DDSPO: DIRECT DIFFUSION SCORE PREFERENCE OPTIMIZATION VIA STEPWISE CONTRASTIVE POLICY-PAIR SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have achieved impressive results in generative tasks such as text-to-image synthesis, yet they often struggle to fully align outputs with nuanced user intent and maintain consistent aesthetic quality. Existing preference-based training methods like Diffusion Direct Preference Optimization help address these issues but rely on costly and potentially noisy human-labeled datasets. In this work, we introduce Direct Diffusion Score Preference Optimization (DDSPO), which—when winning/losing policies are accessible—directly derives per-timestep supervision from these policies. Unlike prior methods that operate solely on final samples, DDSPO provides dense, transition-level signals across the denoising trajectory. In practice, we avoid reliance on labeled data by automatically generating preference signals using a pretrained reference model: we contrast its outputs when conditioned on original prompts versus semantically degraded variants. This practical strategy enables effective score-space preference supervision without explicit reward modeling or manual annotations. Empirical results demonstrate that DDSPO improves text-image alignment and visual quality, outperforming or matching existing preference-based methods while requiring significantly less supervision.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Rombach et al., 2022) have achieved impressive results across a variety of generative tasks, particularly in text-to-image synthesis (Rombach & Esser, 2022b; Podell et al., 2024; Saharia et al., 2022; Ramesh et al., 2022). Despite this progress, they often struggle to fully align generated outputs with nuanced user intent and to consistently produce aesthetically high-quality images. Addressing these shortcomings typically requires task-specific training data, which can be difficult and costly to obtain. As an alternative, human preference annotations provide a lightweight yet expressive means of encoding such qualitative information in the form of ranked comparisons. Leveraging this, recent approaches have incorporated human preferences into the training process, such as through reinforcement learning with human feedback (RLHF). A notable example is Diffusion DPO (Wallace et al., 2024), which extends the Direct Preference Optimization (DPO) framework (Rafailov et al., 2023)—originally developed for language and vision-language models (Rafailov et al., 2023; Xing et al., 2025; Xie et al., 2024b)—to diffusion models. DPO enables models to learn directly from preference comparisons without requiring reward models. While effective, this approach relies heavily on labeled preference datasets, which are expensive to curate and susceptible to noise and inconsistency, limiting their scalability and reliability.

To overcome these limitations, we propose **Direct Diffusion Score Preference Optimization (DDSPO)**, a generalized variant of Diffusion DPO that formulates preference supervision in score space rather than in the final sample space. Unlike prior approaches that define preferences over final samples $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l)$, DDSPO incorporates preference signals into the denoising process at each timestep using denoising scores from winning and losing policies. This is achieved by extending preference labels to tuples $((\boldsymbol{x}_t^w, \boldsymbol{x}_{t-1}^w), (\boldsymbol{x}_t^l, \boldsymbol{x}_{t-1}^l))$, drawn from preferred and dispreferred denoising policies. We define a score-space loss that optimizes the model's denoising predictions toward targets $\boldsymbol{\epsilon}_\star^w$ and $\boldsymbol{\epsilon}_\star^l$, reflecting preferred and dispreferred behavior. By deriving per-timestep targets

from the winning and losing policies , DDSPO enables richer supervision across denoising steps and reduces reliance on final outputs $\boldsymbol{x}_0$, improving robustness to imperfect preference data.

In practice, we define preferred and dispreferred denoising behavior using the outputs of a pretrained reference model conditioned on original and perturbed prompts, respectively. Given an input prompt $\boldsymbol{c}$ and its semantically degraded variant $\boldsymbol{c}^-$, we construct pseudo preference signals as follows: the reference model's score prediction $\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w, t, \boldsymbol{c})$ is treated as the target for preferred denoising behavior, while $\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c}^-)$ represents a dispreferred direction. The model is then trained to match the preferred score and to actively avoid the dispreferred score This stepwise contrastive signal allows DDSPO to steer the model toward desirable generation behavior without requiring explicit preference annotations or reward modeling.

Despite the absence of labeled preference data, DDSPO consistently improves generation quality. It enhances text-image alignment by better capturing user intent and produces more coherent and visually appealing outputs. With only minimal supervision, DDSPO achieves results competitive with or superior to existing preference-based approaches, demonstrating its practicality and broad applicability. Our main contributions are threefold:

- We propose DDSPO, a preference optimization framework that generalizes Diffusion DPO by enabling direct supervision over intermediate denoising steps in score space.
- Along with introducing timestep-level preference supervision, we propose a practical method for constructing such signals using prompt perturbation and a reference model.
- Through extensive experiments, we demonstrate that DDSPO improves semantic alignment and visual quality, without relying on human-labeled preference data.

2 Related work

Improving Diffusion Model Several lines of research (Kirstain et al., 2023a; Xu et al., 2023; Black et al., 2024; Lee et al., 2025b; Fan et al., 2023; Yang et al., 2024) have sought to improve diffusion models by incorporating external signals derived from human preferences. These approaches often rely on pretrained or learned reward models (Kirstain et al., 2023a; Xu et al., 2023; Black et al., 2024; Lee et al., 2025b), and formulate the generation process as a reinforcement learning (RL) problem (Yang et al., 2024; Black et al., 2024; Fan et al., 2023; Gu et al., 2024). By treating denoising as a sequential decision-making task, such methods aim to align model behavior with complex user intent, leveraging richer supervision than simple binary preferences. More recent work also explores combining multiple reward models or iteratively refining preferences (Zhang et al., 2025; Lee et al., 2025b; Zhao et al., 2025; Clark et al., 2023; Hao et al., 2023; Prabhudesai et al., 2023). However, these methods still depend heavily on explicitly labeled preference data or pretrained reward models, incurring substantial annotation costs and computational overhead. In contrast, our work introduces a scalable alternative that requires neither reward models nor preference supervision, instead aligning diffusion models via unsupervised self-consistency across multiple prompt variants.

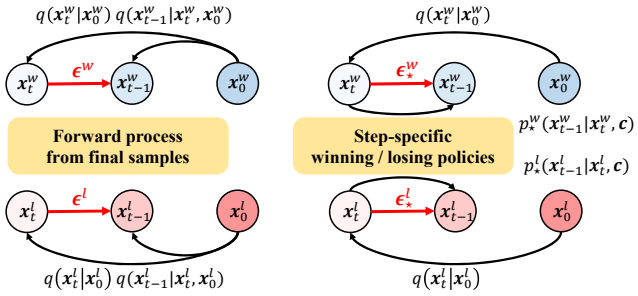
Direct Preference Optimization (DPO) Direct Preference Optimization (DPO)(Rafailov et al., 2023) has recently gained traction as a more direct and efficient way to align generative models with human preferences. Originally proposed for aligning large language models and vision-language models, DPO fine-tunes models using only paired preference data, bypassing the need for explicitly trained reward models (Xing et al., 2025; Liu et al., 2024). This leads to simpler training pipelines and improved alignment stability compared to reinforcement learning from human feedback. Building on this foundation, researchers have adapted DPO to the diffusion model framework (Croitoru et al., 2024; Wallace et al., 2024; Zhu et al., 2025; Zhang et al., 2024; Liang et al., 2025). In Diffusion DPO (Wallace et al., 2024), preference information is incorporated throughout the denoising trajectory. Zhu et al. (2025) reinterprets this via score matching to align pretraining and fine-tuning objectives; however, because its targets are approximated via the forward process as well, the supervision effectively comes only from final samples and does not provide timestep-level signals. Furthermore, Liang et al. (2025) selects win–lose pairs at each denoising step using a timestep-aware preference model. However, it requires a separately trained intermediate reward model and is incompatible with deterministic flow-matching, yielding only limited improvements in text-to-image alignment. Distinct from prior approaches, our method introduces preference supervision

over denoising transitions by modeling preferences across $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$ pairs from winning and losing policies, thereby capturing alignment signals throughout the generation trajectory—while requiring no human-labeled preference data. See Appendix for extended related-work discussion

Self-Training Recent advances have shown that generative models can be effectively aligned with human preferences through self-training, without relying on paired preference data or externally trained reward models (Zhu et al., 2024; Lee et al., 2025a; Yuan et al., 2024; Chen et al., 2024; Majumder et al., 2024; Deng et al., 2024; He et al., 2019; Xie et al., 2020; Wei et al., 2020; Zoph et al., 2020; Sohn et al., 2020; Ghiasi et al., 2021; Kang et al., 2023). Especially for diffusion models, these methods typically involve generating synthetic preference signals from the model's own outputs—contrasting higher-quality generations with intentionally degraded ones (Zhu et al., 2024; Deng et al., 2024; Majumder et al., 2024) or leveraging iterative refinements across model checkpoints (Yuan et al., 2024). By treating the model's stronger outputs as preferred examples and weaker ones as negatives, they enable preference-aware fine-tuning through supervised objectives such as Direct Preference Optimization (Majumder et al., 2024; Yuan et al., 2024). In some cases, contrastive learning (Lee et al., 2025a) or prompt relabeling (Chen et al., 2024) further reinforces semantic consistency and corrects misalignments. Our work follows this direction by using prompting techniques to automatically construct both positive and negative samples—not only final image pairs but also intermediate noise pairs—enabling the generalized DDSPO framework to operate without any human-labeled preference data.

3 METHOD

Our objective is to propose a novel optimization formulation that aligns pretrained diffusion models with desirable generation behavior—such as improved text-image alignment or aesthetic quality. Through this formulation, we guide pretrained diffusion models toward desired generation behaviors without relying on human-annotated preferences. We first review Direct Preference Optimization (DPO)(Rafailov et al., 2023) and its extension to diffusion models in Sec. 3.1, which assume access to preference-labeled final samples $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l \mid \boldsymbol{c})$. In Sec. 3.2, we introduce Direct Diffusion Score Preference Optimization (DDSPO), which directly optimizes denoising behavior at each timestep using preference-labeled denoising scores derived from winning and losing policies. It can be seen as a generalization of Diffusion DPO(Wallace et al., 2024), which approximates such policies via the forward diffusion process from preferred and dispreferred samples Fig. 1 illustrates the difference in supervision targets—$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ vs. $p_\star(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c})$. Sec. 3.3 describes a practical method for constructing such score-level preferences by leveraging a frozen pretrained reference model conditioned on perturbed prompts, eliminating the need for explicit labels.



(a) Baseline (b) DDSPO

Figure 1: **Final-sample vs. Policy-derived supervision.**

3.1 PRELIMINARY

Direct Preference Optimization Direct Preference Optimization (DPO) is a learning framework that aligns model outputs with human preferences without requiring explicit reward supervision. Given a conditioning input $\boldsymbol{c}$, we assume access only to preference-labeled pairs $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l \mid \boldsymbol{c})$, where we write $\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l$ to indicate that $\boldsymbol{x}_0^w$ is preferred over $\boldsymbol{x}_0^l$. The interpretation of $\boldsymbol{c}$ and $\boldsymbol{x}_0$ depends on the task, where $\boldsymbol{c}$ is a text caption and $\boldsymbol{x}_0$ is a synthesized image in text-to-image generation. Such preferences may reflect various notions of quality—such as semantic alignment, aesthetic appeal, or factual correctness—depending on the task context.

These preferences can be formalized using the Bradley–Terry model, which defines the following distribution:

$$\mathbb{P}(\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l \mid \boldsymbol{c}) = \sigma\left(r(\boldsymbol{c}, \boldsymbol{x}_0^w) - r(\boldsymbol{c}, \boldsymbol{x}_0^l)\right), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $r(\boldsymbol{c}, \boldsymbol{x}_0)$ is a latent reward function that is difficult to access directly.

DPO reparameterizes this setup to directly optimize the generation model distribution. Starting from the RLHF-style KL-regularized objective, which maximizes reward while constraining the learned distribution to stay close to a fixed reference distribution $p_{\text{ref}}(\boldsymbol{x}_0 \mid \boldsymbol{c})$ where β is a parameter that balances reward maximization and closeness to the reference distribution.:

$$\max_\theta \mathbb{E}_{\boldsymbol{x}_0 \sim p_\theta(\boldsymbol{x}_0|\boldsymbol{c})}\left[r(\boldsymbol{c}, \boldsymbol{x}_0)\right] - \beta\,\mathrm{KL}\left(p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{c}) \,\|\, p_{\text{ref}}(\boldsymbol{x}_0 \mid \boldsymbol{c})\right), \tag{2}$$

This objective admits a closed-form solution where the optimal distribution $p_\theta^*(\boldsymbol{x}_0 \mid \boldsymbol{c})$ is proportional to the reference distribution $p_{\text{ref}}(\boldsymbol{x}_0 \mid \boldsymbol{c})$ scaled by an exponential of the reward: $p_\theta^*(\boldsymbol{x}_0 \mid \boldsymbol{c}) \propto p_{\text{ref}}(\boldsymbol{x}_0 \mid \boldsymbol{c}) \cdot \exp(r(\boldsymbol{c}, \boldsymbol{x}_0)/\beta)$. This reformulation reveals that the reward function can be implicitly captured by the ratio between the optimized policy and the reference distribution, eliminating the need to model $r(\boldsymbol{c}, x_0)$ explicitly. This leads to the DPO training objective:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l} \log \sigma\left(\beta \log \frac{p_\theta(\boldsymbol{x}_0^w \mid \boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_0^w \mid \boldsymbol{c})} - \beta \log \frac{p_\theta(\boldsymbol{x}_0^l \mid \boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_0^l \mid \boldsymbol{c})}\right). \tag{3}$$

DPO for Diffusion Models Applying DPO to diffusion models introduces a unique challenge: directly computing the log-likelihood ratio $\log \frac{p_\theta(x_0|\boldsymbol{c})}{p_{\text{ref}}(x_0|\boldsymbol{c})}$ is intractable due to the need to marginalize over all possible diffusion trajectories $x_{1:T}$ that generate x_0. To circumvent this, (Wallace et al., 2024) reformulate the objective over entire denoising paths and approximate the reverse process $p_\theta(x_{1:T} \mid x_0)$ using the forward noising process $q(x_{1:T} \mid x_0)$. This reparameterization enables training in the *score space*, where models learn to predict the noise (or equivalently, the score function (Ho et al., 2020; Song et al., 2021)) that guides the denoising process:

$$\begin{aligned}\mathcal{L}(\theta) = -\,\mathbb{E}_{(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0,T), \boldsymbol{x}_t^w \sim q(\boldsymbol{x}_t^w|\boldsymbol{x}_0^w), \boldsymbol{x}_t^l \sim q(\boldsymbol{x}_t^l|\boldsymbol{x}_0^l)} \log \sigma\Big(-\beta \cdot \big[\\ \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t, \boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w, t, \boldsymbol{c})\|_2^2 - (\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2)\big] \Big)\end{aligned} \tag{4}$$

where $\boldsymbol{x}_t^* = \alpha_t \boldsymbol{x}_0^* + \sigma_t \boldsymbol{\epsilon}^*$, with $\boldsymbol{\epsilon}^* \sim \mathcal{N}(0, I)$ sampled from the forward process $q(\boldsymbol{x}_t^* \mid \boldsymbol{x}_0^*)$. This objective encourages the student to better denoise preferred samples than dispreferred ones, aligning its behavior with human preferences while staying close to the reference model. In particular, the term $\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t, \boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w, t, \boldsymbol{c})\|_2^2$ encourages the student to outperform the reference model on *winning samples*, while the term $\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2$ encourages it to underperform the reference model on *losing samples*, thereby amplifying the relative preference signal. Full derivations of both the standard and diffusion DPO objectives are provided in the Appendix.

3.2 Direct Diffusion Score Preference Optimization

We now consider a setting in which preference is defined not over final generated samples, but over a transition between intermediate denoising steps. Crucially, rather than assuming access only to preference-labeled final samples $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l \mid \boldsymbol{c})$, we also assume access to preference-labeled denoising transitions at intermediate steps. Concretely, we define the winning and losing denoising policies $p_\star^w(\boldsymbol{x}_{t-1,t} \mid \boldsymbol{c})$ and $p_\star^l(\boldsymbol{x}_{t-1,t} \mid \boldsymbol{c})$; by sampling from these policies, we obtain tuples $((\boldsymbol{x}_t^w, \boldsymbol{x}_{t-1}^w), (\boldsymbol{x}_t^l, \boldsymbol{x}_{t-1}^l) \mid \boldsymbol{c})$. This enables direct preference of denoising behavior at each timestep of the diffusion process.

Following the Bradley–Terry formulation in Eq. (1), we extend the preference supervision from final outputs $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l \mid \boldsymbol{c})$ to denoising transitions at intermediate steps, using tuples of the form $((\boldsymbol{x}_t^w, \boldsymbol{x}_{t-1}^w), (\boldsymbol{x}_t^l, \boldsymbol{x}_{t-1}^l) \mid \boldsymbol{c})$. The reward function is accordingly redefined as $r(\boldsymbol{c}, \boldsymbol{x}_t, \boldsymbol{x}_{t-1})$, and preferences are modeled over entire denoising transitions—assigning higher preference to the trajectory $(\boldsymbol{x}_t^w, \boldsymbol{x}_{t-1}^w)$ over $(\boldsymbol{x}_t^l, \boldsymbol{x}_{t-1}^l)$.

By applying the same derivation steps as in Eqs. (1) to (3), and extending the supervision to denoising transitions at arbitrary diffusion timesteps $t \sim \mathcal{U}(0, T)$ the loss can be analogously formulated as:

$$\mathcal{L}_{\text{DDSPO}}(\theta) = -\mathbb{E}_{\boldsymbol{c}\sim\mathcal{D}(c),\, t\sim\mathcal{U}(0,T),\, (\boldsymbol{x}^w_{t-1},\boldsymbol{x}^w_t)\sim p^w_\star(\boldsymbol{x}^w_{t-1,t}|\boldsymbol{c}),\, (\boldsymbol{x}^l_{t-1},\boldsymbol{x}^l_t)\sim p^l_\star(\boldsymbol{x}^l_{t-1,t}|\boldsymbol{c})} \log\sigma\left(\beta\log\frac{p_\theta(\boldsymbol{x}^w_{t-1}\mid\boldsymbol{x}^w_t,\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}^w_{t-1}\mid\boldsymbol{x}^w_t,\boldsymbol{c})} - \beta\log\frac{p_\theta(\boldsymbol{x}^l_{t-1}\mid\boldsymbol{x}^l_t,\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}^l_{t-1}\mid\boldsymbol{x}^l_t,\boldsymbol{c})}\right) \tag{5}$$

Since the joint transition distribution $p^w_\star(\boldsymbol{x}^w_{t-1,t} \mid \boldsymbol{c})$ is generally intractable, we approximate it as $p^w_\star(\boldsymbol{x}^w_{t-1,t} \mid \boldsymbol{c}) \approx q(\boldsymbol{x}^w_t \mid \boldsymbol{x}^w_0)\, p_\star(\boldsymbol{x}^w_{t-1} \mid \boldsymbol{x}^w_t, \boldsymbol{c})$, where $\boldsymbol{x}^w_0 \sim \mathcal{D}$, $\boldsymbol{x}^w_t \sim q(\boldsymbol{x}^w_t \mid \boldsymbol{x}^w_0)$, and $\boldsymbol{x}^w_{t-1} \sim p_\star(\boldsymbol{x}^w_{t-1} \mid \boldsymbol{x}^w_t, \boldsymbol{c})$. This provides a practical sampling scheme in which the preferred transition is constructed by forward noising followed by preference-guided denoising. An analogous approximation is applied to the dispreferred transition $p^l_\star(\boldsymbol{x}^l_{t-1}, \boldsymbol{x}^l_t \mid \boldsymbol{c})$. Under this approximation, we derive the following score-space objective (Song et al., 2021);see Appendix for the full derivation:

$$\mathcal{L}_{\text{DDSPO}}(\theta) = -\mathbb{E}_{(\boldsymbol{x}^w_0,\boldsymbol{x}^l_0)\sim\mathcal{D},\, \boldsymbol{c}\sim\mathcal{D}(c),\, t\sim\mathcal{U}(0,T),\, \boldsymbol{x}^w_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}^w_0),\, \boldsymbol{x}^l_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}^l_0)} \log\sigma\Bigg(-\beta\cdot\Big[\|\boldsymbol{\epsilon}^w_\star - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^w_t,t,\boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^w_\star - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c})\|_2^2 - \big(\|\boldsymbol{\epsilon}^l_\star - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^l_t,t,\boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^l_\star - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^l_t,t,\boldsymbol{c})\|_2^2\big)\Big]\Bigg). \tag{6}$$

Here, $\boldsymbol{\epsilon}^w_\star$, $\boldsymbol{\epsilon}^l_\star$ denote denoising score targets for preferred and dispreferred directions, respectively.

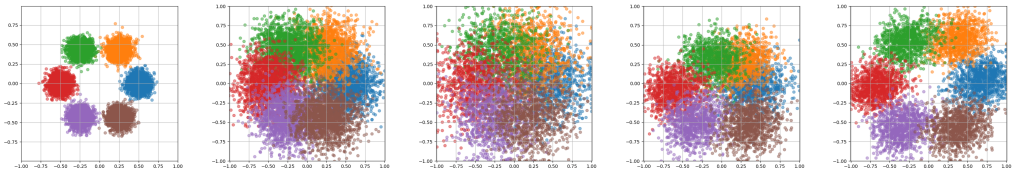
Diffusion DPO as a Special Case of DDSPO Here, we show that Diffusion DPO is a special case of DDSPO that approximates the preference-conditioned transition distribution $p^w_\star(\boldsymbol{x}^w_{t-1,t} \mid \boldsymbol{c})$ using only final preference-labeled samples. While DDSPO uses the approximation $p^w_\star(\boldsymbol{x}^w_{t-1,t} \mid \boldsymbol{c}) \approx q(\boldsymbol{x}^w_t \mid \boldsymbol{x}^w_0)\, p_\star(\boldsymbol{x}^w_{t-1} \mid \boldsymbol{x}^w_t, \boldsymbol{c})$, Diffusion DPO further replaces $p_\star$ with q, yielding $p^w_\star(\boldsymbol{x}^w_{t-1,t} \mid \boldsymbol{c}) \approx q(\boldsymbol{x}^w_t \mid \boldsymbol{x}^w_0)\, q(\boldsymbol{x}^w_{t-1} \mid \boldsymbol{x}^w_t, \boldsymbol{x}^w_0)$. This shifts the notion of preference toward the final outputs, making the model heavily reliant on $\boldsymbol{x}^w_0$ and $\boldsymbol{x}^l_0$. Consequently, supervision is broadcast from these final samples to all timesteps and, especially at large t, can be broad and unspecific. In contrast, DDSPO supervises each local transition $\boldsymbol{x}_t \to \boldsymbol{x}_{t-1}$ under the winning or losing denoising policies $p_\star(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c})$, yielding more informative, dense, and localized timestep-specific signals and reducing sensitivity to noise in the final samples.

3.3 Constructing Preference Pairs for DDSPO Training

Score Preference Pairs In practice, direct supervision over intermediate transition preferences, in the form of denoising score pairs $(\boldsymbol{\epsilon}^w_\star, \boldsymbol{\epsilon}^l_\star)$, is rarely available. To address this, we introduce a practical and effective strategy for generating such signals based on prompt perturbation. Specifically, we treat the denoising score predicted from the original prompt $\boldsymbol{c}$ as the preferred signal, and construct a dispreferred counterpart by generating a corrupted version of the prompt $\boldsymbol{c}^-$ and computing the corresponding denoising score. Note that the type of perturbation applied to the prompt may vary depending on the specific application of preference optimization (*e.g.*, text-to-image alignment or aesthetic quality). These signals can be obtained using a pretrained reference model. In this setup, the noise $\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t, t, \boldsymbol{c})$ guided by $\boldsymbol{c}$ is considered positively aligned with the intended semantics, while the noise $\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^l_t, t, \boldsymbol{c}^-)$ from $\boldsymbol{c}^-$ is dispreferred, as it reflects guidance from an incomplete or misleading prompt. By plugging these noises as supervision targets into the DDSPO objective in Eq. (6), we get

$$\mathcal{L}(\theta) = -\mathbb{E}_{(\boldsymbol{x}^w_0,\boldsymbol{x}^l_0)\sim\mathcal{D},\, (\boldsymbol{c},\boldsymbol{c}^-)\sim\mathcal{D}(c),t\sim\mathcal{U}(0,T),\, \boldsymbol{x}^w_t\sim q(\boldsymbol{x}^w_t|\boldsymbol{x}^w_0),\, \boldsymbol{x}^l_t\sim q(\boldsymbol{x}^l_t|\boldsymbol{x}^l_0)} \log\sigma\Bigg(-\beta\cdot\Big[\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^w_t,t,\boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c}) - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c})\|_2^2 - \big(\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^l_t,t,\boldsymbol{c}^-) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^l_t,t,\boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^l_t,t,\boldsymbol{c}^-) - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^l_t,t,\boldsymbol{c})\|_2^2\big)\Big]\Bigg) \tag{7}$$

Here, the term $\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^w_t,t,\boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c}) - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t,t,\boldsymbol{c})\|_2^2$ simplifies to a standard distillation loss, as the second term vanishes. This encourages the student model



(a) GT Distribution (b) Training Dataset (c) Reference Model (d) Diffusion DPO (e) DDSPO

Figure 2: **Toy Experiments for Comparison between Diffusion DPO (D-DPO) and DDSPO.** (a) and (b) show samples from the ground-truth distribution and its noisy variant used for training. (c) is generated by the reference diffusion model trained on (b). (d) and (e) are distributions learned by the models finetuned with D-DPO and DDSPO, respectively.

to closely follow the reference without degrading its original capabilities. In contrast, the term $\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c}^-) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c}^-) - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2$ penalizes the model when its prediction moves closer to the degraded direction $\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c}^-)$ than the reference model's own output does. Together, these two terms encourage the model to retain useful denoising capabilities while avoiding alignment with poor or corrupted guidance signals.

Training Pipeline We first consist an image preference dataset $\mathcal{D} = \{(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l, \boldsymbol{c}, \boldsymbol{c}^-)\}$ by generating $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l)$ from the reference model for each prompt pair $(\boldsymbol{c}, \boldsymbol{c}^-)$. During training, we sample $(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l, \boldsymbol{c}, \boldsymbol{c}^-) \sim \mathcal{D}$, draw $t \sim \mathcal{U}\{1, \dots, T\}$, obtain $\boldsymbol{x}_t^w \sim q(\boldsymbol{x}_t \mid \boldsymbol{x}_0^w)$ and $\boldsymbol{x}_t^l \sim q(\boldsymbol{x}_t \mid \boldsymbol{x}_0^l)$ via the forward process, compute the required denoising signals at timestep t, and then optimize Eq. (7).

Efficient DDSPO without Image Preference Pair Dataset Motivated by recent findings (Kim et al., 2025) that effective score distillation does not require strict pairing between the noised image $\boldsymbol{x}_t$ and the conditioning prompt $\boldsymbol{c}$ (*e.g.*, at large t the distribution of $\boldsymbol{x}_t$ is approximately standard Gaussian regardless of $\boldsymbol{x}_0$, and reverse generation is driven by the conditioning prompt, so strict prompt–image pairing is unnecessary to obtain meaningful signals), we introduce an efficient variant of DDSPO that avoids explicitly synthesizing dispreferred images. Instead of generating $\boldsymbol{x}_0^l$ for each $(\boldsymbol{c}, \boldsymbol{c}^-)$, we reuse a randomly sampled positive image $\boldsymbol{x}_0^+$ (from another prompt) as a surrogate for $\boldsymbol{x}_0^l$, obtain $\boldsymbol{x}_t^l \sim q(\boldsymbol{x}_t \mid \boldsymbol{x}_0^+)$ via the forward process, and estimate the losing direction using the perturbed prompt $\boldsymbol{c}^-$ (*e.g.*, compute $\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c}^-)$). Although this introduces a prompt–image mismatch on the losing branch, the conditioning still provides a meaningful negative direction—especially at larger timesteps—and aggregation across timesteps and prompts yields informative supervision. In practice, this design cuts the cost of generating additional dispreferred samples, highlighting DDSPO's flexibility under unpaired supervision.

4 Experiments

We assess the effectiveness of DDSPO on a range of conditional generation tasks, focusing on its ability to improve alignment and perceptual quality. Particularly, our approach relies on automatically generated preference signals in the absence of human-annotated preference data. In this section, we begin with a controlled 2D toy experiments (Sec. 4.1) to validate DDSPO's core mechanism under minimal conditions. We then apply DDSPO to two practical text-to-image tasks: improving prompt-image alignment (Sec. 4.2) and enhancing aesthetic quality (Sec. 4.3). In each case, we compare against Diffusion DPO (D-DPO) and include ablation study and comparisons to state-of-the art methods. Finally, we explore the different prompt perturbation methods together with its efficient variant in Sec. 4.4.

4.1 Preliminary Experiments

We first conduct toy experiments to evaluate the effectiveness of DDSPO in a controlled 2D setting. Specifically, we simulate a simplified conditional generation task where each condition corresponds to a distinct mode of a multi-modal Gaussian distribution, as illustrated in Fig. 2a. A reference model is trained on a noisy dataset shown in Fig. 2b, mimicking real-world scenarios with imperfect supervision. The learned distribution from this reference model, shown in Fig. 2c, closely resembles the noisy training distribution. To construct preference data, we sample $N = 2$ preference pairs per

Table 1: **Results in Improving Text-to-Image Alignment.** (a) We compare D-DPO and DDSPO on GenEval, T2I-CompBench, FID and IS using the same perturbation-based preference data. (b) We further evaluate DDSPO across backbones (SD 1.4, SDXL, SANA).(c) Finally, following CaPO's evaluation protocol, we assess SDXL alongside its Itercomp, CaPO, and DDSPO variants. Rows highlighted in gray denote baselines; full results are provided in the appendix.

(a) Comparisons to D-DPO

Model	GE↑	CB↑	FID↓	IS↑
SD-1.4	.4245	.3150	13.05	36.76
+D-DPO	.4841	.3723	18.02	36.64
+DDSPO	.5045	.3823	16.39	38.10

(b) DDSPO on diverse backbones

Model	GE↑	CB↑
SD-1.4	.4245	.3150
+DDSPO	.5045	.3723
SDXL	.5229	.4034
+DDSPO	.6049	.4857
SANA	.6812	.4846
+DDSPO	.7266	.5255

(c) Comparisons to SOTA

Model	GE↑	CB↑
SDXL	.5229	.4185
+Itercomp	.6108	.4644
+CaPO	.5900	.4652
+DDSPO	.6049	.5064

class, each consisting of a preferred sample $\boldsymbol{x}_0^w$ from the target class $\boldsymbol{c}$ and a dispreferred sample $\boldsymbol{x}_0^l$ from a neighboring class $\boldsymbol{c}^-$, simulating our perturbed prompting strategy. This dataset is then used to finetune the model using both D-DPO and DDSPO, where DDSPO additionally leverages intermediate score preferences at each timestep t.

As shown in Fig. 2d, the model finetuned with D-DPO often fails to maintain clear separation between modes, resulting in overlapping or distorted output distributions. This issue becomes more pronounced under limited supervision (see Appendix for extended results). The degradation arises because D-DPO relies on denoising targets derived from potentially misaligned final samples, which can provide misleading learning signals as D-DPO always provides noise towards this particular final sample, as discussed in Sec. 3.2.

In contrast, DDSPO enables the model to learn well-separated, condition-specific outputs, as shown in Fig. 2e. Rather than relying on a final samples (which can induce global, coarse signals), DDSPO directly models preference at each t by contrasting the winning and losing denoising policies $p_\star^w$ and $p_\star^l$ over the local transition $(\boldsymbol{x}_t \to \boldsymbol{x}_{t-1})$. This leads to more robust learning signals across the trajectory of denoising. Note that in the Fig. 2e, slight shifts of the centers away from zero can occur because DDSPO separates distributions across conditions by following the "winning" direction while pushing away from the "losing" (neighboring) direction in score space, yielding clearer boundaries between adjacent conditional distributions. In more complex, real-world tasks such as image generation, the condition space (e.g., text space) is far more densely packed, providing supervision that contrasts across a wider range of diverse directions.

4.2 Improving Text-to-Image Alignment

Target Task In this section, we evaluate DDSPO for improving text-to-image alignment in diffusion models, enabling more faithful generation with respect to the input prompt.

Experimental Setup We use Stable Diffusion v1.4 (SD-1.4) (Rombach & Esser, 2022a) as the baseline model and apply D-DPO and DDSPO for finetuning to improve text-to-image alignment. To construct a preference dataset, we sample 200K prompts from DiffusionDB (Wang et al., 2022), which serve as the original conditions $\boldsymbol{c}$. For each prompt $\boldsymbol{c}$, we generate a perturbed caption $\boldsymbol{c}^-$ by randomly removing 40% to 70% of the tokens, thereby reducing semantic specificity and content richness. To assess text-to-image alignment, we evaluate each model on the GenEval (Ghosh et al., 2023) and T2I-CompBench (Huang et al., 2023) benchmarks, which are designed to measure compositional and semantic consistency between text prompts and generated images. Additionally, we compute the Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS)(Salimans et al., 2016) on 30K images generated from MS-COCO(Lin et al., 2014) validation prompts to assess output quality and diversity.

Comparisons to Diffusion DPO We compare DDSPO and D-DPO on text–image alignment for a pretrained diffusion model (Table 1a). Both methods increase alignment over the baseline, evidencing the effectiveness of prompt-perturbation supervision. Leveraging dense timestep-level signals, DDSPO attains higher alignment than D-DPO on GenEval and CompBench (*e.g.*, +34.2% and +17.5% larger gains) while also yielding better FID↓/IS↑. Although our FID is higher (worse) than the reference model's, we attribute this to distributional shift induced by preference optimization

Figure 3: **Qualitative comparison between SDXL and DDSPO.** Images are generated from the same prompts using identical random seeds.

Table 2: **Aesthetic Quality Comparison with State-of-the-arts methods** We compare our DDSPO with two other methods—Diffusion DPO (D-DPO), Diffusion KTO (D-KTO) that are trained on the human-annotated Pick-a-Pic dataset. Unlike D-DPO and D-KTO, DDSPO is trained without any human-labeled data. The models are evaluated on HPSv2 and PickScore.

(a) Comparisons to SOTA methods with SD-1.5

Model	Dataset	HPSv2↑	PickScore↑
SD-1.5	-	26.95	21.14
+D-DPO	Pickapic	27.25	21.34
+D-KTO	Pickapic	27.89	21.39
+DDSPO	None	27.46	21.35

(b) Comparisons to SOTA methods with SDXL

Model	Dataset	HPSv2↑	PickScore↑
SDXL	-	27.89	22.27
+D-DPO	Pickapic	28.55	22.61
+DDSPO	None	28.78	22.70

rather than degraded sample quality; notably, human-labeled preference methods exhibit similar behavior, with reported FIDs of 15.05 (Diffusion DPO) and 20.31 (Diffusion KTO).

DDSPO with Various Model Architectures We apply DDSPO to models with three different architectures, including both U-Net-based models (SD-1.4 and SDXL) and a DiT-based flow-matching model (SANA (Xie et al., 2024a)), as shown in Table 1b. The results clearly demonstrate that DDSPO significantly improves text-to-image alignment across all three models on both benchmarks, highlighting its effectiveness and broad applicability across diverse diffusion architectures.

Comparisons to SOTA Methods In Table 1c, we compare our method against current SOTA approaches, Itercomp (Zhang et al., 2025) and CaPO (Lee et al., 2025c), for improving text-to-image alignment. All three methods show substantial improvements over the SDXL baseline on both benchmarks. Notably, IterComp and CaPO rely on human-annotated datasets for training. In contrast, DDSPO achieves comparable performance on GenEval and outperforms all methods on CompBench, despite not using any human annotations, demonstrating its effectiveness in a data efficient setup. Qualitative results comparing the baseline SDXL model and the DDSPO-finetuned model can be found in Fig. 3.

4.3 Improving Aesthetic Quality

Target Task In this section, we apply DDSPO to improve the aesthetic quality of images generated by diffusion models, promoting visually appealing and artistically coherent outputs.

Experimental Setup We evaluate the effectiveness of DDSPO in enhancing aesthetic quality with SD-1.5 and SDXL. Following the same training setup as in Sec. 4.2, we construct supervision pairs using sampled prompt pairs $(\boldsymbol{c}, \boldsymbol{c}^-)$, where the negative prompts $\boldsymbol{c}^-$ are generated by prompting LLaMA3-8B (Grattafiori et al., 2024) to intentionally degrade the aesthetic quality of the original prompt. The prompt template used for generation is provided in the appendix. To evaluate aesthetic quality, we use the HPSv2 (Wu et al., 2023) and PickScore (Kirstain et al., 2023b) metrics, both of

Table 3: **Effects of Various Prompt Perturbation Strategies and Negative Image Generation.** Ablation study comparing (a) different strategies for constructing dispreferred prompts to generate negative samples, and (b) standard DDSPO with variants that avoid explicit negative image (NI) generation by reusing preferred samples with unpaired or perturbed prompts.

(a) Comparisons of different perturbed prompting methods

Method	T2I Alignment		Aesthetic Quality	
	GE↑	CB↑	HPSv2↑	PickScore↑
SD-1.4	.4245	.3150	26.88	21.11
Rand-removal	.5045	.3823	27.28	21.36
LLaMA (TA)	.4854	.3807	27.25	21.36
LLaMA (AQ)	.4758	.3616	27.51	21.39

(b) Comparison of DDSPO with and without explicit negative image generation

Method	NI generation	GE↑	CB↑
SD-1.4	-	.4245	.3150
Rand-removal	✓	.5045	.3823
Rand-positive	✗	.4866	.3763
Not-paired	✗	.4891	.3842

which are measured by a model trained to reflect human aesthetic preferences using Human Preference Dataset v2 and Pick-a-Pic (Kirstain et al., 2023b), respectively. For evaluation, we generate images by sampling from prompts in the HPSv2 (Wu et al., 2023) and PartiPrompts (Yu et al., 2022) benchmarks, and assess them using the corresponding metrics.

Results In Table 2, we compare DDSPO with SOTA methods for aesthetic quality improvement. DDSPO significantly improves both HPSv2 and PickScore compared to the two baselines reported in Table 2a and Table 2b. Notably, DDSPO is trained solely on automatically constructed preference data using perturbed prompts, while all competing methods rely on human-annotated preference labels. Despite this, DDSPO achieves comparable performance across all evaluation metrics, highlighting its effectiveness in the absence of human supervision.

4.4 Preference Data Construction by Perturbed Prompting

Effects of Different Perturbed Prompts Table 3a presents an exploration of different prompt perturbation strategies for constructing dispreferred scores across the two previously studied tasks. We evaluate three strategies: *Rand-removal*, which drops 40% to 70% of tokens from the original prompt; and *LLaMA*, which uses an LLM (Grattafiori et al., 2024) to rewrite the prompt in a way that degrades task performance. The LLaMA-based strategy includes two task-specific variants: LLaMA (TA) for text-to-image alignment and LLaMA (AQ) for aesthetic quality improvement. When trained with DDSPO, we observe that Rand-removal yields the best performance on text-to-image alignment, while LLaMA (AQ) performs best for aesthetic quality improvement, as it is specifically tailored to that task.

Efficient DDSPO We next evaluate the effectiveness of DDSPO in a setup where dispreferred images $\boldsymbol{x}_0^l$ are not explicitly generated. Instead of using an image conditioned on the corrupted prompt $\boldsymbol{c}^-$, we obtain $\boldsymbol{x}_t^l$ by applying the forward diffusion process to a randomly selected preferred image $\hat{\boldsymbol{x}}_0^w$, originally sampled from an unrelated prompt. Based on this setup, we evaluate two variants: Random-positive, which uses the original prompt $\hat{\boldsymbol{c}}$ paired with the randomly selected image $\hat{\boldsymbol{x}}_0^w$ as the negative prompt $\boldsymbol{c}^-$; and Not-paired, which uses a perturbed version of the original prompt (via Rand-removal), even though the resulting prompt is not strictly paired with the noised image $\boldsymbol{x}_t^l$. As shown in Table 3a, both Rand-positive and Not-paired perform comparably well to the setup where dispreferred samples are explicitly generated (Rand-removal). Interestingly, Not-paired outperforms Random-positive, despite the latter using strictly paired image-prompt pairs. These results demonstrate that the negative direction can be modeled effectively from unpaired prompt–image instances, highlighting DDSPO's flexibility in preference-signal modeling via stepwise supervision.

5 Conclusion

We introduce DDSPO, a novel preference optimization framework that extends Diffusion DPO by supervising each denoising step with preferred/dispreferred targets derived from winning and losing policies. In practice, we construct these targets label-free using prompt perturbations and a pre-trained reference model, eliminating manual annotations and reward models. Our experiments show that DDSPO consistently improves text-image consistency and visual fidelity, offering a scalable and practical approach within the broader landscape of preference-based training methods.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. URL https://arxiv.org/abs/2305.13301.

Xinyan Chen, Jiaxin Ge, Tianjun Zhang, Jiaming Liu, and Shanghang Zhang. Learning from mistakes: Iterative prompt relabeling for text-to-image diffusion model training. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2937–2952, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. *arXiv preprint arXiv:2405.13637*, 2024.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 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. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8836–8845. IEEE Computer Society, 2021.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. The dialog must go on: Improving visual dialog via generative self-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6746–6756. IEEE Computer Society, 2023.

Dohyun Kim, Sehwan Park, Geonhee Han, Seung Wook Kim, and Paul Hongsuck Seo. Random conditioning with distillation for data-efficient diffusion model compression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18607–18618, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023a.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf.

Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. Aligning text to image in diffusion models is easier than you think. *arXiv preprint arXiv:2503.08250*, 2025a.

Kyungmin Lee, Xiaohang Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. *arXiv preprint arXiv:2502.02588*, 2025b.

Kyungmin Lee, Xiaohang Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models, 2025c. URL https://arxiv.org/abs/2502.02588.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization, 2025. URL https://arxiv.org/abs/2406.04314.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*, 2024.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generative models through direct preference optimization. In *ACM Multimedia 2024*, 2024. URL https://openreview.net/forum?id=7lqptq5dLG.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. URL https://api.semanticscholar.org/CorpusID:248097655.

Robin Rombach and Patrick Esser. Stable diffusion v1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4, 2022a.

Robin Rombach and Patrick Esser. Stable diffusion v1-5. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5, 2022b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

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

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. URL https://arxiv.org/abs/2210.14896.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

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.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024a.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024b.

Shuo Xing, Yuping Wang, Peiran Li, Ruizheng Bai, Yueqi Wang, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. *arXiv preprint arXiv:2502.13146*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model, 2024. URL https://arxiv.org/abs/2311.13231.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

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*, 2024.

Xinchen Zhang, Ling Yang, Guohao Li, YaQi Cai, xie jiake, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin CUI. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4w99NAikOE.

Ziyi Zhang, Li Shen, Sen Zhang, Deheng Ye, Yong Luo, Miaojing Shi, Bo Du, and Dacheng Tao. Aligning few-step diffusion models with dense reward difference learning. *arXiv preprint arXiv:2411.11727*, 2024.

Hanyang Zhao, Haoxian Chen, Yucheng Guo, Genta Indra Winata, Tingting Ou, Ziyu Huang, David D Yao, and Wenpin Tang. Fine-tuning diffusion generative models via rich preference optimization. *arXiv preprint arXiv:2503.11720*, 2025.

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

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 291–300, 2024.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33: 3833–3845, 2020.