Diffusion Model Alignment Using Direct Preference Optimization

Bram Wallace[1*] Meihua Dang[2] Rafael Rafailov[2] Linqi Zhou[2] Aaron Lou[2]
Senthil Purushwalkam[1] Stefano Ermon[2] Caiming Xiong[1] Shafiq Joty[1]
Nikhil Naik[1]
[1]Salesforce AI, [2]Stanford University

bram@openai.com {spurushwalkam,cxiong,sjoty}@salesforce.com naik@alum.mit.edu
{mhdang,rafailov,lzhou907,aaronlou}@stanford.edu ermon@cs.stanford.edu

Abstract

Large language models (LLMs) are fine-tuned using human comparison data with Reinforcement Learning from Human Feedback (RLHF) methods to make them better aligned with users' preferences. In contrast to LLMs, human preference learning has not been widely explored in text-to-image diffusion models; the best existing approach is to fine-tune a pretrained model using carefully curated high quality images and captions to improve visual appeal and text alignment. We propose Diffusion-DPO, a method to align diffusion models to human preferences by directly optimizing on human comparison data. Diffusion-DPO is adapted from the recently developed Direct Preference Optimization (DPO) [36], a simpler alternative to RLHF which directly optimizes a policy that best satisfies human preferences under a classification objective. We re-formulate DPO to account for a diffusion model notion of likelihood, utilizing the evidence lower bound to derive a differentiable objective. Using the Pick-a-Pic dataset of 851K crowdsourced pairwise preferences, we fine-tune the base model of the state-of-the-art Stable Diffusion XL (SDXL)-1.0 model with Diffusion-DPO. Our fine-tuned base model significantly outperforms both base SDXL-1.0 and the larger SDXL-1.0 model consisting of an additional refinement model in human evaluation, improving visual appeal and prompt alignment. We also develop a variant that uses AI feedback and has comparable performance to training on human preferences, opening the door for scaling of diffusion model alignment methods.

1. Introduction

Text-to-image diffusion models have been the state-of-the-art in image generation for the past few years. They are typically trained in a single stage, using web-scale datasets of text-image pairs by applying the diffusion objective. This stands in contrast to the state-of-the-art training methodology for Large Language Models (LLMs). The best performing LLMs [31, 51] are trained in two stages. In the first ("pretraining") stage, they are trained on large web-scale data. In the second ("alignment") stage, they are fine-tuned to make them better aligned with human preferences. Alignment is typically performed using supervised fine-tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) using preference data. LLMs trained with this two-stage process have set the state-of-the-art in language generation tasks and have been deployed in commercial applications such as ChatGPT and Bard.

Despite the success of the LLM alignment process, most text-to-image diffusion training pipelines do not incorporate learning from human preferences. [11, 38, 39] perform two-stage training, following large-scale pretraining with fine-tuning on a high-quality text-image pair dataset. This approach is much more rudimentary than the final-stage alignment methods of LLMs. [7, 13] develop more advanced alignment methods, but have not demonstrated the ability to stably generalize to a fully open-vocabulary setting across an array of feedback. Other methods use the pixel-level gradients from reward models on generations to tune diffusion models, but suffer from mode collapse and can only incorporate a relatively narrow set of feedback types [9, 34].

We address this gap in diffusion model alignment for the first time, developing a method to directly optimize diffusion models on human preference data. We generalize Direct Preference Optimization (DPO) [36], where a generative model is trained on paired human preference data to implicitly estimate a reward model. We define a notion of data likelihood under a diffusion model in a novel formulation and derive a simple but effective loss resulting in stable and efficient preference training, dubbed Diffusion-DPO. We connect this formulation to a multi-step RL approach in the same setting as existing work [7, 13].

We demonstrate the efficacy of Diffusion-DPO by fine-tuning state-of-the-art text-to-image diffusion models, such

[*]Work done while at Salesforce AI

Figure 1. We develop Diffusion-DPO, a method based on Direct Preference Optimization (DPO) [36] for aligning diffusion models to human preferences by directly optimizing the model on user feedback data. After fine-tuning on the state-of-the-art SDXL-1.0 model, our method produces images with exceptionally high visual appeal and text alignment, samples above.

as Stable Diffusion XL (SDXL)-1.0 [33]. Human evaluators prefer DPO-tuned SDXL images over the SDXL-(base + refinement) model 69% of the time on the PartiPrompts dataset, which represents the state-of-the-art in text-to-image models as measured by human preference. Example generations shown in Fig. 1. Finally, we show that learning from AI feedback (instead of human preferences) using the Diffusion-DPO objective is also effective, a setting where previous works have been unsuccessful [9]. In sum, we introduce a novel paradigm of learning from human preferences for diffusion models and present the resulting state-of-the-art model.

2. Related Work

Aligning Large Language Models LLMs are typically aligned to human preferences using supervised fine-tuning on demonstration data, followed by RLHF. RLHF consists of training a reward function from comparison data on model outputs to represent human preferences and then using reinforcement learning to align the policy model. Prior work [5, 29, 32, 50] has used policy-gradient methods [30, 41] to this end. These methods are successful, but expensive and require extensive hyperparameter tuning [37, 63], and can be prone to reward hacking [12, 14, 44]. Alternative approaches sample base model answers and select based on predicted rewards [4, 6, 16] to use for supervised training [3, 18, 53]. Methods that fine-tune the policy model directly on feedback data [2, 12], or utilize a ranking loss on preference data to directly train the policy model [36, 52, 60, 62] have emerged. The latter set of methods match RLHF in performance. We build on these fine-tuning methods in this work, specifically, direct preference optimization [36] (DPO). Finally, learning from AI feedback, using pretrained reward models, is promising for efficient scaling of alignment [5, 25].

Aligning Diffusion Models Alignment of diffusion models to human preferences has so far been much less explored than in LLMs. Multiple approaches [33, 39] finetune on datasets scored as highly visually appealing by an aesthetics classifier [40], to bias the model to visually appealing generations. Emu [11] finetunes a pretrained model using a small, curated image dataset of high quality photographs with manually written detailed captions to improve visual appeal and text alignment. Other methods [17, 42] recaption existing web-scraped image datasets to improve text fidelity. Caption-aware human preference scoring models are trained on generation preference datasets [24, 55, 58], but the impact of these reward models to the generative space has been limited. DOODL [54] introduces the task of aesthetically improving a single generation iteratively at inference time. DRAFT [9] and AlignProp [34], incorporate a similar approach into training: tuning the generative model to directly increase the reward of generated images. These methods perform well for simple visual appeal criteria, but lack stability and do not work on more nuanced rewards such as text-image alignment [9]. DPOK and DDPO [7, 13] are RL-based approaches to maximize the scored reward (with distributional constraints) over a relatively limited vocabulary set; the performance of these methods degrades as the number of train/test prompts increases. Diffusion-DPO is unique among alignment approaches in effectively increasing measured human appeal across an open vocabulary (DPOK, DDPO), without increased inference time (DOODL) while maintaining distributional guarantees and improving generic text-image alignment in addition to visual appeal (DRAFT, AlignProp). See Tab. 1, Supp. S1).

Methods	Open Vocabulary	Equal Inference Cost	Divergence Control
DPOK [13]	✗	✓	✓
DDPO [7]	✗	✓	✗
DOODL [54]	✓	✗	✗
DRaFT [9], AlignProp [34]	✓	✓	✗
Diffusion-DPO (ours)	✓	✓	✓

Table 1. Method class comparison. Existing methods fail in one or more of: Generalizing to an open vocabulary, maintaining the same inference complexity, avoiding mode collapse/providing distributional guarantees. Diffusion-DPO addresses these issues.

3. Background

3.1. Diffusion Models

Given samples from a data distribution $q(\boldsymbol{x}_0)$, noise scheduling function α_t and σ_t (as defined in [39]), denoising diffusion models [19, 45, 49] are generative models $p_\theta(\boldsymbol{x}_0)$ which have a discrete-time reverse process with a Markov structure $p_\theta(\boldsymbol{x}_{0:T}) = \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ where

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t), \sigma^2_{t|t-1}\frac{\sigma^2_{t-1}}{\sigma^2_t}\boldsymbol{I}). \quad (1)$$

Training is performed by minimizing the evidence lower bound (ELBO) associated with this model [23, 48]:

$$L_{\text{DM}} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},t,\boldsymbol{x}_t}\left[\omega(\lambda_t)\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2\right], \quad (2)$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$, $t \sim \mathcal{U}(0, T)$, $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \alpha_t\boldsymbol{x}_0, \sigma_t^2\boldsymbol{I})$. $\lambda_t = \alpha_t^2/\sigma_t^2$ is a signal-to-noise ratio [23], $\omega(\lambda_t)$ is a pre-specified weighting function (typically chosen to be constant [19, 47]).

3.2. Direct Preference Optimization

Our approach is an adaption of *Direct Preference Optimization (DPO)* [36], an effective approach for learning from human preference for language models. Abusing notation, we also use $\boldsymbol{x}_0$ as random variables for language.

Reward Modeling Estimating human partiality to a generation $\boldsymbol{x}_0$ given conditioning $\boldsymbol{c}$, is difficult as we do not have access to the latent reward model $r(\boldsymbol{c}, \boldsymbol{x}_0)$. In our setting, we assume access only to ranked pairs generated from some conditioning $\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l|\boldsymbol{c}$, where $\boldsymbol{x}_0^w$ and $\boldsymbol{x}_0^l$ denoting the "winning" and "losing" samples. The Bradley-Terry (BT) model stipulates to write human preferences as:

$$p_{\text{BT}}(\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l|\boldsymbol{c}) = \sigma(r(\boldsymbol{c}, \boldsymbol{x}_0^w) - r(\boldsymbol{c}, \boldsymbol{x}_0^l)) \quad (3)$$

where σ is the sigmoid function. $r(\boldsymbol{c}, \boldsymbol{x}_0)$ can be parameterized by a neural network ϕ and estimated via maximum likelihood training for binary classification:

$$L_{\text{BT}}(\phi) = -\mathbb{E}_{\boldsymbol{c},\boldsymbol{x}_0^w,\boldsymbol{x}_0^l}\left[\log\sigma\left(r_\phi(\boldsymbol{c}, \boldsymbol{x}_0^w) - r_\phi(\boldsymbol{c}, \boldsymbol{x}_0^l)\right)\right] \quad (4)$$

where prompt $\boldsymbol{c}$ and data pairs $\boldsymbol{x}_0^w, \boldsymbol{x}_0^l$ are from a static dataset with human-annotated labels.

RLHF RLHF aims to optimize a conditional distribution $p_\theta(\boldsymbol{x}_0|\boldsymbol{c})$ (conditioning $c \sim \mathcal{D}_c$) such that the latent reward model $r(\boldsymbol{c}, \boldsymbol{x}_0)$ defined on it is maximized, while regularizing the KL-divergence from a reference distribution p_{ref}

$$\max_{p_\theta} \mathbb{E}_{\boldsymbol{c}\sim\mathcal{D}_c,\boldsymbol{x}_0\sim p_\theta(\boldsymbol{x}_0|\boldsymbol{c})}\left[r(\boldsymbol{c}, \boldsymbol{x}_0)\right] - \beta\mathbb{D}_{\text{KL}}\left[p_\theta(\boldsymbol{x}_0|\boldsymbol{c})\|p_{\text{ref}}(\boldsymbol{x}_0|\boldsymbol{c})\right] \quad (5)$$

where the hyperparameter β controls regularization.

DPO Objective In Eq. (5) from [36], the unique global optimal solution p^*_θ takes the form:

$$p^*_\theta(\boldsymbol{x}_0|\boldsymbol{c}) = p_{\text{ref}}(\boldsymbol{x}_0|\boldsymbol{c})\exp\left(r(\boldsymbol{c}, \boldsymbol{x}_0)/\beta\right)/Z(\boldsymbol{c}) \quad (6)$$

where $Z(\boldsymbol{c}) = \sum_{\boldsymbol{x}_0} p_{\text{ref}}(\boldsymbol{x}_0|\boldsymbol{c}) \exp\left(r(\boldsymbol{c}, \boldsymbol{x}_0)/\beta\right)$ is the partition function. Hence, the reward function is rewritten as

$$r(\boldsymbol{c}, \boldsymbol{x}_0) = \beta \log \frac{p^*_\theta(\boldsymbol{x}_0|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_0|\boldsymbol{c})} + \beta \log Z(\boldsymbol{c}) \tag{7}$$

Using Eq. (4), the reward objective becomes:

$$L_{\text{DPO}}(\theta) = -\mathbb{E}_{\boldsymbol{c},\boldsymbol{x}^w_0,\boldsymbol{x}^l_0}\left[\log\sigma\left(\beta \log \frac{p_\theta(\boldsymbol{x}^w_0|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}^w_0|\boldsymbol{c})} - \beta \log \frac{p_\theta(\boldsymbol{x}^l_0|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}^l_0|\boldsymbol{c})}\right)\right] \tag{8}$$

By this reparameterization, instead of optimizing the reward function r_ϕ and then performing RL, [36] directly optimizes the optimal conditional distribution $p_\theta(\boldsymbol{x}_0|\boldsymbol{c})$.

4. DPO for Diffusion Models

In adapting DPO to diffusion models, we consider a setting where we have a fixed dataset $\mathcal{D} = \{(\boldsymbol{c}, \boldsymbol{x}^w_0, \boldsymbol{x}^l_0)\}$ where each example contains a prompt $\boldsymbol{c}$ and a pairs of images generated from a reference model p_{ref} with human label $\boldsymbol{x}^w_0 \succ \boldsymbol{x}^l_0$. We aim to learn a new model p_θ which is aligned to the human preferences, with preferred generations to p_{ref}. The primary challenge we face is that the parameterized distribution $p_\theta(\boldsymbol{x}_0|\boldsymbol{c})$ is not tractable, as it needs to marginalize out all possible diffusion paths $(\boldsymbol{x}_1, ...\boldsymbol{x}_T)$ which lead to $\boldsymbol{x}_0$. To overcome this challenge, we utilize the evidence lower bound (ELBO). Here, we introduce latents $\boldsymbol{x}_{1:T}$ and define $R(\boldsymbol{c}, \boldsymbol{x}_{0:T})$ as the reward on the whole chain, such that we can define $r(\boldsymbol{c}, \boldsymbol{x}_0)$ as

$$r(\boldsymbol{c}, \boldsymbol{x}_0) = \mathbb{E}_{p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0,\boldsymbol{c})}\left[R(\boldsymbol{c}, \boldsymbol{x}_{0:T})\right]. \tag{9}$$

As for the KL-regularization term in Eq. (5), following prior work [19, 45], we can instead minimize its upper bound joint KL-divergence $\mathbb{D}_{\text{KL}}\left[p_\theta(\boldsymbol{x}_{0:T}|\boldsymbol{c}) \| p_{\text{ref}}(\boldsymbol{x}_{0:T}|\boldsymbol{c})\right]$. Plugging this KL-divergence bound and the definition of $r(\boldsymbol{c}, \boldsymbol{x}_0)$ (Eq. (9)) back to Eq. (5), we have the objective

$$\max_{p_\theta} \mathbb{E}_{\boldsymbol{c}\sim\mathcal{D}_c,\boldsymbol{x}_{0:T}\sim p_\theta(\boldsymbol{x}_{0:T}|\boldsymbol{c})}\left[r(\boldsymbol{c}, \boldsymbol{x}_0)\right] - \beta\mathbb{D}_{\text{KL}}\left[p_\theta(\boldsymbol{x}_{0:T}|\boldsymbol{c}) \| p_{\text{ref}}(\boldsymbol{x}_{0:T}|\boldsymbol{c})\right]. \tag{10}$$

This objective has a parallel formulation as Eq. (5) but defined on path $\boldsymbol{x}_{0:T}$. It aims to maximize the reward for reverse process $p_\theta(\boldsymbol{x}_{0:T})$, while matching the distribution of the original reference reverse process. Paralleling Eqs. (6) to (8), this objective can be optimized directly through the conditional distribution $p_\theta(\boldsymbol{x}_{0:T})$ via objective:

$$L_{\text{DPO-Diffusion}}(\theta) = -\mathbb{E}_{(\boldsymbol{x}^w_0,\boldsymbol{x}^l_0)\sim\mathcal{D}} \log\sigma\Bigg(\beta\mathbb{E}_{\substack{\boldsymbol{x}^w_{1:T}\sim p_\theta(\boldsymbol{x}^w_{1:T}|\boldsymbol{x}^w_0)\\ \boldsymbol{x}^l_{1:T}\sim p_\theta(\boldsymbol{x}^l_{1:T}|\boldsymbol{x}^l_0)}}\left[\log \frac{p_\theta(\boldsymbol{x}^w_{0:T})}{p_{\text{ref}}(\boldsymbol{x}^w_{0:T})} - \log \frac{p_\theta(\boldsymbol{x}^l_{0:T})}{p_{\text{ref}}(\boldsymbol{x}^l_{0:T})}\right]\Bigg) \tag{11}$$

We omit $\boldsymbol{c}$ for compactness (details included in Supp. S2). To optimize Eq. (11), we must sample $\boldsymbol{x}_{1:T} \sim p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$. Despite the fact that p_θ contains trainable parameters, this sampling procedure is both (1) *inefficient* as T is usually large ($T = 1000$), and (2) *intractable* since $p_\theta(\boldsymbol{x}_{1:T})$ represents the reverse process parameterization $p_\theta(\boldsymbol{x}_{1:T}) = p_\theta(\boldsymbol{x}_T)\prod_{t=1}^T p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. We solve these two issues next.

From Eq. (11), we substitute the reverse decompositions for p_θ and p_{ref}, and utilize Jensen's inequality and the convexity of function $-\log\sigma$ to push the expectation outside. With some simplification, we get the following bound

$$L_{\text{DPO-Diffusion}}(\theta) \leq -\mathbb{E}_{\substack{(\boldsymbol{x}^w_0,\boldsymbol{x}^l_0)\sim\mathcal{D}, t\sim\mathcal{U}(0,T),\\ \boldsymbol{x}^w_{t-1,t}\sim p_\theta(\boldsymbol{x}^w_{t-1,t}|\boldsymbol{x}^w_0),\\ \boldsymbol{x}^l_{t-1,t}\sim p_\theta(\boldsymbol{x}^l_{t-1,t}|\boldsymbol{x}^l_0)}} \log\sigma\left(\beta T \log \frac{p_\theta(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_t)}{p_{\text{ref}}(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_t)} - \beta T \log \frac{p_\theta(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_t)}{p_{\text{ref}}(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_t)}\right) \tag{12}$$

Efficient training via gradient descent is now possible. However, sampling from reverse joint $p_\theta(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{c})$ is still intractable and r of Eq. (9) has an expectation over $p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$. So we approximate the reverse process $p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ with the forward $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ (an alternative scheme in Supp. S2). With some algebra, this yields:

$$\begin{aligned}L(\theta) = -\mathbb{E}_{(\boldsymbol{x}^w_0,\boldsymbol{x}^l_0)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), \boldsymbol{x}^w_t\sim q(\boldsymbol{x}^w_t|\boldsymbol{x}^w_0), \boldsymbol{x}^l_t\sim q(\boldsymbol{x}^l_t|\boldsymbol{x}^l_0)}\\ \log\sigma(-\beta T(\\ + \mathbb{D}_{\text{KL}}(q(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_{0,t}) \| p_\theta(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_t))\\ - \mathbb{D}_{\text{KL}}(q(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_{0,t}) \| p_{\text{ref}}(\boldsymbol{x}^w_{t-1}|\boldsymbol{x}^w_t))\\ - \mathbb{D}_{\text{KL}}(q(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_{0,t}) \| p_\theta(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_t))\\ + \mathbb{D}_{\text{KL}}(q(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_{0,t}) \| p_{\text{ref}}(\boldsymbol{x}^l_{t-1}|\boldsymbol{x}^l_t))).\end{aligned} \tag{13}$$

Using Eq. (1) and algebra, the above loss simplifies to:

$$\begin{aligned}L(\theta) = -\mathbb{E}_{(\boldsymbol{x}^w_0,\boldsymbol{x}^l_0)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), \boldsymbol{x}^w_t\sim q(\boldsymbol{x}^w_t|\boldsymbol{x}^w_0), \boldsymbol{x}^l_t\sim q(\boldsymbol{x}^l_t|\boldsymbol{x}^l_0)}\\ \log\sigma\left(-\beta T\omega(\lambda_t)\right(\\ \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^w_t, t)\|^2_2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^w_t, t)\|^2_2\\ - \left(\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^l_t, t)\|^2_2 - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}^l_t, t)\|^2_2\right)))\end{aligned} \tag{14}$$

where $\boldsymbol{x}^*_t = \alpha_t\boldsymbol{x}^*_0 + \sigma_t\boldsymbol{\epsilon}^*, \boldsymbol{\epsilon}^* \sim \mathcal{N}(0, I)$ is a draw from $q(\boldsymbol{x}^*_t|\boldsymbol{x}^*_0)$ (Eq. (2)). $\lambda_t = \alpha^2_t/\sigma^2_t$ is the signal-to-noise ratio, $\omega(\lambda_t)$ a weighting function (constant in practice [19, 23]). We factor the constant T into β. This loss encourages $\boldsymbol{\epsilon}_\theta$ to improve more at denoising x^w_t than $\boldsymbol{x}^l_t$, visualization in Fig. 2. We also derive Eq. (14) as a multi-step RL approach in the same setting as DDPO and DPOK [7, 13] (Supp. S3) but as an off-policy algorithm, which justifies our sampling choice in Eq. 13. A noisy preference model perspective yields the same objective (Supp. S4).

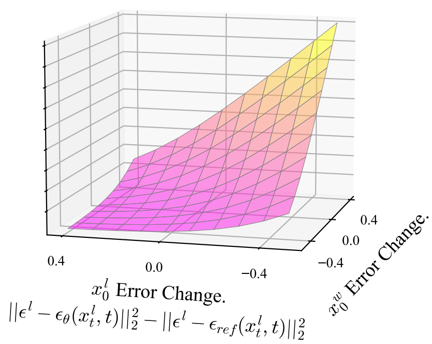

Figure 2. Loss surface visualization. Loss can be decreased by improving at denoising x_0^w and *worsening* for x_0^l. A larger β increases surface curvature.

5. Experiments

5.1. Setting

Models and Dataset: We demonstrate the efficacy of Diffusion-DPO across a range of experiments. We use the objective from Eq. (14) to fine-tune Stable Diffusion 1.5 (SD1.5) [39] and the state-of-the-art open-source model Stable Diffusion XL-1.0 (SDXL) [33] base model. We train on the **Pick-a-Pic [24] dataset**, which consists of pairwise preferences for images generated by SDXL-beta and Dreamlike, a fine-tuned version of SD1.5. The prompts and preferences were collected from users of the Pick-a-Pic web application (see [24] for details). We use the larger Pick-a-Pic v2 dataset. After excluding the $\sim$12% of pairs with ties, we end up with 851,293 pairs, with 58,960 unique prompts.

Hyperparameters We use AdamW [27] for SD1.5 experiments, and Adafactor [43] for SDXL to save memory. An effective batch size of 2048 (pairs) is used; training on 16 NVIDIA A100 GPUs with a local batch size of 1 pair and gradient accumulation of 128 steps. We train at fixed square resolutions. A learning rate of $\frac{2000}{\beta}2.048 \cdot 10^{-8}$ is used with 25% linear warmup. The inverse scaling is motivated by the norm of the DPO objective gradient being proportional to β (the divergence penalty parameter) [36]. For both SD1.5 and SDXL, we find $\beta \in [2000, 5000]$ to offer good performance (Supp. S5). We present main SD1.5 results with $\beta = 2000$ and SDXL results with $\beta = 5000$.

Evaluation We automatically validate checkpoints with the 500 unique prompts of the Pick-a-Pic validation set: measuring median PickScore reward of generated images. Pickscore [24] is a caption-aware scoring model trained on Pick-a-Pic v1 to estimate human-perceived image quality. For final testing, we generate images using the baseline and Diffusion-DPO-tuned models conditioned on captions from the Partiprompt [59] and HPSv2 [55] benchmarks (1632 and 3200 captions respectively). While DDPO [7] is a related method, we did not observe stable improvement when training from public implementations on Pick-a-Pic. We employ labelers on Amazon Mechanical Turk to compare generations under three different criteria: Q1: General Preference (*Which image do you prefer given the prompt?*), Q2: Visual Appeal (prompt not considered) (*Which image is more visually appealing?*) Q3: Prompt Alignment (*Which image better fits the text description?*). Five responses are collected for each comparison with majority vote (3+) being considered the collective decision.

5.2. Primary Results: Aligning Diffusion Models

First, we show that the outputs of the Diffusion-DPO-finetuned SDXL model are significantly preferred over the baseline SDXL-base model. In the Partiprompt evaluation (Fig. 3-top left), DPO-SDXL is preferred 70.0% of the time for General Preference (Q1), and obtains a similar win-rate in assessments of both Visual Appeal (Q2) and Prompt Alignment (Q3). Evaluation on the HPS benchmark (Fig. 3-top right) shows a similar trend, with a General Preference win rate of 64.7%. We also score the DPO-SDXL HPSv2 generations with the HPSv2 reward model, achieving an average reward of 28.16, topping the leaderboard [56].

We display qualitative comparisons to SDXL-base in Fig. 3 (bottom). Diffusion-DPO produces more appealing imagery, with vivid arrays of colors, dramatic lighting, good composition, and realistic people/animal anatomy. While all SDXL images satisfy the prompting criteria to some degree, the DPO generations appear superior, as confirmed by the crowdsourced study. We do note that preferences are not universal, and while the most common shared preference is towards energetic and dramatic imagery, others may prefer quieter/subtler scenes. The area of personal or group preference tuning is an exciting area of future work.

After this parameter-equal comparison with SDXL-base, we compare SDXL-DPO to the complete SDXL pipeline (base + refiner) in Fig. 4. The refinement model is an image-to-image diffusion model that improves visual quality of generations, and is especially effective on detailed backgrounds and faces. In our experiments with PartiPrompts and HPSv2, SDXL-DPO (3.5B parameters, SDXL-base architecture only), handily beats the complete SDXL model (6.6B parameters). In the General Preference question, it has a benchmark win rate of 69% and 64% respectively, comparable to its win rate over SDXL-base alone. This is explained by the ability of the DPO-tuned model (Fig. 4, bottom) to generate fine-grained details and its strong performance across different image categories. While the refinement model is especially good at improving the generation of human details, the win rate of Diffusion-DPO on the *People* category in Partiprompt dataset over the base + re-

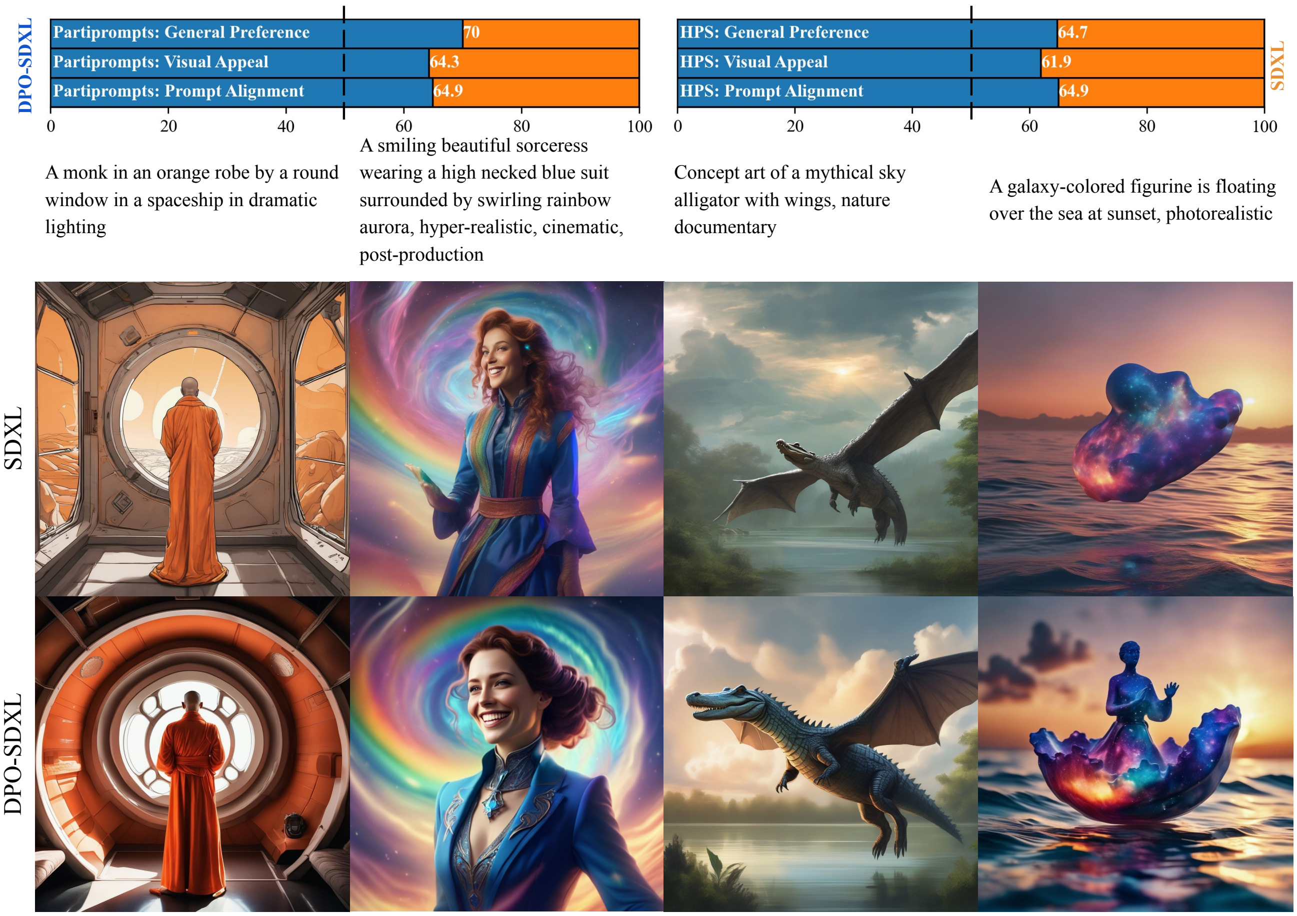

Figure 3. (Top) DPO-SDXL significantly outperforms SDXL in human evaluation. (L) PartiPrompts and (R) HPSv2 benchmark results across three evaluation questions, majority vote of 5 labelers. (Bottom) Qualitative comparisons between SDXL and DPO-SDXL. DPO-SDXL demonstrates superior prompt following and realism. DPO-SDXL outputs are better aligned with human aesthetic preferences, favoring high contrast, vivid colors, fine detail, and focused composition. They also capture fine-grained textual details more faithfully.

finer model is still an impressive 67.2% (compared to 73.4% over the base). Further evaluations, including comparisons to Emu [11] and DDPO [61] are in Supp. S1.

5.3. Image-to-Image Editing

Image-to-image translation performance also improves after Diffusion-DPO tuning. We test DPO-SDXL on TEdBench [22], a text-based image-editing benchmark of 100 real image-text pairs, using SDEdit [28] with noise strength 0.6. Labelers are shown the original image and SDXL/DPO-SDXL edits and asked *"Which edit do you prefer given the text?"* DPO-SDXL is preferred 65% of the time, SDXL 24%, with 11% draws. We show qualitative SDEdit results on color layouts (strength 0.98) in Fig. 5.

5.4. Learning from AI Feedback

In LLMs, learning from AI feedback has emerged as a strong alternative to learning from human preferences [25]. Diffusion-DPO can admit learning from AI feedback by directly ranking generated pairs into (y_w, y_l) using a pretrained scoring network. We use HPSv2 [55] for an alternate prompt-aware human preference estimate, CLIP (OpenCLIP ViT-H/14) [21, 35] for text-image alignment, Aesthetic Predictor [40] for non-text-based visual appeal, and PickScore. We run all experiments on SD 1.5 ($\beta = 5000$, 1000 steps, 2048 batch size). Training on PickScore and HPS rankings increase the win rate for both raw visual appeal and prompt alignment (Fig. 6). We note that PickScore feedback is interpretable as pseudo-labeling the Pick-a-Pic dataset—a form of data cleaning [57, 64]. Training for Aesthetics and CLIP improves those capabilities more specifically, in the case of Aesthetics at the expense of CLIP. The ability to train for text-image alignment via CLIP is a noted improvement over prior work [9]. Moreover, training SD1.5 on the pseudo-labeled PickScore dataset ($\beta = 5000$, 2000 steps, batch size 2048) outperforms

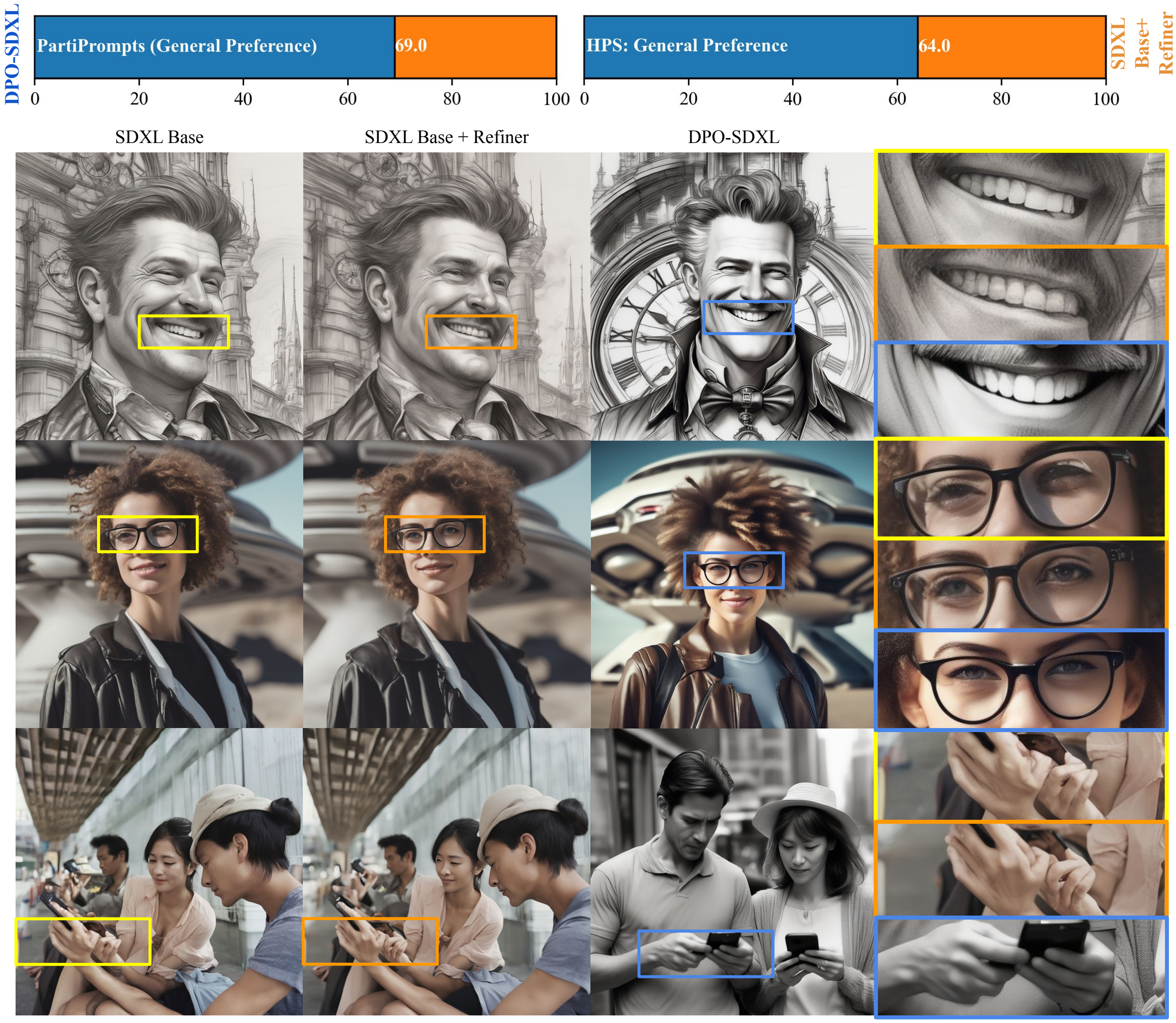

Figure 4. DPO-SDXL (base only) significantly outperforms the much larger SDXL-(base+refinement) model pipeline in human evaluations on the PartiPrompts and HPS datasets. While the SDXL refinement model is used to touch up details from the output of SDXL-base, the ability to generate high quality details has been naturally distilled into DPO-SDXL by human preference. Among other advantages, DPO-SDXL shows superior generation of anatomical features such as teeth, hands, and eyes. Prompts: *close up headshot, steampunk middle-aged man, slick hair big grin in front of gigantic clocktower, pencil sketch / close up headshot, futuristic young woman with glasses, wild hair sly smile in front of gigantic UFO, dslr, sharp focus, dynamic composition / A man and woman using their cellphones, photograph*

training on the raw labels. On the General Preference Partiprompt question, the win-rate of DPO increases from 59.8% to 63.3%, indicating that learning from AI feedback can be a promising direction for diffusion model alignment.

5.5. Analysis

Implicit Reward Model As a consequence of the theoretical framework, our DPO scheme implicitly learns a reward model and can estimate the differences in rewards between two images by taking an expectation over the inner term of Eq. (14) (details in Supp. S4.1). We estimate over 10 random $t \sim \mathcal{U}\{0, 1\}$ Our learned models (DPO-SD1.5 and DPO-SDXL) perform well at binary preference classification (Tab. 2), with DPO-SDXL exceeding all existing recognition models on this split. These results shows that the implicit reward parameterization in the Diffusion-DPO objective has comprable expressivity and generalization as the classical reward modelling objective/architecture.

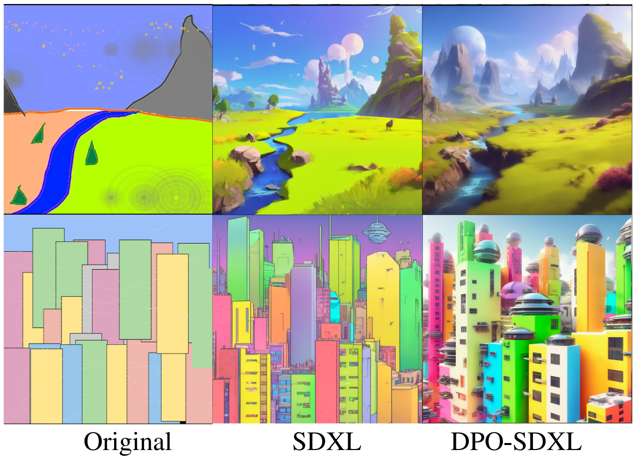

Figure 5. Diffusion-DPO generates more visually appealing images in the downstream image-to-image translation task. Comparisons of using SDEdit [28] from color layouts. Prompts are *"A fantasy landscape, trending on artstation"* (top) , *"High-resolution rendering of a crowded colorful sci-fi city"* (bottom).

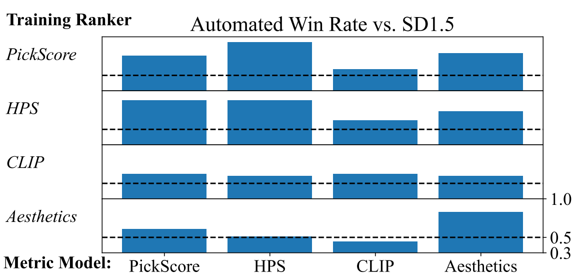

Figure 6. Automated head-to-head win rates under reward models (x labels, columns) for SD1.5 DPO-tuned on the "preferences" of varied scoring networks (y labels, rows). Example: Tuning on *Aesthetics* preferences (bottom row) achieves high Aesthetics scores but has lower text-image alignment as measured by CLIP.

Model	PS	HPS	CLIP	Aes.	DPO-SD1.5	DPO-SDXL
Acc.	64.2	59.3	57.1	51.4	60.8	72.0

Table 2. Preference accuracy on the Pick-a-Pic (v2) validation set. The v1-trained PickScore has seen the evaluated data.

Training Data Quality Fig. 7 shows that despite SDXL being superior to the training data (including the y_w), as measured by Pickscore, DPO training improves its performance substantially. In this experiment, we confirm that Diffusion-DPO can improve on in-distribution preferences as well, by training ($\beta = 5k$, 2000 steps) the Dreamlike model on a subset of the Pick-a-Pic dataset generated by the Dreamlike model alone. This subset represents 15% of the original dataset. Dreamlike-DPO improves on the baseline model, though the performance improvement is limited, perhaps because of the small size of the dataset.

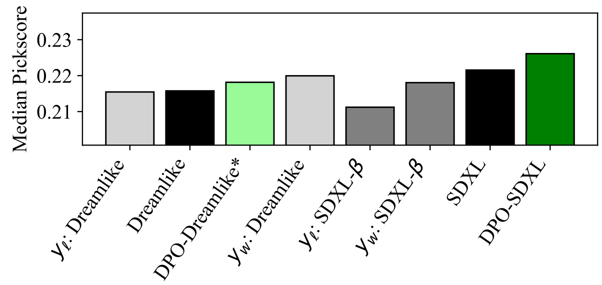

Figure 7. Diffusion-DPO improves on the baseline Dreamlike and SDXL models, when finetuned on both in-distribution data (in case of Dreamlike) and out-of-distribution data (in case of SDXL). y_l and y_w denote the Pickscore of winning and losing samples.

Supervised Fine-tuning (SFT) is beneficial in the LLM setting as initial pretraining prior to preference training. To evaluate SFT in our setting, we fine-tune models on the preferred (x, y_w) pairs of the Pick-a-Pic dataset. We train for the same length schedule as DPO using a learning rate of $1e-9$ and observe convergence. While SFT improves vanilla SD1.5 (55.5% win rate over base model), any amount of SFT deteriorates the performance of SDXL, even at lower learning rates. This contrast is attributable to the much higher quality of Pick-a-Pic generations vs. SD1.5, as they are obtained from SDXL-beta and Dreamlike. In contrast, the SDXL-1.0 base model is superior to the Pick-a-Pic dataset models. See Supp. S6 for further discussion.

6. Conclusion

In this work, we introduce Diffusion-DPO: a method that enables diffusion models to directly learn from human feedback in an open-vocabulary setting for the first time. We fine-tune SDXL-1.0 using the Diffusion-DPO objective and the Pick-a-Pic (v2) dataset to create a new state-of-the-art for open-source text-to-image generation models as measured by generic preference, visual appeal, and prompt alignment. We additionally demonstrate that DPO-SDXL outperforms even the SDXL base plus refinement model pipeline, despite only employing 53% of the total model parameters. Dataset cleaning/scaling is a promising future direction as we observe preliminary data cleaning improving performance (Sec. 5.4). While DPO-Diffusion is an offline algorithm, we anticipate online learning methods to be another driver of future performance. There are also exciting application variants such as tuning to the preferences of individuals or small groups.

Finally, any effort in text-to-image generation presents ethical risks, particularly when data are web-collected. We discuss these risks in detail in Supp. S7

References

[1] https://huggingface.co/carperai/sd-2-1-pickscore-450epochs. 2

[2] Model index for researchers, 2023. 2

[3] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Neural Information Processing Systems*, 2017. 2

[4] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. 2

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. 2

[6] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. *Neural Information processing systems*, 2022. 2

[7] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 1, 3, 4, 5

[8] Kevin Black et al. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2

[9] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2023. 1, 2, 3, 6

[10] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance, 2022. 1

[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. 1, 3, 6, 2

[12] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023. 2

[13] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023. 1, 3, 4, 5

[14] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. 2

[15] Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, Matthieu Geist, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. *Neural Information Processing Systems*, 2021. 5

[16] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022. 2

[17] Gabriel Goh, James Betker, Li Jing, Aditya Ramesh, Tim Brooks, Jianfeng Wang, Lindsey Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Prafulla Dhariwal, Casey Chu, Joy Jiao, Jong Wook Kim, Alex Nichol, Yang Song, Lijuan Wang, and Tao Xu. Improving image generation with better captions. 2023. 3, 11, 13

[18] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling, 2023. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. pages 6840–6851, 2020. 3, 4

[20] Kaiyi Huang et al. T2i-compbench: A comprehensive benchmark for open-world compositional image generation, 2023. 9

[21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 6

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6007–6017, 2023. 6

[23] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. 2021. 3, 4

[24] 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. *arXiv preprint arXiv:2305.01569*, 2023. 3, 5, 9

[25] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback, 2023. 2, 6

[26] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review, 2018. 4

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 6, 8

[29] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, 2022. 2

[30] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016. 2

[31] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1

[32] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 2

[33] 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, 2023. 2, 3, 5, 1

[34] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 1, 3

[35] 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. 6, 1

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. 1, 2, 3, 4, 5, 8

[37] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization, 2022. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable diffusion 2. `https://huggingface.co/stabilityai/stable-diffusion-2`. Accessed: 2023 - 11 - 16. 1

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 3, 5

[40] Christoph Schuhmann. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022. Accessed: 2023 - 11- 10. 3, 6, 1

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 2

[42] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation, 2023. 3

[43] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 5

[44] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *Neural Information Processing Systems*, 2022. 2

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015. 3, 4

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

[47] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[48] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Neural Information Processing Systems*, 2021. 3

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[50] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *Neural Information Processing Systems*, 18, 2020. 2

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer,

Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[52] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023. 2

[53] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. 2

[54] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. *arXiv preprint arXiv:2303.13703*, 2023. 3, 1

[55] 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. 3, 5, 6

[56] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Hpsv2 github. `https://github.com/tgxs002/HPSv2/tree/master`, 2023. Accessed: 2023 - 11 - 15. 5

[57] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 6

[58] 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. *arXiv preprint arXiv:2304.05977*, 2023. 3

[59] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 5

[60] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *Neural Information Processing Systems*, 2023. 2

[61] Yinan Zhang et al. Large-scale reinforcement learning for diffusion models, 2024. 6, 1

[62] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slic-hf: Sequence likelihood calibration with human feedback, 2023. 2

[63] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. Secrets of rlhf in large language models part i: Ppo, 2023. 2

[64] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020. 6