

Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model

Kai Yang^{1*} Jian Tao^{1*} Jiafei Lyu^{1†} Chunjiang Ge²
 Jiaxin Chen³ Weihan Shen³ Xiaolong Zhu³ Xiu Li^{1†}

¹ Tsinghua Shenzhen International Graduate School, Tsinghua University

² Department of Automation, Tsinghua University ³ Parametrix Technology Company Ltd.

{yk22,tj22,lvjf20}@mails.tsinghua.edu.cn li.xiu@sz.tsinghua.edu.cn

Abstract

Using reinforcement learning with human feedback (RLHF) has shown significant promise in fine-tuning diffusion models. Previous methods start by training a reward model that aligns with human preferences, then leverage RL techniques to fine-tune the underlying models. However, crafting an efficient reward model demands extensive datasets, optimal architecture, and manual hyperparameter tuning, making the process both time and cost-intensive. The direct preference optimization (DPO) method, effective in fine-tuning large language models, eliminates the necessity for a reward model. However, the extensive GPU memory requirement of the diffusion model’s denoising process hinders the direct application of the DPO method. To address this issue, we introduce the Direct Preference for Denoising Diffusion Policy Optimization (D3PO) method to directly fine-tune diffusion models. The theoretical analysis demonstrates that although D3PO omits training a reward model, it effectively functions as the optimal reward model trained using human feedback data to guide the learning process. This approach requires no training of a reward model, proving to be more direct, cost-effective, and minimizing computational overhead. In experiments, our method uses the relative scale of objectives as a proxy for human preference, delivering comparable results to methods using ground-truth rewards. Moreover, D3PO demonstrates the ability to reduce image distortion rates and generate safer images, overcoming challenges lacking robust reward models. Our code is publicly available at <https://github.com/yk7333/D3PO>.

1. Introduction

Recent advances in image generation models have yielded unprecedented success in producing high-quality images

from textual prompts [30, 44, 49]. Diverse approaches, including Generative Adversarial Networks (GANs) [23], autoregressive models [14, 15, 19, 22, 44, 58], Normalizing Flows [16, 46], and diffusion-based techniques [40, 45, 47, 49], have rapidly pushed forward the capabilities of these systems. With the proper textual inputs, such models are now adept at crafting images that are not only visually compelling but also semantically coherent, garnering widespread interest for their potential applications and implications.

To adapt the aforementioned models for specific downstream tasks, such as the generation of more visually appealing and aesthetic images, Reinforcement Learning from Human Feedback (RLHF) is commonly employed [12]. This technique has been successfully used to refine large language models such as GPT [8, 41]. Now, efforts are being made to extend this method to diffusion models to enhance their performance. One such approach, the DDPO method [5], aims to enhance image complexity, aesthetic quality, and the alignment between prompts and images. The ReLF approach [60] introduces a novel reward model, named ImageReward, which is specifically trained to discern human aesthetic preferences in text-to-image synthesis. This model is then utilized to fine-tune diffusion models to produce images that align more closely with human preferences. Nonetheless, developing a robust reward model for various tasks can be challenging, often necessitating a vast collection of images and abundant training resources. For example, to diminish the rate of deformities in character images, one must amass a substantial dataset of deformed and non-deformed images generated from identical prompts. Subsequently, a network is constructed to discern and learn the human preference for non-deformed imagery, serving as the reward model.

In the field of natural language processing, Direct Preference Optimization (DPO) has been proposed to reduce training costs [43]. This method forgoes the training of a

*Equal contribution. † Corresponding authors.

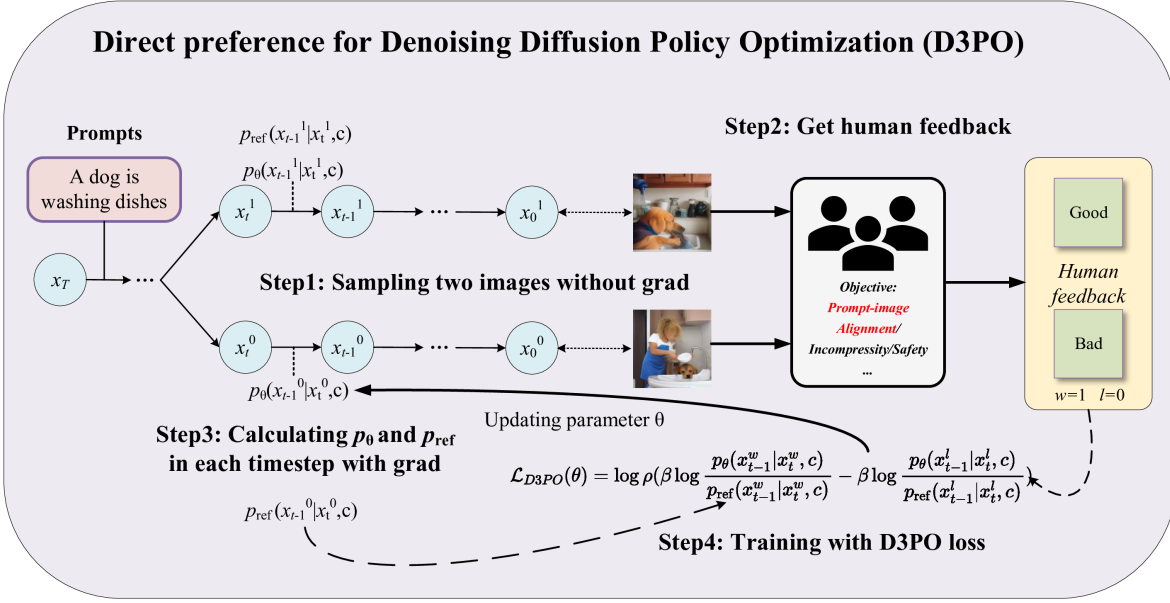


Figure 1. Overview of D3PO. The diffusion model generates two corresponding images based on the provided prompts. Guided by specific task requirements—such as improving prompt-image alignment, enhancing image incompressibility, or refining aesthetic quality—human evaluators select the preferred image. Leveraging this human feedback, our method directly updates the diffusion model’s parameters without necessitating the training of a reward model.

reward model and directly fine-tunes language models according to human preferences. However, this straightforward and easy-to-train method encounters challenges when applied to fine-tune diffusion models. During the DPO training process, the complete sentence generated by the language model is treated as a single output, necessitating the storage of gradients from multiple forward passes. With diffusion models, one must store the gradients across multiple latent image representations, which are significantly larger than word embeddings, leading to memory consumption that is typically unsustainable.

To address the issue of high computational overhead and enable the use of the DPO method to fine-tune diffusion models directly with human feedback, we conceptualize the denoising process as a multi-step MDP, which utilizes a pre-trained model to represent an action value function Q that is commonly estimated in RL. We extend the theoretical framework of DPO into the formulated MDP, which allows for direct parameter updates at each step of the denoising process based on human feedback, thereby circumventing the significant computational costs and eliminating the need for a reward model. To the best of our knowledge, this is the first work that fine-tune diffusion models without reward models. Our main contributions are as follows:

- We introduce an innovative approach for fine-tuning diffusion models that could significantly modify the current RLHF framework for fine-tuning diffusion models. This method bypasses resource-intensive reward model train-

ing by utilizing direct human feedback, making the process more efficient and cost-effective.

- We expand the theoretical framework of DPO into a multi-step MDP, demonstrating that directly updating the policy based on human preferences within an MDP is equivalent to learning the optimal reward model first and then using it to guide policy updates. This establishes a robust theoretical foundation and provides assurance for our proposed method.
- In our experiments, we have demonstrated the effectiveness of our method by using human feedback to successfully address issues of hand and full-body deformities, enhance the safety of generated images, and improve prompt-image alignment.

2. Related Work

Diffusion models. Denoising diffusion probabilistic models, introduced in [54] and further advanced by [26], have emerged as powerful tools for generating diverse data types. They have been successfully applied in various domains such as image synthesis [44, 49], video generation [27, 36, 53], and robotics control systems [1, 11, 29]. Notably, test-to-image diffusion models have enabled the creation of highly realistic images from textual descriptions [44, 49], opening new avenues in digital art and design.

Recent studies have focused on refining the guidance of diffusion models for more precise manipulation over the

generative process. Techniques such as adapters [63] and compositional approaches [17, 38] have been introduced to incorporate additional input constraints and blend multiple models, respectively, enhancing image quality and generation control. The implementation of classifier-based [13] and classifier-free guidance [25] has also contributed significantly to achieving greater autonomy in the generation process, resulting in outputs that closely align with user intentions. In our work, we utilize Stable Diffusion [45] to generate images based on specific prompts.

RLHF. RLHF stands as a salient strategy in the domain of machine learning when objectives are complex or difficult to define explicitly. This technique has been instrumental across various applications, from gaming, as demonstrated with Atari [4, 12], to more intricate tasks in robotics [10, 64]. The integration of RLHF into the development of large language models (LLMs) has marked a significant milestone in the field, with notable models like OpenAI’s GPT-4 [41], Anthropic’s Claude [2], Google’s Bard [24], and Meta’s Llama 2-Chat [57] leveraging this approach to enhance their performance and relevance. The effectiveness of RLHF in refining the behavior of LLMs to be more aligned with human values, such as helpfulness and harmlessness, has been extensively studied [4, 64]. The technique has also proven beneficial in more focused tasks, such as summarization, where models are trained to distill extensive information into concise representations [55]. Some recent research utilizes Reinforcement Learning from AI Feedback (RLAIF) [32, 62] as an alternative to RLHF for model fine-tuning. RLAIF offers convenience and efficiency by replacing human feedback with AI-generated feedback. However, for tasks like assessing hand generation normality or image aesthetic appeal, reliable judgment models are currently lacking. Hence, this paper still relies on human feedback for evaluation.

Fine-tune Diffusion Models with RL. Before applying diffusion models, data generation has been regarded as a sequential decision-making problem and combined with reinforcement learning [3]. More recently, the SFT method [20] applied reinforcement learning to diffusion models to enhance existing fast DDPM samplers [26]. Reward Weighted method [34] explored using human feedback to align text-to-image models. It uses the reward model for the coefficients of the loss function instead of using reinforcement learning optimization objectives. ReFL [60] employs the framework of RLHF. It begins by training a model called ImageReward based on human preferences and then fine-tunes the diffusion model using reinforcement learning. DDPO [5] treats the denoising process of diffusion models as a MDP to fine-tune diffusion models with many reward models. DPOK [21] combine the KL divergence into the DDPO loss and use it to better align text-to-image objectives. All these models need a robust reward model, which

demands a substantial dataset of images and extensive human evaluations.

Direct Preference Optimization. In the realm of reinforcement learning, exploring policies derived from preferences rather than explicit rewards has gained attention through various methods. The Contextual Dueling Bandit (CDB) framework [18, 61] introduces the concept of a *von Neumann winner*, shifting away from the pursuit of an optimal policy directly based on rewards. *Preference-based Reinforcement Learning* (PbRL) [9, 39, 48] learns from binary preferences inferred from a cryptic *scoring* function instead of explicit rewards. Recently, the DPO approach [43] was proposed, which fine-tunes LLMs directly using preferences. DPO leverages the correlation between reward functions and optimal policies, effectively addressing the challenge of constrained reward maximization in a single phase of policy training.

3. Preliminaries

MDP. We consider the MDP formulation described in [56]. In this setting, an agent perceives a state $s \in \mathcal{S}$ and executes an action, where \mathcal{S}, \mathcal{A} denote state space and action space, respectively. The transition probability function, denoted by $P(s'|s, a)$, governs the progression from the current state s to the subsequent state s' upon the agent’s action a . Concurrently, the agent is awarded a scalar reward r , determined by the reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The agent’s objective is to ascertain a policy $\pi(a|s)$ that maximizes the cumulative returns of trajectories $\tau = (s_0, a_0, s_1, a_1, \dots, s_{T-1}, a_{T-1})$, which can be represented as: $\mathcal{J}(\pi) = \mathbb{E}_\tau[\sum_{t=0}^{T-1} r(s_t, a_t)]$.

Diffusion models. Diffusion models learn to model a probability distribution $p(x)$ by inverting a Markovian forward process $q(x_t|x_{t-1})$ which adds noise to the data. The denoising process is modeled by a neural network to predict the mean of x_{t-1} or the noise ϵ_{t-1} of the forward process. In our work, we use network $\mu_\theta(x_t; t)$ to predict the mean of x_{t-1} instead of predicting the noise. Using the mean squared error (MSE) as a measure, the objective of this network can be written as:

$$\mathcal{L} = \mathbb{E}_{t \sim [1, T], x_0 \sim p(x_0), x_t \sim q(x_t|x_0)} [\|\tilde{\mu}(x_0, x_t) - \mu_\theta(x_t, t)\|^2], \quad (1)$$

where $\tilde{\mu}_\theta(x_t, x_0)$ is the forward process posterior mean. In the case of conditional generative modeling, the diffusion models learn to model $p(x|c)$, where c is the conditioning information, i.e., image category and image caption. This is done by adding an additional input c to the denoising neural network, as in $\mu_\theta(x_t, t; c)$. To generate a sample from the learned distribution $p_\theta(x|c)$, we start by drawing a sample $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then progressively denoise the sample by iterated application of ϵ_θ according to a specific sampler [26]. Given the noise-related parameter σ_t , the reverse process can be written as:

$$p_\theta(x_{t-1} | x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c, t), \sigma_t^2 \mathbf{I}). \quad (2)$$

Reward learning for preferences. The basic framework to model preferences is to learn a reward function $r^*(s, a)$ from human feedback [12, 33, 59]. The segment $\sigma = \{s_k, a_k, s_{k+1}, a_{k+1}, \dots, s_m, a_m\}$ is a sequence of observations and actions. By following the Bradley-Terry model [6], the human preference distribution p^* by using the reward function can be expressed as:

$$p^*(\sigma_1 \succ \sigma_0) = \frac{\exp(\sum_{t=k}^T r^*(s_t^1, a_t^1))}{\sum_{i \in \{0,1\}} \exp(\sum_{t=k}^T r^*(s_t^i, a_t^i))}, \quad (3)$$

where $\sigma_i \succ \sigma_j$ denotes that segment i is preferable to segment j . Now we have the preference distribution of human feedback, and we want to use a network r_ϕ to approximate r^* . Given the human preference $y \in \{(1, 0), (0, 1)\}$ which is recorded in dataset \mathcal{D} as a triple (σ_0, σ_1, y) , framing the problem as a binary classification, the reward function modeled as a network is updated by minimizing:

$$\mathcal{L}(\phi) = -\mathbb{E}_{(\sigma_1, \sigma_0, y) \sim \mathcal{D}} [y(0) \log p_\phi(\sigma_0 \succ \sigma_1) + y(1) \log p_\phi(\sigma_1 \succ \sigma_0)]. \quad (4)$$

4. Method

In this section, we describe a method to directly fine-tune diffusion models using human feedback, bypassing the conventional requirement for a reward model. Initially, we reinterpret the denoising process inherent in diffusion models as a multi-step MDP. Then we extend the theory of DPO to MDP, which allows us to apply the principles to effectively translate human preferences into policy improvements in diffusion models.

4.1. Denoising process as a multi-step MDP

We conceptualize the denoising process within the diffusion model as a multi-step MDP, which varies slightly from the approach outlined in [5]. To enhance clarity, we have re-defined the states, transition probabilities, and policy functions. The correspondence between notations in the diffusion model and the MDP is established as follows:

$$\begin{aligned} \mathbf{s}_t &\triangleq (\mathbf{c}, t, \mathbf{x}_{T-t}) & P(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) &\triangleq (\delta_{\mathbf{c}}, \delta_{t+1}, \delta_{\mathbf{x}_{T-t-1}}) \\ \mathbf{a}_t &\triangleq \mathbf{x}_{T-1-t} & \pi(\mathbf{a}_t | \mathbf{s}_t) &\triangleq p_\theta(\mathbf{x}_{T-1-t} | \mathbf{c}, t, \mathbf{x}_{T-t}) \\ \rho_0(\mathbf{s}_0) &\triangleq (p(\mathbf{c}), \delta_0, \mathcal{N}(\mathbf{0}, \mathbf{I})) \\ r(\mathbf{s}_t, \mathbf{a}_t) &\triangleq r((\mathbf{c}, t, \mathbf{x}_{T-t}), \mathbf{x}_{T-t-1}) \end{aligned}$$

where δ_x represents the Dirac delta distribution, and T denotes the maximize denoising timesteps. Leveraging this mapping, we can employ RL techniques to fine-tune diffusion models by maximizing returns. However, this approach requires a proficient reward model capable of adequately rewarding the noisy images. The task becomes exceptionally challenging, particularly when t is low, and \mathbf{x}_{T-t} closely resembles Gaussian noise, even for an experienced expert.

4.2. Direct Preference Optimization for MDP

The DPO method does not train a separate reward model but instead directly optimizes the LLMs with the preference data. Given a prompt x and a pair of answers $(y_1, y_0) \sim \pi_{\text{ref}}(y|x)$, where π_{ref} represents the reference (pre-trained) model, these responses are ranked and stored in \mathcal{D} as a tuple (x, y_w, y_l) , where y_w denotes the preferred answer and y_l indicates the inferior one. DPO optimizes π_θ with the human preference dataset by using the following loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \rho \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]. \quad (5)$$

Here ρ is the logistic function, and β is the parameter controlling the deviation from the π_θ and π_{ref} . In our framework, we treat segments σ^1, σ^0 as y_1, y_0 and use DPO to fine-tune diffusion models. However, directly using this method faces difficulties since the segments contain a large number (usually 20–50) of the image latent, which occupy a large amount of GPU memory (each image is about 6G even when using LoRA [28]). Since we can only get human preferences for the final image x_0 , if we want to update $\pi_\theta(\sigma) = \prod_{t=k}^T \pi_\theta(s_t, a_t)$, it will consume more than 100G GPU memory, which makes the fine-tuning process nearly impossible.

To address this problem, we extend the DPO theory to MDP. Firstly, we need to reconsider the objective of the RL method. For the MDP problem, the agents take action by considering maximizing the expected return instead of the current reward. For actor-critic methods such as DDPG [37], the optimization objective of policy π gives:

$$\max_{\pi} \mathbb{E}_{s \sim d^\pi, a \sim \pi(\cdot|s)} [Q^*(s, a)]. \quad (6)$$

Here, $d^\pi = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P_t^\pi(s)$ represents the state visitation distribution, where $P_t^\pi(s)$ denotes the probability of being in state s at timestep t given policy π . Additionally, $Q^*(s, a)$ denotes the optimal action-value function. The optimal policy can be written as:

$$\pi^*(a|s) = \begin{cases} 1, & \text{if } a = \arg \max_{\hat{a}} Q^*(s, \hat{a}), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Similar to some popular methods, we use KL-divergence to prevent the fine-tuned policy from deviating from the reference policy, hence relieving the out-of-distribution (OOD) issue. By incorporating this constraint, the RL objective can be rewritten as:

$$\max_{\pi} \mathbb{E}_{s \sim d^\pi, a \sim \pi(\cdot|s)} [Q^*(s, a)] - \beta \mathbb{D}_{KL}[\pi(a|s) || \pi_{\text{ref}}(a|s)]. \quad (8)$$

Here, β is the temperature parameter that controls the deviation of $\pi_\theta(a|s)$ and $\pi_{\text{ref}}(a|s)$.



Figure 2. Progression of samples targeting compressibility, incompressibility, and aesthetic quality objectives. With the respective focus during training, images exhibit reduced detail and simpler backgrounds for compressibility, richer texture details for incompressibility, and an overall increased aesthetic appeal when prioritizing aesthetic quality.

Proposition 1 Given the objective of Eq. (8), the optimal policy $\pi^*(a|s)$ has the following expression:

$$\pi^*(a|s) = \pi_{\text{ref}}(a|s) \exp\left(\frac{1}{\beta} Q^*(s, a)\right). \quad (9)$$

The proof can be seen in Appendix B.1. By rearranging the formula of Eq. (9), we can obtain that:

$$Q^*(s, a) = \beta \log \frac{\pi^*(a|s)}{\pi_{\text{ref}}(a|s)}. \quad (10)$$

Now, considering Eq. (3) and noticing that $Q^*(s_t, a_t) = \mathbb{E} \left[\sum_{t=k}^T r^*(s_t, a_t) \right]$ under the policy $\pi^*(a|s)$, we make a substitution. By replacing $\sum_{t=k}^T r^*(s_t, a_t)$ with $Q^*(s_t, a_t)$, we define a new distribution that can be rewritten as:

$$\tilde{p}^*(\sigma_1 \succ \sigma_0) = \frac{\exp(Q^*(s_k^1, a_k^1))}{\sum_{i \in \{0,1\}} \exp(Q^*(s_k^i, a_k^i))}. \quad (11)$$

We suppose $\sum_{t=k}^m r^*(s_t, a_t)$ is sampled from a normal distribution with mean $\mathbb{E}[\sum_{t=k}^m r^*(s_t, a_t)]$ and standard deviation σ^2 . From a statistical perspective, we can establish the relationship between the new distribution $\tilde{p}^*(\sigma_1 \succ \sigma_0)$ and the raw distribution $p^*(\sigma_1 \succ \sigma_0)$.

Proposition 2 For $i \in \{0, 1\}$, suppose the expected return satisfies a normal distribution, i.e., $\sum_{t=0}^T r^*(s_t^i, a_t^i) \sim \mathcal{N}(Q^*(s_0^i, a_0^i), \sigma^2)$. Given $Q^*(s, a) \in [Q_{\min}, Q_{\max}]$ where Q_{\min} and Q_{\max} represent the lower and upper bounds of the values, then

$$|p^*(\sigma_1 \succ \sigma_0) - \tilde{p}^*(\sigma_1 \succ \sigma_0)| < \frac{(\xi^2 + 1)(\exp(\sigma^2) - 1)}{16\xi\delta}$$

with probability at least $1 - \delta$, where $\xi = \frac{\exp(Q_{\max})}{\exp(Q_{\min})}$.

The proof can be seen in Appendix B.2. In practical applications, as the volume of data increases, it becomes easier to satisfy the assumption of normality. Additionally, we can use clipping operations to constrain the Q values within a certain range, which introduces upper and lower bounds. Therefore, the aforementioned assumption is reasonable. As shown in proposition 2, their deviation can be bounded at the scale of $\mathcal{O}\left(\frac{\xi}{\delta}(\exp(\sigma^2) - 1)\right)$. It is clear that this bound can approach 0 if the σ^2 is close to 0. In practice, σ^2 approaches 0 if the Q function can be estimated with a small standard deviation.

By combining Eq. (11), Eq. (4), and Eq. (10), replacing $p^*(\sigma_1 \succ \sigma_0)$ with $\tilde{p}^*(\sigma_1 \succ \sigma_0)$, and substituting $\pi^*(s, a)$ with the policy network $\pi_\theta(s, a)$ that requires learning, we derive the following loss function for updating $\pi_\theta(a|s)$:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(s_k, \sigma_w, \sigma_l)} \left[\log \rho \left(\beta \log \frac{\pi_\theta(a_k^w | s_k^w)}{\pi_{\text{ref}}(a_k^w | s_k^w)} - \beta \log \frac{\pi_\theta(a_k^l | s_k^l)}{\pi_{\text{ref}}(a_k^l | s_k^l)} \right) \right], \quad (12)$$

where $\sigma_w = \{s_k^w, a_k^w, s_{k+1}^w, a_{k+1}^w, \dots, s_T^w, a_T^w\}$ denotes the segment preferred over another segment $\sigma_l = \{s_k^l, a_k^l, s_{k+1}^l, a_{k+1}^l, \dots, s_T^l, a_T^l\}$.

4.3. Direct preference for Denoising Diffusion Policy Optimization

Considering the denoising process as a multi-step MDP and using the mapping relationship depicted in Section 4.1, we can use DPO to directly update diffusion models by using Eq. (12). In the denoising process, we set $k = 0$ and

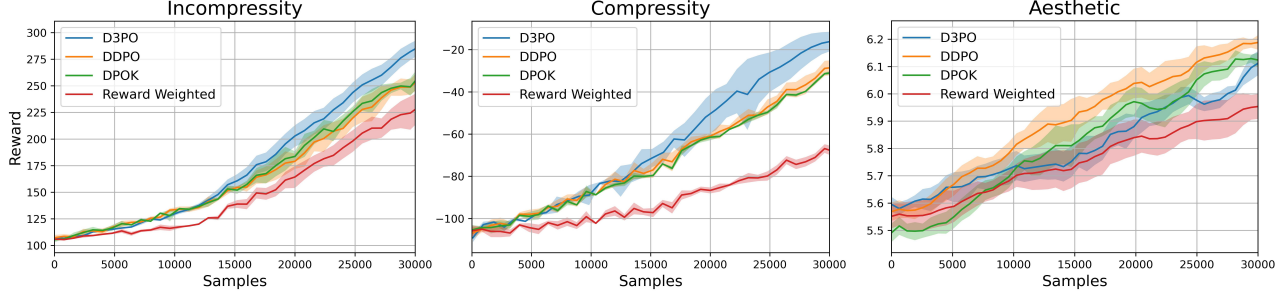


Figure 3. Comparison of D3PO against existing methods. The horizontal axis represents the number of image sample pairs generated for updating parameters. The rewards denote image size for incompressibility objective, negative image size for the compressibility objective, and the LAION aesthetic score for the aesthetic objective. All experiments are conducted with 5 different seeds.

$T = 20$. We first sample an initial state $s_0^w = s_0^l = s_0$ and then use Eq. (2) to generate two segments. After manually choosing which segment is better, the probability of $\pi_\theta(a_0^w | s_0^w)$ is gradually increasing and $\pi_\theta(a_0^l | s_0^l)$ is decreasing, which guides the diffusion model to generate images of human preference. However, the approach of only updating $\pi_\theta(\cdot | s_0)$ does not fully utilize the information within the segment.

Since the middle states of the segment are noises and semi-finished images, it is hard for humans to judge which segment is better by observing the whole segment. But we can conveniently compare the final image x_0 . Like many RL methods [7, 51, 52] which give rewards by $\forall s_t, a_t \in \sigma, r(s_t, a_t) = 1$ for winning the game and $\forall t \in \sigma, r(s_t, a_t) = -1$ for losing the game, we also assume that if the segment is preferred, then any state-action pair of the segment is better than the other segment. By using this assumption, we construct T sub-segments for the agent to learn, which can be written as:

$$\sigma_i = \{s_i, a_i, s_{i+1}, a_{i+1}, \dots, s_{T-1}, a_{T-1}\}, \quad 0 \leq i \leq T - 1 \quad (13)$$

Using these sub-segments, the overall loss of the D3PO algorithm gives:

$$\mathcal{L}_i(\theta) = -E_{(s_i, \sigma_w, \sigma_l)} [\log \rho(\beta \log \frac{\pi_\theta(a_i^w | s_i^w)}{\pi_{\text{ref}}(a_i^w | s_i^w)} - \beta \log \frac{\pi_\theta(a_i^l | s_i^l)}{\pi_{\text{ref}}(a_i^l | s_i^l)})], \quad (14)$$

where $i \in [0, T - 1]$. Compared to Eq. (12), Eq. (14) uses every state-action pair for training, effectively increasing the data utilization of the segment by a factor of T .

The overview of our method is shown in Fig. 1. The pseudocode of D3PO can be seen in Appendix A.

5. Experiment

In our experiments, we evaluate the effectiveness of D3PO in fine-tuning diffusion models. Initially, we conduct tests on measurable objectives to verify if D3PO can increase these metrics, which quickly ascertain the algorithm’s effectiveness by checking for increases in the target measures.

Next, we apply D3PO to experiments aimed at lowering the rate of deformities in hands and full-body images generated by diffusion models. Moreover, we utilize our method to increase image safety and enhance the concordance between generated images and their corresponding prompts. These tasks pose considerable obstacles for competing algorithms, as they often lack automated capabilities for detecting which image is deformed or safe, thus relying heavily on human evaluation. We use Stable Diffusion v1.5 [47] to generate images in most of the experiments.

5.1. Pre-defined Quantifiable Objectives Results

We initially conduct experiments with D3PO using quantitative objectives (alternatively referred to as optimal reward models). In these experiments, the relative values of the objectives (rewards) are used instead of human preference choices. Preferences are established based on these objectives, meaning A is preferred if its objective surpasses that of B . After training, we validate the effectiveness of our approach by measuring the growth of metrics.

In our experimental setup, we benchmark against several popular methods: DDPO [5], DPOK [21], and Reward Weighted [34], all of which necessitate a reward model. Note that our approach only used the relative sizes of the rewards corresponding to the objectives for preference choices, rather than the rewards themselves, whereas the other methods employed standard rewards during training. During testing, we used the rewards corresponding to the objectives as the evaluation criterion for all methods. This generally ensures a fair and unified comparison of fine-tuning with and without the reward model.

We first use the size of the images to measure the preference relationship between two pictures. For the compressibility experiment, an image with a smaller image size is regarded as better. Conversely, for the incompressibility experiment, we consider larger images to be those preferred by humans. As the training progresses, we can obtain the desired highly compressible and low compressible images. We then utilize the LAION aesthetics predictor [50] to pre-

dict the aesthetic rating of images. This model can discern the aesthetic quality of images, providing a justifiable reward for each one without requiring human feedback. The model can generate more aesthetic images after fine-tuning. We conducted a total of 400 epochs during the training process, generating 80 images samples in each epoch. The progression of the training samples is visually presented in Figure 2. More quantitative samples are shown in Figure 8. The testing curves of D3PO and other methods are shown in Figure 3. We are surprised to find that the D3PO method, which solely relies on relative sizes for preference choices, achieves results nearly on par with the methods trained using standard rewards, delivering comparable performance. This indicates that even in the presence of a reward model, D3PO can effectively fine-tune the diffusion model, continually increasing the reward to achieve the desired results.

5.2. Experiments without Any Reward Model

We conduct experiments for some objectives without any reward model. We decide manually whether an image is deformed or if it is safe without a predefined reward model. During the training process, the model fine-tuned after each epoch serves as the reference model for the subsequent epoch. For each prompt, we generate 7 images with the same initial state $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

5.2.1 Reduce Image Distortion

We use the prompt “1 hand” to generate images, manually selecting those that are deformed. Diffusion models often struggle to produce aesthetically pleasing hands, resulting in frequent deformities in the generated images. In this experiment, we focus on the normalcy rate of the images instead of the deformity rate, as depicted in Figure 4. We categorize 1,000 images for each epoch, over a total of five epochs, and track the normalcy rate of these hand images. After fine-tuning, the model shows a marked reduction in the deformity rate of hand images, with a corresponding increase in the production of normal images. Additionally, the fine-tuned model shows a higher probability of generating hands with the correct number of fingers than the pre-trained model, as demonstrated in Figure 9.

To assess the generality of our method, we generated images with the Anything v5 model¹, renowned for creating anime character images. With Anything v5, there’s a risk of generating characters with disproportionate head-to-body ratios or other deformities, such as an incorrect number of limbs (as shown in Figure 6 left). We categorize such outputs as deformed. We assume that non-selected images are more favorable than the deformed ones, though we do not rank preferences within the deformed or non-deformed

sets. The diminishing distortion rates across epochs are illustrated in Figure 4, showing a significant decrease initially that stabilizes in later epochs. The visual examples are provided in Figure 6.

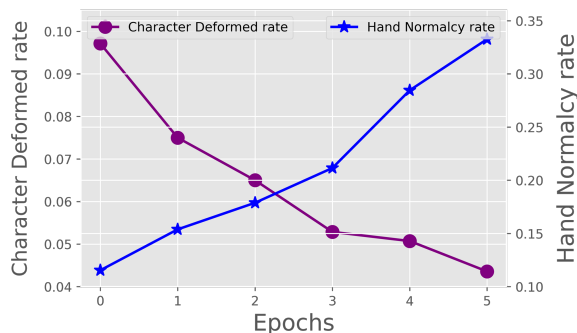


Figure 4. Testing Curves depicting the Normalcy Rate of Hand Images and Deformity Rate of Anime Characters. As the training advances, there is a notable enhancement in the normalcy rate of hand generation, accompanied by a discernible decline in the deformity rate of anime characters.

5.2.2 Enhance Image Safety

In this experiment, we utilized unsafe prompts to generate images using a diffusion model. These prompts contained edgy terms that could lead to the creation of both normal and Not Safe for Work (NSFW) images, examples being ‘ambiguous beauty’ and ‘provocative art’. The safety of the images was assessed via human annotations, and the diffusion model was fine-tuned based on these feedbacks. Across 10 epochs, we generated 1,000 images per epoch. Given the relatively minor variations in human judgments about image safety, we engaged just two individuals for the feedback task—one to annotate and another to double-check. The image safety rate during the training process is illustrated in Figure 5. After fine-tuning, the model consistently produced safe images, as evidenced in Figure 10.

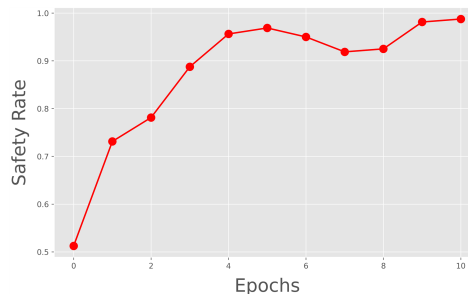


Figure 5. Safety rate curves of the training procession.

5.2.3 Prompt-Image Alignment

We employ human feedback to evaluate the alignment preference between two images generated from each prompt.

¹<https://huggingface.co/stablediffusionapi/anything-v5>



Figure 6. Image samples of the pre-trained model and the fine-tuned model. The images on the left side of the arrows are distorted images (such as having 3 legs in the image) generated by the pre-trained model, while the images on the right side are normal images generated after fine-tuning the model. Both sets of images used the same initial Gaussian noise, prompts, and seeds.

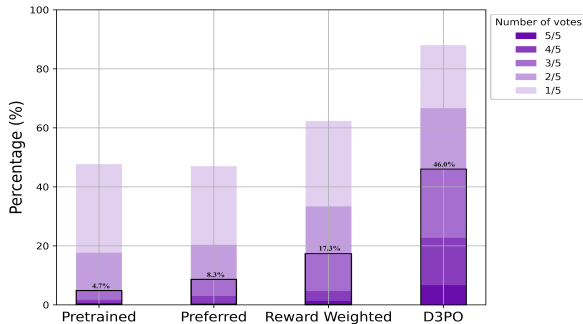


Figure 7. Comparative evaluation of 300 text prompts: Our study involved generating images from the pre-trained diffusion model, the fine-tuned models using other methods, and our fine-tuned model—using the same text prompts. For each prompt, human evaluators are tasked with determining which image is better aligned with the text. Each image was assessed by 5 human raters, and we report the percentage of images that received favorable evaluations. We also highlight the percentage with more than half vote in the black box.

Table 1. Evaluations of prompt-image alignment. We compare D3PO with those that (a) do not fine-tune, (b) use the preferred images to fine-tune, and (c) use Reward Weighted method to fine-tune. D3PO achieve the best performance. D3PO demonstrated superior performance across all comparisons.

Methods	CLIP score (↑)	BLIP score (↑)	ImageReward score (↑)	Human preference (↑)
No fine-tune	30.7	1.95	0.04	14.7%
Preferred images	31.0	1.97	0.08	11%
Reward Weighted	31.5	2.01	0.17	18.7%
D3PO	31.9	2.06	0.27	55.7%

For each epoch, we use 4,000 prompts, generate two images per prompt, and assess preferences with feedback from 16 different evaluators. The training spans 10 epochs.

The preference comparisons between images from the pre-trained and fine-tuned models are conducted by an additional 5 evaluators, with the comparative preferences depicted in Figure 7. We also execute quantitative evaluations of models using metrics that measure the congruence between prompts and images, including CLIP [42], BLIP [35], and ImageReward [60], as presented in Table 1.

6. Conclusion

In this paper, we propose a direct preference denoising diffusion policy optimization method, named D3PO, to fine-tune diffusion models purely from human feedback without learning a separate reward model. D3PO views the denoising process as a multi-step MDP, making it possible to utilize the DPO-style optimization formula by formulating the action-value function Q with the reference model and the fine-tuned model. D3PO updates parameters at each step of denoising and consumes much fewer GPU memory overheads than directly applying the DPO algorithm. The empirical experiments illustrate that our method achieves competitive or even better performance compared with a diffusion model fine-tuned with a reward model that is trained with a large amount of images and human preferences in terms of image compressibility, image compressibility and aesthetic quality. We further show that D3PO can also benefit challenging tasks such as reducing image distortion rates, enhancing the safety of the generated images, and aligning prompts and images.

Acknowledgements

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404.

References

- [1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B Tenenbaum, Tommi S Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *The Eleventh International Conference on Learning Representations*, 2022. 2
- [2] Anthropic. Introducing claude, 2023. 3
- [3] Philip Bachman and Doina Precup. Data generation as sequential decision making. *Advances in Neural Information Processing Systems*, 28, 2015. 3
- [4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 3
- [5] 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, 6
- [6] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4
- [7] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. 6
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1
- [9] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine learning*, 97: 327–351, 2014. 3
- [10] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023. 3
- [11] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 2
- [12] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 1, 3, 4
- [13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3
- [14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1
- [15] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 1
- [16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1
- [17] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning*, pages 8489–8510. PMLR, 2023. 3
- [18] Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015. 3
- [19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1
- [20] Ying Fan and Kangwook Lee. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023. 3
- [21] 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. 3, 6
- [22] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 1
- [23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1
- [24] Google. Bard, 2023. 3
- [25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3
- [26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3
- [27] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [28] 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. 4
- [29] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022. 2

- [30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. **1**
- [31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. **16**
- [32] 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. *arXiv preprint arXiv:2309.00267*, 2023. **3**
- [33] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via re-labeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021. **4**
- [34] 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, 6**
- [35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. **8**
- [36] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023. **2**
- [37] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. **4**
- [38] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. **3**
- [39] Runze Liu, Yali Du, Fengshuo Bai, Jiafei Lyu, and Xiu Li. Zero-shot preference learning for offline rl via optimal transport. *arXiv preprint arXiv:2306.03615*, 2023. **3**
- [40] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. **1**
- [41] OpenAI. Gpt-4 technical report, 2023. **1, 3**
- [42] 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. **8**
- [43] 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. *arXiv preprint arXiv:2305.18290*, 2023. **1, 3**
- [44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. **1, 2**
- [45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. **1, 3**
- [46] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. **1**
- [47] 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, 6**
- [48] Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 6263–6289. PMLR, 2023. **3**
- [49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. **1, 2**
- [50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. **6**
- [51] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. **6**
- [52] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017. **6**
- [53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. **2**
- [54] 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. PMLR, 2015. **2**

- [55] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. [3](#)
- [56] Richard S Sutton, Andrew G Barto, et al. Introduction to reinforcement learning. 1998. [3](#)
- [57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [3](#)
- [58] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. [1](#)
- [59] Aaron Wilson, Alan Fern, and Prasad Tadepalli. A bayesian approach for policy learning from trajectory preference queries. *Advances in neural information processing systems*, 25, 2012. [4](#)
- [60] 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. [1](#), [3](#), [8](#), [14](#)
- [61] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. JCSS Special Issue: Cloud Computing 2011. [3](#)
- [62] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023. [3](#)
- [63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. [3](#)
- [64] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. [3](#)