Aesthetic Post-Training Diffusion Models from Generic Preferences with Step-by-step Preference Optimization

Zhanhao Liang[1], Yuhui Yuan[5], Shuyang Gu[5], Bohan Chen[2], Tiankai Hang[3],
Mingxi Cheng[4], Ji Li[4], Liang Zheng[1]
[1]The Australian National University [2]University of Liverpool
[3]Southeast University [4]Microsoft [5]Microsoft Research Asia

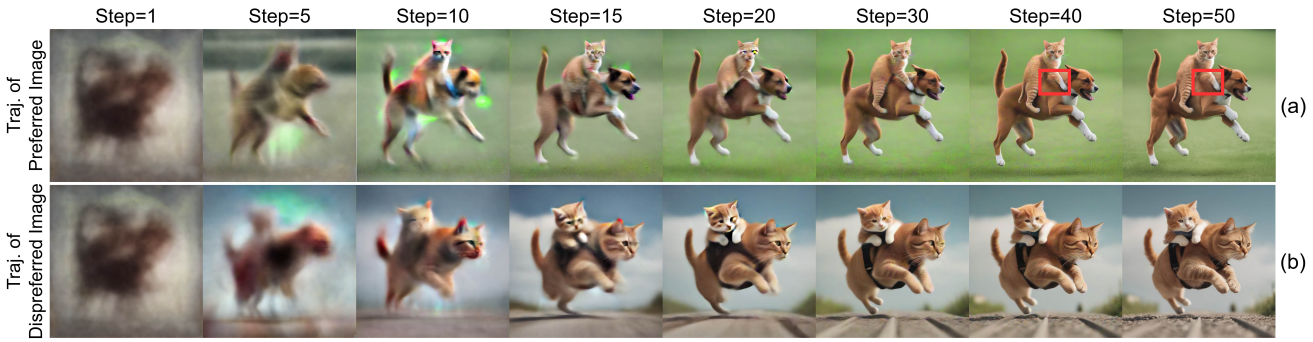

Figure 1. Problem illustration of existing DPO methods. We show two denoising trajectories of a preferred image (a) and a dispreferred image (b) generated from prompt "A cat jumps on a dog". (1) *Disagreement between generic preferences and aesthetic preference*. While (a) is preferred due to its better layout, its details are poorer than (b). Red boxes show the erroneous fusion of the cat's leg and the dog's body. (2) *Large sample differences between trajectories.* This makes it difficult to identify subtle aesthetic-related differences at each step.

Abstract

Generating visually appealing images is fundamental to modern text-to-image generation models. A potential solution to better aesthetics is direct preference optimization (DPO), which has been applied to diffusion models to improve general image quality including prompt alignment and aesthetics. Popular DPO methods propagate preference labels from clean image pairs to all the intermediate steps along the two generation trajectories. However, preference labels provided in existing datasets are blended with layout and aesthetic opinions, which would disagree with aesthetic preference. Even if aesthetic labels were provided (at substantial cost), it would be hard for the two-trajectory methods to capture nuanced visual differences at different steps. To improve aesthetics economically, this paper uses existing generic preference data and introduces step-by-step preference optimization (SPO) that discards the propagation strategy and allows fine-grained image details to be assessed. Specifically, at each denoising step, we 1) sample a pool of candidates by denoising from a shared noise latent, 2) use a step-aware preference model to find a suitable win-lose pair to supervise the diffusion model, and 3) randomly select one from the pool to initialize the next denoising step. This strategy ensures that diffusion models focus on the subtle, fine-grained visual differences instead of layout aspect. We find that aesthetics can be significantly enhanced by accumulating these improved minor differences. When fine-tuning Stable Diffusion v1.5 and SDXL, SPO yields significant improvements in aesthetics compared with existing DPO methods while not sacrificing image-text alignment compared with vanilla models. Moreover, SPO converges much faster than DPO methods due to the use of more correct preference labels provided by the step-aware preference model. Code and models are available at https://github.com/RockeyCoss/SPO.

1. Introduction

This paper aims to improve the ability of diffusion models in generating visually appealing images based on human preference data. That is, given a pool of image pairs and human preference within each pair, we fine-tune diffusion models so that they are more likely to generate images consistent with human aesthetic preference[1].

[1]We fine-tune models using crowd-sourced preference data. We do not discuss individual, cultural or political impact here.

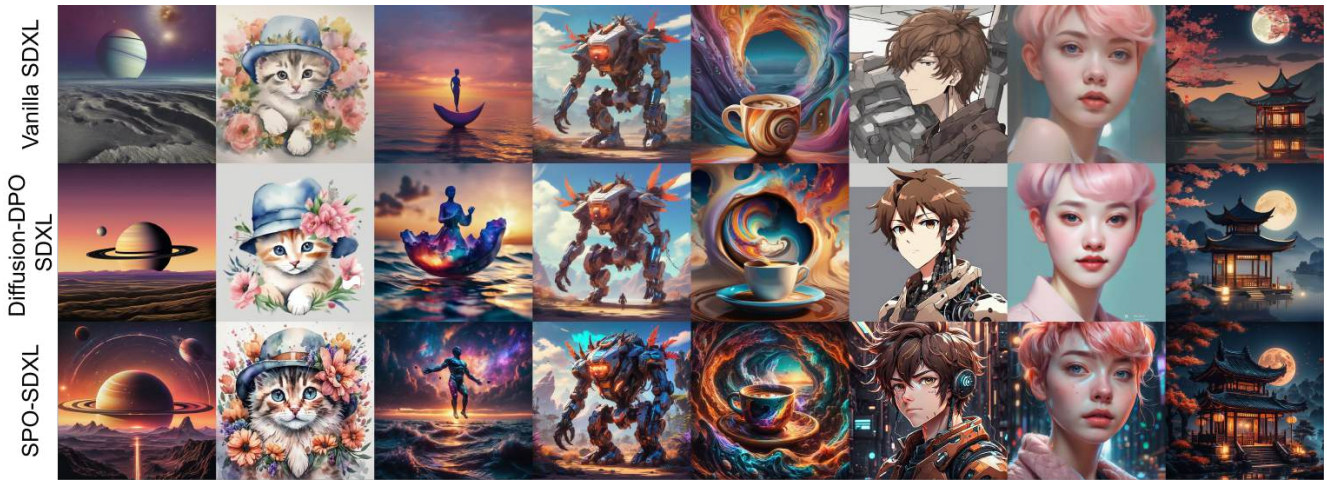

Figure 2. Qualitative comparison between Vanilla SDXL, Diffusion-DPO-SDXL and SPO-SDXL. SPO-SDXL exhibits very strong image aesthetics including more visual details and appealing styles. Prompts are provided in the supplementary material.

Direct preference optimization (DPO) is shown to be effective for aligning diffusion models with human preferences in general aspects such as image-text alignment [29]. Given a pair of images generated from the same prompt, DPO methods encourage predictions to align with the preferred image while discouraging resemblance to the dispreferred one. A few existing studies use human preferences at timestep 0 with clean images $\boldsymbol{x}_0^w$ and $\boldsymbol{x}_0^l$, where w and l are win and lose preference labels, respectively [21, 29, 32]. The labels are directly propagated to intermediate samples along the two trajectories, assuming all intermediate samples along the win trajectory are also preferred.

However, for two problems existing systems are not best effective for aesthetic alignment. First, generic preferences provided in public datasets like Pick-a-Pic [11] are not aesthetics specific, and they often disagree. In Fig. 1, (a) is generally preferred because of the correctly generated dog, cat, and their spatial arrangement, but in terms of aesthetics, (a) should be dispreferred. More examples are provided in the supplementary material. This introduces noisy supervision signals and compromises the model improvement towards better aesthetics. Technically, this problem can be fixed by manually annotating an aesthetics-only preference dataset, but it is very expensive, because of layout influence and complexity of aesthetics. Second, in two-trajectory methods, the images within each pair at each denoising step look very different, as shown in Fig. 1. Even if accurate aesthetic preference was provided at each step (at great cost), it would still be very non-trivial to learn, because the large layout discrepancy would dominate over aesthetic nuances.

To better align diffusion models with aesthetic preference, we introduce step-by-step preference optimization (SPO). SPO is new in that it completely discards the current preference propagation strategy, pushing for evaluating image details. Specifically, at each step beginning with a noisy image $\boldsymbol{x}_t$, we generate a pool of $\boldsymbol{x}_{t-1}$ samples. We evaluate their quality using a step-aware preference model (SPM) and assign win/lose labels to the pair showing the largest quality difference. We then randomly select an image from the pool to initialize timestep $t-1$. Because the win-lose pair 1) comes from the same image and 2) is generated in one or very few steps, the two samples would only exhibit small differences in details. SPM allows us to capture such detailed differences and guide the diffusion model to generate more visually pleasing images.

We use SPO to fine-tune Stable Diffusion v1.5 (SD-1.5) [23] and SDXL [18]. Although we use a training set with generic human perferences [11], we demonstrate significant improvement in aesthetics compared with those fine-tuned by popular DPO methods. Moreover, SPO converges much faster than Diffusion-DPO. This is because the step-by-step design makes it easier to focus on fine-grained visual details, and SPM produces more accurate preference labels. We summarize key points of this paper below.

- We aim to improve image aesthetics of diffusion models.
- We point out that generic human preferences are often different from pure aesthetic preference and that obtaining aesthetic-only preference data is very expensive.
- We design SPO, where we determine win-lose pairs at each step in an online manner. We make sure win-lose pairs come from the same noisy sample, so after one or very few steps of denoising their differences are small and in fine-grained details, and can be captured by a preference model trained with generic preference data.
- When fine-tuning SD-1.5 and SDXL, SPO is more effective in enhancing image aesthetics compared with DPO methods, and converges faster than Diffusion-DPO [29].

2. Related Work

Recently, inspired by post-training methods that improve LLMs, *e.g.*, reinforcement learning from human feedback (RLHF) [16, 35], various post-training methods are proposed to align pre-trained diffusion models with human preferences. For example, Chen et al. [2] leverage the PPO loss [26] to fine-tune the text encoder of diffusion models. AligningT2I [13] develops a reward model to evaluate the quality of generated images and fine-tunes the diffusion model using image-text pairs, weighted by assessment of the reward model. DPOK [7] and DDPO [1] use policy gradient to fine-tune diffusion models, aiming at maximizing reward signals. Furthermore, ReFL [31], DRaFT [3], and AlignProp [19] directly propagate the gradients through differentiable reward models to fine-tune the denoising steps. Recent methods are inspired by direct preference optimization (DPO) [21], which eliminates the need for explicit reward models when post-training LLMs. Diffusion-DPO [29] fine-tunes diffusion models on the Pick-a-Pic [11] dataset that contains image preference pairs. D3PO [32] generates pairs of images from the same prompt and uses a preference model to identify preferred and dispreferred images. DenseReward [33] improves the DPO scheme with a temporal discounting approach to emphasize initial denoising steps. These methods optimize the trajectory-level preference, where the accumulated image differences are too large to allow the network to focus on aesthetics subtleties. In comparison, SPO by its step-by-step mechanism can focus on nuanced visual differences in just a single or few steps.

3. DPO Revisit and Diagnosis

General formulation. Given a generation model $\pi_\theta(\cdot)$ and a condition c, the probability of generating output o is $\pi_\theta(o \mid c)$. We use $\pi_\theta(\cdot)$ to generate a set of output pairs $\mathcal{S}$, where each pair comes from the same condition c. Human or preference model is employed to label the preference order of the output pairs as (o^w, o^l, c), where o^w is the preferred output and o^l is dispreferred. According to Rafailov et al. [21], the DPO loss used to fine-tune π_θ is defined as:

$$\mathcal{L}_{\mathcal{DPO}} = -\mathbb{E}_{(o^w,o^l,c)\sim\mathcal{S}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(o^w|c)}{\pi_{\text{ref}}(o^w|c)} - \beta\log\frac{\pi_\theta(o^l|c)}{\pi_{\text{ref}}(o^l|c)}\right)\right]. \tag{1}$$

$\pi_{\text{ref}}(\cdot)$ and $\sigma(\cdot)$ refer to the reference model and the sigmoid function, respectively. β is the strength of regularization.

DPO for diffusion model post-training. We denote the text-to-image diffusion model with parameters θ as p_θ and text prompt as c. The denoising process generates intermediate states $\{\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, ..., \boldsymbol{x}_1, \boldsymbol{x}_0\}$. Existing works including Diffusion-DPO [29] and D3PO [32] measure preference based on the final image $\boldsymbol{x}_0$ and assign the preference for $\boldsymbol{x}_0$ directly to the entire generation trajectory, or all the intermediate states. Let $\mathcal{T}_w$ and $\mathcal{T}_l$ denote the denoising trajectories which generate the preferred and dispreferred images, respectively. Using the Markov property of diffusion models and Jensen's inequality, they reformulate the general DPO loss in Eq. 1 into the following step-wise form:

$$\mathcal{L}_{\mathcal{DPO}-\mathcal{D}} = -\mathbb{E}_{\substack{(\boldsymbol{x}^w_{t-1},\boldsymbol{x}^w_t)\sim\mathcal{T}_w \\ (\boldsymbol{x}^l_{t-1},\boldsymbol{x}^l_t)\sim\mathcal{T}_l}}\left[\log\sigma\left(\beta\log\frac{p_\theta(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_t,c)}{p_{\text{ref}}(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_t,c)} - \beta\log\frac{p_\theta(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_t,c)}{p_{\text{ref}}(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_t,c)}\right)\right], \tag{2}$$

where $\mathcal{L}_{\mathcal{DPO}-\mathcal{D}}$ encourages denoising steps to progress towards the preferred image and away from dispreferred one.

Diagnosis on aesthetic alignment. Aligning diffusion models with human aesthetic preference is very challenging. From architecture, win-lose pairs in the two-trajectory methods usually differ significantly (primarily in layout), rendering it hard to focus on image detail comparisons. From data, existing aesthetic scoring datasets [15, 25] do not provide paired image data coming from the same prompt. Collecting a dedicated aesthetic preference dataset would be more costly compared with tasks like classification or prompt alignment preference. So a cost-efficient choice is existing generic preference datasets where annotators were asked to score images based on holistic opinions like prompt alignment and aesthetics. However as pointed out in Section 1, generic preferences may be contradictory to aesthetic preference. If we directly propagate them to all diffusion steps like Diffusion-DPO and D3PO do (see Fig. 3), noisy preferences will compromise fine-tuning.

4. Proposed Approach

4.1. Framework Overview

To align diffusion models with aesthetic preference while still using generic preferences, we propose step-by-step preference optimization (SPO), an online reinforcement learning method. Fig. 3 (c) depicts its workflow. Given an intermediate $\boldsymbol{x}_t$, at timestep t, we sample a pool of denoised samples $\{\boldsymbol{x}^1_{t-1}, \boldsymbol{x}^2_{t-1}, \cdots, \boldsymbol{x}^k_{t-1}\}$. We apply a step-aware preference model (Section 4.2) to compare the quality of these candidate samples, and select the highest-quality sample and the lowest-quality sample as the win-sample and the lose-sample, respectively. SPO ensures the win-lose pair comes from the same $\boldsymbol{x}_t$ and thus have small differences in image details to reflect aesthetics. We then randomly select a sample from the candidate pool (Section 4.3), which is used to initialize the next iteration. SPO is optimized by a revised DPO loss function (Section 4.4) and can be extended to the SDXL model (Section 4.5).

4.2. Step-Aware Preference Model (SPM)

Overall use. SPM predicts preference order among the k sampled denoised samples $\{\boldsymbol{x}^1_{t-1}, \boldsymbol{x}^2_{t-1}, \cdots, \boldsymbol{x}^k_{t-1}\}$. Thus, SPM takes timestep $t-1$, intermediate sample $\boldsymbol{x}_{t-1}$, and

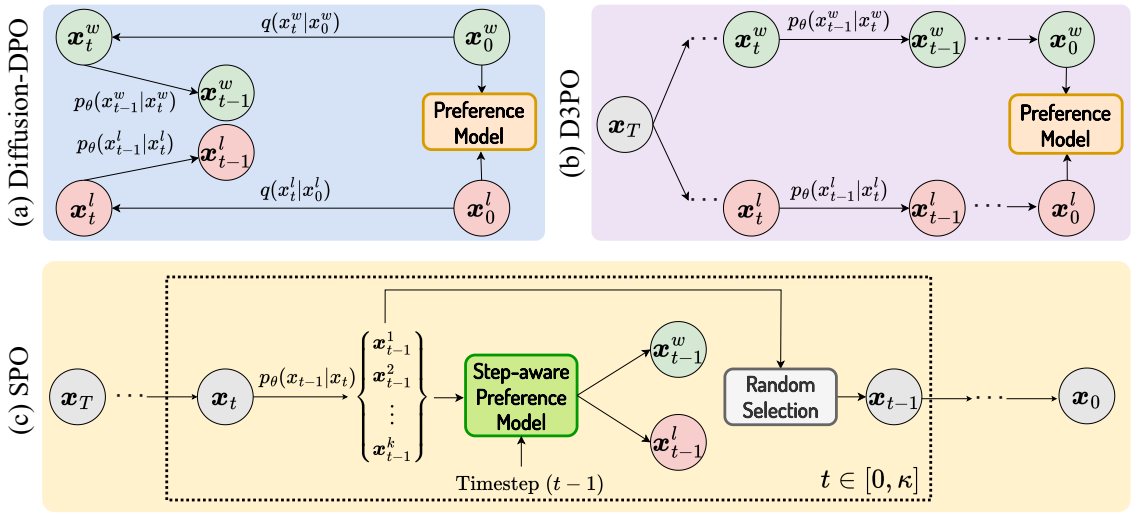

Figure 3. Comparing frameworks of SPO, Diffusion-DPO, and D3PO approaches. SPO does not adopt direct preference propagation as other DPO methods do. In SPO, a pool of samples are generated at each step, from which a proper win/lose pair is selected and used to fine-tune the diffusion model. Then, a single sample is randomly selected to initialize the next iteration.

prompt c as input, and outputs a quality score. SPM is different from existing preference models. The latter uses clean images $\boldsymbol{x}_0$ and prompt c as input without time condition, because they are designed for assessing clean images. Following [11], we construct SPM based on CLIP [20].

For **SPM training**, we initialize the model with PickScore [11] and fine-tune it following [4], making the model useful for noisy images. Specifically, we add the same amount of noise to a pair of clean images, assuming that their preference order can be kept. During training, for each pair of images and their human-labeled preference, we randomly sample a timestep t and add the same noise to both images to obtain $\boldsymbol{x}_t^w$ and $\boldsymbol{x}_t^l$. Then we feed the noisy intermediate pair $\{\boldsymbol{x}_t^w, \boldsymbol{x}_t^l\}$ and timestep t to SPM and train the model to correctly predict the preference, using loss function $\mathcal{L}_{\text{pref}} = (\log 1 - \log \hat{p}_w)$, where $\hat{p}_w$ is the probability of the win image being the preferred one, following [11]. $\hat{p}_w$ is computed using the following equation:

$$\hat{p}_w = \frac{\exp\left(\tau \cdot f_{\text{CLIP-V}}(\boldsymbol{x}_t^w, t) \cdot f_{\text{CLIP-T}}(c)\right)}{\exp\left(\tau \cdot f_{\text{CLIP-V}}(\boldsymbol{x}_t^w, t) \cdot f_{\text{CLIP-T}}(c)\right) + \exp\left(\tau \cdot f_{\text{CLIP-V}}(\boldsymbol{x}_t^l, t) \cdot f_{\text{CLIP-T}}(c)\right)}, \quad (3)$$

where c represents the text prompt, $\tau \in \mathcal{R}$ is a temperature. $f_{\text{CLIP-V}}(\cdot)$ and $f_{\text{CLIP-T}}(\cdot)$ are the vision and text encoders of CLIP, respectively. To allow for timestep-conditional preference prediction, we modify the CLIP vision encoder using time-conditional adaptive layernorm [17]. To alleviate the domain gap between the noised image at the t-th timestep and images used to train the PickScore model, we estimate $\hat{\boldsymbol{x}}_0$ from the noisy sample directly, following DDIM [28].

4.3. Random Selection of $\boldsymbol{x}_t$ to Start Next Step

In SPO, we randomly sample $\boldsymbol{x}_t$ from the candidate pool $\{\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \cdots, \boldsymbol{x}_t^k\}$, which is then used to start the next denoising step. This ensures every win-lose pair comes from the same latent. Since $\boldsymbol{x}_t$ is very noisy when t is large, we use the standard diffusion sampling process when t is greater than the threshold κ. Only when $t \leq \kappa$ do we sample candidate pools and apply random selection. This random selection is shown in Fig. 3 (c): selection of $\boldsymbol{x}_{t-1}$ is depicted and is consistent with the above discussion.

4.4. Objective Function of SPO

At the t-th denoising timestep, we sample a pool of candidates $\{\boldsymbol{x}_{t-1}^1, \boldsymbol{x}_{t-1}^2, \cdots, \boldsymbol{x}_{t-1}^k\}$ and use the step-aware preference model to construct a preference pair $(\boldsymbol{x}_{t-1}^w, \boldsymbol{x}_{t-1}^l)$, where $\boldsymbol{x}_{t-1}^w$ and $\boldsymbol{x}_{t-1}^l$ are the most and least preferred in the pool. Using various prompts, we can obtain many preference pairs at t-th timestep. Using the general form of DPO loss in Eq. 1, the SPO objective at the t-th timestep is:

$$\mathcal{L}_t(\theta) = -\mathbb{E}_{c\sim p(c), \boldsymbol{x}_{t-1}^w, \boldsymbol{x}_{t-1}^l \sim p_\theta(\boldsymbol{x}_{t-1}|c,t,\boldsymbol{x}_t)} \left[\log \sigma \left(\beta \log \frac{p_\theta\left(\boldsymbol{x}_{t-1}^w|c,t,\boldsymbol{x}_t\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{t-1}^w|c,t,\boldsymbol{x}_t\right)} - \beta \log \frac{p_\theta\left(\boldsymbol{x}_{t-1}^l|c,t,\boldsymbol{x}_t\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{t-1}^l|c,t,\boldsymbol{x}_t\right)}\right)\right], \quad (4)$$

where c is the prompt and $p(c)$ is the distribution of prompts. By combining the SPO objectives across all T

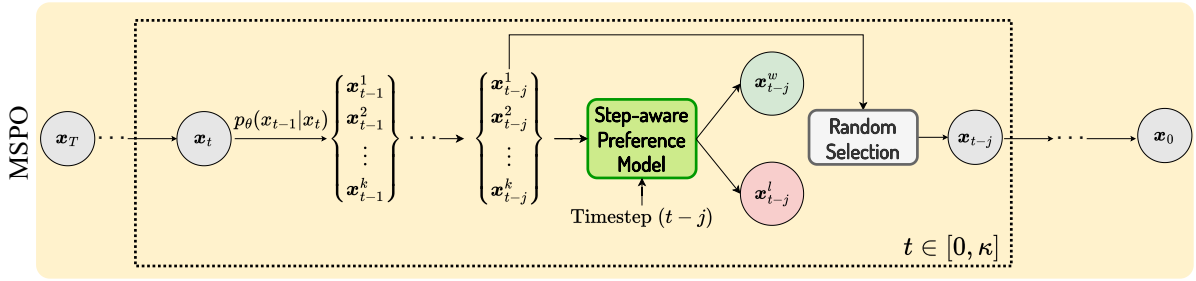

Figure 4. Framework of multi-step-by-step preference optimization (MSPO). From $\boldsymbol{x}_t$, we first sample k latents $\{\boldsymbol{x}_{t-1}^1, \boldsymbol{x}_{t-1}^2, \cdots, \boldsymbol{x}_{t-1}^k\}$ as SPO does. For each latent, we perform multiple (*i.e.*, j) denoising steps to obtain $\{\boldsymbol{x}_{t-j}^1, \boldsymbol{x}_{t-j}^2, \cdots, \boldsymbol{x}_{t-j}^k\}$, from which a win-lose pair is selected by SPM. Then we apply random selection and iterations in the same manner as SPO.

timesteps, we obtain the final SPO objective:

$$\mathcal{L}(\theta) = -\mathbb{E}_{t\sim\mathcal{U}[1,T-\kappa],c\sim p(c),x_T\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\boldsymbol{x}_{t-1}^w,\boldsymbol{x}_{t-1}^l\sim p_\theta(\boldsymbol{x}_{t-1}|c,t,\boldsymbol{x}_t)} \left[\log\sigma\left(\beta\log\frac{p_\theta\left(\boldsymbol{x}_{t-1}^w|c,t,\boldsymbol{x}_t\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{t-1}^w|c,t,\boldsymbol{x}_t\right)} - \beta\log\frac{p_\theta\left(\boldsymbol{x}_{t-1}^l|c,t,\boldsymbol{x}_t\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{t-1}^l|c,t,\boldsymbol{x}_t\right)}\right)\right], \tag{5}$$

where $\mathcal{U}$ and $\mathcal{N}$ are the uniform distribution and Gaussian distribution, respectively.

4.5. Extension to Multi-step Preference Optimization for SDXL

For stronger models like SDXL, we observe that the difference between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ is *too small*, so is the difference between the selected candidates $\boldsymbol{x}_{t-1}^w$ and $\boldsymbol{x}_{t-1}^l$. While a relatively small difference allows SPM to focus on image details, that difference is too small would create ambiguities and confuse fine-tuning. To address this issue, we expand step-by-step preference optimization to multi-step preference optimization (MSPO), which uses multiple denoising steps to increase the diversity of the candidate set (see Fig. 4). This simple extension allows us to select more different samples and more clear preference signals.

4.6. Discussions and Insights

DPO for diffusion model vs. language model. Diffusion models involve many intermediate steps, each producing a latent feature/image. In contrast, language models typically predict the final result at each position with a single step without iterative refinement. This distinction necessitates specific DPO design in diffusion models.

SPO is an implicit aesthetic optimizer and distillator. We do not use a dedicated aesthetic preference dataset. Aesthetic optimization is done implicitly through SPM, which is trained to perceive image quality from generic opinions. When the two images to be compared are relatively similar, the output from SPM thus mainly describes image details instead of significant layout differences. Sample win-lose pairs selected by the SPM are shown in Fig. 5. In this way, we are able to distill aesthetic details from generic data.

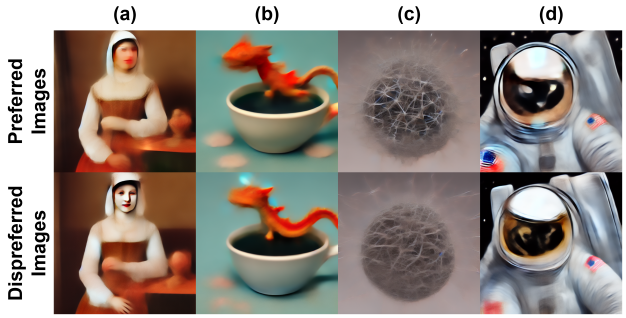

Figure 5. Win-lose pairs identified by the SPM during fine-tuning. Top: Preferred images. Bottom: Dispreferred images. In SPO, these pairs look similar so image details can be focused on. From these pairs, SPM favors images with fewer artifacts in (a) and (b) and more refined details in (c) and (d). These images appear blurry because for visualization purpose we directly predict the clean image from the intermediate noisy latents.

How often does generic preferences and aesthetic preference disagree? It is hard to quantify because it is non-trivial to annotate a validation set. Fig. 1 and supplementary material depict some disagreement scenarios based on careful manual inspection. In fact, because of the nuanced nature of aesthetics, an image may be superior in certain aesthetic aspects while its paired image is better in other aspects. So there are probably more disagreement scenarios than we can think of now.

Limitations. First, SPO is not applicable to recent flow matching models, such as SD3 [6] and Flux [12] because SPO requires the intermediate steps to be stochastic which flow matching models do not satisfy. Second, SPO is specifically designed to improve aesthetics and offers limited help for improving image-text alignment. Third, we also limit ourselves to learning from crowd-sourced data without tapping into the subjective, political, and historical aspects of aesthetics. It is interesting to study these problems in future.

5. Experiments

5.1. Experimental Setup

Datasets. For SPO, we train the step-aware preference model (SPM) on the Pick-a-Pic V1 dataset. This dataset has over 580k labeled image preference pairs, each generated by the same text prompt with various diffusion models [18, 22, 24]. Human annotators were asked to rate the *general* quality of each image, forming win-lose pairs. When fine-tuning diffusion models with SPO, DDPO or D3PO, we use a subset of 4k prompts (without images) randomly selected from Pick-a-Pic V1, where win-lose pairs are generated online. Note that these three methods do not use images or preference labels but use text prompts only. For DDPO and D3PO, PickScore trained on Pick-a-Pic V1 is used as their reward model to provide guidance. For Diffusion-DPO and MAPO [27], we use their online-available checkpoints for evaluation, which were trained on Pick-a-Pic V2 dataset with over 800k image preference pairs. So overall, SPO and the competitive models are trained on similar datasets and can be fairly compared.

If not specified, we report quantitative results based on the 500 validation prompts, *i.e.*, `validation_unique` split of Pick-a-Pic [11], which is adopted in [29]. We also use GenEval [8] to evaluate image-text alignment, including rendering of single and two objects, counting, colors, position and attribute binding. There are 553 test prompts.

Evaluation Protocol. This paper evaluates image quality with four automatic metrics. We use PickScore [11], HPSV2 [30] and ImageReward [31] for prompt-aware human preference estimation. These models are trained on human preference datasets and learn to replicate human decisions about which images are more favorable. We use Aesthetic score [25] to evaluate visual appeal. This score is prompt agnostic and employs a linear estimator on top of the vision encoder of CLIP [20] to predict the aesthetic quality of images. Note that PickScore, HPSV2, and ImageReward all assess aesthetics to some extent. Apart from these automatic metrics, we also use human studies. We invite 10 people to assess 300 pairs of images generated by two methods of interest. Their preferences are summarized into winning percentage from 0 to 100%. The text prompts are randomly selected from PartiPrompts (100 prompts) [34] and HPSV2 benchmark (200 prompts) [30].

Implementation details. We apply DDIM [28] with $\eta = 1.0$ and 20 timesteps as the sampler and use classifier free guidance [9] with scale 5.0 for sampling during online training. We use LoRA [10] for both SD-1.5 and SDXL, fine-tuning the models for 10 epochs. The LoRA rank is 4 and 64 for SD-1.5 and SDXL, respectively. We set the strength of regularization $\beta = 10$. For SD-1.5, we set the batch size as 40 and learning rate as $6e^{-5}$. For SDXL, we set the batch size as 8, gradient accumulation as 2, and learning rate as $1e^{-5}$. Since very noisy images are difficult to compare, we do not use SPM to very early stages. That is, we only compute the preference of $\boldsymbol{x}_t$ when $t \leq \kappa$ and consider all $\boldsymbol{x}_t$ with $t > \kappa$ as tied. We empirically set κ as 750 and will evaluate κ in Section 5.4. When fine-tuning SDXL with the MSPO (Section 4.5), we set the number of inner steps to 4. We do not apply SDXL refiner [18] to ensure fair comparison. For SPM training, we adopt learning rates of 3e-6 and 1e-6 for SD-1.5 and SDXL, respectively. We use DDIM scheduler with classifier free guidance scale of 5 and 20 steps to perform inference on validation prompts.

Table 1. Method comparison on SDXL. SPO overall yields the **best** fine-tuning performance especially in aesthetics. Note PickScore, HPSV2, and ImageReward partially assess aesthetics.

Method	PickScore	HPSV2	ImageReward	Aesthetic
SDXL	21.95	26.95	0.5380	5.950
Diff.-DPO	22.64	29.31	0.9436	6.015
MAPO	22.11	28.22	0.7165	6.096
SPO	**23.06**	**31.80**	**1.0803**	**6.364**

Table 2. Comparing different methods on SD-1.5.

Method	PickScore	HPSV2	ImageReward	Aesthetic
SD-1.5	20.53	23.79	-0.1628	5.365
DDPO	21.06	24.91	0.0817	5.591
D3PO	20.76	23.97	-0.1235	5.527
Diff.-DPO	20.98	25.05	0.1115	5.505
SPO	**21.43**	**26.45**	**0.1712**	**5.887**

Table 3. Method comparison on GenEval based on SDXL. †: results are reproduced with a classifier-free guidance scale of 5.0 and 50 inference steps. GenEval evaluates image-text alignment.

Method	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall
SDXL†	97.50	71.21	39.06	84.04	11.00	17.75	53.43
Diff.-DPO	**99.06**	**80.81**	**46.56**	**88.30**	**13.25**	**29.50**	**59.58**
SPO	97.81	73.48	41.25	85.64	13.00	20.00	55.20

5.2. Main Evaluation on Aesthetic Alignment

Automatic metrics. Using the four automatic scores (Section 5.1), we compare SPO with Diffusion-DPO, D3PO, etc., in Table 1 and Table 2 based on SDXL and SD-1.5, respectively. Note that ImageReward, PickScore, and HPSV2 assess overall image quality including aesthetics.

We have two observations. First, compared with vanilla SDXL and SD-1.5, the DPO methods yield consistent improvement in these metrics, demonstrating their effectiveness. Second, we observe that **SPO yields the best scores across the four metrics for both SDXL and SD-1.5**. For example, for SDXL, we achieve 23.06, 31.80, 1.0803, and 6.364 in PickScore, HPSV2, ImageReward, and Aesthetics, respectively. The improvement over Diffusion-DPO

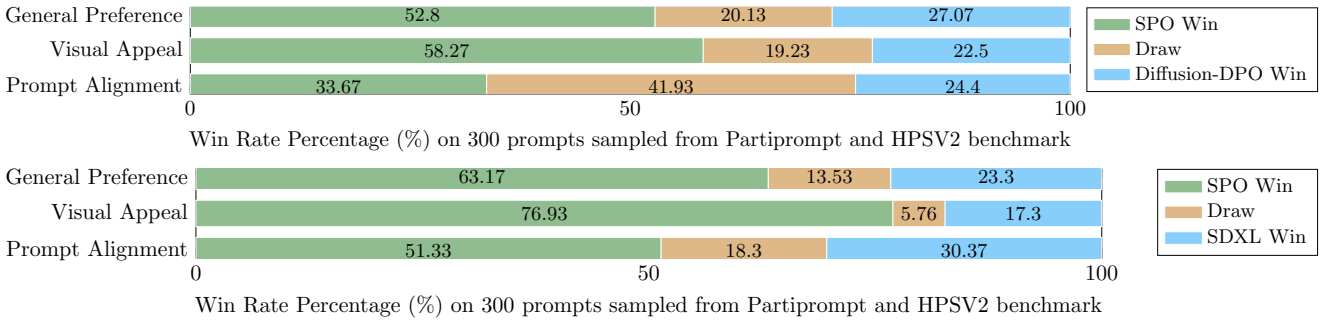

Figure 6. User study results comparing SPO with Diffusion-DPO and Vanilla SDXL. We sampled 100 and 200 prompts for evaluation from Partiprompts [34] and HPSV2 benchmark [30], respectively. SPO yields clear improvement in visual appeal.

Table 4. Comparing random sampling with other sampling strategies.

Initial.	P-S	HPSV2	I-R	AE
$\boldsymbol{x}_{t-1}^{w}$	17.87	11.31	-2.2692	3.963
$\boldsymbol{x}_{t-1}^{l}$	19.36	18.63	-1.3743	5.338
random	**21.43**	**26.45**	**0.1712**	**5.887**

Table 5. Comparing SPM with variants: no time condition or the PickScore model [11].

Prefer. model	P-S	HPSV2	I-R	AE
SPM	**21.43**	**26.45**	**0.1712**	**5.887**
w/o step con.	21.19	25.84	0.1365	5.678
PickScore	20.28	23.09	-0.2982	5.410

Table 6. Impact of number of sampled images k at each step. We use $k = 4$.

#samples k	P-S	HPSV2	I-R	AE
2	21.37	26.56	0.3235	5.714
4	**21.43**	26.45	0.1712	**5.887**
8	21.19	**27.62**	**0.4199**	5.605

Table 7. Impact of #inner steps j when fine-tuning SDXL with MSPO. We set $j = 4$.

#inner steps j	P-S	HPSV2	I-R	AE
1	22.85	31.37	1.0071	6.359
2	22.84	31.17	1.0118	6.268
3	22.94	31.55	**1.0847**	6.380
4	**23.06**	**31.80**	1.0803	6.364
5	23.03	31.23	0.9656	**6.423**
6	22.95	30.57	0.9770	6.390

Table 8. Impact of timestep range.

Timestep Range	P-S	HPSV2	I-R	AE
[0-250]	20.61	23.34	-0.1823	5.413
[0-500]	20.69	25.67	0.0810	5.399
[0-750]	**21.43**	**26.45**	0.1712	**5.887**
[0-1000]	19.77	22.72	-0.4529	5.111
[250-750]	21.19	26.23	**0.2658**	5.581
[500-750]	20.43	24.91	-0.1553	5.582
[250-500]	20.60	25.60	0.1037	5.336

Table 9. Comparing win-lose pair choices. Choosing images of the highest and lowest quality is generally better than random selection. Note both strategies allow win-lose preference to align with aesthetic preference.

win-lose sample	P-S	HPSV2	I-R	AE
best & worst	**21.43**	26.45	**0.1712**	**5.887**
random	21.21	**26.51**	0.1656	5.796

is +0.42, +2.49, +0.1367, and +0.349 on the four metrics, respectively. This again demonstrates the effectiveness of SPO in aesthetic improvement.

User studies. We compare SPO with Diffusion-DPO and Vanilla SDXL through user studies. We ask annotators to compare three aspects: general preference, visual appeal, and prompt alignment and summarize the winning percentage in Fig. 6. Results indicate that SPO has consistently more winning votes from users in general preference and visual appeal, and it is apparent that its visual appeal has a greater winning margin. For example, in general preference alignment, SPO wins 52.8% of all cases, while in visual appeal, it wins 58.27% of the cases. Similar observation can be made in SPO vs. vanilla SDXL.

Qualitative comparison. We show sample images in Fig. 2. Images generated by SPO are more visually appealing than those generated by SDXL and Diffusion-DPO.

5.3. Evaluation on Image-Text Alignment

SPO slightly improves vanilla models. From Table 1 and Table 2, PickScore metric indicates some improvement over SDXL and SD-1.5. User studies (Fig. 6), on prompt alignment has similar results. On GenEval, Table 3 shows SPO improves prompt alignment score by 1.77% over SDXL.

SPO has mixed results compared with Diffusion-DPO. Table 1 and Table 2 demonstrate improvement over Diffusion-DPO, and yet our user studies support a performance tie or SPO's slight win in prompt alignment. Further, GenEval results indicate that SPO is not as good as Diffusion-DPO in improving image-text alignment, which we tend to agree with, because the SPO design does not fully capture layout changes. But an interesting insight is that image-text alignment and aesthetics are probably often blended. So while SPO is designed for aesthetic alignment, it still yields some prompt alignment improvement, but the improvement is not as much as Diffusion-DPO.

5.4. Further Analysis

If not specified, we use SD-1.5 and Pick-a-Pic *val.* set.

Effectiveness of step-aware preference model (SPM). An important design of SPM is the addition of timestep conditioning. To verify its usefulness, we remove the timestep conditioning in SPM (SPM w/o step). A second variant is to simply use the PickScore model [11] which has no time condition and is trained on clean images only. Results in Table 5 show that both variants lead to performance drop,

validating the SPM design.

Effectiveness of random selection for initializing next iteration. At the end of each SPO iteration, we need to initialize the next iteration, where random selection is an option. Here we compare random selection with using the win sample $\boldsymbol{x}_{t-1}^w$ or the lose sample $\boldsymbol{x}_{t-1}^l$ for initialization. Results are presented in Table 4. We clearly find that random selection is better. If we only select $\boldsymbol{x}_{t-1}^w$ or $\boldsymbol{x}_{t-1}^l$, training becomes biased towards the intermediate samples that are more preferred or more dispreferred, respectively. This prevents the network from learning from more diverse trajectories, deteriorating generalization ability.

Impact of number of candidates sampled at each denoising step. To find useful win-lose pairs, we obtain a set of candidates $\{\boldsymbol{x}_{t-1}^1, \boldsymbol{x}_{t-1}^2, \cdots, \boldsymbol{x}_{t-1}^k\}$ at each step, drawn from the conditional distribution $p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c}, t\right)$. Table 6 presents results by varying k. We have two observations.

First, when increasing k, the discrepancy within sampled pairs becomes larger (but still small enough to only reflect image detail differences). The larger contrast between the preferred and dispreferred samples helps the model learn human preferences. Second, when k is too large, quality of the most dispreferred sample would be lower than average of samples generated by the base model. As a result, the "push away" effect of the dispreferred image is weakened, causing performance degradation. We choose $k = 4$.

Impact of the number of inner steps j in MSPO. A larger j allows the generated images to have higher diversity, which would be useful for strong diffusion models. In Table 7, we study how j impacts SPO. We observe that fine-tuning performance increases with j in the beginning and then becomes saturated. When $j = 1$, MSPO reduces to SPO. When j goes to infinite, MSPO would essentially reduce to Diffusion-DPO because there will only be one step. From the table, we choose $j = 4$.

Impact of timestep range. SPO is only applied to timestep range [0-κ]. Table 8 summarizes the results of applying SPO to various timestep ranges. We have the following observations.

First, discarding very large timesteps, *i.e.*, [750-1000], yields better performance as these timesteps barely generate image details and are very noisy. *Second*, if we remove [0-250] and only use [250-750], there is a considerate performance drop, indicating that [0-250] is useful. Similarly, [0-500] is also useful. *Third*, if we compare [500-750], [250-500] and [0-250], we find that applying SPO to [250-500] achieves slightly better overall performance. We speculate [250-500] is a critical timestep range for SPO. Compared to larger timesteps, timesteps in [250-500] focus more on detail refinement. Moreover, compared to smaller timesteps, the denoising steps in [250-500] sample $\boldsymbol{x}_{t-1}$ with a sufficiently large variance to construct win-loss pairs for training. Based on these findings, we set $\kappa = 750$ and apply SPO to timestep range [0-750].

Comparing ways of choosing win-lose pairs. In Table Table 9, we compare two ways. The proposed method uses a sample with highest quality and another sample with lowest quality. The other way is to randomly select two samples and use SPM to decide their win-lose preference. Results show that the proposed way is generally better. That said, both options allow for proper assessment of aesthetic preference because they ensure a sample pair comes from the same sample and has relatively small differences.

Computational cost. We use $4\times$ A100 GPUs, which take 12 and 29.5 hours to fine-tune SD-1.5 and SDXL, respectively. We also spend 8 and 29 hours training SPM for SD-1.5 and SDXL, respectively. In comparison, the GPU hours used for fine-tuning SD-1.5 and SDXL using Diffusion-DPO are 384 and $4,800$, respectively. As a result, the total training GPU hours of SPO-SD-1.5 and SPO-SDXL is about 20.8% and 4.9% of DPO-SD-1.5 and DPO-SDXL, respectively. This significant efficiency gain is probably because of the SPO design, where the image details are properly highlighted for the network to learn.

6. Conclusion

This paper studies how to align diffusion models with human aesthetic preference. This problem is challenging for two reasons. First, existing two-trajectory methods exhibit large image discrepancies, preventing models from focusing on aesthetic nuances. Second, it is very non-trivial to collect aesthetic-only preference data, while existing datasets record generic preferences that may conflict with image aesthetic preference. To improve aesthetic alignment, we propose an aesthetic alignment solution that distills aesthetics from generic preference data. Specifically, the proposed step-by-step preference optimization method allows a pair of samples to originate from the same image, so their differences are relatively small after one or very few steps, which would reflect their image details or aesthetics. Our preference model captures these differences, enabling the model to improve towards generating better image details. In experiments, we demonstrate that SPO better aligns SD-1.5 and SDXL with human aesthetic preference compared with other DPO methods and is efficient to train.

Acknowledgments

We gratefully acknowledge the support of the ARC Future Fellowship (FT240100820), awarded to Liang Zheng.

References

[1] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3

[2] Chaofeng Chen, Annan Wang, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Enhancing diffusion models with text-encoder reinforcement learning. *arXiv preprint arXiv:2311.15657*, 2023. 3

[3] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 3

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[6] 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

[7] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[11] 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, 2024. 2, 3, 4, 6, 7

[12] Black Forest Labs. Flux.1 model family. 2024. 5

[13] 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. 3

[14] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. *arXiv preprint arXiv:2403.09622*, 2024. 2

[15] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 3

[16] 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. 3

[17] 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, 1

[18] 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. 2, 6

[19] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 3

[20] 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. 4, 6

[21] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[22] Patrick Esser Robin Rombach. Model card for stable diffusion v2.1, 2023. Hugging Face Models Repository. 6

[23] 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

[24] RunwayML. Model card for stable diffusion v1.5, 2023. Hugging Face Models Repository. 6

[25] Christoph Schuhmann. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2023 - 11- 10. 3, 6

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[27] Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*, 2024. 6

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4, 6

[29] 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. *arXiv preprint arXiv:2311.12908*, 2023. 2, 3, 6

[30] 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. 6, 7

[31] 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, 2024. 3, 6

[32] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023. 2, 3

[33] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024. 3

[34] 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. 6, 7

[35] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 4, 2024. 3