ITERATIVE DPO WITH AN IMPROVEMENT MODEL FOR FINE-TUNING DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) has been proven as an effective solution in aligning generative models with human preferences. However, as shown in recent works, DPO could suffer from constraints from the offline preference dataset. This paper introduces a novel improvement approach for online iterative optimization of the diffusion models without introducing extra annotation of the online data. We propose to learn a preference improvement model to extract the implicit preference from the preference dataset. The learned improvement model is then used to generate winning images from the images generated by the current diffusion model. We can construct new pairs of preference data by using images generated by the current diffusion model as losing images, and its corresponding improved images as winning images. The diffusion model can therefore be optimized via iteratively applying online preference datasets. This method enables online improvement beyond offline DPO training without requiring additional human labeling or risking overfitting the reward model. Results demonstrate improvements in preference alignment with higher diversity compared with other fine-tuning methods. Our work bridges the gap between offline preference learning and online improvement, offering a promising direction for enhancing diffusion models in image generation tasks with limited preference data.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful paradigm for aligning generative models with human preferences, showing remarkable success in both language models (Ouyang et al., 2022) and diffusion models for image generation (Black et al., 2024; Fan et al., 2023). Traditional RLHF approaches, often implemented using Proximal Policy Optimization (PPO) (Schulman et al., 2017), have faced significant challenges, including training instability, overfitting the reward model, and high computational costs.

Direct Preference Optimization (DPO) was introduced as an alternative that simplifies the training process by directly optimizing the model based on preference data (Rafailov et al., 2024; Wallace et al., 2024). While DPO offers more efficient training and improved stability, it is inherently limited to offline examples, potentially constraining its performance: Offline DPO, which relies solely on a fixed dataset of preferences, often exhibits suboptimal performance due to the lack of on-policy data (Xu et al., 2024; Tajwar et al., 2024). Recent studies have investigated online iterative DPO methods, such as online annotated preference data from LLMs (Rosset et al., 2024), reward models (Xu et al., 2024), or human feedbacks (Xiong et al., 2024). However, online labeling can be prohibitively expensive, and runs the risk of reward hacking Zhang et al. (2024).

On the other hand, recent research has explored the concept of self-improvement in generative models especially LLMs, including self-rewarding models (Yuan et al., 2024b) and self-improving language models (Choi et al., 2024). These approaches offer a promising direction to achieve self-improvement without extra labeled data, which could be a natural remedy for data constraints in DPO. However, such self-improvement capability remains under-explored in text-to-image diffusion models. Recently, Yuan et al. (2024a) proposed a self-play approach for diffusion models. However, their optimization target is equivalent to aligning with the winning data distribution. The performance is thus still upper-bounded by the offline dataset. The self-improvement approaches have been widely studied for LLMs since the base LLM can be re-purposed in a natural way to

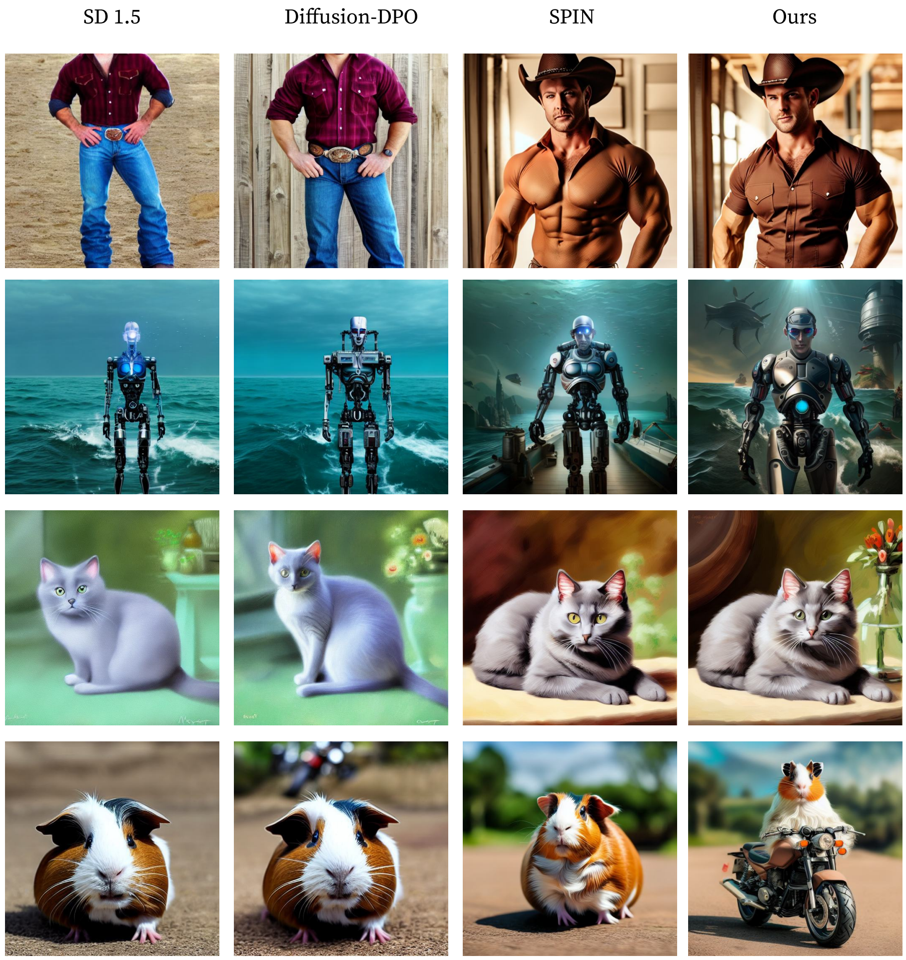

Figure 1: Visualization of sampled images from the baseline and fine-tuned diffusion models. Prompts (from top to bottom): *1. Beefy cowboy, tucked in shirt; 2. A cyborg on the ocean; 3. Cute grey cat, digital oil painting by Monet; 4. A guinea pig riding a motorcycle.* The samples on the same row are sampled from the same seed.

provide a self-improvement signal. However, it is not straightforward to apply the self-improvement approach to a mixed-modality T2I model.

In this paper, we aim to answer the following research question: *With a fixed offline preference dataset, can we achieve online improvement for diffusion models without extra annotations?* To solve the challenges of limited offline datasets, we introduce a novel multi-task formulation to train a model to learn the generic *improvement directions* from the preference dataset.

Specifically, we learn an improvement diffusion model to generate a winning image given a losing image and a prompt. The learned improvement diffusion model can then be applied to generate online preference pairs depending on the output of the current diffusion model, enabling continuous improvement in iterative DPO training. This approach offers several key advantages:

- Leveraging the advantages of DPO while mitigating its limitations in offline settings;
- Enabling online learning without the need for extra annotations;
- Providing a mechanism for continuous improvement for diffusion models with fixed preference datasets.

Experimental results demonstrate that our iterative training for diffusion models with the learned improvement model leads to improvements over DPO baselines including Diffusion-DPO (Wallace et al., 2024) and SPIN (Yuan et al., 2024a). Specifically, we observe consistently higher scores on PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), and Aesthetic score (Schuhmann et al., 2022), indicating improved image quality and better alignment with human preferences. Moreover, we observe improved Vendi scores (Friedman & Dieng, 2023), suggesting that our method enhances quality without sacrificing diversity in the generated images, even improving the diversity compared with SPIN and Diffusion DPO.

We summarize our contributions as follows: **(1)** We introduce a novel method that learns an improvement direction from an offline preference dataset. **(2)** Using the improvement model to generate online training data, we address the critical challenge of learning from limited offline preference data, allowing training beyond the initial dataset. **(3)** Our experiments demonstrate improvements in preference alignment with better diversity compared with baseline fine-tuning methods.

2 Preliminaries

2.1 Diffusion models

Let $x_0 \in \mathbb{R}^n$ be a data sample, and q_0 be the data distribution, i.e., $x_0 \sim q_0(x_0)$. Diffusion models approximate q_0 with $p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T}$, where $p_\theta(x_{0:T}) = p_T(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$ is a Markov chain with the following dynamics:

$$p(x_T) = \mathcal{N}(0, I), \qquad p_\theta(x_{t-1}|x_t) = \mathcal{N}\big(\mu_\theta(x_t, t), \Sigma_t\big). \tag{1}$$

The *forward* or *diffusion process* $q(x_{1:T}|x_0)$ is a Markov chain that adds Gaussian noise to the data according to a variance schedule $\beta_1, \dots, \beta_T$:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \qquad q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I). \tag{2}$$

Let $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\Big)$. The training of diffusion models is performed by optimizing a variational bound on the negative log-likelihood $\mathbb{E}_q[-\log p_\theta(x_0)]$, which is equivalent to optimizing:

$$\mathbb{E}_{x_t,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big], \tag{3}$$

where $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, $x_0 \sim q_0(x_0)$, $\epsilon \sim \mathcal{N}(0, I)$.

2.2 DPO and Diffusion-DPO

DPO. Assume that we have access to a general preference dataset $\mathcal{D} = \{c, x_w, x_l\}$ where c is the text prompt, x_l is the losing response and x_w is the winning response. Given a conditional generative model $p_\theta(x|c)$ and a reference model $p_{\text{ref}}(x|c)$, we can align the model with the preference using the DPO loss (Rafailov et al., 2024):

$$-\mathbb{E}_{c,x_w,x_l\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{p_\theta(x_w|c)}{p_{\text{ref}}(x_w|c)} - \frac{p_\theta(x_l|c)}{p_{\text{ref}}(x_l|c)}\right)\right]. \tag{4}$$

Diffusion-DPO. For diffusion models, since $p_\theta(x|c)$ is not generally tractable, Wallace et al. (2024) proposes an approximation by finding an upper-bound of the original DPO objective:

$$\begin{aligned} -\,\mathbb{E}_{c,x_l,x_w\sim\mathcal{D},t}\big[\log\sigma\big(-\beta T\big(&\|\epsilon^w - \epsilon_\theta(x_t^w, t, c)\|_2^2 + \|\epsilon^l - \epsilon_{\text{ref}}(x_t^l, t, c)\|_2^2 \\ &-\|\epsilon^w - \epsilon_{\text{ref}}(x_t^w, t, c)\|_2^2 - \|\epsilon^l - \epsilon_\theta(x_t^l, t, c)\|_2^2\big)\big)\big]. \end{aligned} \tag{5}$$

Drawbacks of DPO. The interpretation of DPO training is straightforward: it aims to pull up the probability of the winning response and pull down the losing one. During training, all the responses

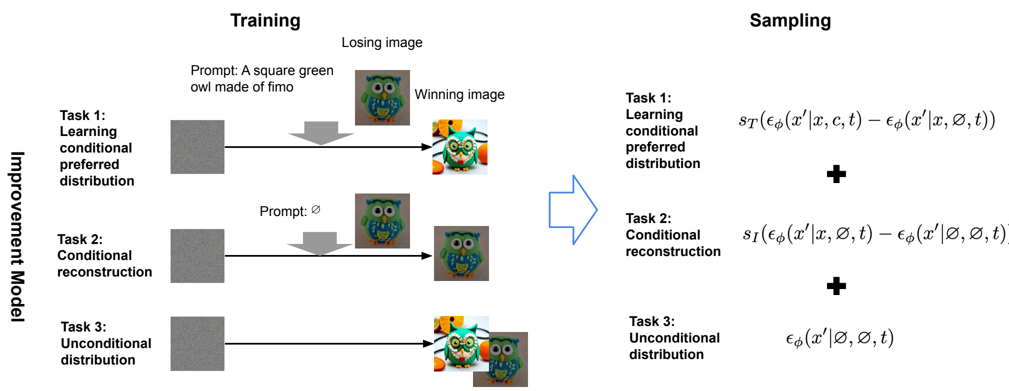

Figure 2: The overview of the training and sampling pipeline of the proposed improvement model. The left side diagram demonstrates the three tasks used for training the improvement model. The three tasks are co-trained together to make the model both learn the generic capability of improving image toward the preferred distribution represented in the offline dataset and the retain generalized image generation capability without losing diversity. The right side diagram shows the sampling strategy of the improvement model where the diffusion score of the improvement images are combined from the three tasks learned before.

are from the preference dataset, and the actual output of the model is never checked. The quality of the learned policy in DPO can be compromised by a biased distribution towards unseen responses. This bias arises when the offline preference dataset lacks diversity or is not readily accessible. This phenomenon has been observed in Xu et al. (2024).

Given the downsides of using an offline dataset in DPO, the recent work (Xiong et al., 2024) has explored augmenting training datasets through online training, incorporating online samples that enhance performance in preference learning (Tajwar et al., 2024). However, annotating these samples requires extra effort, and optimizing with a reward model could risk reward over-optimization and hacking. This paper explores whether DPO-based training without extra annotations can be further improved.

3 Method

We consider a scenario where *only a fixed offline preference dataset is available, without access to additional annotation sources.* We propose to build an *improvement model* from the preference dataset that generates improved images (for a given prompt) when given images generated by the current diffusion model as input. The input (image condition) and output (improved image) of the improvement model therefore correspond to a losing/winning preference pair that can be used for iterative DPO training without extra annotation.

The intuition behind iterative DPO training with an improvement model is straightforward. Recall that DPO training pulls up the probability of the winning response, which is the output of the improvement model in our case. It also pulls down the probability of the losing response, which is the output of the current diffusion model. Thus, if we can successfully train an improvement model, we can continuously improve the current diffusion model using the improvement model with iterative DPO training till convergence.

We introduce how to train such an improvement model in Section 3.1, the sampling from the improvement model in Section 3.2, and iterative training of the improvement model in Section 3.3.

3.1 Training an improvement diffusion model

The objective of the improvement diffusion model ϕ is to predict a conditional distribution over improved images, $p^\dagger_\phi(x_w|x_l, c)$ i.e. It learns to generate a winning image x_w given a text prompt c and a losing image x_l. This can be accomplished by the ability of diffusion models to condition

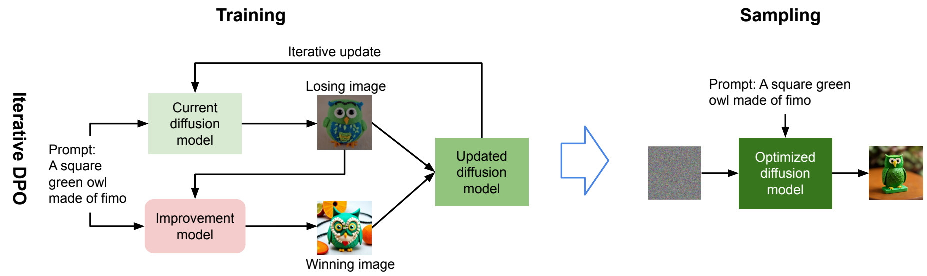

Figure 3: The overview of the training pipeline of iterative DPO with the improvement model. The diagram on the left side demonstrates the iterative DPO algorithm with the improvement model. The current diffusion model generates a losing image and passes it to the improvement model to improve to a winning image. Both images are paired as the preference dataset to fine-tune the diffusion model. The diffusion model is optimized iteratively until it converges. The optimized diffusion model is then deployed for inference as shown in the diagram on the right side.

the denoising trajectory on arbitrary additional signals. For example, InstructPix2Pix Brooks et al. (2023) is an image-editing model that takes an original image and an editing instruction (in text) as its input conditions by encoding the image with additional channels in the first convolutional layer of the UNet.

Multi-task training. In order to train a model with a generic improvement capability to map any given image to higher quality ones without sacrificing diversity, we design a multi-task training algorithm that takes different text-image condition combinations as input (See the left side of Figure 2). The model is trained on a mixture of the following tasks:

1. Learning the conditional winning distribution: Given both text c and a losing image condition x_l, we learn the target distribution of x_w:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),x_w,x_l,c,t}[\|\epsilon - \epsilon_\phi(x_t|x_l, c, t)\|^2], \tag{6}$$

where $x_t = \sqrt{\bar{\alpha}_t}x_w + \sqrt{1-\bar{\alpha}_t}\epsilon$.

2. Reconstruction: Given only the image condition $x \in \{x_w, x_l\}$, the model is encouraged to reconstruct x:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),x,c,t}[\|\epsilon - \epsilon_\phi(x'_t|x, \varnothing, t)\|^2], \tag{7}$$

$x'_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon$.

3. Unconditional distribution: Without conditioning input from either x_w or x_l, the model generates images drawn from a distribution encompassing both winning and losing images:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),x,c,t}[\|\epsilon - \epsilon_\phi(x''_t|\varnothing, \varnothing, t)\|^2], \tag{8}$$

$x''_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon$, $x, c \sim \mathcal{D}$.

Interpretation. The design of the improvement model is inspired by InstructPix2Pix (Brooks et al., 2023). However, their setting cannot be directly applied here because prompts in our datasets lack specific improvement/editing instructions. Furthermore, applying their objective function to our setting could cause the sampled distribution to collapse. Consider a single-task training that solely focuses on learning the conditional distribution of $p(x_w|x_l, c)$. Without an objective to force the model to utilize x_l, it might learn to ignore x_l and instead learn an unconditional distribution $p(x_w|c)$. This is likely to happen when we are fine-tuning from a pre-trained diffusion model where image condition weights are initialized to 0. This necessitates the additional reconstruction task which aims to capture the information from the image condition. Further, the difference between $\epsilon_\phi(\cdot|x_l, c, t)$ and $\epsilon_\phi(\cdot|x_l, \varnothing, t)$ provides a natural "improvement direction" for the main improvement task. Moreover, learning the unconditional score is crucial for achieving both high conditional generation accuracy and sample diversity (Ho & Salimans, 2022).

Algorithm 1 Iterative DPO training with improvement model.

Input: Improvement model $p^{\dagger}_{\phi}$, prompt set $\mathcal{D}_c$, initialized model p_θ, number of iterations T_{iter}, number of samples n, training batch size b, text guidance weight s_T, image guidance weight s_I, number of training steps per iteration T_{train}

for $t_{\text{iter}} \in [1, T_{\text{iter}}]$ **do**

  Randomly sample n images from p_θ conditioned on $\mathcal{D}_c$, and construct $\mathcal{D}_l$

  Randomly sample n images from $p^{\dagger}_{\phi}$ conditioned on $\mathcal{D}_c$ and $\mathcal{D}_l$. With guidance weights s_T and s_I, construct $\mathcal{D}_w$

  for $t_{\text{train}} \in [1, T_{\text{train}}]$ **do**

    Compute an estimation of gradient using Equation (10) with batch size b, and update θ

  end for

end for

Output: Fine-tuned model p_θ

3.2 Sampling from an improvement diffusion model

Double classifier-free guidances. For conditional sampling from the improvement model, we adapt the double classifier-free guidance technique introduced in InstructPix2Pix (Brooks et al., 2023) and design the sampling algorithm as:

$$\begin{aligned}\bar{\epsilon}_\phi(x'|x,c,t) = \epsilon_\phi(x'|\varnothing,\varnothing,t) + s_I(\epsilon_\phi(x'|x,\varnothing,t) - \epsilon_\phi(x'|\varnothing,\varnothing,t))\\ + s_T(\epsilon_\phi(x'|x,c,t) - \epsilon_\phi(x'|x,\varnothing,t)),\end{aligned} \tag{9}$$

where x' is the output, s_T is the text guidance weight and s_I is the image guidance weight. The first term $\epsilon_\phi(x'|\varnothing,\varnothing,t)$ is to sample without any condition as the standard diffusion model. The second term $s_I(\epsilon_\phi(x'|x,\varnothing,t) - \epsilon_\phi(x'|\varnothing,\varnothing,t))$ is to sample from the image only condition to reconstruct the input images. It helps to regularize the divergence of the output from the input images. The last term $s_T(\epsilon_\phi(x'|x,c,t) - \epsilon_\phi(x'|x,\varnothing,t))$ is to sample from both the image and text condition to improve from losing images to winning ones. The overall sampling algorithm of the improvement model is illustrated on the right side of Figure 2.

Roles of the guidance weights. To further refine the sampling process, we utilize two guidance weights: text guidance weight s_T and image guidance weight s_I. The text guidance weight s_T determines the strength of the improvement direction - a larger s_T value leads to more significant alignment with text prompt. Meanwhile, the image guidance weight s_I controls how closely the output image resembles the input condition image, i.e., increasing s_I enforces greater similarity of input and output images.

3.3 Iterative DPO with an improvement diffusion model

Objective function for iterative DPO. Building on the improvement diffusion model $p^{\dagger}_{\phi}(x_w|x_l, c)$, we can sample pairs of preference images x_w and x_l, where x_l are generated from the current diffusion model as the losing image and x_w output from the improvement model as the winning image. These online sampled pairs provide valuable data for optimizing the diffusion model using the DPO objective function below:

$$\begin{aligned}-\mathbb{E}_{x\sim\mathcal{D},x_l\sim p_\theta(\cdot|x),x_w\sim p^{\dagger}(\cdot|x_l,c),t}\Big[\log\sigma\Big(-\beta\big(\|\epsilon^w - \epsilon_\theta(x^w_t,t)\|^2_2 + \|\epsilon^l - \epsilon_{\text{ref}}(x^l_t,t)\|^2_2\\ -\|\epsilon^w - \epsilon_{\text{ref}}(x^w_t,t)\|^2_2 - \|\epsilon^l - \epsilon_\theta(x^l_t,t)\|^2_2\big)\Big)\Big],\end{aligned} \tag{10}$$

where the losing images x_l are from the current output of the model, and the winning images x_w are from the improvement model conditioning on x_l and c, constructing a new preference dataset. After one iteration of optimization, we can regenerate new preference data from the current diffusion model and the improvement model, and perform DPO training iteratively (details in Algorithm 1). An illustration of the iterative training pipeline is in Figure 3.

Comparison with SPIN. Here we compare our method with SPIN, a self-play method that can be applied to diffusion models (Yuan et al., 2024a). The iterative objective function of SPIN aims to move the model's output distribution closer to a target distribution. However, a key limitation of this approach is its strong reliance on the quality of the preferred responses. SPIN uses these preferred responses, along with the prompt set, to construct an SFT dataset, while discarding the losing responses. This strategy assumes that the preferred distribution is near-optimal. If this assumption doesn't hold, the model risks falling into a suboptimal area. Furthermore, the output distribution is still constrained by the available preferred responses in the training dataset, potentially limiting the outputs' diversity. In contrast, we learn the improvement direction from the preference dataset while retaining information from the losing distribution. By iteratively applying this learned improvement direction, we can optimize the model towards better performances. Thus, our method could surpass SPIN models, achieving both higher alignment and higher diversity.

4 RELATED WORK

Variants of DPO. Direct preference optimization (DPO) (Rafailov et al., 2024) is developed to optimize the generation policy with the offline preference dataset. It eliminates the dependency on the explicit reward model. However, the optimal solution derived from the Bradley-Terry (BT) model makes DPO prone to weakening the regularization and overfitting to the offline training dataset. Azar et al. (2024) propose the IPO by introducing the identity function into the generic ΨPO framework and derive an efficient optimization process and achieve improved performance than DPO. Meng et al. (2024) argue for the effectiveness of the reference model regularization in DPO. They therefore propose the simple preference optimization (SimPO) method that bypasses the reference model regularization and introduces a reward margin to the optimization objective to better approximate the noisy preference dataset. Their approach also shows improved performance over DPO. Hong et al. (2024) also argue about impediments in optimizing the reference model under distributional discrepancy and propose the margin-aware preference optimization (MaPO) method to replace KL regularization on the reference model with an amplification factor defined by the trained policy's likelihood estimation. These DPO variants explore the challenges of distribution discrepancy between the reference model and model under optimization. They optimize with the offline sampled dataset which is verified to be less efficient than on-policy sampling (Tajwar et al., 2024).

Iterative DPO and self-play methods. To understand and address the limitations of offline training associated with DPO, recent works have investigated the performance gap between online and offline training methods (Tajwar et al., 2024; Tang et al., 2024). Their findings indicate that online training can lead to better generation, and is beneficial when high-reward responses have a low likelihood under the pretrained model. Accordingly, several works have proposed iterative DPO methods that train DPO using online samples generated by the improved policy (Guo et al., 2024; Xu et al., 2023b; Xiong et al., 2023). However, they require a reward model to label the online samples. To eliminate the dependence on reward models, researchers have developed self-play or self-improvement methods. For example, Yuan et al. (2024b) use the language model itself to provide the reward signal, and Chen et al. (2024) treat self-generated responses as losing to human demonstrations for iterative improvement. More recently, Choi et al. (2024); Wu et al. (2024) reformulate these ideas under the constant-sum two-player game framework (Munos et al., 2023; Swamy et al., 2024), and propose algorithms to find the approximate Nash equilibrium. Our work proposes a self-improvement method for text-to-image diffusion models, which has been under-explored.

Aligning diffusion models with human preferences. Inspired by the success of RLHF and DPO in fine-tuning language models, recent works have explored applications in aligning diffusion models with human preferences. RLHF-based methods maximize a reward score given by a separately trained reward model (Radford et al., 2021; Lee et al., 2023; Xu et al., 2023a; Wu et al., 2023; Kirstain et al., 2023). For differentiable rewards, reward maximization can be done by backpropagating the reward function gradient through the denoising process (Clark et al., 2024; Prabhudesai et al., 2023). For black-box reward functions, DDPO (Black et al., 2024) and DPOK (Fan et al., 2023) propose PPO-based RL fine-tuning. PRDP (Deng et al., 2024) further improves training stability on large-scale datasets by converting reward maximization to an equivalent reward difference prediction objective. However, RLHF-based methods generally have a complicated pipeline involving reward model training, and are prone to reward hacking. These issues can be partially mitigated by DPO-based methods, such as Diffusion-DPO (Wallace et al., 2024) and SPIN-Diffusion (Yuan et al., 2024a), which directly fine-tune the diffusion model from offline preference datasets without

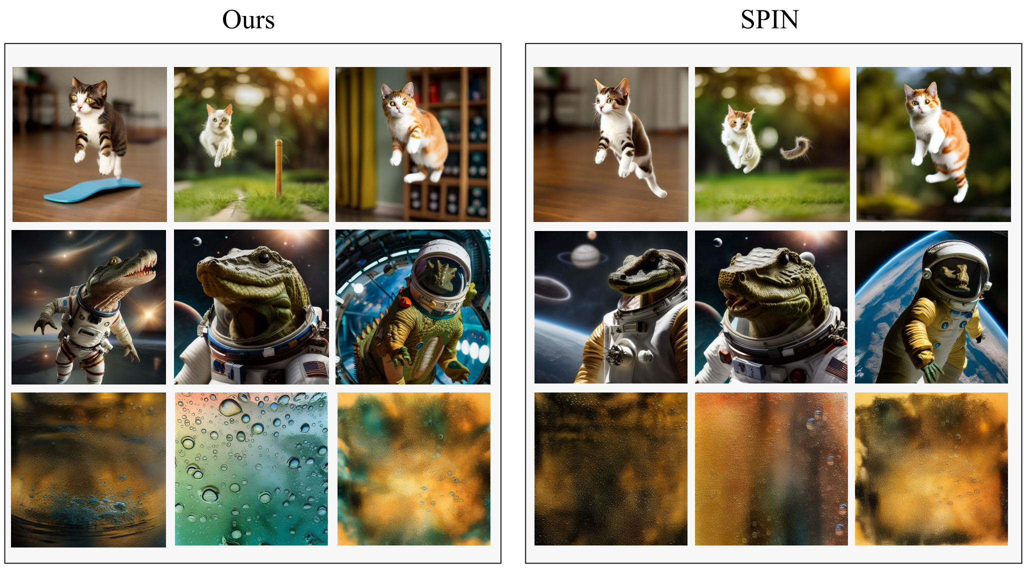

Figure 4: Prompts (from top to bottom): *1. A cat jumping for a toy. 2. A crocodile in a space suit. 3. An abstract print of water and oil mixing, bubbles, textual.* Samples on the same row are from the same prompt. For each prompt, the examples are from the same set of random seeds for both SPIN and our model. Our model generates more diverse outputs than SPIN in terms of output backgrounds, colors, etc.

requiring reward models. However, their performance can be limited due to a lack of online training. Our approach combines the benefits of online training from RLHF and the simplicity from DPO.

5 Experiments

5.1 Experimental setup

5.1.1 Training

Model and Dataset. We use the Pick-a-pic (Kirstain et al., 2023) training dataset as the offline preference dataset, following Diffusion-DPO (Wallace et al., 2024). For the improvement model, we add 4 channels to the first convolutional layer of the UNet, and initialize the weights from Stable Diffusion 1.5 (Rombach et al., 2022) following Brooks et al. (2023). For iterative DPO training, we fine-tune the model initialized from the third iteration in SPIN (Yuan et al., 2024a).

Hyperparameters. For the improvement model training, we use AdamW with a learning rate 10^{-4}, and train up to 200K steps with batch size 2048, and sample from it with $s_T = 3.5, s_I = 3.0$. For iterative DPO training, we train for 3 iterations, and for each iteration, we first generate 38400 pairs of preference data, and train for 5k steps for each iteration with the batch size 2048 and learning rate 10^{-4}, $\beta = 2000$, with SD 1.5 as the reference model.

5.1.2 Evaluation

Prompt sets. We use two prompt sets for evaluation: We randomly sample 500 unique prompts from the training set to reflect the model's performance on the training set. We also use the 500 unique prompts from the test dataset in Pick-a-pic v2 as the test set. We sample 64 random images from each prompt.

Metrics. We evaluate our method against DPO and other baselines using a comprehensive set of metrics. For quality assessment, we employ PickScore (Kirstain et al., 2023), Human Preference

Table 1: Evaluation of Pickscore, HPSv2, Aesthetic score and Vendi score. Our model achieves improved reward scores without sacrificing diversity compared with SPIN and Diffusion-DPO.

Score	Method	Training subset	Test set
Pickscore (↑)	SD 1.5	20.46	20.74
	Diffusion-DPO	20.80	21.05
	SPIN	<u>21.15</u>	<u>21.41</u>
	Iterative (Ours)	**21.20**	**21.46**
HPSv2 (↑)	SD 1.5	26.65	26.90
	Diffusion-DPO	26.96	27.19
	SPIN	<u>27.39</u>	<u>27.57</u>
	Iterative (Ours)	**27.40**	**27.59**
Aesthetic (↑)	SD 1.5	5.48	5.42
	Diffusion-DPO	5.55	5.49
	SPIN	<u>5.92</u>	<u>5.86</u>
	Iterative (Ours)	**5.94**	**5.88**
Vendi score (↑)	SD 1.5	**2.61**	**2.64**
	Diffusion-DPO	2.44	2.47
	SPIN	2.43	2.48
	Iterative (Ours)	<u>2.51</u>	<u>2.58</u>

Score v2 (HPSv2) (Wu et al., 2023) and Aesthetic score (Schuhmann et al., 2022), which capture different aspects of image quality and alignment with human preferences. To ensure that our method not only improves quality but also maintains diversity in generated images, we utilize the Vendi score (Friedman & Dieng, 2023) to measure the output diversity.

Baseline methods. We compare our method with the base model SD 1.5, and DPO-based methods: Diffusion-DPO and SPIN. Notice that Diffusion-DPO, SPIN, and our method all share the same data assumption: using the offline preference dataset only without the need for extra annotations or feedback from reward models.

5.2 Reward evaluation

We report the results of Pickscore, HPSv2, and Aesthetic score in Table 1, and the win-rates against SD 1.5 in Table 4 in Appendix C. Our iterative training can further improve the SPIN model with prompts in both training and test sets in terms of the surrogate metrics of human preferences. It further proves that our iterative training with an improvement model can surpass the upper bound in SPIN training.

We also visualize samples from the fine-tuned model in Figure 1 and Figure 5 in Appendix A, where our model output can generate the samples more aligned with the prompt than the SPIN model.

5.3 Diversity evaluation

We present the evaluation of the Vendi score in Table 1, where we calculate the diversity of 64 images per prompt, and report the average over all test prompts.

Our model shows improvement from the SPIN model in terms of diversity, which is only slightly lower than SD 1.5 on the test set (notice that the diversity score from SD 1.5 could be a reference). We visualize the sample sets using the same prompt from both SPIN and our fine-tuned model in Figure 4 which shows that our model tends to generate samples with more diverse colors, styles, and backgrounds with more details. We also show more samples to illustrate improved diversity compared with SPIN in Appendix B.

Table 2: Evaluation on the effect of the online samples. The values presented are Pickscore.

Number of online samples	Training set	Test set
2560	21.10	21.27
12800	21.13	21.36
38400	**21.20**	**21.46**

Table 3: Improvement model evaluation. The values presented are Pickscore.

Method	Training set	Test set
SPIN	21.15	21.41
Improvement model	**21.30**	21.38
Iterative (Ours)	21.20	**21.46**

5.4 EFFECT OF THE ONLINE SAMPLES

Here we present the effect of the number of online samples used for iterative training. From Table 2, we find that more online samples can lead to higher Pickscore from prompts in both training and test sets. This verifies that the key to successful iterative training is the online samples generated from our improvement model: more online samples would lead to better results.

5.5 EVALUATION OF THE IMPROVEMENT MODEL.

In this section, we provide the evaluation of the improvement model, by using the SD 1.5 baseline to generate the losing images as the image condition for the improvement model. We find that the improvement model can achieve significant improvements on the training set, but the generalization ability on the test set is worse than the iterative model, which implies why we do not consider using it as an inference-time model. The improvement model modifies the original architecture of SD 1.5 and is trained with different tasks than text-to-image generation. Thus the generalization ability on test prompts may not be as good as fine-tuned diffusion model. In the iterative training, we reuse the same training prompts and do not use the improvement model on unseen prompts. The iteratively trained model with the improvement model can therefore achieve better generalization ability on the test set.

6 DISCUSSION AND LIMITATION

Note that the gap between SPIN and our method depends on the specific structure of the preferences dataset. If all winning images in the preference set are near-optimal, there is little space for improvement with our improvement model and iterative training. However, if the winning images contain a diverse range from sub-optimal to optimal, SPIN can only get mediocre quality at best. In contrast, our method that learns the improvement direction can outperform SPIN. Due to a lack of resources, we use the open-source benchmark dataset instead of creating a more diverse dataset for losing images that could potentially lead to larger improvements from SPIN.

7 CONCLUSION

This paper introduces a novel approach for diffusion models to overcome the limitations of directly optimizing on the offline preference datasets. By learning a preference improvement model and using it to generate online preference pairs, the method allows for iterative model enhancement without additional human labeling. The results show improved preference alignment with high diversity, offering a promising direction for advancing image generation tasks with limited preference data while effectively bridging offline preference learning and online improvement.

REFERENCES

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*, pp. 4447–4455. PMLR, 2024.

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

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Oilvier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. *arXiv preprint arXiv:2406.01660*, 2024.

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

Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. PRDP: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *CVPR*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2023.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*, 2023.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. *arXiv preprint arXiv:2406.06424*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2023.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

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, 2024.

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*, pp. 10684–10695, 2022.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning*, 2024.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

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 *CVPR*, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable KL-constrained framework for RLHF. *arXiv preprint arXiv:2312.11456*, 2023.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023a.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023b.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. In *Forty-first International Conference on Machine Learning*, 2024.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *arXiv preprint arXiv:2402.10210*, 2024a.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024b.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 4, 2024.

A COMPARISON WITH BASELINES

We present more samples from SD 1.5, Diffusion DPO, SPIN and our fine-tuned models in Figure 5, where our model shows better alignment and image quality compared with the baselines.

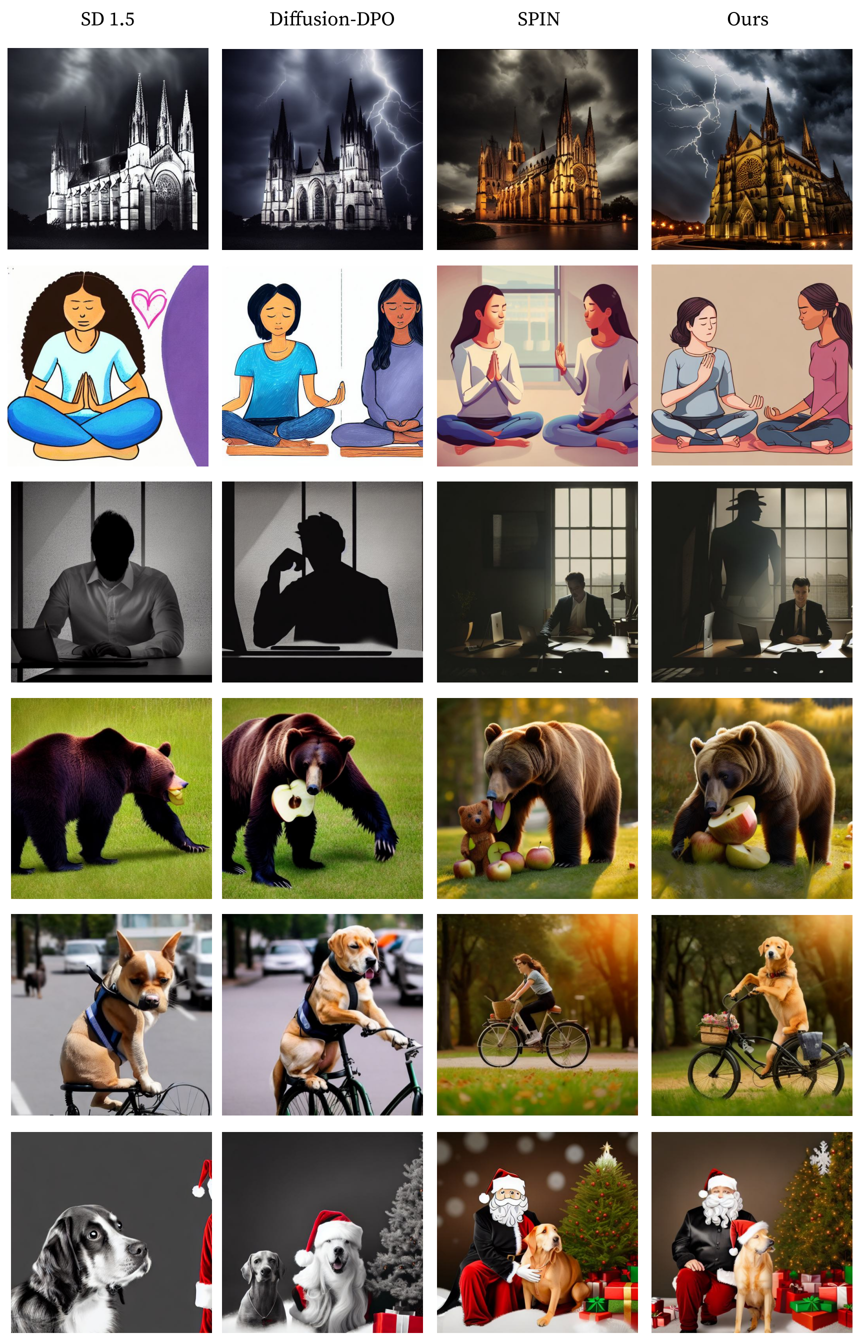

Figure 5: Prompts: *1. Gothic cathedral in a stormy night. 2. Illustration Amanda a 14-year-old girl practicing meditation and mindfulness with her mother. 3. A man sitting at his desk, with silhouettes of his inner demon behind him. 4. Bear eating apple. 5. A dog riding bicycle. 6. A dog and Santa Claus. Christmas trees in background. Black and white background.*

B Samples for Diversity visualization

We present more samples to show the improved diversity compared with SPIN in Figure 6, 7, 8.

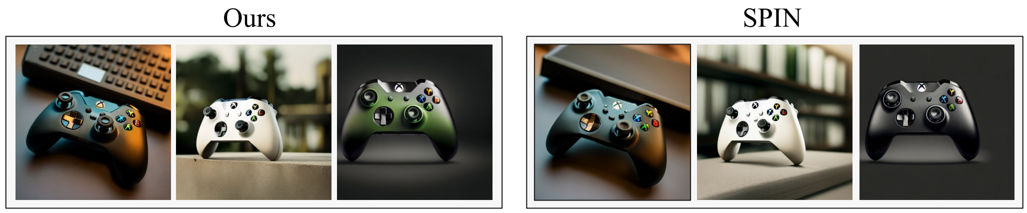

Figure 6: Prompt: *An xbox controller*. The colors and backgrounds from our model are more diverse, Examples are from the same set of seeds.

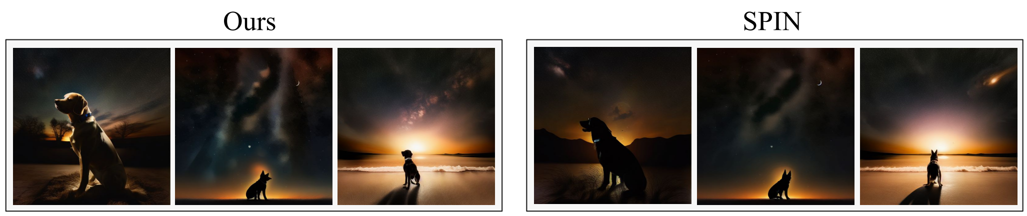

Figure 7: Prompt: *A silhouette of a dog looking at the stars.* The output backgrounds are more diverse from our model. Examples are from the same set of seeds.

Figure 8: Prompt: *A house made of cards*. The output backgrounds from our model are more diverse than SPIN. Examples are from the same set of seeds.

C Win-rate

We present the win-rates from Diffusion DPO, SPIN, and our iterative model against SD 1.5 in Table 4, where our model achieves consistent improvement from SPIN.

Table 4: Win-rate against SD 1.5.

Reward	Diffusion-DPO	SPIN	Iterative (Ours)
Pickscore	69.0%	78.5%	**79.4**%
HPSv2	64.8%	74.7%	**75.4**%
Aesthetic	59.2%	85.5%	**86.4**%