DEPO: Enhancing E-commerce Image Background Generation with Short Trajectory Direct Expected Preference Optimization

Shikun Sun*
Tsinghua University
Beijing, China
ssk21@mails.tsinghua.edu.cn

Chengrui Wang
Taobao & Tmall Group of Alibaba
Beijing, China
wangchengrui.wcr@alibaba-inc.com

Min Zhou
Taobao & Tmall Group of Alibaba
Beijing, China
yunqi.zm@alibaba-inc.com

Zixuan Wang
Tsinghua University
Beijing, China
wangzixu21@mails.tsinghua.edu.cn

Xiaoyu Qin
Tsinghua University
Beijing, China
xyqin@tsinghua.edu.cn

Tiezheng Ge
Taobao & Tmall Group of Alibaba
Beijing, China
tiezheng.gtz@alibaba-inc.com

Bo Zheng
Taobao & Tmall Group of Alibaba
Beijing, China
bozheng@alibaba-inc.com

Jia Jia†
BNRist,Tsinghua University. Key Laboratory of Pervasive Computing, Ministry of Education.
Beijing, China
jjia@tsinghua.edu.cn

Figure 1: E-commerce images synthesized by our fine-tuned DEPO model based on provided foregrounds.

Abstract

Generating high-quality, user-preferred backgrounds for e-commerce product images poses unique challenges for diffusion models, particularly in aligning outputs with human visual preferences. While Direct Preference Optimization (DPO) has shown promise in aligning generative models with human feedback, its application to diffusion models faces key limitations, including the trade-off between reward sparsity and supervision quality, mode collapse, and training instability. To tackle these issues, we propose Direct Expected Preference Optimization (DEPO), a novel framework that adapts DPO to diffusion models through redesigned training and sampling strategies. Specifically, DEPO introduces a DEPO loss combined with trajectory segmentation to enable more frequent and informative reward feedback, employs Langevin MCMC to broaden the exploration space and mitigate mode collapse, and leverages masks to effectively constrain the search space while incorporating targeted engineering designs to improve training stability. By directly linking image-domain evaluations to expected log probabilities and incorporating adversarial training, DEPO achieves better alignment with user preferences while maintaining high image fidelity. Experimental results demonstrate that DEPO surpasses existing methods in both the diversity and quality of background generation.

*Work done when Shikun Sun was an intern at Taobao & Tmall Group of Alibaba.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755499

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Diffusion Models, Generative Models, Preference Optimization

ACM Reference Format:

Shikun Sun, Chengrui Wang, Min Zhou, Zixuan Wang, Xiaoyu Qin, Tiezheng Ge, Bo Zheng, and Jia Jia. 2025. DEPO: Enhancing E-commerce Image Background Generation with Short Trajectory Direct Expected Preference Optimization. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3746027.3755499

1 Introduction

Although diffusion models have demonstrated their strong capability to generate product images in e-commerce scenarios [6, 33], aligning the generated images with human visual preferences remains a significant challenge. This alignment is crucial for enhancing user satisfaction, improving engagement, and driving commercial success. Recent studies, such as direct preference optimization (DPO) [25], have shown considerable promise for aligning generative models with human preferences across both text and image domains by leveraging direct supervision from comparative human feedback. However, existing DPO-based approaches [15, 32, 37, 38] face several limitations when applied to diffusion models. A key challenge is the inherent trade-off between reward sparsity and supervision quality: high-quality reward signals, typically provided at the image level, are inherently sparse, while denser rewards tend to be heuristic and lack deep alignment with the diffusion process. This tension limits the ability to provide consistent and effective guidance throughout the generation pipeline. Additionally, these methods often suffer from reduced output diversity and unstable training dynamics, both of which impede the development of robust and preference-aligned image generation systems.

To address these challenges, we introduce **DEPO** (Direct Expected Preference Optimization), an innovative framework that combines DPO strategies with diffusion models, specifically optimized for e-commerce product background generation. Our approach designs new training and sampling processes to overcome the limitations of conventional DPO and diffusion models. By incorporating trajectory segmentation and DEPO loss, DEPO facilitates more frequent and efficient reward feedback, enhancing the sampling process and promoting a smoother learning curve. We further integrate Langevin MCMC [21] to broaden the exploration space and mitigate mode collapse. To ensure training stability, we design targeted exploration restrictions for backgrounds, directly link evaluations in the image domain to expected log probabilities, and implement a policy gradient selection mechanism. Additionally, we apply adversarial training to reinforce theoretical and empirical stability improvements, ensuring consistent and high-quality training outcomes. Our main contributions are as follows:

1) We are the first to demonstrate that image-domain evaluations can be effectively connected to expected log probabilities within the DPO framework, introducing a novel and stable human preference fine-tuning approach for diffusion models. This key insight enables consistent optimization that not only adheres to the DPO training paradigm but also achieves a more precise alignment between user preferences and reward signals.

2) We are the first to integrate advanced exploration techniques, such as Langevin MCMC [21], into diffusion-based policy search. This integration substantially expands the exploration space and mitigates mode collapse, leading to more diverse and robust image generation.

3) We propose a targeted strategy for e-commerce product background generation by leveraging masks to effectively constrain the search space of DEPO. This focused approach improves training efficiency and enables the model to achieve state-of-the-art performance, demonstrating its practical effectiveness in real-world applications.

Extensive experiments validate that DEPO enhances training stability, increases image diversity, and surpasses existing methods in generating high-quality product backgrounds. These advancements set a new benchmark for the automated generation of e-commerce visual content.

2 Preliminary

2.1 General RL Formulation of Diffusion Models

Following prior work, we can represent the generation process of a diffusion model as a Markov Decision Process (MDP). Consider a sequence of distributions $p(\mathbf{x}_t)$ with increasing noise levels, for $t \in [0, 1, 2, \ldots, T]$. Here, $\mathbf{x}_0$ represents the image distribution, while $\mathbf{x}_T$ is a pure Gaussian distribution. The process of image generation can then be formulated as an MDP defined as follows:

$$\begin{aligned} \mathbf{s}_t &\triangleq (\mathbf{c}, t, \mathbf{x}_t), \\ \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) &\triangleq p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}), \\ P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) &\triangleq \left(\delta_{\mathbf{c}}, \delta_{t-1}, \delta_{\mathbf{x}_{t-1}}\right), \\ \mathbf{a}_t &\triangleq \mathbf{x}_{t-1}, \\ \rho_0(\mathbf{s}_0) &\triangleq (p(\mathbf{c}), \delta_T, \mathcal{N}(\mathbf{0}, \mathbf{I})), \\ R(\mathbf{s}_t, \mathbf{a}_t) &\triangleq \begin{cases} r_\phi(\mathbf{x}_0, \mathbf{c}) & \text{if } t = 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{1}$$

where

- $\mathbf{s}_t$ denotes the state at time step t, incorporating the control condition $\mathbf{c}$, the timestep t, and the latent variable $\mathbf{x}_t$. In our application for product background generation, $\mathbf{c}$ typically comprises the text prompt $\mathbf{c}_{\text{txt}}$ and product image $\mathbf{c}_{\text{img}}$ with the corresponding product mask $\mathbf{c}_{\text{mask}}$.
- The policy π_θ is parameterized as a diagonal Gaussian policy. The mean is determined by the diffusion model, while the standard deviation is set according to the parameters of the diffusion process, θ represents the parameters of the diffusion models.
- The state transition function $P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ is deterministic, dependent solely on the action $\mathbf{a}_t$.
- $\rho_0(\mathbf{s}_0)$ denotes the distribution of initial states.
- $R(\mathbf{s}_t, \mathbf{a}_t)$ defines the reward function, which assigns a reward $r_\phi(\mathbf{x}_0, \mathbf{c})$ when $t = 0$ and 0 at other times. In our case, ϕ represents the parameters of a judging function that scores images.

This MDP framework conceptualizes the image generation process as a sequence of transitions, each driven by a Gaussian policy through targeted action selection to generate meaningful outputs at every stage. By organizing the process in this manner, it facilitates the formulation of the reward function $r_\phi(\mathbf{x}_0, \mathbf{c})$ and the construction of positive and negative samples, thereby enabling the application of various reinforcement learning (RL) algorithms.

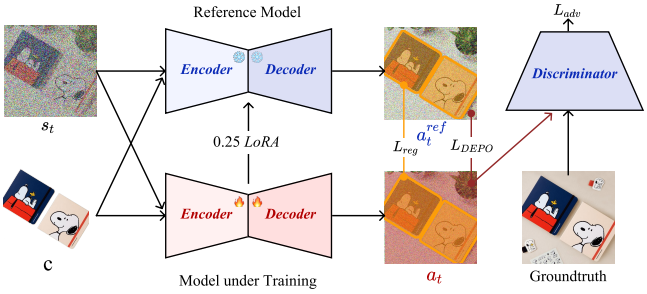

Figure 2: Overview of our DEPO.

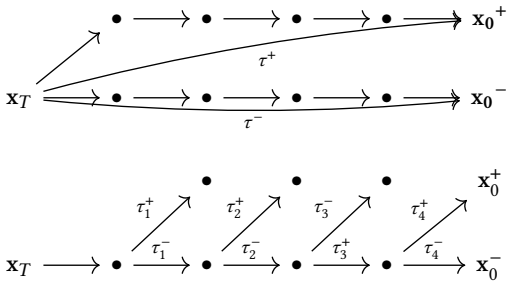

Figure 3: Our approach will yield much more frequent returns compared to the original diffusion trajectory.

A particularly effective technique for refining diffusion models is Direct Preference Optimization (DPO). This approach utilizes user preferences or predefined criteria to directly influence the policy, effectively tuning the model to produce outcomes that are closely aligned with the desired results.

2.2 DPO Formulation of Diffusion Models

With ahead MDP formulation of the generation process of diffusion models, We can directly borrow the DPO formulation from Natural Language Processing. From the same $\mathbf{x}_T$, we follow the MDP process to generate two images, $\mathbf{x}_0^+$ and $\mathbf{x}_0^-$, with the preference of the first one. After collection of a lot of such image pairs, we can train the diffusion model with the following loss:

$$\mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = -\mathbb{E}_{\left(\mathbf{x}_T, \mathbf{x}_0^+, \mathbf{x}_0^-\right)\sim\mathcal{D}} \left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(\mathbf{x}_0^+ \mid \mathbf{x}_T\right)}{\pi_{\text{ref}}\left(\mathbf{x}_0^+ \mid \mathbf{x}_T\right)} - \beta\log\frac{\pi_\theta\left(\mathbf{x}_0^- \mid \mathbf{x}_T\right)}{\pi_{\text{ref}}\left(\mathbf{x}_0^- \mid \mathbf{x}_T\right)}\right)\right], \tag{2}$$

where $\pi(\mathbf{x}_0|\mathbf{x}_T) = \prod_{t=T}^{1}\pi(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Substitute the formula of π in Equation 2, we have that

$$\mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = -\mathbb{E}_{\left(\mathbf{x}_T, \mathbf{x}_0^+, \mathbf{x}_0^-\right)\sim\mathcal{D}} \log\sigma \left[\sum_{i=1}^{T}\left(\beta\log\frac{p_\theta\left(\mathbf{x}_{t-1}^+ \mid \mathbf{x}_t^+\right)}{p_{\text{ref}}\left(\mathbf{x}_{t-1}^+ \mid \mathbf{x}_t^+\right)} - \beta\log\frac{p_\theta\left(\mathbf{x}_{t-1}^- \mid \mathbf{x}_t^-\right)}{p_{\text{ref}}\left(\mathbf{x}_{t-1}^- \mid \mathbf{x}_t^-\right)}\right)\right], \tag{3}$$

where $\{\mathbf{x}_i^+\}$ and $\{\mathbf{x}_i^-\}$ are positive and negative trajectories. and $p(\cdot \mid \cdot)$ indicades diagonal Gaussian policy defined by DDPM. What we want is such a formulation:

$$-\mathbb{E}_{\left(\mathbf{x}_T, \mathbf{x}_0^+, \mathbf{x}_0^-\right)\sim\mathcal{D}} \sum_{i=1}^{T}\log\sigma \left(\beta\log\frac{p_\theta\left(\mathbf{x}_{t-1}^+ \mid \mathbf{x}_t^+\right)}{p_{\text{ref}}\left(\mathbf{x}_{t-1}^+ \mid \mathbf{x}_t^+\right)} - \beta\log\frac{p_\theta\left(\mathbf{x}_{t-1}^- \mid \mathbf{x}_t^-\right)}{p_{\text{ref}}\left(\mathbf{x}_{t-1}^- \mid \mathbf{x}_t^-\right)}\right). \tag{4}$$

However, the sum formulation inside $\log\sigma$ cannot be easily separated. The main challenge is to ensure that the expanded expression within $\log\sigma$ can be decomposed, allowing for the calculation of the loss without the need to store all gradients at each diffusion step. Recent works [32, 37] introduce specific assumptions to argue that it is possible to forcefully decompose the terms in Equation (3) into step pairs as in Equation (4), resulting in the difference between $\mathbf{x}_t^+$ and $\mathbf{x}_t^-$ and a decrease in optimization efficiency. An alternative strategy involves decomposing long trajectories into shorter segments and utilizing trajectory pairs of reduced length. SPO [15] has presented an initial empirical exploration.

3 Methodology Overview

Our methodology follows the fundamental paradigm of online RL algorithms, which consists of an iterative loop: sampling by the current model, followed by model optimization. The overview of our DEPO is shown in Figure 2.

During the sampling stage, we follow SPO and employ a similar sampling procedure to generate shorter trajectories, as illustrated in Figure 3. However, unlike conventional approaches, we incorporate Langevin MCMC at branching timesteps, enhancing sampling flexibility beyond that of DDPM.

During the training stage, unlike other methods, we do not decompose each short trajectory into step pairs containing different states, as this would apply Equation (4) as Equation (3). Instead, we directly employ a new, stable DEPO loss and reformulate it into an easy-to-apply expression tailored for diffusion models, requiring only a reward model in the image domain.

Additionally, we incorporate several engineering techniques to enhance performance, including constrained exploration for background consistency, a policy gradient selection mechanism, and adversarial training strategies.

Table 1: Comparison of Different Methods Based on Key Features. The table illustrates the presence (✓) or absence (✗) of specific features across various algorithms, such as discretization error, exact evaluation, extra exploration, shared state utilization, and high-frequency reward application. A detailed explanation is in Appendix A.

Method	No Discretization Error	Exact Evaluation	Extra Exploration	Shared State	High Frequency Reward
Diffusion-DPO [31]	✓	✓	✗	✗	✓
D3PO [36]	✗	✓	✗	✗	✗
SPO [15]	✗	✗	✗	Partially	✓
DEPO (Ours)	✗	✓	✓	✓	✓

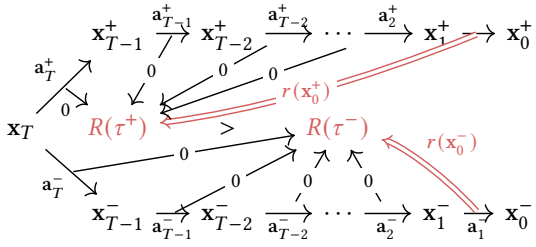

Figure 4: A pair of trajectories τ^+, τ^- from the diffusion process. In this type of sampling algorithm, the reward is inherently sparse, as it is only obtained every T steps.

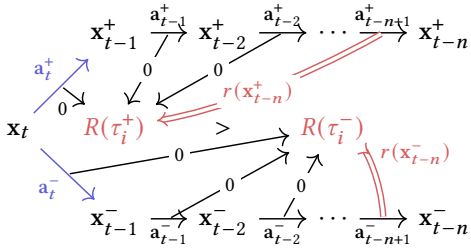

Figure 5: A pair of short trajectories, τ_i^+, τ_i^-, from our diffusion process. In this sampling algorithm, the reward is frequent, being obtained every n steps.

4 Efficient Short Trajectory Sampling

There are two main challenges in the current sampling procedure. Firstly, **the reward is too sparse**: in practice, we can only evaluate the quality of an image in the $\mathbf{x}_0$ domain, which means generating the entire sequence is necessary to obtain a single reward at $\mathbf{x}_0$. Secondly, **there are potential mode collapse issues**: during training, the joint probability distribution between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ becomes increasingly similar across different timesteps, which can lead to mode collapse.

Segmentation of Original Trajectories. One simple approach to addressing the sparse reward problem is to shorten the trajectory of the RL sampling process. Directly increasing the step size, however, can compromise the generation quality of the diffusion model. Due to the inherent structure of diffusion models, **policies for different timesteps are often trained independently using score matching or flow matching algorithms.** This implies that the policy does not rely heavily on dependencies between different timesteps.

To leverage this property, as shown in Figure 3, **we can strategically introduce branching at specific timesteps during image generation, segmenting the entire trajectory into multiple shorter sub-trajectories for training without any systematic errors.** This method enhances the efficiency of the training process and helps mitigate the sparse reward issue. To support this, we train a reward model to assess the quality of the final state of each short trajectory. However, this approach may introduce additional variance, making it necessary to implement a selection mechanism to filter and retain the most promising trajectories, which will be introduced in Section 6.

The comparison between whole trajectories and short trajectories is illustrated in Figure 4 and Figure 5. For the former, the reward is inherently sparse as it is only obtained every T steps. To make the loss function feasible, we need to better accommodate the error introduced in Equation (4). For the latter, we select an appropriate n and directly utilize the pair $\{\mathbf{x}_{t-n}^+, \mathbf{x}_{t-n}^- | \mathbf{x}_t\}$, focusing solely on this pair. In this way, we avoid the systematic error introduced by the interchange between $\log \sigma$ and $\sum$, and the discretization error of the SDE can be further mitigated through the application of Langevin MCMC, which will be introduced as follows.

Expanding Exploration with Langevin MCMC. An additional enhancement in our sampling process is the integration of Langevin MCMC [21]. After performing the initial state transition from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ using the DDPM sampling algorithm (as depicted in blue in Figure 5), where $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ may become increasingly similar across different timesteps thus inducing mode collapse, an exploration step is necessary. To address this, we introduce Langevin MCMC sampling at these timesteps as an exploration mechanism. This approach effectively explores regions within $p_\theta(\mathbf{x}_{t-1})$ and the vicinity of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which, in our evaluation, provides a robust means of enhancing model diversity and alleviating mode collapse. We provide detailed information about Langevin MCMC in Appendix C.

5 Direct Expected Preference Optimization

Given that human preferences are inherently tied to the image domain, there is a lack of corresponding data in the domain of $\mathbf{x}_t$ for $t \neq 0$. However, diffusion models provide predictive estimates of $\mathbf{x}_0$ at each timestep, and these predictions remain reasonably accurate even at relatively small timesteps. This enables the direct application of preference models designed for the image domain. Based on this observation, we make Assumption 5.1.

Assumption 5.1. The reward model in the image domain serves as a reliable reward model for the diffusion model's one-step prediction from a relatively small timestep.

Building on Assumption 5.1, we reconsider the scores for $\mathbf{x}_{t-n}$ typically used by other methods, which are obtained from the diffusion model's direct prediction of $\widehat{\mathbf{x}_0}$, and subsequently evaluated in the $\mathbf{x}_0$ domain. Under Proposition 5.2, we have Lemma 5.3.

Proposition 5.2. *According to DDIM [29],*

$$\mathbb{E}\left[\mathbf{x}_{t-n-1} \,\middle|\, \mathbf{x}_{t-n}\right] = \alpha(t-n)\,\widehat{\mathbf{x}_0}(\mathbf{x}_{t-n}) + \beta(t-n)\epsilon_\theta^{(t-n)}(\mathbf{x}_{t-n})$$

where $\alpha(t-n)$ and $\beta(t-n)$ are deterministic scalars that depend only on $t-n$ and can represent the evolution of $\mathbf{x}_{t-n-1}$, $\epsilon_\theta^{(t-n)}$ is the predicting noise function.

Lemma 5.3 (Accurate Evaluation for Diffusion Models)**.** *The reward model applied to the diffusion model's one-step prediction, $\widehat{\mathbf{x}_0}$, aligns more closely with $\mathbb{E}\left[\mathbf{x}_{t-n-1}|\mathbf{x}_{t-n}\right]$ than with samples from $\mathbf{x}_{t-n-1} \sim p_\theta(\mathbf{x}_{t-n-1}|\mathbf{x}_{t-n})$, and more so than using $\mathbf{x}_{t-n}$ alone.*

Following Lemma 5.3, we propose replacing $p(\mathbf{x}_{t_\tau - n}|\mathbf{x}_{t_\tau})$ with $p(\mathbf{x}_{t_\tau - n - 1}|\mathbf{x}_{t_\tau})$ and introducing an expectation over $\mathbf{x}_{t_\tau - n - 1}$to better align with the distribution level evaluation. Based on this, we

define the **DEPO** Loss:

$$\mathcal{L}_{\text{DEPO}} = -\mathbb{E}_{\tau\sim\mathcal{D}} \log \sigma \mathbb{E}_{\mathbf{x}^{\pm}_{t_\tau-n-1}\sim p_{sg[\theta]}(\mathbf{x}_{t_\tau-n-1}|\mathbf{x}^{\pm}_{t_\tau-n})} \left(\beta \log \frac{p^{\mathbf{m}}_\theta\left(\mathbf{x}^{+}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)}{p_{\text{ref}}\left(\mathbf{x}^{+}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)} - \beta \log \frac{p^{\mathbf{m}}_\theta\left(\mathbf{x}^{-}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)}{p_{\text{ref}}\left(\mathbf{x}^{-}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)} \right), \tag{5}$$

where $p^{\mathbf{m}}_\theta$ denotes that the mean μ of the Gaussian policy is detached within the mask $\mathbf{m}$, which will be introduced later, and sg represents the stop-gradient operation. However, due to the expectation inside $\log\sigma$, we still encounter the computational efficiency issue discussed in the Preliminary section. To address this, we further derive an efficient formula for DEPO that directly computes the term inside $\log\sigma$ in the **DEPO** loss, as stated in Proposition 5.4.

Proposition 5.4 (Efficient Formula for DEPO). *We can directly obtain the expected* log *probability within the* log σ *by*

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}^{\pm}_{t_\tau-n-1}\sim p_{sg[\theta]}(\mathbf{x}_{t_\tau-n-1}|\mathbf{x}^{\pm}_{t_\tau-n})} \log \frac{p^{\mathbf{m}}_\theta\left(\mathbf{x}^{+}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)}{p^{\mathbf{m}}_\theta\left(\mathbf{x}^{-}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)} \\ &= \log \frac{p^{\mathbf{m}}_\theta\left(\mu_{sg[\theta],t_\tau-n-1}\left(\mathbf{x}^{+}_{t_\tau-n}\right) \mid \mathbf{x}_{t_\tau}\right)}{p^{\mathbf{m}}_\theta\left(\mu_{sg[\theta],t_\tau-n-1}\left(\mathbf{x}^{-}_{t_\tau-n}\right) \mid \mathbf{x}_{t_\tau}\right)}. \end{aligned} \tag{6}$$

The proof of Proposition 5.4 is in Appendix D.1. In this way, we can calculate the **DEPO** loss more efficiently as shown in Equation (15).

6 Techniques for Stabilizing Training

Constrained Exploration for Backgrounds. As illustrated in Figure 2, our fine-tuning task focuses on generating the background. Consequently, gradients passing through the product area do not contribute meaningful information and may introduce instability during training. To address this, we enforce gradient backpropagation exclusively through the background region, ensuring that the fine-tuning process concentrates on the relevant areas. For constraints within the product area, we adopt a direct approach by mimicking the original reference model's behavior and applying direct restrictions within that region. We have the following simple Loss:

$$\mathcal{L}_{reg} = \mathbb{E}\left[(\mathbf{a}_t - \mathbf{a}^{ref}_t)\cdot\mathbf{m}\right]^2, \tag{7}$$

where $\mathbf{m}$ is the mask of the product. This method helps maintain the integrity of the product area while enabling efficient and stable training of the background generation.

Policy Gradient Selection Mechanism. Despite reductions in variance, issues with training stability continue to pose challenges. To address these, we have implemented a Policy Gradient Selection Mechanism aimed at maximizing training efficiency and stability by refining the selection and utilization of pivotal samples. This mechanism comprises two primary components:

a) Sample Selection: Our selection process differs from SPO by utilizing a relative gap for filtering rather than an absolute gap. This approach is first applied during trajectory bifurcation at the same timestep and later revisited after sampling across different timesteps to ensure sample quality.

b) Computing DEPO Loss: While calculating the DEPO loss, we identify the sample points requiring optimization based on the probabilities predicted by the reference model. This method enhances our focus on key samples poised to significantly boost model performance without straying too far from the reference model. As the probability space shown in Figure 6, we will detach the gradient propagation under the following situations:

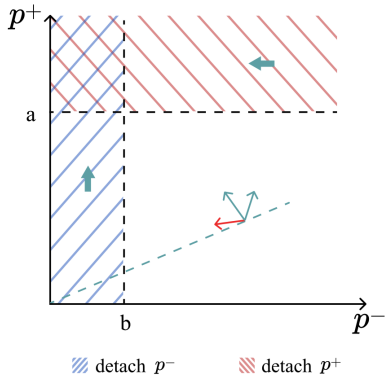

Figure 6: The probability space of $p^+ = p_\theta(\mathbf{x}^+_{t-1}|\mathbf{x}_t)$ and $p^- = p_\theta(\mathbf{x}^-_{t-1}|\mathbf{x}_t)$. $a = \max(p^+_{ref}, p^-_{ref})$ and $b = \min(p^+_{ref}, p^-_{ref})$.

$$\begin{aligned} \log p^-_\theta &\le \lambda_{detach}\cdot\log b,\ \ detach\ p^-_\theta, \\ \log p^+_\theta &\ge \frac{1}{\lambda_{detach}}\cdot\log a,\ \ detach\ p^+_\theta. \end{aligned} \tag{8}$$

The worst range lies in the blue area rather than the red area, because in this case, although $\mathcal{L}_{\text{DEPO}}$ is decreasing, p^+ also decreases undesirably, being influenced by p^-. A detailed explanation is in Appendix B.

Adversarial Training. Preference optimization improves the aesthetics of generated images to better match user preferences. In the e-commerce scene, it is essential for generated images to align seamlessly with the style of manually crafted e-commerce images. To achieve the goal, we employ adversarial loss to reduce the stylistic differences between generated images and selected aesthetic e-commerce images. Following [18], we combine the encoder and mid-block of the pre-trained diffusion model with a trainable prediction head as the discriminator d. Based on Hinge loss function [16], the optimization objective can be formulated as follows:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\tau\sim\mathcal{D}_\tau, x_{fake}\sim p_{sg[\theta]}(\mathbf{x}_{t_{\tau-n-1}}|\mathbf{x}^+_{t_\tau-n})} d\left(x_{fake}, t_{\tau-n-1}, \mathbf{c}\right). \tag{9}$$

In summary, our final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{DEPO}} + c_{reg}\mathcal{L}_{reg} + c_{adv}\mathcal{L}_{adv}. \tag{10}$$

7 Dataset and Reward Model

Pairwise User Preference Dataset. Our online e-commerce image background generation platform enables users to specify the position and size of their products within an image and utilize custom prompts to simultaneously generate a group of images with different backgrounds. Users can then select and download their preferred images. To create the pairwise preference dataset $\mathcal{D}_p$, we collect 1,122,624 images from the platform, and construct 884,649 samples for training and 1,123 samples for evaluation. Each sample

Figure 7: The visual comparisons between different methods. It is clear that the impact of DPO and SPO on image quality is negligible compared to our DEPO in e-commerce image background generation. Without $\mathcal{L}_{adv}$, the background style closely resembles that of the base. Without $\mathcal{L}_{DEPO}$, the background lacks aesthetic appeal and seems unnecessarily complicated. In conclusion, $\mathcal{L}_{DEPO}$ primarily focused on optimizing the image aesthetic, whereas $\mathcal{L}_{adv}$ mainly concentrated on optimizing the image style. Our DEPO framework produces the best visual outcomes.

in the dataset contains a text prompt $\mathbf{c}_{\text{txt}}$ and a pair of generated images x^+, x^-, where image x^+ is preferred over image x^-.

E-commerce Preference Score. We develop a reward model called E-commerce Preference Score (EPS) to predict user preferences for images generated by custom prompts. Following the architecture of CLIP [12, 24], given a prompt $\mathbf{c}_{\text{txt}}$ and an image x, our reward model computes the preference score $s(\mathbf{c}_{\text{txt}}, x)$ with text encoder E_{txt} and image encoder E_{img}:

$$s(\mathbf{c}_{\text{txt}}, x) = E_{txt}(\mathbf{c}_{\text{txt}}) \cdot E_{img}(x) \cdot T, \tag{11}$$

where T is the learned scalar temperature parameter of CLIP. The model is trained on the Pairwise User Preference Dataset, and the loss function can be formulated as:

$$\mathcal{L}_{rm} = -\mathbb{E}_{(\mathbf{c}_{\text{txt}}, x^+, x^-)\sim\mathcal{D}_p}[\log \sigma(s(\mathbf{c}_{\text{txt}}, x^+) - s(\mathbf{c}_{\text{txt}}, x^-))]. \tag{12}$$

We find that the resulting scoring function achieves 69.8% accuracy rate on the evaluation dataset, which is close to the performance of PickScore [12] on Pickaic dataset (70.5% accuracy rate).

Aesthetic E-commerce Dataset. We collect 2,456 well-designed aesthetic images from an e-commerce platform, and obtain the captions of the images using GPT-4o [20]. Each sample in the dataset $\mathcal{D}_a$ contains a control condition $\mathbf{c}$ and an aesthetic e-commerce image x. We randomly select 200 samples from the dataset for evaluation, and use the remaining samples for training.

8 Experiments

8.1 Implementation Details

For our best-performing implementation, we use the following configurations: LoRA with rank 64 for training, a learning rate of 10^{-4}, and a GAN-specific learning rate of 10^{-5}. The sampling process consists of 25 steps, with a short trajectory length of 4. We apply a 50% filter rate between trajectories and a 50% filter rate across different timesteps. Additional hyperparameters include $c_{\text{detach}} = 3.0$, $c_{\text{reg}} = 10^{-4}$, and $c_{\text{adv}} = 0.1$.

During the training of DEPO, we run 3,000 steps, saving the LoRA parameters every 50 steps. The model is trained with an effective batch size of 16 using eight H20 GPUs. While this batch size is significantly smaller than that of Diffusion-DPO, it aligns with recent works such as SPO, which have demonstrated strong visual results. The optimal checkpoint is typically selected around the 1,000-step mark within the 3,000 training steps.

8.2 Evaluation Metrics

E-commerce Preference Score (EPS). The primary quantitative metric in our evaluation is the E-commerce Preference Score (EPS), as our main objective is to align with preferences specific to our commercial context.

General Quality Evaluation Metrics. In addition to EPS, we consider two general evaluation metrics: PickScore and CLIP Score. PickScore is trained on a different preference dataset with an emphasis on the main object, which may not fully align with our specific context. CLIP Score measures the alignment between the generated image and the text prompt but has only limited relevance to our scenario.

Diversity Metrics. We quantify diversity using the trace of the covariance of Inception-V3 [30] latent features, similar to FID [11]. This method aligns with the underlying assumption in FID, where

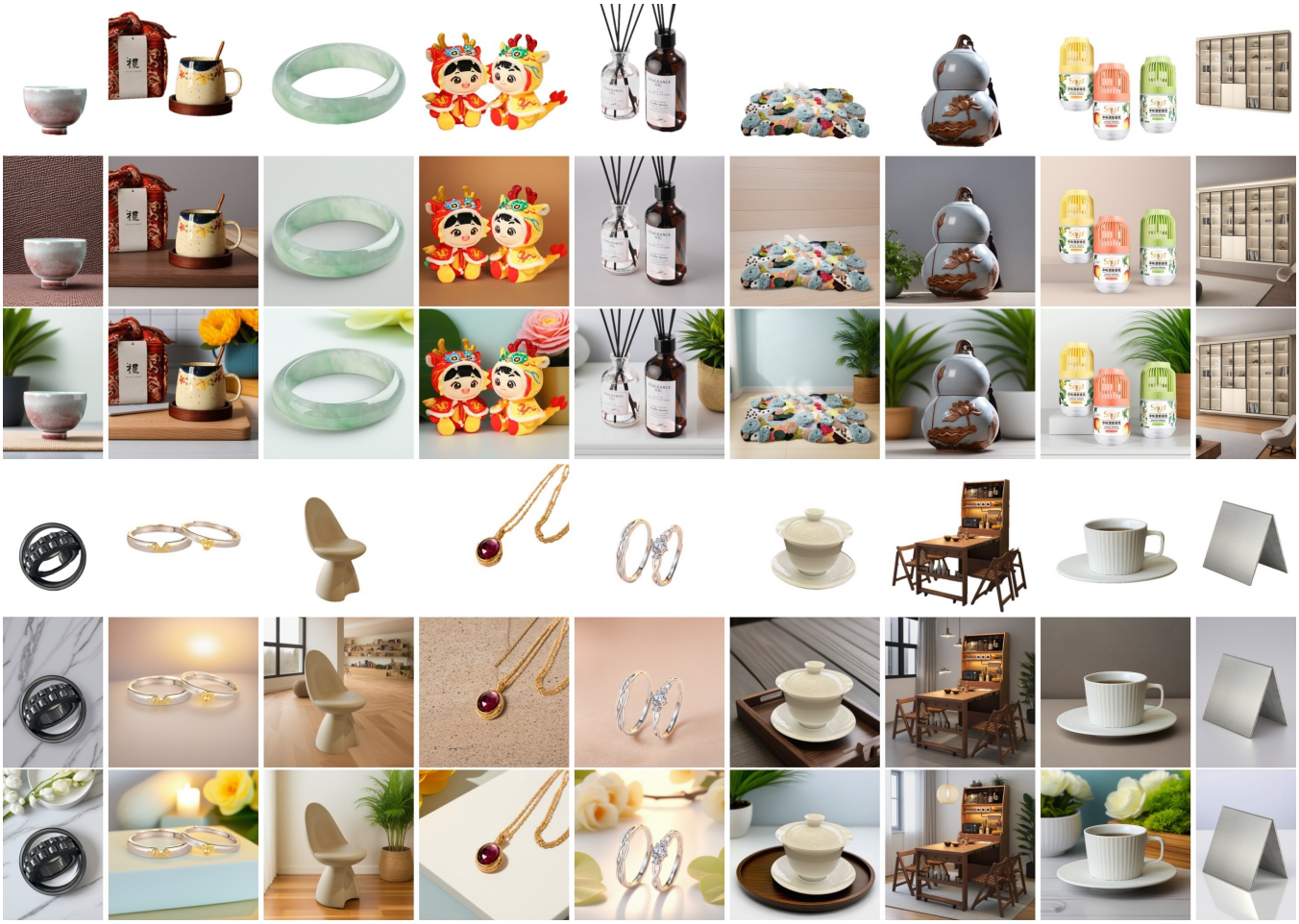

Figure 8: Visualization of foreground images, images generated by baseline, and images generated by our DEPO.

the latent features are treated as samples from a multivariate Gaussian distribution.

8.3 Comparision with Other Methods

Although directly applying Diffusion-DPO to our scenario is unsuccessful, we successfully adapt it by incorporating simple gradient clipping. Since the training code for the reference model of SPO has not been released for now, we use their model solely for comparison purposes. The visual results are presented in Figure 7. It is evident that our DEPO framework produces the best visual outcomes.

We also conducted a user study in Figure 9 to compare our method against three baselines: the original Baseline, Diffusion DPO, and SPO. For each comparison, we collected 76 preference pairs for each comparison.

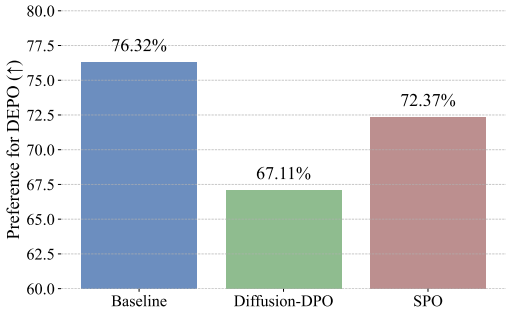

Figure 9: User preference for DEPO compared with other methods. These results demonstrate that our approach is more aligned with human preferences in visual quality and overall presentation compared to existing methods.

8.4 Ablation Study

Langevin MCMC Exploration. The Langevin MCMC algorithm introduces additional exploration, potentially leading to improved performance. As shown in Table 2, Langevin MCMC contributes a 0.31 increase in EPS and a 0.26 increase in CLIP Score, demonstrating its effectiveness.

Online or Offline. Online training offers broader exploration but comes with a higher likelihood of deviating from the original distribution. In our experiments, online training contributes only a 0.05 increase in EPS but shows a 0.35 improvement in CLIP Score. This indicates that, for our specific task, expanding the exploration space through online training has an overall positive impact. For

Table 2: Quantitative metrics (mean ± standard deviation) for EPS, PickScore, and CLIP Score across more baselines and ablations. We perform the generation process five times and report the mean and standard deviation of the results.

Method	EPS↑ (Mean ± Std)	PickScore↑ (Mean ± Std)	CLIP↑ (Mean ± Std)
Baseline	17.31 ± 0.07	21.03 ± 0.03	26.63 ± 0.21
Diffusion-DPO	17.42 ± 0.04	21.07 ± 0.01	26.90 ± 0.12
SPO	17.59 ± 0.05	21.17 ± 0.03	27.19 ± 0.15
DEPO (Ours)	**17.86 ± 0.06**	**21.24 ± 0.02**	**27.64 ± 0.11**
w random filter	17.45 ± 0.10	21.01 ± 0.04	26.87 ± 0.19
w/o sg on product mask	17.78 ± 0.08	21.07 ± 0.01	25.92 ± 0.17
w/o L_{reg} & sg on product mask	17.59 ± 0.05	21.08 ± 0.02	26.89 ± 0.15
w/o L_{adv}	17.81 ± 0.06	21.15 ± 0.03	26.90 ± 0.21
w/o detach on $p^{\pm}$	17.73 ± 0.08	21.10 ± 0.03	26.74 ± 0.22
w/o Langevin MCMC	17.55 ± 0.07	21.16 ± 0.02	27.40 ± 0.14
w/o online sampling	17.81 ± 0.07	21.14 ± 0.04	27.29 ± 0.12
w/o DEPO Loss	16.99 ± 0.04	20.96 ± 0.01	26.69 ± 0.15

Table 3: Diversity Score across different methods. The results align with our visual observations: DEPO achieves the highest Diversity Score among all methods.

Method	Diversity Score↑
Baseline	174.6
Diffusion-DPO	177.5
SPO	176.5
DEPO (Ours)	**183.8**

models trained over the long term, offline training demonstrates better stability.

Constrained Exploration. Although applying the stop-gradient operation on the product mask resulted in a modest improvement of 0.08 in EPS, when combined with L2 regularization, it collectively contributes to a 0.27 increase in EPS. This indicates that such a targeted design is highly effective for our task. Additionally, it yields a significant 0.75 improvement in CLIP Score.

Policy Gradient Selection Mechanism. We primarily compare the filtering rules applied at different timesteps. Compared to a random filter, our high-score filter shows an improvement of 0.41 in EPS. This indicates that, within the range of timesteps, obtaining higher-quality samples is more beneficial for performance than obtaining uniformly distributed samples.

GAN Loss. Although the GAN loss provides only a minor improvement in EPS, it helps address certain problematic cases and improve visual results as shown in Figure 7. Moreover, the GAN loss significantly enhances the text alignment metric, yielding a 0.74 increase.

9 Related Work

Background Generation. Early methods generated backgrounds by composing foreground and background images [17, 19], requiring multiple steps such as image matching [39], foreground placement [1], and harmonization [7, 8]. These methods depend on background libraries, limiting diversity. With advances in text-to-image generation [22, 26], backgrounds can now be generated directly from prompts. Subject-driven methods like DreamBooth [5, 14, 27] improve harmony but struggle with subject fidelity. Inpainting-based approaches [3, 9, 34] maintain the foreground but risk improper subject extension or disharmony. For e-commerce images, subject fidelity is critical. We build on inpainting techniques and introduce preference optimization to enhance image quality.

Preference Optimization. Recent diffusion-based methods optimize image generation to better align with human preferences. Chen et al. [4] refine diffusion models via PPO [28], while AligningT2I [13] leverages a reward model to weight training data. Policy gradient methods like DPOK [10] and DDPO [2] further improve model alignment. ReFL [35] and AlignProp [23] propagate gradients through differentiable reward models.

Inspired by the success of DPO [25] in LLMs, Diffusion-DPO [31] trains on image preference pairs, while D3PO [36] incorporates human-labeled preferences. SPO [15] enhances supervision with step-aware preference models. However, directly applying these methods to background generation often results in unstable training and suboptimal quality. To address this, we introduce improved sampling and training techniques for better performance.

10 Conclusion

We introduce **DEPO**, a novel framework tailored for generating e-commerce product backgrounds by addressing the trade-off between reward sparsity and supervision quality. By incorporating the DEPO loss, Langevin MCMC for expanded exploration, targeted exploration constraints, and adversarial training, DEPO ensures consistent, high-quality outputs aligned with human preferences. Extensive experiments demonstrate that DEPO significantly enhances both the quality and diversity of visually appealing product background generation, outperforming existing methods.

Acknowledgement

This work is supported by the National Key R&D Program of China under Grant No.2024QY1400, and the National Natural Science Foundation of China No. 62425604. This work is supported by Tsinghua University Initiative Scientific Research Program and the Institute for Guo Qiang at Tsinghua University. This work is also supported by Alibaba Group through Alibaba Innovative Research Program. We sincerely thank Qixin Wang for her assistance in creating the figures for this paper.

References

[1] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. 2020. Compositional GAN: Learning Image-Conditional Binary Composition. *Int. J. Comput. Vis.* 128, 10 (2020), 2570–2585. doi:10.1007/S11263-020-01336-9

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training Diffusion Models with Reinforcement Learning. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.

[3] Tingfeng Cao, Junsheng Kong, Xue Zhao, Wenqing Yao, Junwei Ding, Jinhui Zhu, and Jiandong Zhang. 2024. Product2IMG: Prompt-Free E-commerce Product Background Generation with Diffusion Model and Self-Improved LMM. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 10774–10783. doi:10.1145/3664647.3680753

[4] Chaofeng Chen, Annan Wang, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Enhancing Diffusion Models with Text-Encoder Reinforcement Learning. *arXiv preprint arXiv:2311.15657* (2023).

[5] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023. Subject-driven Text-to-Image Generation via Apprenticeship Learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.).

[6] Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, Yuanjie Shao, Xinge You, Changxin Gao, and Nong Sang. 2025. CTR-Driven Advertising Image Generation with Multimodal Large Language Models. *CoRR* abs/2502.06823 (2025). doi:10.48550/ARXIV.2502.06823 arXiv:2502.06823

[7] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. DoveNet: Deep Image Harmonization via Domain Verification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 8391–8400. doi:10.1109/CVPR42600.2020.00842

[8] Xiaodong Cun and Chi-Man Pun. 2020. Improving the Harmony of the Composite Image by Spatial-Separated Attention Module. *IEEE Trans. Image Process.* 29 (2020), 4759–4771. doi:10.1109/TIP.2020.2975979

[9] Zhenbang Du, Wei Feng, Haohan Wang, Yaoyu Li, Jingsen Wang, Jian Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junsheng Jin, et al. 2024. Towards Reliable Advertising Image Generation Using Human Feedback. *arXiv preprint arXiv:2408.00418* (2024).

[10] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2024. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[12] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569* (2023).

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

[14] Dongxu Li, Junnan Li, and Steven C. H. Hoi. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.).

[15] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. 2024. Step-aware Preference Optimization: Aligning Preference with Denoising Performance at Each Step. *arXiv preprint arXiv:2406.04314* (2024).

[16] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* (2017).

[17] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. 2018. ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 9455–9464. doi:10.1109/CVPR.2018.00985

[18] Shanchuan Lin, Anran Wang, and Xiao Yang. 2024. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929* (2024).

[19] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. 2021. Making Images Real Again: A Comprehensive Survey on Deep Image Composition. *CoRR* abs/2106.14490 (2021). arXiv:2106.14490

[20] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).

[21] Giorgio Parisi. 1981. Correlation functions and computer simulations. *Nuclear Physics B* 180, 3 (1981), 378–384.

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[23] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. 2023. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739* (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. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. doi:10.1109/CVPR52688.2022.01042

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

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

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[31] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2023. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908* (2023).

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

[33] Haohan Wang, Wei Feng, Yaoyu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, and Jingping Shao. 2025. Generate e-commerce product background by integrating category commonality and personalized style. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[34] Haohan Wang, Wei Feng, Yang Lu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junjie Shen, Zhangang Lin, Lixing Bo, and Jingping Shao. 2023. Generate E-commerce Product Background by Integrating Category Commonality and Personalized Style. *CoRR* abs/2312.13309 (2023). doi:10.48550/ARXIV.2312.13309 arXiv:2312.13309

[35] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. 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, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. 2023. Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model. *arXiv preprint arXiv:2311.13231* (2023).

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

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

[39] Yinan Zhao, Brian L. Price, Scott Cohen, and Danna Gurari. 2019. Unconstrained Foreground Object Search. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2030–2039. doi:10.1109/ICCV.2019.00212

A A Detailed Explanation for Table 1

The table compares different diffusion-based policy optimization methods across key aspects.

- No Discretization Error indicates whether the method avoids trajectory construction, which introduces discretization artifacts; Diffusion-DPO achieves this (✓), while others construct trajectories (✗).
- Exact Evaluation checks if the reward function is directly applied in the image domain ($\mathbf{x}_0$); all methods except SPO ensure this.
- Extra Exploration refers to additional techniques beyond standard sampling; only DEPO incorporates this (✓).
- Shared State assesses whether the method preserves trajectory-level information instead of breaking it into independent step pairs; only DEPO fully maintains shared state, while SPO does so partially.
- High-Frequency Reward determines whether the method applies frequent reward feedback; both Diffusion-DPO and DEPO optimize with high-frequency signals, whereas others do not.

Overall, DEPO stands out as the most comprehensive approach, integrating extra exploration, shared state preservation, and high-frequency reward, while maintaining exact evaluation despite discretization.

B Further Explanation about Figure 6

Here, we provide additional insights into the optimization process. Focusing on the terms containing θ, and considering the formulation inside the log in Equation (14), we observe that the optimization direction aims to maximize $\frac{p^+}{p^-}$.

As illustrated in Figure 6, all the depicted directions are possible. **However, in the unfavorable case indicated by the red arrow, p^+ is dragged down by p^-.** To mitigate the tendency of the results **moving toward the origin**, we detach the corresponding p, as also shown in Figure 6.

C Langevin MCMC Details

Langevin MCMC algorithm follows the following formulation:

$$\mathbf{x}_t^{i+1} = \mathbf{x}_t^i + \tau \cdot \mathbf{s}_\theta(\mathbf{x}_t^i, t) + \sqrt{2\tau} \cdot \mathbf{z}, \tag{13}$$

where $\tau = 0.02$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and we iterate it for 3 times for all divert timesteps at first 40% and in this way we can get best performance. Following such Langevin MCMC at t will change $\mathbf{x}_t^i$ but $\mathbf{x}_t^i$ still in the distribution at t.

D Proofs

D.1 Proof of Proposition 5.4

We provide the following proof of Proposition 5.4.

Proposition D.1. *We can directly obtain the expected* log *probability within the* log σ *by*

$$\mathbb{E}_{\mathbf{x}^\pm_{t_\tau-n-1}\sim p_{sg[\theta]}(\mathbf{x}_{t_\tau-n-1}|\mathbf{x}^\pm_{t_\tau-n})} \log \frac{p_\theta^{\mathbf{m}}\left(\mathbf{x}^+_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)}{p_\theta^{\mathbf{m}}\left(\mathbf{x}^-_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)} = \log \frac{p_\theta^{\mathbf{m}}(\mu_{sg[\theta],t_\tau-n-1}\left(\mathbf{x}^+_{t_\tau-n}\right) \mid \mathbf{x}_{t_\tau}\right)}{p_\theta^{\mathbf{m}}(\mu_{sg[\theta],t_\tau-n-1}\left(\mathbf{x}^-_{t_\tau-n}\right) \mid \mathbf{x}_{t_\tau}\right)}. \tag{14}$$

Proof. We want to show that

$$\mathbb{E}_{\mathbf{x}^\pm_{t_\tau-n-1}\sim p_{sg[\theta]}(\mathbf{x}_{t_\tau-n-1}|\mathbf{x}^\pm_{t_\tau-n})} \log \frac{p_\theta^{\mathbf{m}}\left(\mathbf{x}^+_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)}{p_\theta^{\mathbf{m}}\left(\mathbf{x}^-_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\right)} = \log \frac{p_\theta^{\mathbf{m}}\left(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^+_{t_\tau-n}) \mid \mathbf{x}_{t_\tau}\right)}{p_\theta^{\mathbf{m}}\left(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^-_{t_\tau-n}) \mid \mathbf{x}_{t_\tau}\right)}.$$

By hypothesis,

$$\mathbf{x}^\pm_{t_\tau-n-1} \sim \mathcal{N}\Big(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^\pm_{t_\tau-n}),\, \sigma^2_{t_\tau-n-1}\,\mathbf{I}\Big).$$

Hence we can write

$$\mathbf{x}^\pm_{t_\tau-n-1} = \mu_{sg[\theta],t_\tau-n-1}\big(\mathbf{x}^\pm_{t_\tau-n}\big) + \sigma_{t_\tau-n-1}\,\mathbf{z}^\pm, \quad \mathbf{z}^\pm \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

which leads to

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^\pm_{t_\tau-n-1}\sim p_{sg[\theta]}} \log \frac{p_\theta^{\mathbf{m}}(\mathbf{x}^+_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau})}{p_\theta^{\mathbf{m}}(\mathbf{x}^-_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau})} \\
&= \mathbb{E}_{\mathbf{z}^+,\mathbf{z}^-\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \log \frac{p_\theta^{\mathbf{m}}\Big(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^+_{t_\tau-n}) + \sigma_{t_\tau-n-1}\,\mathbf{z}^+ \,\Big|\, \mathbf{x}_{t_\tau}\Big)}{p_\theta^{\mathbf{m}}\Big(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^-_{t_\tau-n}) + \sigma_{t_\tau-n-1}\,\mathbf{z}^- \,\Big|\, \mathbf{x}_{t_\tau}\Big)}.
\end{aligned}$$

Assume that $p_\theta^{\mathbf{m}}\big(\mathbf{x}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\big)$ is Gaussian with mean $\bar{\mu}(\mathbf{x}_{t_\tau})$ and covariance $\bar{\Sigma}$, i.e.,

$$p_\theta^{\mathbf{m}}\big(\mathbf{x}_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}\big) = \mathcal{N}\Big(\mathbf{x}_{t_\tau-n-1};\, \bar{\mu}(\mathbf{x}_{t_\tau}),\, \bar{\Sigma}\Big).$$

Then, up to a constant in $\mathbf{x}_{t_\tau-n-1}$,

$$\log p_\theta^{\mathbf{m}}(\mathbf{x}_{t_\tau-n-1} \,|\, \mathbf{x}_{t_\tau}) \;=\; -\tfrac{1}{2}\,(\mathbf{x}_{t_\tau-n-1}-\bar{\mu})^\top \bar{\Sigma}^{-1}\,(\mathbf{x}_{t_\tau-n-1}-\bar{\mu}) \;+\; \text{const}(\mathbf{x}_{t_\tau}).$$

Let us define

$$\mu^+ := \mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^+_{t_\tau-n}), \qquad \mu^- := \mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^-_{t_\tau-n}), \qquad \sigma := \sigma_{t_\tau-n-1}.$$

Then $\mathbf{x}^\pm_{t_\tau-n-1} = \mu^\pm + \sigma\,\mathbf{z}^\pm$, and

$$\log p_\theta^{\mathbf{m}}(\mathbf{x}^\pm_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau}) \;=\; -\tfrac{1}{2}\,((\mu^\pm+\sigma\,\mathbf{z}^\pm)-\bar{\mu})^\top\bar{\Sigma}^{-1}((\mu^\pm+\sigma\,\mathbf{z}^\pm)-\bar{\mu}) \;+\; C,$$

where C is a constant that does not depend on $\mathbf{z}^\pm$. Denote $\Delta^\pm := \mu^\pm - \bar{\mu}$. Then

$$(\mu^\pm+\sigma\,\mathbf{z}^\pm-\bar{\mu}) \;=\; \Delta^\pm \;+\; \sigma\,\mathbf{z}^\pm,$$

which expands as

$$(\Delta^\pm+\sigma\,\mathbf{z}^\pm)^\top\bar{\Sigma}^{-1}(\Delta^\pm+\sigma\,\mathbf{z}^\pm) \;=\; (\Delta^\pm)^\top\bar{\Sigma}^{-1}\Delta^\pm \;+\; 2\,(\Delta^\pm)^\top\bar{\Sigma}^{-1}\,\sigma\,\mathbf{z}^\pm \;+\; \sigma^2\,(\mathbf{z}^\pm)^\top\bar{\Sigma}^{-1}\mathbf{z}^\pm.$$

Taking the expectation over $\mathbf{z}^\pm \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, we have $\mathbb{E}[\mathbf{z}^\pm]=\mathbf{0}$ and $\mathbb{E}[\mathbf{z}^\pm\mathbf{z}^{\pm\top}]=\mathbf{I}$. Thus,

$$\mathbb{E}_{\mathbf{z}^\pm}\Big[(\Delta^\pm+\sigma\,\mathbf{z}^\pm)^\top\bar{\Sigma}^{-1}(\Delta^\pm+\sigma\,\mathbf{z}^\pm)\Big] \;=\; (\Delta^\pm)^\top\bar{\Sigma}^{-1}\,\Delta^\pm \;+\; \sigma^2\,\mathrm{Tr}(\bar{\Sigma}^{-1}).$$

Consequently,

$$\mathbb{E}_{\mathbf{z}^\pm}\big[\log p_\theta^{\mathbf{m}}(\mu^\pm+\sigma\,\mathbf{z}^\pm \mid \mathbf{x}_{t_\tau})\big] \;=\; -\tfrac{1}{2}\,(\Delta^\pm)^\top\bar{\Sigma}^{-1}\,\Delta^\pm \;-\; \tfrac{1}{2}\,\sigma^2\,\mathrm{Tr}(\bar{\Sigma}^{-1}) \;+\; C.$$

(Notice the second term above is just a constant in $\mu^\pm$, so what really matters is the $(\Delta^\pm)^\top\bar{\Sigma}^{-1}\Delta^\pm$ part, which depends on $\mu^\pm$.)

Now we take the difference for "+" versus "−":

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}^+}\Big[\log p_\theta^{\mathbf{m}}(\mu^+ + \sigma\,\mathbf{z}^+ \mid \mathbf{x}_{t_\tau})\Big] \;-\; \mathbb{E}_{\mathbf{z}^-}\Big[\log p_\theta^{\mathbf{m}}(\mu^- + \sigma\,\mathbf{z}^- \mid \mathbf{x}_{t_\tau})\Big] \\
&\quad= \;-\tfrac{1}{2}\,(\Delta^+)^\top\bar{\Sigma}^{-1}\,\Delta^+ \;-\; \tfrac{1}{2}\,\sigma^2\,\mathrm{Tr}(\bar{\Sigma}^{-1}) \;+\; C \;-\; \Big[-\tfrac{1}{2}\,(\Delta^-)^\top\bar{\Sigma}^{-1}\,\Delta^- \;-\; \tfrac{1}{2}\,\sigma^2\,\mathrm{Tr}(\bar{\Sigma}^{-1}) \;+\; C\Big] \\
&\quad= \;-\tfrac{1}{2}\,(\Delta^+)^\top\bar{\Sigma}^{-1}\,\Delta^+ \;+\; \tfrac{1}{2}\,(\Delta^-)^\top\bar{\Sigma}^{-1}\,\Delta^- \;=\; \log\frac{p_\theta^{\mathbf{m}}(\mu^+ \mid \mathbf{x}_{t_\tau})}{p_\theta^{\mathbf{m}}(\mu^- \mid \mathbf{x}_{t_\tau})},
\end{aligned}$$

where the last equality follows from the fact that

$$\log p_\theta^{\mathbf{m}}(\mathbf{y} \mid \mathbf{x}_{t_\tau}) \;=\; -\tfrac{1}{2}\,(\mathbf{y}-\bar{\mu})^\top\bar{\Sigma}^{-1}(\mathbf{y}-\bar{\mu}) \;+\; \text{const in } \mathbf{y}.$$

Hence

$$\log\frac{p_\theta^{\mathbf{m}}(\mu^+ \mid \mathbf{x}_{t_\tau})}{p_\theta^{\mathbf{m}}(\mu^- \mid \mathbf{x}_{t_\tau})} \;=\; -\tfrac{1}{2}\,(\mu^+-\bar{\mu})^\top\bar{\Sigma}^{-1}(\mu^+-\bar{\mu}) \;+\; \tfrac{1}{2}\,(\mu^--\bar{\mu})^\top\bar{\Sigma}^{-1}(\mu^--\bar{\mu}),$$

which exactly matches

$$\mathbb{E}_{\mathbf{z}^+,\mathbf{z}^-}\big[\log p_\theta^{\mathbf{m}}(\mu^+ + \sigma\,\mathbf{z}^+ \mid \mathbf{x}_{t_\tau}) - \log p_\theta^{\mathbf{m}}(\mu^- + \sigma\,\mathbf{z}^- \mid \mathbf{x}_{t_\tau})\big].$$

Therefore, returning to the notation $\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^\pm_{t_\tau-n})$ for $\mu^\pm$, we have shown that

$$\mathbb{E}_{\mathbf{x}^\pm_{t_\tau-n-1}}\log\frac{p_\theta^{\mathbf{m}}(\mathbf{x}^+_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau})}{p_\theta^{\mathbf{m}}(\mathbf{x}^-_{t_\tau-n-1} \mid \mathbf{x}_{t_\tau})} \;=\; \log\frac{p_\theta^{\mathbf{m}}(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^+_{t_\tau-n}) \mid \mathbf{x}_{t_\tau})}{p_\theta^{\mathbf{m}}(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^-_{t_\tau-n}) \mid \mathbf{x}_{t_\tau})},$$

□

With the claim above, we have:

$$\mathcal{L}_{\text{DEPO}} = -\mathbb{E}_{\tau\sim\mathcal{D}}\log\sigma\left(\beta\log\frac{p_\theta^{\mathbf{m}}\left(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^+_{t_\tau-n}) \mid \mathbf{x}_{t_\tau}\right)}{p_{\text{ref}}\left(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^+_{t_\tau-n}) \mid \mathbf{x}_{t_\tau}\right)} - \beta\log\frac{p_\theta^{\mathbf{m}}\left(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^-_{t_\tau-n}) \mid \mathbf{x}_{t_\tau}\right)}{p_{\text{ref}}\left(\mu_{sg[\theta],t_\tau-n-1}(\mathbf{x}^-_{t_\tau-n}) \mid \mathbf{x}_{t_\tau}\right)}\right), \tag{15}$$

where the $p^{\mathbf{m}}(\cdot\mid\cdot)$ and $p_{\text{ref}}(\cdot\mid\cdot)$ are the probability of $\mathbf{x}_{t_\tau-n-1}$ given $\mathbf{x}_{t_\tau}$.

D.2 Proof of Proposition 5.2

Proposition D.2. *According to DDIM [29],*

$$\mathbb{E}\left[\mathbf{x}_{t-1} \,\middle|\, \mathbf{x}_t\right] \;=\; \alpha(t)\,\widehat{\mathbf{x}_0}(\mathbf{x}_t) + \beta(t)\mathbf{x}_t$$

where $\alpha(t)$ and $\beta(t)$ are deterministic scalars that depend only on t and can represent the evolution of $\mathbf{x}_{t-1}$.

Proof. According to the Equation (12) in DDIM paper [29], we have:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\alpha_t}}\right)}_{\text{" predicted } \mathbf{x}_0\text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(\mathbf{x}_t)}_{\text{"direction pointing to } \mathbf{x}_t\text{"}} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}} \tag{16}$$

We define $\alpha(t) = \sqrt{\alpha_{t-1}}$, $\beta(t) = \sqrt{1-\alpha_{t-1}-\sigma_t^2}$, then have

$$\mathbb{E}\left[\mathbf{x}_{t-1} \,\middle|\, \mathbf{x}_t\right] \;=\; \alpha(t)\,\widehat{\mathbf{x}_0}(\mathbf{x}_t) + \beta(t)\epsilon_\theta^{(t)}(\mathbf{x}_t)$$

□

D.3 Proof of Lemma 5.3

Lemma D.3 (Accurate Evaluation for Diffusion Models)**.** *The reward model applied to the diffusion model's one-step prediction, $\widehat{\mathbf{x}_0}$, aligns more closely with $\mathbb{E}\left[\mathbf{x}_{t-1}|\mathbf{x}_t\right]$ than with samples from $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, and more so than using $\mathbf{x}_t$ alone.*

Proof. According to the Equation (12) in DDIM paper [29], we have:

$$\mathbf{x}_t = \sqrt{\alpha_t}\widehat{\mathbf{x}_0}(\mathbf{x}_t) + \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\mathbf{x}_t) \tag{17}$$

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\widehat{\mathbf{x}_0}(\mathbf{x}_t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta^{(t)}(\mathbf{x}_t) + \sigma_t\epsilon_t \tag{18}$$

$$\mathbb{E}\left[\mathbf{x}_{t-1} \,\middle|\, \mathbf{x}_t\right] = \sqrt{\alpha_{t-1}}\widehat{\mathbf{x}_0}(\mathbf{x}_t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta^{(t)}(\mathbf{x}_t) \tag{19}$$

Apparently that $\mathbb{E}\left[\mathbf{x}_{t-1} \,\middle|\, \mathbf{x}_t\right]$ is better than $\mathbf{x}_{t-1}$ due to smaller variance, and $\mathbf{x}_{t-1}$ is better than $\mathbf{x}_t$ because $\alpha_{t-1} > \alpha_t$. □

E Sampling Process

The sampling process is shown in Algorithm 1. For brevity, we omit the case where there are insufficient timesteps for an n-length trajectory, resulting in early termination, where we will directly use such a shorter trajectory. In this work, $r_{tra} = 0.5, r_{time} = 0.5$.

Algorithm 1 Sampling Process

Require: Divert timesteps range t_1 and t_2
Require: Trajectory length n
Require: Filter Rate over Trajectories r_{tra}
Require: Filter Rate over Different Timesteps r_{time}
Require: Control condition $\mathbf{c}$
Require: Reward Model $s(\cdot)$
Sample $t_s \in [t_1, t_2]$, $l_{tra} = []$
Following DDPM, get $\mathbf{x}_{t_s}$
repeat
  Following DDPM and Langevin MCMC, get $\frac{2}{r_{tra}}$ samples of $\mathbf{x}_{t_s-1}$
  Following DDPM n timesteps, get $\frac{2}{r_{tra}}$ trajectories
  Random select $\mathbf{x}_{s_t-n}$ from the trajectories
  Predict $\mathbf{x}_0$ for each trajectories, calculate $s(x_0)$ and then select the best one and worst one as τ^+, τ^-
  Append $\{\tau^+, \tau^-\}$ to l_{tra}
  $t_s = t_s - n$
until $t_s <= 0$
Sort l_{tra} based on the descending order of $|s(x_0^+) - s(x_0^-)|$ and then filter them with ratio r_{time}.
Return l_{tra}

Algorithm 2 Training Algorithm

Require: Collected data l_{tra}
Require: Model parameters on training θ
Require: Model parameters for reference model θ_{ref}
repeat
  Pop $\tau \in l_{tra}$
  Calculate Equation (15) as $\mathcal{L}_{depo}$
  Calculate Equation (9) as $\mathcal{L}_{adv}$
  Calculate Equation (7) as $\mathcal{L}_{reg}$
  $\mathcal{L} = \mathcal{L}_{\text{DEPO}} + c_{reg}\mathcal{L}_{reg} + c_{adv}\mathcal{L}_{adv}$
  Optimize θ with $\mathcal{L}$
until l_{tra} is empty

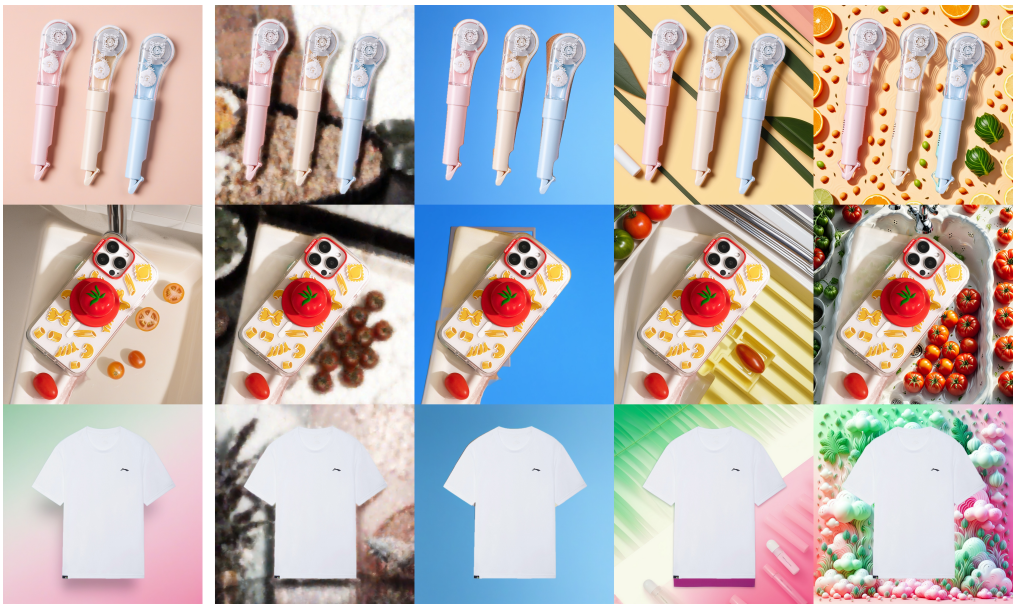

Figure 10: Visualization of mode collapse.

F Training Algorithm

The sampling process is shown in Algorithm 2. For brevity, the optimization process of the Discriminator is omitted.

G More Implementation Details

As shown in Section 7, we use CLIP as the backbone of our reward model, and train the vision and text encoder on the Pairwise User Preference Dataset. To build the discriminator, we use the frozen UNet of the diffusion model as the backbone, followed by several trainable convolutional and pooling layers. The product mask is produced by a private segmentation model, which cat segment the foreground product from the product image provided by the user. We use a slightly fine-tuned SDXL as the base model, with specific design choices tailored to this setting.

H Mode Collapse

In Figure 10, we demonstrate the impact of mode collapse on the background generation. The first column presents normal background generation, serving as a baseline for comparison. The subsequent four columns exhibit various mode collapses. The second column features backgrounds marred by significant noise. The third column displays images with solid blue backgrounds. In the fourth column, the images have backgrounds with diagonal lines extending from the bottom right corner to the top left corner. The fifth column includes images with backgrounds filled with numerous elements. Our method can effectively alleviate mode collapse.

Table 4: Evaluation metrics for collapsed and non-collapsed patterns. Lower scores indicate degraded quality under mode collapse. Figure references denote corresponding visual examples in Figure 10.

Pattern Type	EPS↑	PickScore↑	CLIP Score↑
Blurred & Chaotic (2nd column)	14.50	20.09	23.01
Solid Color (3rd column)	13.61	19.92	21.49
Stripped (4th column)	16.39	20.78	25.76
Dotted (5th column)	15.81	20.85	26.66
Not Collapsed (DEPO)	**17.86**	**21.24**	**27.64**

The image is softly lit with a warm, glowing ambiance, suggesting a serene and inviting atmosphere.	a section of a living room.	a framed display standing on a light-colored wooden surface, such as a table or shelf.	two cylindrical objects placed on a natural background of dried leaves and twigs.	A set of wooden drawers is positioned against a white wall.	The product is situated on a round wooden coaster, which is placed on a striped yellow and white cloth.	a gray and yellow bathroom floor mat placed on a gray tiled floor.
a setting with a mug and a wrapped item placed on a tabletop.	a well-lit indoor setting.	a rectangular object with an attached tassel featuring intricate knotwork and beads.	The product is placed on top of an open book, with part of a grayscale image of hands and an arm visible on the pages.	a set of cylindrical-shaped, vertically ribbed containers in various pastel colors including white, pink, blue, and gray.	The scene features a pair of tall wooden side tables situated against a light-colored wall.	The product features a textured wall with a combination of a vertical wood slat pattern on the left and a tiled pattern on the right.

Figure 11: The prompts of images in Figure 1.

I Additional Visualizations

In Figure 12, we presented the additional images generated by our method. Each example consists of three rows: the first row contains the foreground images, the second row features the images generated by the base model, and the third row displays the images generated by our DEPO.

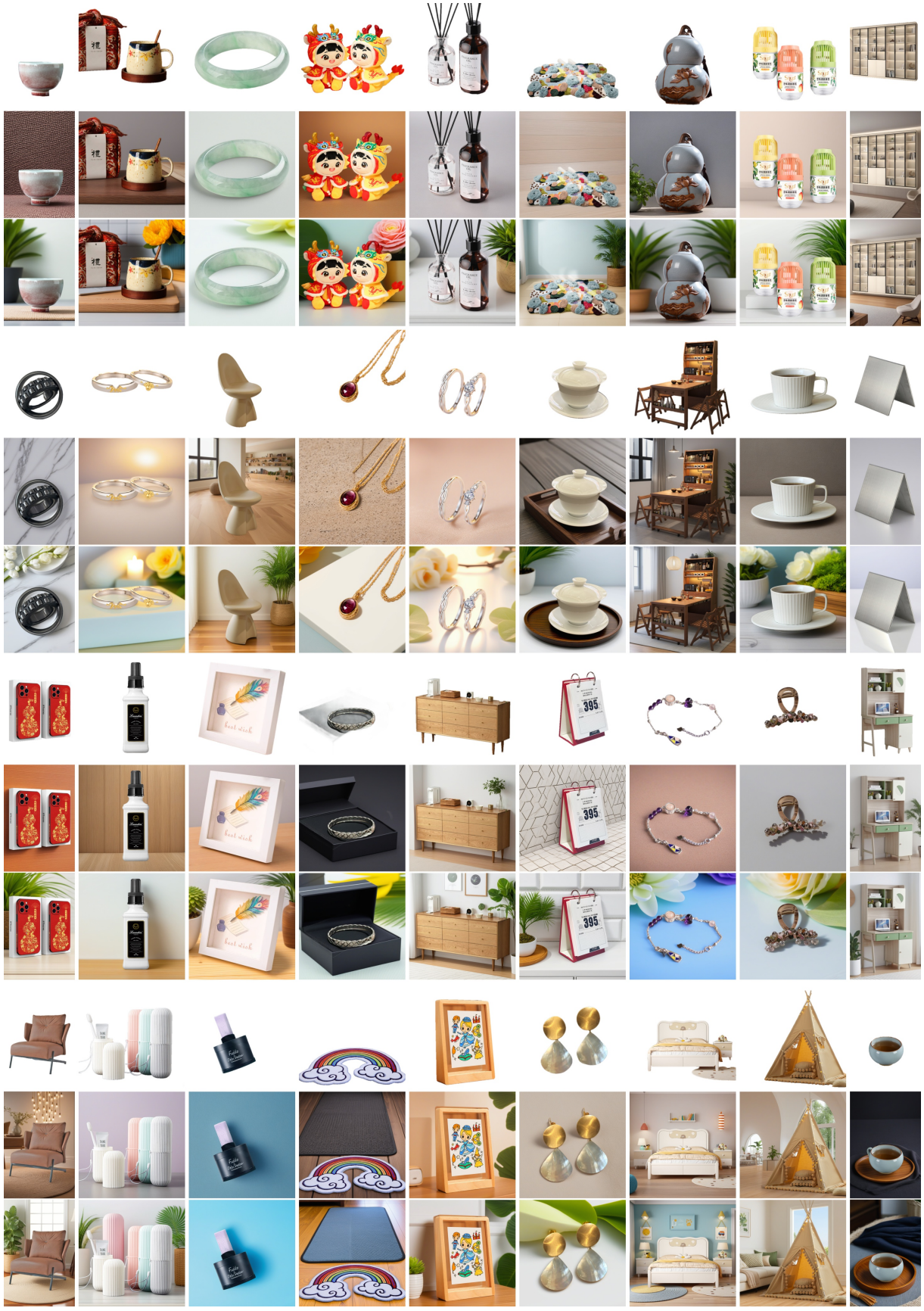

Figure 12: Visualization of foreground images, images generated by baseline, and images generated by our DEPO.