method $\lambda(t)$ against other alternatives, including the square root of the inverse signal-to-noise value (SNR), $1/\sqrt{SNR(t)}$, and the normalized version, norm$(1/\sqrt{SNR(t)})$.

The unnormalized weighting $1/\sqrt{SNR(t)}$ assigns large weights to highly noisy timesteps. Compared to our normalized version, its large value at early timesteps can impose overly excessive regularization,

Model	PickScore	HPSv2	CLIP	Aesthetic	IR	Average
δ=0.001	86.40	75.96	64.48	75.48	71.64	74.79
δ=0.005	**89.96**	**83.84**	**64.56**	**78.04**	**77.76**	**78.83**
δ=0.025	80.40	72.48	60.24	71.04	73.56	71.54

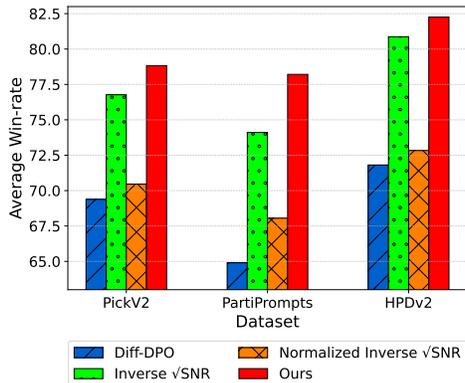Table S5: Ablation study on different monitoring thresholds. Win rates are reported against SD1.5.

Figure S2: Ablation study on different timestep weighting schemes. Win rates are reported against SD1.5.

yielding suboptimal performance. However, with the normalized weighting, we observe a sharp performance drop, as it assigns small regularization weights for later timesteps. The additive offset of value 1 in our method guarantees sufficient regularization for every timestep, while assigning higher weights to early steps. Empirical results in Figure S2 validate the effectiveness of our design, achieving the highest average win rate across datasets.

D Additional Qualitative Results

We provide more qualitative results in Figures S3 – S8. In Figures S3 – S5, we show images generated by SD1.5 using evaluation prompts (Pick-a-Pic v2, PartiPrompts, HPDv2). In particular, we use prompts that involve multiple objects or complex compositional relationships. For example, the prompt *a real flamingo...* describes a complex relationship between objects: the *flamingo* is reading a large open book, and *a stack of books* is placed next to it. While existing methods fail to accurately depict this relationship, our method captures the intended scene described in the prompt. Finally, in Figures S6 – S8, we present results for SDXL using prompts from the Pick-a-Pic v2 test set.

Gothic cathedral in a stormy night

75 years old, grandfather, bodybuilded too much, crush an apple in his hand

A sail boat entering a majestic fjord landscape in winter

Portrait of general with obscure hat

pink eagle

A smooth purple octopus sitting on a rock in the middle of the sea, waves crashing, golden hour, sun reflections, high quality 3d render

   (a) SD1.5 [26]   (b) Diff-DPO [31]   (c) Diff-KTO [16]   (d) DSPO [42]   (e) Ours

Figure S3: Qualitative comparisons on Pick-a-pic test set prompts.

A photograph of a portrait of a statue of a pharaoh wearing steampunk glasses,
white t-shirt and leather jacket.

a real flamingo reading a large open book. a big stack of books is piled up next to it. dslr photograph.

A cozy living room with a painting of a corgi on the wall above a couch
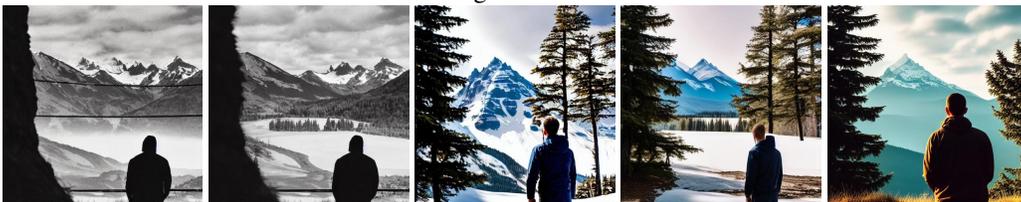and a round coffee table in front of a couch and a vase of flowers on a coffee table

a tree reflected in the hood of a blue car

a black dog jumping up to hug a woman wearing a red sweater
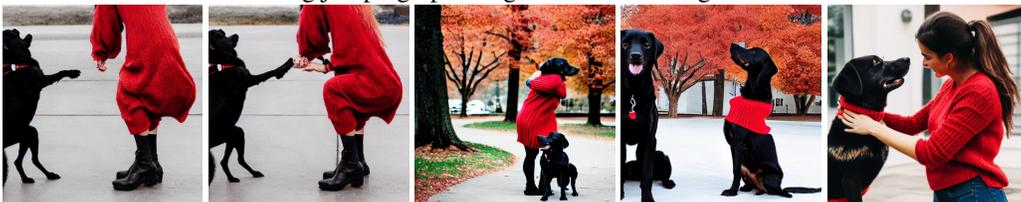
a man looking at a distant mountain

(a) SD1.5 [26]   (b) Diff-DPO [31]   (c) Diff-KTO [16]   (d) DSPO [42]   (e) Ours

Figure S4: Qualitative comparisons on texts from PartiPrompts.

A bear in an astronaut suit sits on a rock on Mars surrounded by flowers under a starry sky.

A portrait of a stylized business cat in sharp focus with a medium shot perspective, resembling boxart.

A white bichon frise puppy dog riding a black motorcycle in Hollywood at sundown with palm trees in the background.

A candy house on the ocean in a fantasy setting.

A monkey wearing a jacket.

An anime-style advertisement featuring a pizza and an explosion.

| (a) SD1.5 [26] | (b) Diff-DPO [31] | (c) Diff-KTO [16] | (d) DSPO [42] | (e) Ours |

Figure S5: Qualitative comparisons on HPDv2 test set prompts.

20

A woman with a pearl earring, blue eyes,in the style of blue and khaki,
smiling and happy, meticulous, solapunk, li - core
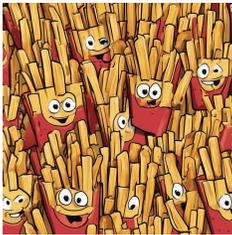

(a) SDXL

(b) Diff-DPO
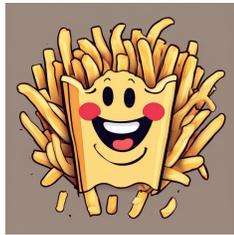
(c) MAPO

(d) DSPO

(e) Ours

smily french fries

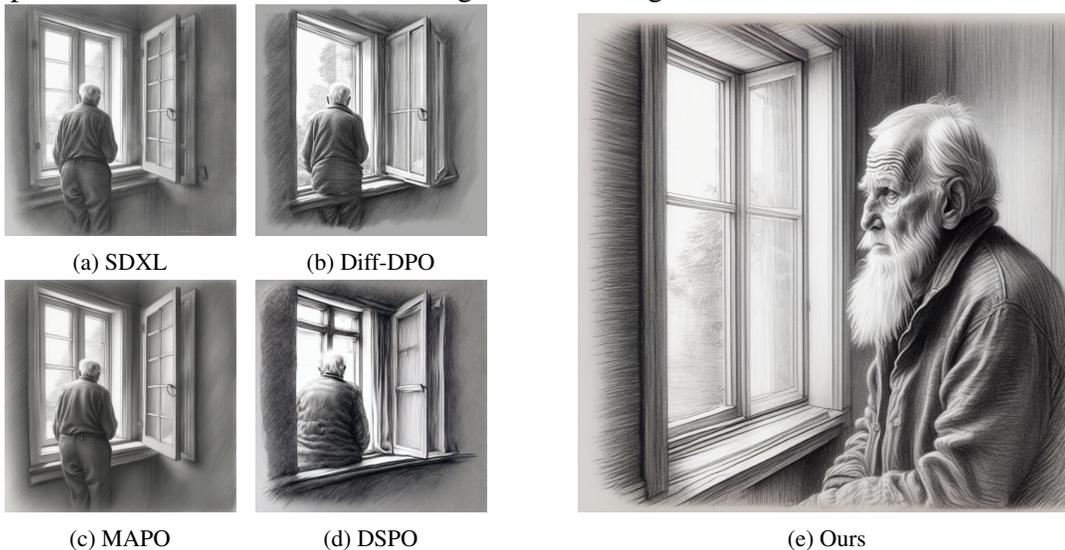(a) SDXL

(b) Diff-DPO

(c) MAPO

(d) DSPO

(e) Ours

Figure S6: Qualitative comparisons with the SDXL model.

pencil sketch of An old man looking outside through the first floor window at home

(a) SDXL

(b) Diff-DPO

(c) MAPO

(d) DSPO

(e) Ours

An evil villain holding a mini Earth, pixelart

(a) SDXL

(b) Diff-DPO

(c) MAPO

(d) DSPO

(e) Ours

Figure S7: Qualitative comparisons with the SDXL model.

a cute tiny bird wondering around water
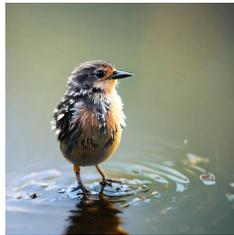

(a) SDXL

(b) Diff-DPO

(c) MAPO

(d) DSPO

(e) Ours

A cute hedgehog holding flowers

(a) SDXL

(b) Diff-DPO

(c) MAPO

(d) DSPO

(e) Ours

Figure S8: Qualitative comparisons with the SDXL model.