Beyond Linear Shortcuts: Rectifying Diffusion Preference Optimization with Intrinsic Generative Geometry

Pin Wang[1,2], Huaibo Huang[1,2⋆], Jiayang Sun[1,2], Hongbo Wang[1,2], Wentao Jiang[3], Tiezheng Ge[3], Jie Cao[1,2], and Ran He[1,2]

[1] MAIS & NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Taobao & Tmall Group of Alibaba
huaibo.huang@nlpr.ia.ac.cn

Abstract. Direct Preference Optimization (DPO) has proven effective for aligning diffusion models with human preferences. However, by relying solely on endpoint supervision, existing DPO fine-tuning practices ignore the curved trajectory of the pretrained generation process. This forces a linear shortcut in the latent space, disregarding the intrinsic geometry and leading to off-trajectory and off-manifold generation. To address this, we propose **T**rajectory-**C**onsistent **P**reference **O**ptimization (TCPO), a fine-tuning paradigm respecting the model's intrinsic generative geometry. TCPO first performs *Curvature-Adaptive Sampling* (CAS) to dynamically select a curvature-aware intermediate latent as a geometric anchor. Then, *Trajectory-Rectified Fusion* (TRF) combines this geometric anchor with the conventional endpoint target which acts as a semantic anchor to form a trajectory-consistent supervision signal. Extensive experiments show that TCPO achieves superior visual fidelity and text-to-image alignment on standard benchmarks, with improved training efficiency. [4]

Keywords: Diffusion model · Direct Preference Optimization · Text-to-Image Alignment

1 Introduction

Text-to-image diffusion models [4,9,27–29,35] have achieved remarkable progress in recent years. Driven by large-scale training data [31] and advanced architectures [1, 5, 30], these models demonstrate an exceptional capability to generate high-fidelity images. To further align pre-trained diffusion models with human intentions, Direct Preference Optimization (DPO) and its variants [3, 23, 24, 38, 40–42, 46, 49] have been widely adopted. By directly utilizing paired preference data for model training, these methods effectively circumvent the complexities

⋆ Corresponding author.
[4] Code is available at https://github.com/wangpin-aier/TCPO

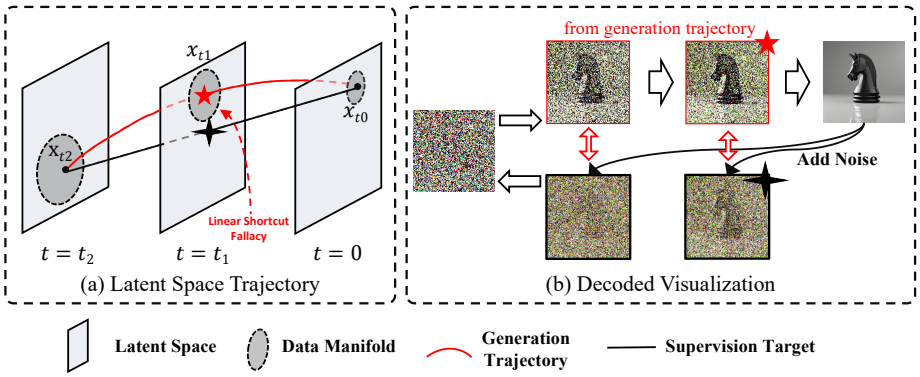

Fig. 1: Motivation of our Trajectory-Consistent Preference Optimization (TCPO). (a) In the latent space, the intrinsic generative trajectory of diffusion models (red curve) is naturally curved. Standard preference optimization methods (e.g., DPO) enforce a linear supervision target (black line) derived from the forward noising process. This linear assumption forces the intermediate states off the data manifold (e.g., the red star deviates from the target black triangle), leading to what we term the **Linear Shortcut Fallacy**. **(b)** Visualizing actual decoded latents (from SDXL) illustrates this discrepancy. The intermediate states from the generation trajectory (top row, red boxes) significantly mismatch the theoretical noisy states (bottom row, black boxes) at the same timestep, as highlighted by the red double arrows.

and reward hacking issues associated with traditional reward models [8, 31, 43], achieving promising results in preference alignment.

However, existing DPO-based methods fundamentally violate the geometric reality of pretrained diffusion models. Governed by a spectral evolution bias—where coarse, low-frequency structures emerge prior to fine, high-frequency details—the true generative trajectory is inherently curved [14, 19, 36]. Despite this, current approaches rely on endpoint-directed supervision, enforcing a linear update direction from the noisy latent to the clean target. We identify this critical flaw as the **Linear Shortcut Fallacy**. Such a forced linear shortcut introduces a severe geometric mismatch, accumulating significant off-trajectory drift (Fig. 1). Consequently, the model is pushed off its pre-trained data manifold, ultimately leading to suboptimal convergence, structural collapse, and semantic degradation.

To address these challenges, we propose **TCPO**, a **T**rajectory-**C**onsistent **P**reference **O**ptimization method, designed to align diffusion models while strictly respecting their intrinsic generative dynamics. First, *Curvature-Adaptive Sampling* (CAS) reconstructs the model's naive generation trajectory via DDIM inversion. It then dynamically identifies a latent state along this inverted trajectory to serve as a *geometric anchor*. Targeting this intermediate anchor rather than the final endpoint x_0 keeps the optimization tightly bound to the naturally curved trajectory. Geometrically, this is because shorter secants across the man-

Fig. 2: Qualitative improvement over the original SDXL. To effectively align the generative capabilities of SDXL with comprehensive human preferences, we propose TCPO for model fine-tuning. This figure illustrates the comparison before and after our optimization: the top row shows the baseline outputs from the original SDXL, whereas the bottom row demonstrates the results from TCPO. Under identical initial noise, our method generates images with superior visual appeal and stricter text alignment.

ifold provide a much tighter approximation of the underlying curved arc than a single long chord spanning the entire generation process.

However, relying solely on the geometric anchor is overly conservative. It primarily encourages the reconstruction of the original, unaligned model's behavior rather than injecting new preference signals. Furthermore, this strict adherence makes the optimization highly susceptible to accumulated approximation errors derived from the DDIM inversion process. To overcome these limitations, we introduce *Trajectory-Rectified Fusion* (TRF). TRF constructs a hybrid supervision signal by fusing the geometric anchor with the clean preference target (acting as the semantic anchor) via *Spherical Linear Interpolation* (SLERP) [33]. This fusion effectively rectifies the optimization direction, pulling the aggressive semantic update back onto the valid data manifold. Consequently, TRF ensures that the model successfully learns human preferences without compromising its intrinsic geometric consistency.

Our comprehensive evaluation demonstrates that TCPO effectively enhances the overall performance of text-to-image models across multiple benchmarks. Specifically, on the HPD v2 benchmark, the SDXL model fine-tuned with TCPO achieves notable improvements across several metrics over the original SDXL base model: the PickScore increases by 0.61, HPS by 0.88. On the parti-prompts and the pick-a-pic v2 test sets, TCPO also delivers outstanding performance, surpassing the state-of-the-art methods InPO and DPO in all evaluated metrics. Furthermore, TCPO achieves training efficiency that is 4.31× higher than DPO and 1.23× higher than InPO. These results collectively validate the versatility and effectiveness of TCPO across different models and datasets.

Our contributions can be summarized as follows:

- We highlight a critical yet largely overlooked bottleneck in diffusion preference optimization: the preservation of intrinsic generative geometry.
- We propose the TCPO framework, introducing CAS to anchor natural trajectories and TRF to integrate human preferences without off-manifold drift.
- Extensive experiments demonstrate that TCPO achieves superior alignment and generation quality with improved training efficiency.

2 Related Work

2.1 Text-to-Image Generative Models

Diffusion models have emerged as a powerful paradigm for high-fidelity image generation, evolving through several key milestones. Inspired by foundational works [2, 4, 9, 22, 35], early systems [27, 29] demonstrated the remarkable potential of these models for text-to-image synthesis. To address the computational demands of operating in pixel space, subsequent developments [25, 28] introduced latent diffusion models built upon VAE [12], significantly reducing inference costs while maintaining generation quality. More recently, the integration of flow matching [16, 17] has spurred architectural innovations, offering alternative formulations for probability flow modeling [1, 5, 30]. Furthermore, diffusion models have been widely extended to various downstream applications [7, 18, 37, 44]. Despite these advances, current approaches still exhibit limitations in generation fidelity and alignment with human preferences, motivating the need for further refinement through fine-tuning strategies.

2.2 Diffusion Models Alignment

The predominant approach for fine-tuning diffusion models remains supervised fine-tuning on curated datasets. Inspired by analogous developments in large language models [6, 10, 26, 34], a family of preference optimization methods has emerged, with Diffusion-DPO [40] serving as the foundational instantiation. Several distinct challenges have motivated subsequent innovations. First, the partition function in the DPO objective poses computational difficulties, necessitating refined approximations for stable training [20,21,49]. Second, as global preference labels inadequately capture the varying quality of intermediate denoising steps, latent-space reward models have been introduced to enable fine-grained, stepwise credit assignment [15,39,48]. Third, acknowledging the practical constraints of data collection, alternative formulations have been explored to reduce annotation burden—ranging from methods that operate solely with positive samples to fully self-supervised paradigms that eliminate the need for pairwise preferences entirely [24, 46].

3 Preliminaries

3.1 Diffusion Models

Diffusion models [9] operate through a dual-phase mechanism: a forward corruption process that injects noise into data, and a learned restoration process that recovers clean samples from Gaussian noise.

Forward Process. Given data $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, the forward chain gradually perturbs it via:

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

where α_t controls the noise intensity at step t. Through reparameterization, one can sample any noisy state directly:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

with $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

Reverse Process. A neural network learns to denoise by predicting the added noise $\boldsymbol{\epsilon}_\theta^t(\boldsymbol{x}_t)$. The training objective minimizes:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\lambda(t)\left\|\boldsymbol{\epsilon}_\theta^t(\boldsymbol{x}_t) - \boldsymbol{\epsilon}\right\|_2^2\right], \tag{3}$$

where $\lambda(t)$ denotes a weighting schedule [9, 36].

Deterministic Sampling. DDIM [35] enables non-Markovian generation by formulating the reverse as an ODE. Its update rule is:

$$\boldsymbol{x}_t = \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t+1}}}\boldsymbol{x}_{t+1} + \sqrt{\bar{\alpha}_t}\left(\sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} - \sqrt{\frac{1-\bar{\alpha}_{t+1}}{\bar{\alpha}_{t+1}}}\right)\boldsymbol{\epsilon}_\theta^{t+1}(\boldsymbol{x}_{t+1}), \tag{4}$$

DDIM Inversion. Based on the deterministic nature of the DDIM ODE, the sampling process can be inverted to map a real image $\boldsymbol{x}_0$ back to its latent noise $\boldsymbol{x}_T$. By reversing the time steps from t to $t+1$, the inversion update rule is defined as:

$$\boldsymbol{x}_{t+1} = \sqrt{\frac{\bar{\alpha}_{t+1}}{\bar{\alpha}_t}}\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t+1}}\left(\sqrt{\frac{1-\bar{\alpha}_{t+1}}{\bar{\alpha}_{t+1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\right)\boldsymbol{\epsilon}_\theta^t(\boldsymbol{x}_t), \tag{5}$$

3.2 Direct Preference Optimization

DPO [40] aligns models with human preferences without explicit reward modeling. Consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l, \boldsymbol{c})\}$ containing pairs of winning (w) and losing (l) outputs for context $\boldsymbol{c}$.

Preference Modeling. The Bradley-Terry model captures pairwise comparisons via:

$$p(\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l|\boldsymbol{c}) = \sigma\big(r(\boldsymbol{x}_0^w, \boldsymbol{c}) - r(\boldsymbol{x}_0^l, \boldsymbol{c})\big), \tag{6}$$

where σ is the sigmoid and r an implicit reward function.

DPO Objective. By reparameterizing the reward through the policy ratio, the loss becomes:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{p_\theta(\boldsymbol{x}_0^w|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_0^w|\boldsymbol{c})} - \beta\log\frac{p_\theta(\boldsymbol{x}_0^l|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_0^l|\boldsymbol{c})}\right)\right], \tag{7}$$

where β controls the deviation from reference model p_{ref}.

Extension to Diffusions. Since marginal likelihoods are intractable, [40] derive a tractable gradient by leveraging the ELBO. This yields a noise-prediction objective:

$$\begin{aligned}\mathcal{L}_{\text{DPO-Diff}}(\theta) = -\mathbb{E}_{(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim\mathcal{D},\, t\sim\mathcal{U}(0,T),\, \boldsymbol{x}_t^{w,l}\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^{w,l})} \\ \log\sigma\Big(-\beta T\omega(\lambda_t)\Big(\left\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w,t)\right\|_2^2 - \left\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w,t)\right\|_2^2 \\ - \left\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l,t)\right\|_2^2 + \left\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l,t)\right\|_2^2\Big)\Big),\end{aligned} \tag{8}$$

where $\omega(\lambda_t)$ incorporates the diffusion weighting, and $\boldsymbol{\epsilon}^{w,l}$ denotes the ground-truth noise added to winning and losing samples respectively.

4 Methodology

We propose Trajectory-Consistent Preference Optimization (TCPO), a framework that aligns diffusion models with human preferences while strictly respecting their naive generative dynamics. As shown in Fig. 3, TCPO rectifies the Linear Shortcut Fallacy via two mechanisms: **Curvature-Adaptive Sampling (CAS)**, which finds a geometric anchor on the trajectory, and **Trajectory-Rectified Fusion (TRF)**, which constructs a hybrid supervision signal to balance preference learning and intrinsic geometric consistency.

4.1 Curvature-Adaptive Sampling

Standard preference optimization typically constructs supervision targets in an endpoint-directed manner, pointing directly towards the clean data x_0. This strategy implicitly relies on a **global straight-line (chord) approximation** of the underlying generative trajectory, implying that $\epsilon_{t_1} = \epsilon_{t_2}$ for any $t_1, t_2 \in \{0, 1, \ldots, T-1\}$, where T is the maximum timestep and ϵ_t denotes the noise prediction. However, this assumption fundamentally contradicts the spectral nature of diffusion models, where coarse, low-frequency structures always emerge significantly earlier than fine, high-frequency details. Consequently, this leads to what we term the **Linear Shortcut Fallacy**. We detail why the intrinsic generative trajectory follows a curved path in the latent space, driven by this spectral evolution bias, in Section A of the supplementary material.

Geometrically, directly optimizing towards the final semantic endpoint implicitly assumes a linear update path in the latent space. However, because the intrinsic generative manifold is highly curved, this linear shortcut inevitably

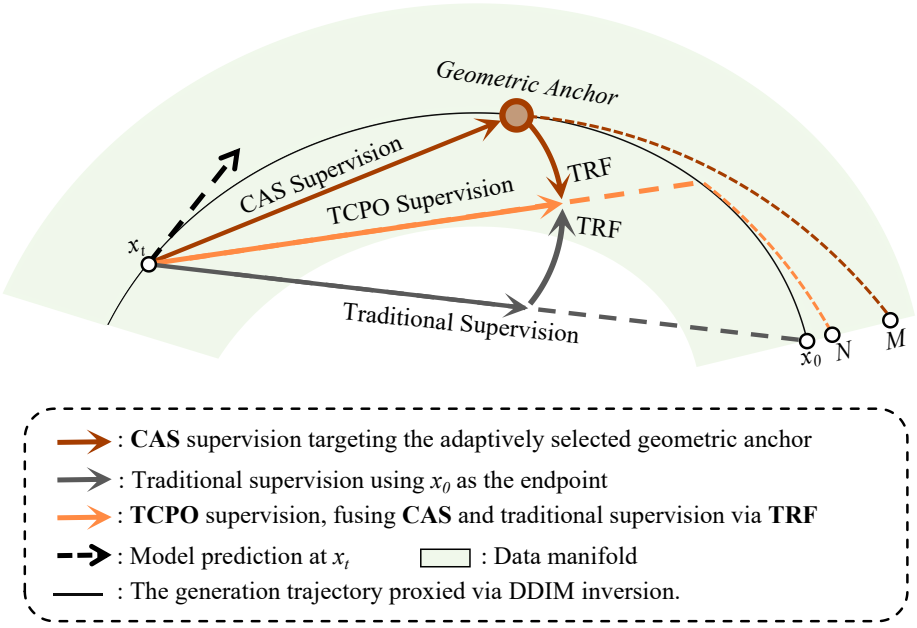

Fig. 3: Overview of TCPO. The traditional supervision target (grey arrow) points directly to x_0, inadvertently pushing the model off the curved data manifold. To maintain geometric consistency, *Curvature-Adaptive Sampling* (CAS) provides an on-manifold target (brown arrow) directed towards the geometric anchor. However, relying solely on the CAS target commits the model to a path (brown dashed curve) that ultimately terminates at a semantically distant endpoint M. To resolve this, *Trajectory-Rectified Fusion* (TRF) integrates the geometric anchor with the semantic anchor (x_0). The resulting TCPO target (orange arrow) rectifies the prediction direction, ensuring the model stays strictly on the manifold while its corresponding trajectory (orange dashed curve) successfully converges to N, closely approximating the clean image x_0.

forces the intermediate latents into low-density, off-manifold regions where the pre-trained score field is unreliable. Therefore, anchoring the optimization to the intrinsic trajectory is a geometric necessity to preserve structural integrity.

To operationalize this anchoring, we must first quantify the local bending of the trajectory. Let $\mathcal{T} = \{x_{t_0}, x_{t_1}, \ldots, x_{t_{n-1}}\}$ denote the discrete sequence of latent states obtained via DDIM inversion, which serves as a proxy for the generative path. Here, the starting index 0 represents the clean image, and the ending index $n-1$ corresponds to the current noisy training state. Geometrically, the generation process can be viewed as a curve in the latent space, where the update direction at the i-th step is governed by the predicted noise vector:

$$\mathbf{v}_i = \epsilon_\theta(x_{t_i}, t_i), \tag{9}$$

To quantify the local curvature of this trajectory, we consider the rate of change of its unit tangent vector:

$$\mathbf{u}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}, \tag{10}$$

In continuous dynamics, the local bending (or angular velocity) is measured by $\left\|\frac{d\mathbf{u}}{dt}\right\|$. Given that our timesteps are uniformly sampled with a constant interval Δt, we can approximate this derivative using finite differences:

$$\frac{d\mathbf{u}}{dt} \approx \frac{\mathbf{u}_i - \mathbf{u}_{i-1}}{\Delta t}, \tag{11}$$

Taking the squared norm of this discrete derivative yields:

$$\left\|\frac{\mathbf{u}_i - \mathbf{u}_{i-1}}{\Delta t}\right\|^2 = \frac{1}{\Delta t^2}\left(\|\mathbf{u}_i\|^2 - 2\mathbf{u}_i^\top \mathbf{u}_{i-1} + \|\mathbf{u}_{i-1}\|^2\right) = \frac{2}{\Delta t^2}\left(1 - \mathbf{u}_i^\top \mathbf{u}_{i-1}\right), \tag{12}$$

Since Δt is a constant across the uniform schedule, the term $(1 - \mathbf{u}_i^\top \mathbf{u}_{i-1})$ is strictly proportional to the squared local curvature. Therefore, we define our curvature proxy κ_i as the cosine distance between consecutive velocity fields:

$$\kappa_i = 1 - \frac{\mathbf{v}_i^\top \mathbf{v}_{i-1}}{\|\mathbf{v}_i\|\|\mathbf{v}_{i-1}\|}, \quad \kappa_i \in [0, 2], \tag{13}$$

This formulation provides a mathematically grounded, scale-invariant metric to dynamically assess the trajectory's non-linearity at any given timestep. Equipped with this curvature metric, we dynamically calculate a geometric anchor to serve as the supervision target, ensuring that the optimization process does not deviate from the intrinsic generative trajectory. Specifically, rather than relying on a fixed lookahead horizon, for a current state x_t, we search for the furthest valid anchor index t_{i^*} (where $i^* < n$) such that the accumulated curvature along the path does not exceed a predefined threshold η:

$$i^* = \min\left\{1 < \tau < n \;\middle|\; \sum_{j=\tau}^{n-1} \kappa_j \leq \eta\right\}, \tag{14}$$

Here, η is an **adaptive curvature budget**. To ensure robustness across varying trajectory complexities, we derive η from the total curvature of the inverted trajectory:

$$\eta = \frac{1}{\alpha}\sum_{j=0}^{n-1} \kappa_j, \tag{15}$$

where α is a hyperparameter controlling the granularity of the linearization (effectively representing the expected number of lookahead segments). This yields the **Geometric Anchor** $x_{\text{geo}} \coloneqq x_{t_{i^*}}$. In regions of high curvature, the horizon $t - t_{i^*}$ naturally shrinks to respect the budget; in flat regions, it expands, allowing for longer-range guidance.

4.2 Trajectory-Rectified Fusion

While the geometric anchor x_{geo} provides geometrically consistent supervision, relying solely on it can be conservative and may lack strong semantic signals for alignment. Conversely, the clean data x_0 (denoted as the **Semantic Anchor**, x_{sem}) offers strong alignment guidance but violates trajectory constraints. TRF bridges this gap by fusing the implicit noise predictions pointing to these two anchors into a unified target.

Given the current state x_t, we first compute the implicit noise vectors (via standard DDIM reparameterization) directly pointing to the respective anchors: $\epsilon_{\text{geo}} = \epsilon(x_{\text{geo}})$ and $\epsilon_{\text{sem}} = \epsilon(x_{\text{sem}})$. We then construct the rectified target noise ϵ_{TRF} by projecting the semantic guidance of ϵ_{sem} onto the geometric frame of ϵ_{geo}. To respect the spherical geometry of the noise space, we employ Spherical Linear Interpolation (SLERP):

$$\epsilon_{\text{TRF}} = \text{SLERP}(\epsilon_{\text{geo}}, \epsilon_{\text{sem}}; \gamma) = \frac{\sin(\gamma\Theta)}{\sin\Theta}\epsilon_{\text{geo}} + \frac{\sin((1-\gamma)\Theta)}{\sin\Theta}\epsilon_{\text{sem}}, \tag{16}$$

where $\Theta = \arccos\left(\frac{\epsilon_{\text{geo}}^{\top}\epsilon_{\text{sem}}}{\|\epsilon_{\text{geo}}\|\|\epsilon_{\text{sem}}\|}\right)$ is the angle between the noise anchors, and $\gamma \in [0, 1]$ is a rectification coefficient. This geometric design is theoretically grounded: due to the Concentration of Measure phenomenon, noise vectors in high-dimensional Gaussian spaces inherently concentrate on the surface of a hypersphere. SLERP ensures that the interpolated noises strictly traverse this spherical manifold, fundamentally avoiding the off-manifold norm collapse toward the origin caused by naive linear interpolation (proof in Section B of the supplementary material). Ultimately, (16) constructs a supervision noise vector ϵ_{TRF} that points towards the semantic endpoint but is explicitly pulled back towards the inverted trajectory. This fused noise implicitly defines our rectified target state y_{TRF}, ensuring the model successfully learns human preferences without compromising its intrinsic geometric consistency.

4.3 Optimization Objective

TCPO optimizes the policy π_θ using a preference dataset $\mathcal{D} = \{x^w, x^l\}$, consisting of winner and loser images. For each pair, we perform inversion to obtain their respective trajectories, identify their anchors, and compute the fused target noises ϵ_w^{TRF} and ϵ_l^{TRF} using Eq. (16). These target noises implicitly correspond to the rectified target states y_w^{TRF} and y_l^{TRF}.

The optimization objective follows the DPO formulation adapted for diffusion models:

$$\begin{aligned}\mathcal{L}_{\text{TCPO}} = -\mathbb{E}_{(x^w,x^l)\sim\mathcal{D},t\sim U[0,T]}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w^{\text{TRF}} \mid x_t^w)}{\pi_{\text{ref}}(y_w^{\text{TRF}} \mid x_t^w)} \\ - \beta\log\frac{\pi_\theta(y_l^{\text{TRF}} \mid x_t^l)}{\pi_{\text{ref}}(y_l^{\text{TRF}} \mid x_t^l)}\Bigg)\Bigg],\end{aligned} \tag{17}$$

Here, $\pi_\theta(y \mid x_t)$ represents the likelihood of transitioning to state y given x_t. In the context of diffusion models trained with Mean Squared Error (MSE), the log-likelihood ratio can be approximated by the difference in L2 reconstruction errors:

$$\log \frac{\pi_\theta(y \mid x_t)}{\pi_{\text{ref}}(y \mid x_t)} \approx -\frac{1}{2\sigma_t^2} \left(\|\epsilon_\theta(x_t) - \epsilon_{\text{target}}(y)\|^2 - \|\epsilon_{\text{ref}}(x_t) - \epsilon_{\text{target}}(y)\|^2 \right), \quad (18)$$

where $\epsilon_{\text{target}}(y)$ is the implicit noise vector pointing to the target y. By substituting Eq. (18) into Eq. (17), TCPO effectively reweights the standard denoising objective based on the preference margin.

The TCPO objective is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{TCPO}} = -\mathbb{E}_{(x^w, x^l)\sim\mathcal{D}, t\sim U[0,T]} \Bigg[\log \sigma \Bigg(-\beta \\ \Big[\|\epsilon_\theta(x_t^w) - \epsilon_w^{TRF}\|^2 - \|\epsilon_\theta(x_t^l) - \epsilon_l^{TRF}\|^2 \\ - \|\epsilon_{\text{ref}}(x_t^w) - \epsilon_w^{TRF}\|^2 + \|\epsilon_{\text{ref}}(x_t^l) - \epsilon_l^{TRF}\|^2 \Big] \Bigg) \Bigg], \end{aligned} \quad (19)$$

In this formulation, TCPO effectively reweights the standard denoising objective based on the preference margin, while the target noises ϵ_w and ϵ_l are rectified by our TRF mechanism to ensure trajectory consistency.

5 Experiments

5.1 Implementation Details

Our experiments are built on Stable Diffusion XL-base-1.0 (SDXL) [25]. Following the setting of [20,40], we conduct all training on the Pick-a-Pic v2 dataset [13] after filtering out approximately 12% of the draw samples, which leaves 851293 image-text pairs and 58960 unique text prompts. For fair comparison, the majority of hyperparameters are inherited from [20, 40]. We employ Adafactor [32] as the optimizer for training. All experiments are conducted on eight NVIDIA A100 GPUs. Unless otherwise specified, we set $\alpha = 2.0$, and $\gamma = 0.2$. DDIM Inversion is done with the classifier-free guidance scale set to 1 and the inversion step number set to 5. Training uses a batch size of 1 image pair per step, with gradient accumulation over 128 steps, resulting in an effective batch size of 1024 pairs. The hyperparameter β is set to 5000. The learning rate is set to $\frac{2000}{\beta} 2.048 * 10^{-8}$, with a linear warmup phase of 200 steps and a total of 400 training steps. All experiments use a fixed random seed of 42 to ensure reproducibility.

5.2 Evaluation Details

We compare TCPO with existing baselines using user studies and automatic preference metrics. Specifically, we compare our method with Diffusion-DPO [40], InPO [20], and MaPO [11].

Table 1: We conduct a comprehensive evaluation of TCPO against baseline methods using prompts from DrawBench, Parti-Prompts, HPDv2, and Pick-a-Pic v2 test set. Performance is reported using both mean and median scores across multiple evaluators, with the highest score in each metric bolded and the second highest underlined.

Datasets	Models	Aesthetic		PickScore		HPS		MPS	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
DrawBench	SDXL	5.5886	5.6442	22.4952	22.6703	28.5959	28.9023	11.4993	12.2265
	DPO	5.6428	5.7216	22.8330	22.9842	29.0824	29.4022	12.1187	**13.1680**
	InPO	5.6435	5.6725	22.9671	23.0606	29.1010	29.3284	12.0145	12.7148
	MaPO	**5.7401**	5.7255	22.5230	22.6339	28.8760	29.0263	11.6293	12.6562
	TCPO(Ours)	5.7302	**5.7858**	**23.0146**	**23.1029**	**29.2670**	**29.6212**	**12.1738**	13.0117
PartiPrompts	SDXL	5.7693	5.7571	22.6274	22.6496	28.4241	28.4723	11.2538	11.6718
	DPO	5.7946	5.8097	22.9307	22.9264	28.9085	28.8873	11.8097	12.1055
	InPO	5.8332	5.8150	22.9952	22.9889	28.8443	28.8621	11.6527	11.9453
	MaPO	5.8979	**5.9017**	22.5946	22.5412	28.5551	28.4649	11.2751	11.6328
	TCPO(Ours)	**5.9001**	5.8871	**23.1618**	**23.1608**	**29.1819**	**29.1982**	**11.8487**	**12.1602**
HPD	SDXL	6.1338	6.1192	22.7835	22.7657	28.6278	28.6167	14.2736	14.4258
	DPO	6.1124	6.1310	23.1330	23.1520	29.1650	29.1740	14.6354	14.7891
	InPO	6.1797	6.1666	23.2761	23.2338	29.2413	29.2557	14.6645	14.8242
	MaPO	**6.2416**	**6.2453**	22.8488	22.8169	28.9939	29.0085	14.3608	14.4453
	TCPO(Ours)	6.1917	6.1906	**23.3950**	**23.3790**	**29.5000**	**29.4886**	**14.8954**	**15.1094**
Pick-a-Pic	SDXL	6.0040	5.9788	22.1655	22.2426	27.9776	28.0080	11.5371	11.8398
	DPO	6.0127	6.0168	22.6397	22.6137	28.5933	28.5017	11.9866	12.1719
	InPO	6.0416	6.0167	22.5820	22.5573	28.5171	28.6445	11.9247	12.1992
	MaPO	**6.2037**	**6.2294**	22.3179	22.3404	28.5381	28.6262	11.6570	11.8242
	TCPO(Ours)	6.0744	6.0883	**22.8045**	**22.8264**	**28.8740**	**28.9950**	**12.2160**	**12.4141**

Image quality is assessed through four evaluators: the LAION aesthetic classifier [31] for overall visual appeal, MPS [47] for comprehensive multi-dimensional evaluation, together with PickScore [13] and HPS [43] for overall human preference prediction on image-text pairs. To validate the effectiveness of our approach, we generated images using both the baseline and our model on DrawBench (200 prompts) [29], Parti-Prompts (1632 prompts) [45], HPDv2 (3200 prompts) [43] and the Pick-a-Pic v2 [13] test set (500 prompts). We present the mean and median for all metrics in each individual dataset.

Additionally, we conduct a user study (20 participants) to compare TCPO with the baselines. We design three questions as in [40]: (1) Which image is your overall preferred choice? (2) Which image is more visually attractive? (3) Which image better matches the text description?

We also compare the training GPU hours (on A100) with baselines. In this work, we adopt the following baselines: base model (SDXL), Diffusion-DPO, and Diffusion-InPO.

5.3 Main Results

Qualitative Results. As illustrated in Fig. 4, our TCPO demonstrates superior text-image alignment and visual fidelity over state-of-the-art baselines.

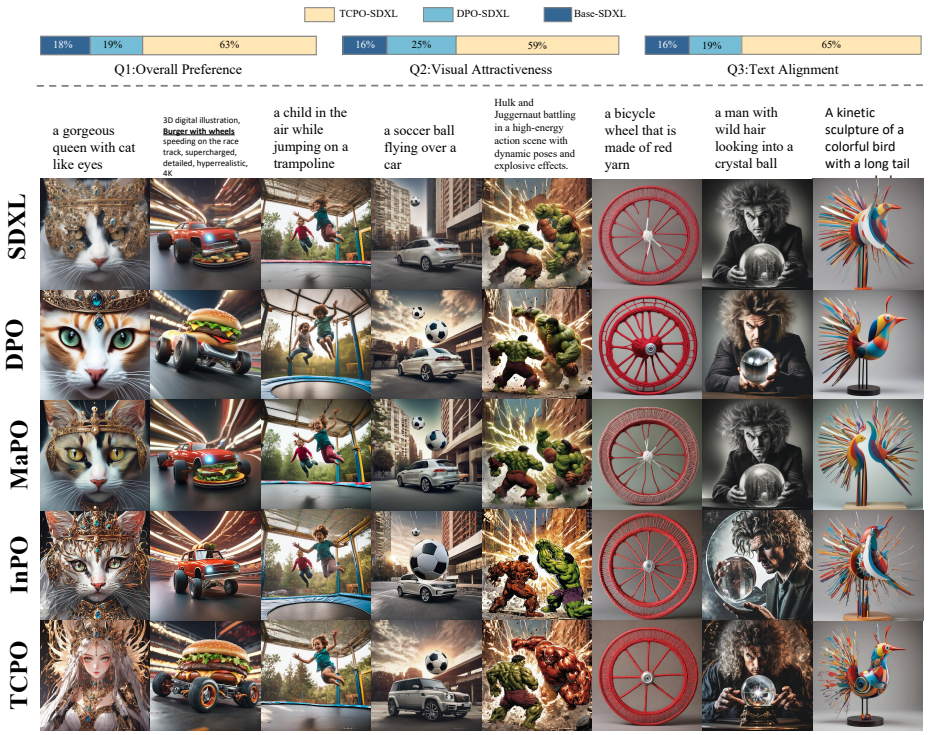

Fig. 4: Qualitative Comparisons. Visual comparison between SDXL, DPO, MaPO, InPO, and our TCPO. Across multiple evaluation dimensions, our method demonstrates superior capability in understanding human intentions, yielding high-quality images with accurate text-image alignment.

Regarding attribute binding (first two columns), baselines often suffer from severe subject hijacking—generating a cat rather than "a gorgeous queen with cat-like eyes". In contrast, TCPO accurately renders a human queen with feline traits and seamlessly integrates a "burger with wheels". For spatial reasoning and numeracy (third and fourth columns), TCPO correctly synthesizes an airborne child and a single soccer ball, avoiding the object duplication seen in other models. Furthermore, TCPO successfully disentangles interacting characters (Hulk and Juggernaut) without feature bleeding, while authentically capturing intricate textures (red yarn wheel) and complex structures (kinetic bird). Overall, TCPO excels in translating complex human intentions into high-fidelity, strictly aligned visual outputs.

Quantitative Results. We conduct a comprehensive quantitative comparison (see Tab. 1) between our proposed TCPO method and current state-of-the-art human preference baselines. As summarized in Tab. 1, after fine-tuning with our TCPO approach, SDXL consistently outperforms all baseline models on nearly all evaluation metrics across four diverse test datasets, clearly validating the ro-

Table 2: Ablation studies of our TCPO on fine-tuning SDXL. In each column of preference evaluators, the highest value is highlighted in bold. For detailed analysis, refer to Sec. 5.4

	Aesthetic		**PickScore**		**HPS**		**MPS**	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Base-SDXL	6.0040	5.9788	22.1655	22.2426	27.9776	28.0080	11.5371	11.8398
InPO-SDXL	6.0416	6.0167	22.5820	22.5573	28.5171	28.6445	11.9247	12.1992
$\gamma = 0.0$	6.0107	6.0071	22.7554	22.7432	28.6238	28.6176	10.2044	10.3554
$\gamma = 0.4$	6.0071	5.9989	22.5730	22.5137	28.5480	28.5672	11.9249	12.0937
w/o SLERP	**6.0942**	6.0855	22.7661	22.8022	28.8214	28.8684	12.2125	12.3359
TCPO-SDXL	6.0744	**6.0883**	**22.8045**	**22.8264**	**28.8740**	**28.9950**	**12.2160**	**12.4141**

Table 3: Performance of our method with different hyperparameter choices (α). Results show consistent improvement over Base-SDXL and InPO in all metrics, demonstrating robustness to parameter variation, the highest value is highlighted in bold. For detailed analysis, refer to Sec. 5.4

	Aesthetic		**PickScore**		**HPS**		**MPS**	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Base-SDXL	6.0040	5.9788	22.1655	22.2426	27.9776	28.0080	11.5371	11.8398
InPO-SDXL	6.0416	6.0167	22.5820	22.5573	28.5171	28.6445	11.9247	12.1992
$\alpha = 1.43$	6.0654	6.0748	22.8019	**22.8399**	28.8297	28.8868	12.1869	12.1484
$\alpha = 3.33$	6.0742	**6.0961**	22.7939	22.7803	28.7977	28.8081	**12.2424**	**12.4218**
TCPO-SDXL	**6.0744**	6.0883	**22.8045**	22.8264	**28.8740**	**28.9950**	12.2160	12.4141

bustness and effectiveness of our proposed method. Specifically, results on the Pick-a-Pic test set show that TCPO achieves notable improvements, such as a 0.2807 increase in HPS over DPO and a 0.2225 boost in PickScore compared to InPO. Notably, on the PartiPrompts dataset, TCPO demonstrates comprehensive superiority by achieving the highest mean scores across all four evaluated metrics, including a remarkable HPS of 29.1819 and an MPS of 11.8487. Furthermore, to evaluate the generalizability of TCPO across different model architectures and parameter scales, we provide additional quantitative results on Flow Matching models (SD3.5) and small-scale models (SSD-1B) in Section E of the supplementary material.

5.4 Ablation Study

We conduct comprehensive ablation studies on Pick-a-Pic v2 test set to evaluate the key components of TCPO, with all quantitative results summarized in Tab. 2, Tab. 3, and Section F of the supplementary material. It is important to note that setting the TRF fusion weight $\gamma = 0$ (which nullifies the geometric anchor) or

setting the anchor position $\alpha = 1$ (which aligns the anchor with the endpoint) mathematically degrades our method to the standard DPO baseline.

Impact of Anchor Position (α). We evaluate the stability of our Curvature Adaptive Sampling (CAS) by varying $\alpha \in \{1.43, 2.0, 3.33\}$ (corresponding to 1/0.7, 1/0.5, and 1/0.3, respectively). As shown in Table 3, TCPO demonstrates robust performance across different α values, consistently outperforming the baseline ($\alpha = 1$). This confirms that targeting an intermediate geometric anchor is fundamentally superior to endpoint-directed optimization.

Impact of TRF Fusion Weight (γ). To assess the trade-off between geometric preservation and preference injection, we test $\gamma \in \{0, 0.2, 0.4\}$, as shown in the third and fourth rows of Tab. 2. The results indicate that an appropriate fusion weight ($\gamma = 0.2$) optimally respects the model's naive generative dynamics while effectively learning human preferences. Conversely, an excessively large weight ($\gamma = 0.4$) makes the optimization overly conservative, leading to insufficient semantic learning, which perfectly aligns with our theoretical expectations.

Effectiveness of SLERP. Finally, we ablate the interpolation function in TRF by comparing Spherical Linear Interpolation (SLERP) against standard Linear Interpolation (LERP). As demonstrated in the fifth row of Tab. 2, SLERP yields consistently better alignment scores. This empirical result validates our geometric hypothesis: SLERP successfully avoids the off-manifold norm collapse toward the origin caused by naive linear interpolation, thereby maintaining the structural integrity of the generated images.

5.5 Training Efficiency Analysis

In Fig. 5 and Tab. 4, we compare our proposed TCPO against baselines in terms of image quality and training efficiency. As shown in Fig. 5, TCPO significantly accelerates convergence, achieving training speeds 4.31× and 1.23× faster than Diffusion-DPO and Diffusion-InPO, respectively. Furthermore, Tab. 4 confirms that this reduced computational overhead does not compromise performance; instead, TCPO consistently generates higher-quality images, demonstrating a superior balance between efficiency and effectiveness.

6 Conclusion

In this paper, we uncover a critical issue in diffusion preference alignment: the Linear Shortcut Fallacy inherent in standard objectives (e.g., DPO). This linear assumption ignores the intrinsically curved generative trajectory, pushing models off the data manifold and degrading generation quality. To address this, we propose TCPO. First, CAS provides an on-manifold target to maintain geometric consistency. Then, TRF integrates this geometric anchor with a semantic anchor, rectifying the prediction direction to prevent semantic deviation. Extensive experiments demonstrate that TCPO effectively mitigates off-manifold drift by strictly aligning with intrinsic trajectories. Ultimately, incorporating such geometric priors provides a new perspective for advancing diffusion alignment.

Table 4: Quantitative comparison of training efficiency and generation quality. We report the training GPU hours (measured on NVIDIA A100 GPUs) and HPS scores of TCPO alongside baseline methods. The best results for each metric are highlighted in bold.

Methods	DPO	InPO	TCPO
HPS↑	28.5017	28.6445	**28.995**
GPU Hours↓	∼776	∼221	**∼180**

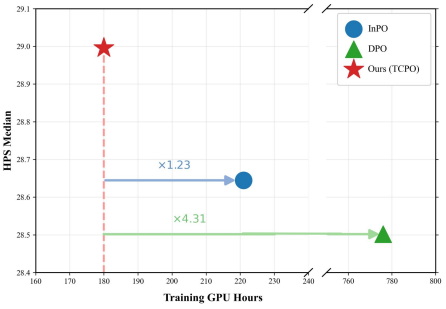

Fig. 5: Comparative evaluation of image quality and training efficiency.

Acknowledgement

This work is supported by New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2025ZD0123505), National Natural Science Foundation of China (Grant Nos. 62425606, 62550062, 62576338, 62576342, 32341009), Beijing Natural Science Foundation (Grant Nos. L252145, L257008, 4252054), and Beijing Nova Program (Grant No. 20240484601). This work is also sponsored by CCF-ALIMAMA TECH Kangaroo Fund (NO. CCF-ALIMAMA OF 2025006).

References

1. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742 (2025)
2. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception Prioritized Training of Diffusion Models. In: CVPR (2022)
3. Dai, L., Wei, X., Wang, H., Wang, S., Tang, J.: Aligning Text-to-Image Diffusion Models to Human Preference by Classification. In: NeurIPS (2025)
4. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: NeurIPS (2021)
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: ICML (2024)
6. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D.: KTO: Model Alignment as Prospect Theoretic Optimization. In: ICML (2024)
7. Ge, S., Huang, C., Ai, Y., Fan, Q., Huang, H., He, R.: Expand and Prune: Maximizing Trajectory Diversity for Effective GRPO in Generative Models. In: CVPR (2026)
8. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2022)

9. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020)
10. Hong, J., Lee, N., Thorne, J.: ORPO: Monolithic Preference Optimization without Reference Model. In: EMNLP (2024)
11. Hong, J., Paul, S., Lee, N., Rasul, K., Thorne, J., Jeong, J.: Margin-aware Preference Optimization for Aligning Diffusion Models without Reference. In: ICLR (2025)
12. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: ICLR (2022)
13. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In: NeurIPS (2023)
14. Lee, S., Kim, B., Ye, J.C.: Minimizing Trajectory Curvature of ODE-based Generative Models (2023)
15. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic Post-Training Diffusion Models from Generic Preferences with Step-by-step Preference Optimization. In: CVPR (2025)
16. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. In: ICLR (2023)
17. Liu, X., Gong, C., Liu, Q.: Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In: ICLR (2022)
18. Liu, X., Liu, J., Wang, H., He, R., Huang, H.: Think-Then-Generate: Structural Chain-of-Thought Reasoning for Consistent 3D Generation. In: CVPR (2026)
19. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In: NeurIPS (2022)
20. Lu, Y., Wang, Q., Cao, H., Wang, X., Xu, X., Zhang, M.: InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment. In: CVPR (2025)
21. Lu, Y., Wang, Q., Cao, H., Xu, X., Zhang, M.: Smoothed Preference Optimization via ReNoise Inversion for Aligning Diffusion Models with Varied Human Preferences. In: ICML (2025)
22. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. In: ICLR (2023)
23. Na, B., Park, M., Sim, G., Shin, D., Bae, H., Kang, M., Kwon, S.J., Kang, W., Moon, I.C.: Diffusion Adaptive Text Embedding for Text-to-Image Diffusion Models. In: NeurIPS (2025)
24. Peng, L., Wu, B., Cheng, H., Zhao, Y., He, X.: SUDO: Enhancing Text-to-Image Diffusion Models with Self-Supervised Direct Preference Optimization. In: NeurIPS (2025)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: ICLR (2023)
26. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In: NeurIPS (2024)
27. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 (2022)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022)
29. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J.,

Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: NeurIPS (2022)
30. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. In: SIGGRAPH ASIA (2024)
31. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
32. Shazeer, N., Stern, M.: Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In: ICML (2018)
33. Shoemake, K.: Animating rotation with quaternion curves. In: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '85 (1985)
34. Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., Wang, H.: Preference Ranking Optimization for Human Alignment. In: AAAI (2024)
35. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: ICLR (2022)
36. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: ICLR (2021)
37. Sun, J., Wang, H., Cao, J., Huang, H.: Marmot: Object-level self-correction via multi-agent reasoning. Machine Intelligence Research (2026)
38. Sun, J., Wang, P., Wang, H., Liu, X., Huang, H., He, R.: Towards fine-grained attribution: Instance-aware preference optimization for aligning diffusion models. In: CVPR (2026)
39. Tee, J.T.J., Yoon, H.S., Syarubany, A.H.M., Yoon, E., Yoo, C.D.: A Gradient Guidance Perspective on Stepwise Preference Optimization for Diffusion Models. In: NeurIPS (2025)
40. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion Model Alignment Using Direct Preference Optimization. In: CVPR (2024)
41. Wang, F.Y., Shui, Y., Piao, J., Sun, K., Li, H.: Diffusion-NPO: Negative Preference Optimization for Better Preference Aligned Generation of Diffusion Models. In: ICLR (2025)
42. Wang, H., Huang, H., Wang, P., Hao, J., Zhou, C., He, R.: Coloring the noise: Adversarial sobolev alignment for faithful image super-resolution. In: International Conference on Machine Learning (ICML) (2026)
43. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. arXiv:2306.09341 (2023)
44. Yang, W., Cao, J., Duan, J., He, R.: Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. In: ICCV (2025)
45. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. TMLR (2022)
46. Yuan, H., Chen, Z., Ji, K., Gu, Q.: Self-Play Fine-Tuning of Diffusion Models for Text-to-Image Generation. In: NeurIPS (2024)

47. Zhang, S., Wang, B., Wu, J., Li, Y., Gao, T., Zhang, D., Wang, Z.: Learning Multi-dimensional Human Preference for Text-to-Image Generation. In: CVPR (2024)
48. Zhang, T., Da, C., Ding, K., Yang, H., Jin, K., Li, Y., Gao, T., Zhang, D., Xiang, S., Pan, C.: Diffusion Model as a Noise-Aware Latent Reward Model for Step-Level Preference Optimization. In: NeurIPS (2025)
49. Zhu, H., Xiao, T., Honavar, V.: DSPO: Direct score preference optimization for diffusion model alignment. In: ICLR (2025)