Self-Supervised Direct Preference Optimization for Text-to-Image Diffusion Models

Liang Peng[1*†] **Boxi Wu**[2*‡] **Haoran Cheng**[2*] **Yibo Zhao**[1,2] **Xiaofei He**[1,2]
[1]FABU Inc. [2]Zhejiang University
{pengliang, wuboxi, chenghaoran}@zju.edu.cn

Abstract

Direct preference optimization (DPO) is an effective method for aligning generative models with human preferences and has been successfully applied to fine-tune text-to-image diffusion models. Its practical adoption, however, is hindered by a labor-intensive pipeline that first produces a large set of candidate images and then requires humans to rank them pairwise. We address this bottleneck with self-supervised direct preference optimization, a new paradigm that removes the need for any pre-generated images or manual ranking. During training, we create preference pairs on the fly through self-supervised image transformations, allowing the model to learn from fresh and diverse comparisons at every iteration. This online strategy eliminates costly data collection and annotation while remaining plug-and-play for any text-to-image diffusion method. Surprisingly, the on-the-fly pairs produced by the proposed method not only match but exceed the effectiveness of conventional DPO, which we attribute to the greater diversity of preferences sampled during training. Extensive experiments with Stable Diffusion 1.5 and Stable Diffusion XL confirm that our method delivers substantial gains.

1 Introduction

Text-to-image diffusion models [1, 2, 3, 4, 5, 6, 7, 8] have emerged as a dominant paradigm for high-quality image generation conditioned on natural language prompts. They have demonstrated the ability to synthesize diverse and visually appealing images across a wide range of prompts and styles. However, while these models are typically pretrained on large-scale datasets, they can fail to align with human preferences, especially in applications requiring text-image alignment or subjective aesthetic quality. To bridge this gap, recent works have explored preference-based post-training strategies that adjust the model's outputs based on human feedback.

One prominent method in this direction is Direct Preference Optimization (DPO) [9, 10], which trains generative models using pairwise human preference data. DPO has shown strong results in aligning text-to-image models with user intent. Nevertheless, its practical adoption is hindered by a costly and rigid training pipeline: it first requires the offline generation of a large set of image candidates, followed by extensive human annotation in the form of pairwise ranking. This process is not only time-consuming and expensive but also inflexible—once the ranking data is collected, it cannot easily adapt to new prompts or domains. Furthermore, the fixed nature of the dataset may limit the diversity of learning signals, potentially affecting generalization.

In this work, we present Self-Supervised Direct Preference Optimization (Self-DPO), a novel framework that eliminates the reliance on pre-generated images and manual rankings. Instead of requiring

*Equal contribution
†Work was done at FABU Inc.
‡Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

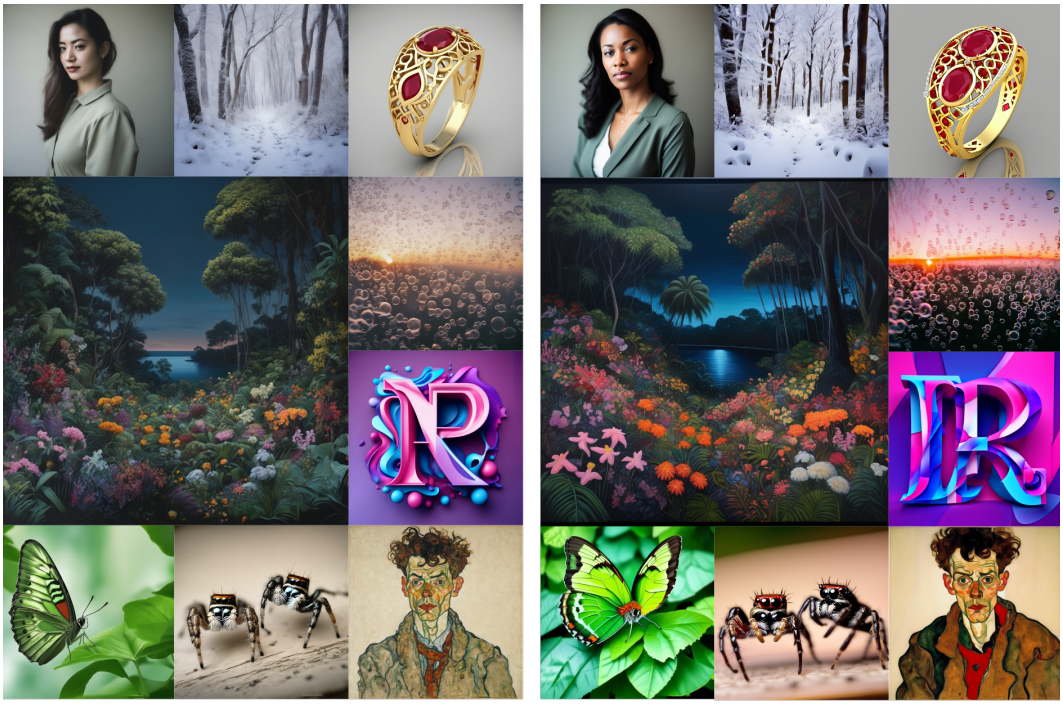

Figure 1: We propose Self-DPO, incorporating direct preference optimization in a self-supervised manner. We provide the comparisons on SDXL base model with the same seed. Our method requires less data, yet generate more visually appealing results. Best viewed in color with zoom in.

external preference data, our method constructs training pairs dynamically using self-supervised image transformations. During each training iteration, the model first identifies a "winning image" that satisfies human-aligned quality criteria. We then generate a corresponding "losing image" by intentionally degrading the winner, either through visual-quality reductions or text–image misalignment. The resulting win–lose pair supplies an immediate preference signal, allowing the model to perform online direct preference optimization. By learning from this continuous stream of synthetically generated preference pairs, Self-DPO achieves effective human alignment.

This self-supervised, on-the-fly approach brings several key advantages. It removes the need for costly pre-generated images and human ranking efforts, enables scalable and dynamic training, and introduces greater diversity into the preference supervision signal. We summarize the data requirements in Table 1. Remarkably, we find that Self-DPO not only matches but surpasses conventional DPO in both qualitative and quantitative metrics (*e.g.*, boosting ImageReward win rate from 61 to 85). We provide some visualization results in Figure 1. Experiments on Stable Diffusion 1.5 [1] and XL [2] demonstrate consistent improvements in visual quality and text-image alignment. Our method is plug-and-play, generalizable across architectures, and easily integrable into existing diffusion model training pipelines.

Data requirements	SFT	Self-DPO	DPO
Text Caption	✓	✓	✓
Image	✓	✓	✓
Extra Preference Image	–	–	✓

Table 1: Post-training data requirements. Self-DPO shares the same data requirements as SFT [11]. DPO [10] requires extra preference images that are collected and annotated by humans, which is highly time-consuming and expensive.

2 Related Work

2.1 Text-to-Image Diffusion Models

Text-to-image diffusion models have recently attracted significant attention in generative modeling due to their ability to produce high-quality and diverse images from textual descriptions. The introduction

of Denoising Diffusion Probabilistic Models (DDPMs) [12, 13] marked a major breakthrough in this field, establishing diffusion models as a powerful generative approach. Diffusion models [12, 14, 15] operate by reversing a gradual diffusion process, where noise is incrementally added to the clean latent and then learned to be removed, ultimately synthesizing a coherent and realistic image/video. Building upon this foundation, diffusion models have become a representative paradigm in text-to-image generation, leading to models like GLIDE [16] for text-guided image editing, Imagen [17] with cascaded diffusion for high-resolution synthesis, and Stable Diffusion [1] using latent diffusion for efficient generation. To enhance image generation quality, researchers typically focus on two main directions. One approach involves architectural improvements, with the Diffusion Transformer (DiT) [3] gaining attention for its improved image fidelity, performance, and diversity. Notable advancements include PixArt-α [18], Hunyuan-DiT [19], and SD3 [4]. Another approach leverages supervised fine-tuning to refine text-to-image diffusion models. These methods curate datasets by integrating various strategies, such as preference models [2], pre-trained image models [20, 21, 22, 23, 24] (e.g., image captioning models), and expert-assisted data filtering [11].

2.2 Preference-Based Optimization Methods

In recent years, preference-based optimization has gained traction, refining models through user feedback or ranked preference pairs. In Large Language Models, Reinforcement Learning from Human Feedback (RLHF) leverages human comparisons to train a reward model that guides policy learning [25, 26]. Alternatively, direct preference optimization (DPO) fine-tunes models directly on preference data, bypassing the need for an explicit reward model while achieving comparable performance [9]. Subsequently, preference-based optimization has been applied to image generation. Some methods enhance image quality by increasing rewards for preferred outputs [27, 28, 29], while others use reinforcement learning [29, 30]. However, training reliable reward models remains challenging and computationally expensive, with over-optimization potentially leading to mode collapse, reducing diversity [31, 28]. Similarly, Direct Preference Optimization has been introduced in text-to-image generation, with Diffusion-DPO [10] demonstrating the effectiveness of optimizing on human comparison data to enhance both visual appeal and text alignment. Additionally, direct score preference optimization [32] refines diffusion models through score matching, providing a novel approach to preference learning. Several recent studies have further explored adapting preference learning techniques from large language models to fine-tune diffusion models [33, 34**?** , 35], highlighting the growing interest in aligning generative models with human preferences.

2.3 Self-Supervised Learning

Self-supervised learning has emerged as a pivotal paradigm in machine learning. Among its most widely used approaches are contrastive self-supervised learning and generative self-supervised learning. Contrastive self-supervised learning distinguishes representations by pulling similar instances closer while pushing dissimilar ones apart. MoCo [36] enhances training efficiency through a momentum encoder, while SimCLR [37] simplifies the process with strong data augmentations. CLIP [38] extends contrastive learning to vision-language tasks, aligning images and text in a shared latent space, enabling zero-shot transfer. Generative approaches inherently follow unsupervised or self-supervised learning principles, training without labeled data to model the underlying distributions of the input. GANs [39] utilize adversarial training to generate realistic data, while VAEs [40] encode data into a structured latent space for controlled synthesis. VQ-VAE [41] introduces discrete latent representations for high-quality generation. MAE [42] leverages masked image modeling to learn rich visual features. More recently, denoising diffusion models [13] have demonstrated impressive results by iteratively adding and removing noise, learning robust representations in a self-supervised manner. Self-supervised learning enables training without manually labeled data, laying the foundation for future advancements in representation learning, generative modeling, and multimodal understanding.

3 Methods

3.1 Preliminary

Diffusion Models. Denoising Diffusion Probabilistic Models [13] represent the image generation process as a Markovian process. Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be the data point and $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ denote the forward

process, where noise is added to the data at each timestep t. The forward process is defined as:

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

where β_t is a schedule that controls the variance of noise added at each timestep, and $\mathcal{N}(\cdot)$ denotes the normal distribution. The forward process gradually adds noise, with $\boldsymbol{x}_T$ being pure noise after T timesteps. The reverse process aims to learn the distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, which represents the process of denoising and generating data from pure noise. The reverse process can be parameterized as:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)), \tag{2}$$

where $\mu_\theta(\boldsymbol{x}_t, t)$ and $\Sigma_\theta(\boldsymbol{x}_t, t)$ are the mean and covariance learned by the model at each timestep t. The model is trained by minimizing the following objective:

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

where ϵ is the noise added at each timestep, and $\epsilon_\theta(\boldsymbol{x}_t, t)$ is the model's predicted noise. The model is trained to predict this noise accurately at each step, enabling it to reverse the diffusion process and generate high-quality data.

Reinforcement Learning from Human Feedback. For diffusion models, human preferences at each diffusion step are modeled using a Bradley-Terry formulation [43], where the probability of preferring a "winning" sample $\boldsymbol{x}_t^w$ over a "losing" sample $\boldsymbol{x}_t^l$ for a given prompt $\boldsymbol{c}$ is defined as:

$$p_{\text{BT}}(\boldsymbol{x}_t^w \succ \boldsymbol{x}_t^l|\boldsymbol{c}) = \sigma\big(r(\boldsymbol{c}, \boldsymbol{x}_t^w) - r(\boldsymbol{c}, \boldsymbol{x}_t^l)\big), \tag{4}$$

with $\sigma, r, \boldsymbol{c}$ representing the sigmoid function, reward model, and prompt, respectively. Subsequently, by conceptualizing the diffusion denoising process as a multi-step Markov Decision Process, the generative model is fine-tuned via reinforcement learning. The training objective [10, 44] is formulated as:

$$\mathcal{L}_{\text{rlhf}} = \mathbb{E}_{\boldsymbol{c}\sim D}\mathbb{E}_{p_\theta(\boldsymbol{x}_{0:T}|\boldsymbol{c})}\sum_{t=0}^{T-1} r(\boldsymbol{c}, \boldsymbol{x}_t) - \lambda\,\mathbb{D}_{\text{KL}}\big(p_\theta(\boldsymbol{x}_{0:T}|\boldsymbol{c}) \,\|\, p_{\text{ref}}(\boldsymbol{x}_{0:T}|\boldsymbol{c})\big), \tag{5}$$

where $p_{\text{ref}}(\boldsymbol{x}_{0:T}|\boldsymbol{c})$ is the distribution from the pretrained diffusion model and λ controls the influence of the KL divergence regularization term.

Direct Preference Optimization (DPO). DPO streamlines RLHF by using the learning policy's log likelihood to imp licitly encode the reward. In text-to-image diffusion models, this leads to a step-wise reward defined as:

$$r(\boldsymbol{c}, \boldsymbol{x}_t) = \lambda \log \frac{p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}, \boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}, \boldsymbol{c})}. \tag{6}$$

Recent works [10, 32] follow this line and adapt it to diffusion models. They optimize the model p_θ based on the Bradley-Terry model [43], leading to an objective function:

$$\mathcal{L}_{\text{Diffusion-DPO}} = -\mathbb{E}\Big[\log\sigma\Big(\lambda\log\frac{p_\theta(\boldsymbol{x}_t^w \mid \boldsymbol{x}_{t+1}^w, \boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_t^w \mid \boldsymbol{x}_{t+1}^w, \boldsymbol{c})} - \lambda\log\frac{p_\theta(\boldsymbol{x}_t^l \mid \boldsymbol{x}_{t+1}^l, \boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_t^l \mid \boldsymbol{x}_{t+1}^l, \boldsymbol{c})}\Big)\Big], \tag{7}$$

where the winning and losing samples $(\boldsymbol{x}_t^w, \boldsymbol{x}_t^l)$ and the prompt $\boldsymbol{c}$ are drawn from the dataset, and timestep t is uniformly sampled from the diffusion process. This formulation effectively aligns the learning policy with human preference signals embedded in the reference model.

3.2 Self-DPO for Text-to-image Diffusion Models

Inspired by DPO's effective alignment with human preferences at the image level [10], our work aims to extend this formulation into the standard post-training process for diffusion models (*e.g.*, SFT). Unfortunately, a direct application of DPO is not feasible because it requires collecting image pairs (typically generated by different models with the same prompt or by using different seeds with the same model—along with their associated manual rankings). This approach does not align with the conventional fine-tuning pipeline and incurs significant additional costs. To overcome this limitation, we propose Self-DPO, which generates preference image pairs in a self-supervised manner.

In each training iteration, we denote the text-image pair as $(\boldsymbol{c}, \boldsymbol{x})$, where the image component is regarded as the wining image $\boldsymbol{x}^w$. Traditional DPO methods [10, 9] require generating image pairs corresponding to the same prompt, followed by manual selection of the preferred (wining) and less preferred (losing) images, denoted as $\boldsymbol{x}^w$ and $\boldsymbol{x}^l$, respectively. Our method removes such cumbersome

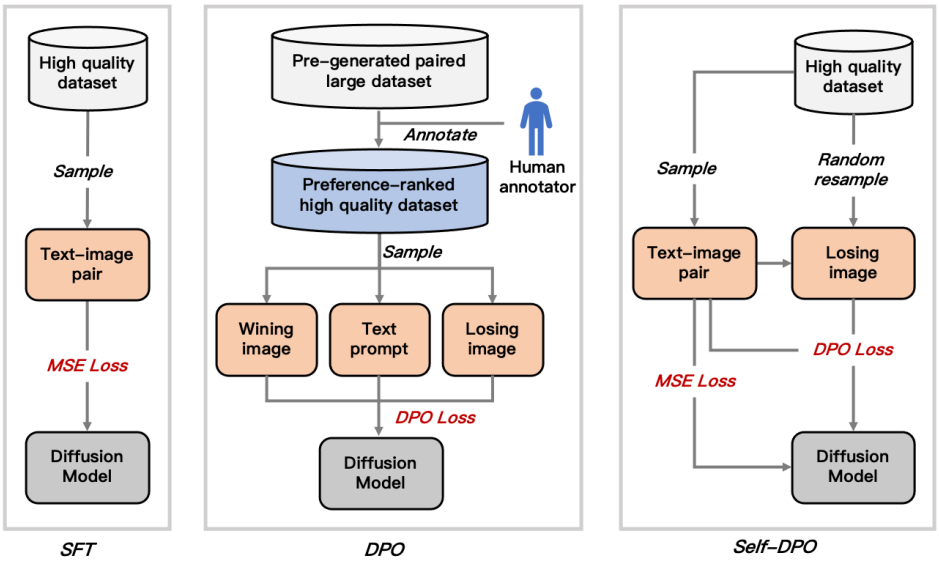

Figure 2: Different post-training processes. We generate the "losing" images self-supervisedly, enabling direct preference optimization without extra collecting and ranking steps. This lightweight procedure eliminates the substantial overhead of conventional DPO while retaining the same data requirements as standard SFT. Best viewed in color with zoom in.

steps. Because every image in the curated high-quality dataset already satisfies human preference, each can be regarded as a winner $\boldsymbol{x}^w$. We obtain a corresponding self-supervised losing sample by deliberately degrading the winner: $\boldsymbol{x}^{sl} = \textbf{Downgrade}(\boldsymbol{x}^w)$. Inspired by [10], the self-supervised direct preference loss is:

$$\begin{aligned}\mathcal{L}_{\text{Self-DPO}} = -\log\sigma\big(C\big(&\left(\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w,t)-\boldsymbol{\epsilon}^w\|_2^2-\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{sl},t)-\boldsymbol{\epsilon}^{sl}\|_2^2\right)\\&-\left(\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w,t)-\boldsymbol{\epsilon}^w\|_2^2-\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^{sl},t)-\boldsymbol{\epsilon}^{sl}\|_2^2\right)\big)\big),\\&\qquad where\quad \boldsymbol{x}^{sl}=\textbf{Downgrade}(\boldsymbol{x}^w)\end{aligned}\tag{8}$$

where C and $\boldsymbol{\epsilon}^{sl}$ refer to a scale factor and the noise corresponding to losing images, and $\boldsymbol{\epsilon}_{\text{ref}}$ is the reference model. In our experiments, the **Downgrade** operation can be simply performed by randomly selecting images from the training dataset. For each self-generated image pair, the winning sample closely aligns with the prompt, whereas the losing sample fails to correspond to the description. Surprisingly, this simple manner brings significant improvements to the model. We also compare different degradation strategies in the experiments. The training process overview is shown in Figure 2. The final loss is as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{MSE}} + \lambda_2\mathcal{L}_{\text{Self-DPO}}\tag{9}$$

We empirically set λ_1 and λ_2 to $0.5, 0.5$, respectively.

4 Experiments

4.1 Setup

Implement Details: Following Diffusion-DPO [10], for the SD1.5 [1] experiments, AdamW [47] is utilized, while SDXL [2] training is conducted with Adafactor [48] to conserve memory. Following the official implementation in Diffusion-DPO [10], C in Equation 8 is set to -2500. For SD 1.5, a batch size of 2048 pairs (resolution: $512 * 512$) is maintained by training across 4 NVIDIA A100 GPU. Each GPU handles 8 pairs locally with gradient accumulation

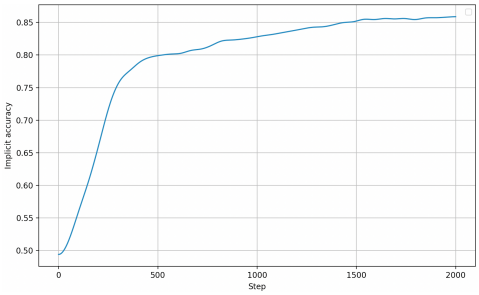

Figure 5: Implicit accuracy during the training stage.

Figure 3: Qualitative comparisons with the SD1.5 base model. All results are generated with the same random seed. Comparing with SFT and DPO [10], the model trained by Self-DPO demonstrates superior text prompt alignment. It also shows more appealing visual quality, especially in layout, colors, and details. Best viewed in color.

over 64 steps. For SDXL, considering the resource limitation, we use the total batch size of 96 pairs (resolution: $1024 * 1024$). Training is performed at fixed square resolutions. We use a learning rate of 1e-6 coupled with a 25% linear warmup.

Training Dataset: Our training data is sourced from the Pick-a-Pic V2 dataset[45], which keeps the same to Diffusion-DPO [10]. It contains pairwise preference annotations for images generated by Dreamlike (a fine-tuned variant of SD1.5), SD2.1, and SDXL. These prompts and preferences were collected from users of the Pick-a-Pic web application. ***Please note that Self-DPO only uses the preference image and associated text in the dataset, instead of using the whole manually annotated image pairs.***

Evaluation: We conduct evaluation on three datasets: Pick-a-Pic V2 [45] validation set (contains 500 prompts), PartiPrompts [46] (contains 1632 prompts, including diverse categories and challenge aspects), and HPDv2 [24] (contains 3200 prompts, including anime, concept art, paintings and photo). We compare Self-DPO with three different type baselines, *i.e.*, base models (Stable Diffusion 1.5 (SD1.5) and Stable Diffusion XL (SDXL)), SFT models, DPO models, where DPO models are the

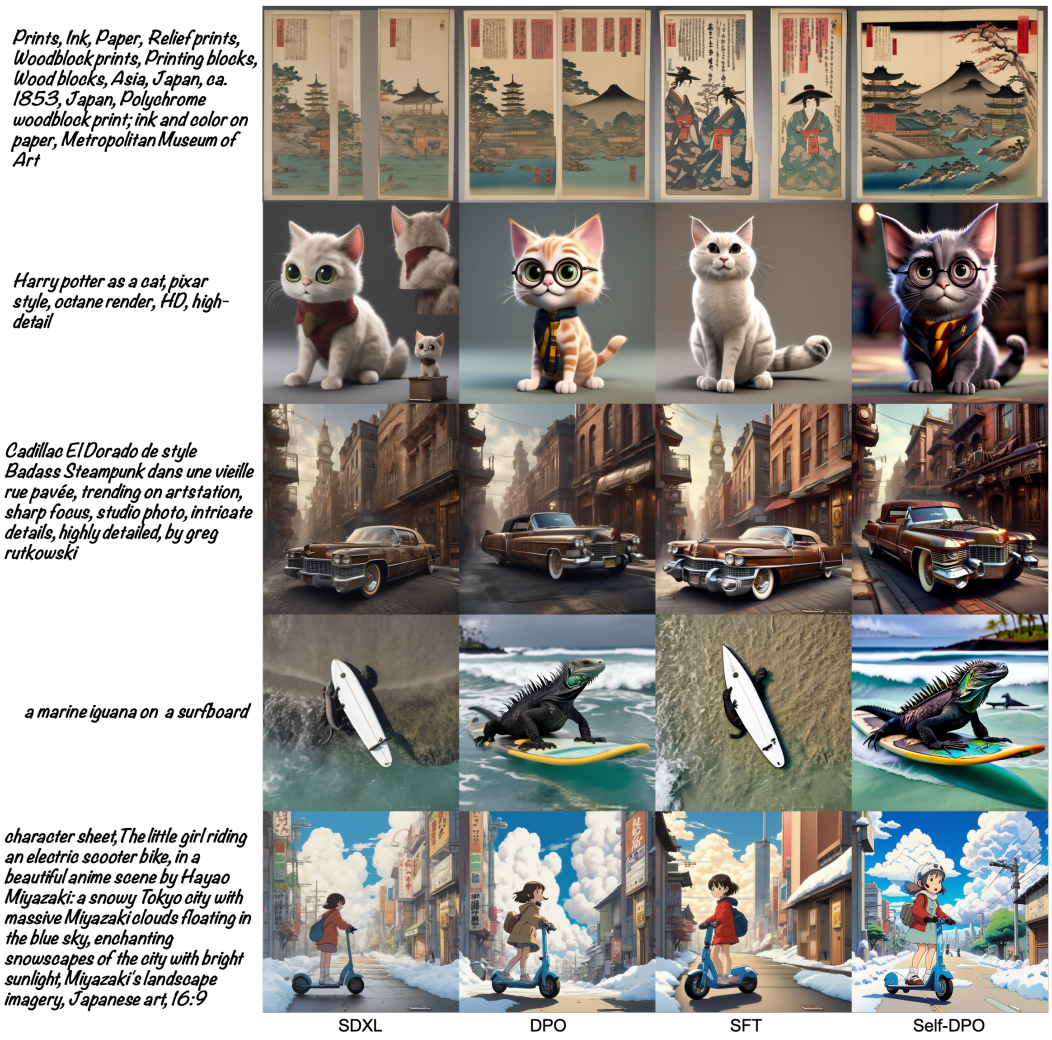

Figure 4: Qualitative comparisons with the SDXL base model. All results are generated using the same random seed. Please note that the training dataset (Pick-a-Pic V2 [45]) used for fine-tuning is obtained from a SD1.5 variant, SD2.1, and SDXL models. Consequently, directly fine-tuning (SFT) on this dataset does not lead to improvements—and may even result in degraded performance (*e.g.*, as shown in the second row). In contrast, DPO [10] optimizes the model by leveraging preference relationships with pre-ranked pairs, thereby avoiding this issue. Self-DPO uses the same data requirements as SFT but yields significant improvements in both text prompt alignment and visual quality, demonstraing its effectiveness. Best viewed in color.

publicly released from Diffusion-DPO [10]. For evaluation metrics, we employ the popular PickScore [45] , Aesthetics [49], CLIP [38], HPS V2 [24] , and ImageAward [50] scores. For convenience of comparison, we scale the scores to fit a similar range ($\times$*100 for PickScore, CLIP, HPS, and ImageAward,* $\times$*10 for Aesthetics*). **PickScore** is a caption-sensitive scoring model, originally trained on Pick-a-Pic (v1), that estimates the perceived image quality by humans. **Aesthetics** assesses the visual appeal of an image, considering factors such as lighting, color harmony, composition, and overall artistic quality. **CLIP** measures the semantic alignment between an image and a corresponding text prompt. By computing the cosine similarity between the image and text embeddings, this score evaluates how well the image content matches the provided textual description. **Human Preference Score (HPS V2)** is a metric designed to align with human judgments of image quality, particularly in the context of text-to-image synthesis. **ImageAward** quantifies the quality, aesthetic appeal, or alignment of a generated image with respect to desired attributes. It is typically derived from a reward model trained on human preference data.

Datasets	Methods		SD1.5					SDXL				
			P.S.	Aes.	CLIP	HPS	I.R.	P.S.	Aes.	CLIP	HPS	I.R.
Pick-a-Pic V2	Base	Avg score	20.57	53.15	32.58	26.17	-14.81	22.10	**60.01**	35.86	26.83	50.62
	SFT		21.10	**56.35**	33.75	27.03	45.03	21.48	57.84	35.67	26.67	30.89
	DPO		20.91	54.07	33.19	26.46	4.13	**22.57**	59.93	37.30	27.30	81.14
	Self-DPO		**21.23**	**56.35**	**34.79**	**27.33**	**71.00**	22.34	59.97	**37.53**	**27.89**	**103.96**
	SFT	Win rate	75.00	77.20	60.40	90.20	80.00	19.40	31.80	47.00	44.60	42.4
	DPO		73.80	60.00	60.00	71.80	61.00	**72.60**	47.20	**63.00**	79.80	69.8
	Self-DPO		**78.60**	**77.80**	**68.40**	**94.20**	**85.20**	60.80	**50.80**	62.40	**93.80**	**79.2**
PartiPrompts	Base	Avg score	21.39	53.13	33.21	26.79	1.48	22.63	57.69	35.77	27.33	69.78
	SFT		21.75	**55.31**	33.93	27.57	50.75	22.02	56.41	35.31	27.13	47.29
	DPO		21.61	53.58	33.88	26.98	21.43	**22.90**	57.85	36.95	27.73	103.36
	Self-DPO		**21.84**	55.09	**35.11**	**27.84**	**75.66**	22.79	**58.69**	**37.00**	**28.30**	**117.50**
	SFT	Win rate	67.28	**70.89**	53.43	85.42	73.35	21.38	38.11	45.10	43.75	40.93
	DPO		67.10	57.17	56.74	61.83	63.05	**63.42**	53.62	**62.32**	73.10	68.44
	Self-DPO		**69.85**	68.50	**63.24**	**89.40**	**81.00**	56.19	**60.48**	60.17	**92.16**	**76.84**
HPD V2	Base	Avg score	20.84	54.32	33.96	26.84	-11.79	22.78	61.34	37.68	27.68	78.27
	SFT		21.57	**57.41**	35.26	27.89	57.74	22.24	60.08	37.39	27.76	66.62
	DPO		21.30	55.80	34.68	27.22	13.24	**23.18**	**61.35**	**38.45**	28.14	102.74
	Self-DPO		**21.58**	57.10	**36.30**	**28.11**	**76.13**	22.98	61.30	38.35	**28.77**	**110.67**
	SFT	Win rate	**79.53**	**75.31**	59.34	90.10	81.16	23.47	37.28	46.63	58.22	45.81
	DPO		75.72	66.28	57.56	72.43	64.69	**72.66**	**50.28**	**58.69**	80.56	69.78
	Self-DPO		**79.53**	74.03	**68.47**	**92.49**	**85.19**	58.78	48.06	55.65	**94.97**	**72.50**

Table 2: Quantitative comparisons. We compare different fine-tuning methods (SFT, DPO [10], and Self-DPO) on SD 1.5 and SDXL base models over three datasets (Pick-a-Pic V2 [45], PartiPrompts [46], and HPDv2 [24]). "P.S." refers to PickScore, "Aes." is Aesthetics, and "I.R." denotes ImageAward. For the SD1.5 base model, our method achieves the best performance across most metrics. In contrast, for the SDXL base model, we observe that SFT clearly degrades performance. This is likely due to the fact that the training dataset (Pick-a-pic V2 [45]) used for fine-tuning is derived from SD1.5 variant, SD2.1, and SDXL models. Interestingly, our method still achieves competitive results compared to DPO, which utilizes the full dataset and optimizes the model by leveraging preference relationships with pre-ranked pairs. These results underscore the effectiveness and robustness of our approach.

4.2 Quantitative Results

We provide quantitative comparisons in Table 2. We compare our method with SFT and DPO [10]. For the SD1.5 base model, our method significantly outperforms the alternatives. For example, Self-DPO achieves a CLIP score of 34.79 on the Pick-a-Pic V2 dataset—an improvement of +2.21 over the base model—whereas SFT and DPO yield improvements of +1.17 and +0.61, respectively. In terms of overall human preference metrics, Self-DPO delivers substantially higher gains, improving the base model from –14.81 to 71.00 on the ImageReward metric, which far exceeds the improvements observed with SFT (45.03) and DPO (4.13). Notably, the win rate of Self-DPO reaches 94.20 on the HPS metric. Results on other datasets (PartiPrompts and HPD v2) further confirm these improvements. In contrast, results on the SDXL base model show a slightly different scenario, particularly with SFT. We observe that SFT substantially degrades the performance of the base model. This degradation is likely due to the fact that, while the training dataset (comprising generations from an SD1.5 variant, SD2.1, and SDXL) is of much higher quality than that used for SD1.5, it does not exhibit clear superiority for SDXL. Nevertheless, the model trained by Self-DPO still demonstrates significantly better performance compared to both SFT and the base model. On the PickScore, Aesthetics, and CLIP metrics, Self-DPO achieves results comparable to DPO, and it attains superior scores on the HPS and ImageRe-

Methods	P.S.	Aes.	CLIP	HPS	I.R.
Base model	20.57	53.15	32.58	26.17	-14.81
DPO	20.91	54.07	33.19	26.46	4.13
w/ Blur	20.80	55.16	33.22	26.57	4.26
w/ Random grid	20.85	55.86	32.71	26.64	24.21
w/ Random image	**21.23**	**56.35**	**34.79**	**27.33**	**71.00**

Table 3: Ablation on different downgrade manners. The "Random image" row achieves most significant improvements, implying that degrading the image quality can be detrimental.

ward metrics. These findings validate our hypothesis that image-level learning benefits text-to-image diffusion models and highlight the superiority of Self-DPO. We also employ UnifiedReward [51] as the VLM evaluation metric, a state-of-the-art reward model designed for multimodal understanding and generation tasks. Built upon strong VLM models [52, 53, 54, 55], it supports pointwise scoring to align model outputs with human preferences. The results are shown in Table 4.4. We further validate our method's effectiveness by using a new real-world T2I evaluation setting. Specifically, we conduct experiments on a subset of the LAION-COCO dataset containing 400k text-image pairs. Unlike carefully curated datasets, we do not manually design the data distribution or employ re-captioning techniques to refine prompts. The results are shown in Table 4.4. While naive SFT degrades performance, our Self-DPO method achieves consistent improvements. Please note that we cannot provide a DPO baseline as it requires additional paired images and annotations for the same prompts, which are not available in regular text-image dataset.

4.3 Qualitative Results

We provide qualitative comparisons in Figure 3 and Figure 4 for SD1.5 and SDXL, respectively. All quantitative and qualitative results are generated using the same random seed. We observe that Self-DPO consistently improves over other models, particularly in terms of producing more vivid colors and better adherence to text prompts. For example, in the third row of Figure 3, only Self-DPO successfully reveals the concept "cyberpunk". In the fourth row of Figure 4, both DPO and Self-DPO capture the intended meaning of the prompt, but Self-DPO exhibits a more appealing visual quality. These quantitative and qualitative results confirm the effectiveness of our approach. We show the implicit win rate to measure the optimization process in Figure 5. Specifically, the implicit win rate during training refers to the probability that the model prefers the winning image over the losing image. In other words, a higher implicit win rate indicates a stronger preference for the winning image. As training progresses, it steadily increases and eventually reaches around 0.85. This demonstrates that the model gradually learns to distinguish between the wsinning and losing images, and prefers to generate the winning image.

4.4 Ablation

We conduct ablation studies on various downgrade approaches, as shown in Table 3. In particular, we examine two additional methods—namely, blur (which involves downsampling and upsampling the image by a factor of 4) and random grid (which divides the image into an 8×8 grid and randomly swaps two grids). We observe that these two downgrade methods lead to worse performance, indicating that significantly degraded image quality can be harmful. Additionally, we remove the MSE loss and report the results in Table 4. Interestingly, the performance does not drop noticeably and consistently performs better than DPO, further confirming the effectiveness of Self-DPO.

Methods	P.S.	Aes.	CLIP	HPS	I.R.
Base model	20.57	53.15	32.58	26.17	-14.81
DPO	20.91	54.07	33.19	26.46	4.13
Self-DPO w/o MSE	**21.26**	56.22	**35.00**	27.31	67.78
Self-DPO	21.23	**56.35**	34.79	**27.33**	**71.00**

Table 4: Ablation on MSE loss. When removing MSE loss, Self-DPO still shows comparable performance and demonstrates the superiority to DPO.

Methods	P.S.	Aes.	CLIP	HPS	I.R.
Base	20.57	53.15	32.58	26.17	-14.81
SFT	20.49	52.63	31.83	26.13	-23.95
Self-DPO	20.68	52.87	34.31	26.52	17.11

Table 5: Results on a real-world image-text dataset.

Base model	Base	SFT	DPO	Self-DPO
SD1.5	2.44	2.62	2.53	2.72
SDXL	2.96	2.73	3.10	3.07

Table 6: Comparisons on UnifiedReward.

5 Limitation and Future Work

In the present implementation, we construct losing samples by randomly sampling images from the training set. This manner can become unreliable when the dataset is small, as the resulting

losers may fail to provide sufficiently informative preference signals. Future work therefore can explore more sophisticated self-supervised strategies for synthesizing losing images, such as more content-aware perturbations or adversarial degradations that better challenge the model. Furthermore, our approach is built on direct preference optimization, which is an offline reinforcement-learning paradigm. Extending the self-supervised pairing concept to online policy-gradient or actor-critic frameworks represents another promising direction.

6 Conclusion

In this paper, we propose Self-Supervised Direct Preference Optimization (Self-DPO), a fully self-supervised framework that aligns generative models with human preferences without requiring pre-generated image pools or manual rankings. By synthesizing win–lose pairs on the fly, Self-DPO not only removes costly data-collection steps but also exposes the model to a broader and more diverse set of preference signals. At every training iteration, it constructs its own preference pairs through controlled degradations of high-quality images and immediately updates the model via preference learning. Extensive experiments across multiple datasets and base architectures demonstrate that Self-DPO consistently delivers superior performance, validating its effectiveness and versatility.

Acknowledgments

This research was supported by The National Nature Science Foundation of China (Grant Nos: 62402417, 62273302,62036009, 61936006), in part by the Key R&D Program of Ningbo (Grant Nos: 2024Z115, 2025Z035), in part by Yongjiang Talent Introduction Programme (Grant No: 2023A-197-G).

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[2] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[5] Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. *arXiv preprint arXiv:2411.15488*, 2024.

[6] 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*, 2024.

[7] Black Forest Labs. Flux, 2024. Black Forest Labs. Flux, 2024.

[8] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. Diffusion model-based video editing: A survey. *arXiv preprint arXiv:2407.07111*, 2024.

[9] 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.

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

[11] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Rong-Cheng Tu, Wenhao Sun, Zhao Jin, Jingyi Liao, Jiaxing Huang, and Dacheng Tao. Spagent: Adaptive task decomposition and model selection for general video generation and editing. *arXiv preprint arXiv:2411.18983*, 2024.

[15] Wenhao Sun, Rong-Cheng Tu, Yifu Ding, Zhao Jin, Jingyi Liao, Shunyu Liu, and Dacheng Tao. Vorta: Efficient video diffusion via routing sparse attention. *arXiv preprint arXiv:2505.18809*, 2025.

[16] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[17] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[18] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[19] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

[20] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[21] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

[22] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[23] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, Zhao Jin, and Dacheng Tao. Asymrnr: Video diffusion transformers acceleration with asymmetric reduction and restoration. In *Forty-second International Conference on Machine Learning*.

[24] 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.

[25] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

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

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

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

[29] 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.

[30] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

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

[32] 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.

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

[34] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *Advances in Neural Information Processing Systems*, 37:24897–24925, 2024.

[35] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *Advances in Neural Information Processing Systems*, 37:73366–73398, 2025.

[36] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[37] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

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

[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[40] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[41] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[43] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[44] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023*. Neural Information Processing Systems Foundation, 2023.

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

[46] 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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[48] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

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

[50] 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.

[51] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

[52] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[54] Rong-Cheng Tu, Yatai Ji, Jie Jiang, Weijie Kong, Chengfei Cai, Wenzhe Zhao, Hongfa Wang, Yujiu Yang, and Wei Liu. Global and local semantic completion learning for vision-language pre-training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[55] Qwen Team. Qwen2.5-vl, January 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:[Yes]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.