Towards Better Alignment: Training Diffusion Models with Reinforcement Learning Against Sparse Rewards

Zijing Hu[1]* Fengda Zhang[2]* Long Chen[3] Kun Kuang[1]† Jiahui Li[1] Kaifeng Gao[1]
Jun Xiao[1] Xin Wang[4] Wenwu Zhu[4]

[1]Zhejiang University, [2]Nanyang Technological University, [3]The Hong Kong University of Science and Technology, [4]Tsinghua University

{zj.hu,fdzhang}@zju.edu.cn, zjuchenlong@gmail.com, {kunkuang,jiahuil,kite_phone}@zju.edu.cn,
junx@cs.zju.edu.cn, {xin_wang,wwzhu}@tsinghua.edu.cn

Abstract

*Diffusion models have achieved remarkable success in text-to-image generation. However, their practical applications are hindered by the misalignment between generated images and corresponding text prompts. To tackle this issue, reinforcement learning (RL) has been considered for diffusion model fine-tuning. Yet, RL's effectiveness is limited by the challenge of sparse reward, where feedback is only available at the end of the generation process. This makes it difficult to identify which actions during the denoising process contribute positively to the final generated image, potentially leading to ineffective or unnecessary denoising policies. To this end, this paper presents a novel RL-based framework that addresses the sparse reward problem when training diffusion models. Our framework, named B^2-DiffuRL, employs two strategies: **B**ackward progressive training and **B**ranch-based sampling. For one thing, backward progressive training focuses initially on the final timesteps of denoising process and gradually extends the training interval to earlier timesteps, easing the learning difficulty from sparse rewards. For another, we perform branch-based sampling for each training interval. By comparing the samples within the same branch, we can identify how much the policies of the current training interval contribute to the final image, which helps to learn effective policies instead of unnecessary ones. B^2-DiffuRL is compatible with existing optimization algorithms. Extensive experiments demonstrate the effectiveness of B^2-DiffuRL in improving prompt-image alignment and maintaining diversity in generated images. The code for this work is available*[1].

1. Introduction

The text-to-image generation task aims to produce images from textual descriptions, holding significant potential for various applications [49, 56]. Recently, diffusion models have garnered widespread attention due to their success in this domain [15, 23, 59]. These models employ a sequential denoising process that transforms random noise into detailed images. However, even the most advanced text-to-image diffusion models, such as DALLE3 [6] and Stable Diffusion [52], often encounter issues with misalignment between the generated images and the textual descriptions [28]. This misalignment limits the practicality and effectiveness of these models in real-world applications.

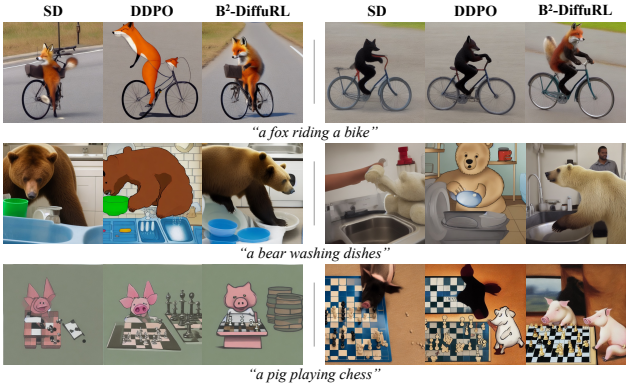


Figure 1. **(Prompt-image Misalignment)** Text-to-image diffusion models (*e.g.*, Stable Diffusion (SD) [52]) may not generate high-quality images that accurately align with prompts. Existing reinforcement learning-based diffusion model fine-tuning methods (*e.g.*, DDPO [8]) have limited effect and loss of image diversity. For each set of images above, we use the same seed for sampling.

To solve this problem, recent studies have explored incorporating reinforcement learning (RL) techniques to fine-tune pre-trained text-to-image diffusion models [8, 17, 32, 46, 63, 65]. By formulating the step-by-step denoising process as a *sequential decision-making problem*, RL enables diffusion models to optimize for specific long-term objectives, beyond merely fitting to static data as done in standard supervised learning [29, 52, 64]. In this formulation, noisy images at different timesteps are viewed as *states* in RL, while denoising at each timestep corresponds to an *action*.

*Equal contribution. †Corresponding author.
[1]https://github.com/hu-zijing/B2-DiffuRL.

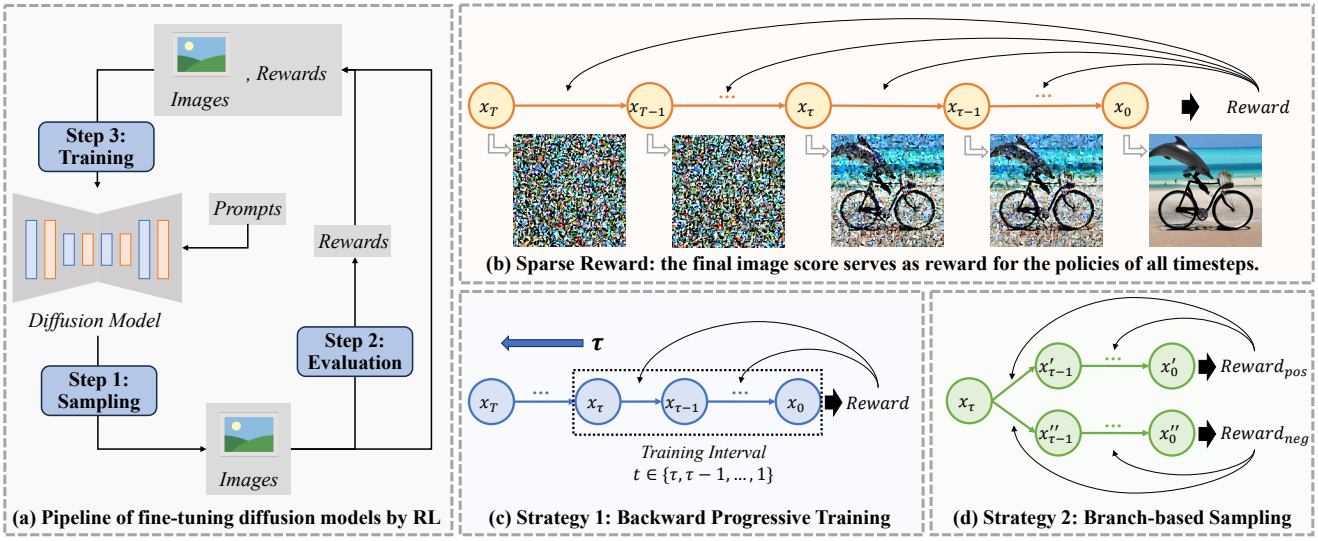


Figure 2. (**Sparse Reward**) When people train diffusion models with reinforcement learning (RL), the reward is only available at the end of the generation process. This sparsity limits the success of RL in diffusion models. We propose B^2-DiffuRL, a new RL framework with two strategies, to mitigate this issue.

The alignment scores between the final generated images and the textual prompts, which can be derived from human preferences or model evaluations, serve as the *rewards*. The pipeline of training diffusion models with RL is illustrated in Figure 2 (a). Researchers first sample images using the diffusion model with given prompts and then calculate the alignment scores as rewards. These sampled trajectories, consisting of images at different timesteps and their corresponding alignment scores, can be used as training data for RL to further enhance the diffusion models [25].

However, RL has so far made limited success in improving prompt-image alignment, primarily due to the key challenge of ***sparse reward***. As shown in Figure 2 (b), reward in this context is sparse because it is only available at the end of the generation process. Sparse rewards are harmful to RL-based diffusion fine-tuning in two ways:

- ***Limited improvement in alignment***. The denoising actions at different timesteps focus on varying levels of semantics (*e.g.*, early timesteps define layout, middle timesteps refine style, and late timesteps enhance detailed objects) and have different impacts on the final image [68, 69]. With sparse rewards, it is difficult to identify which actions during the denoising process contribute positively to the final alignment, so actions at different timesteps receive inappropriate rewards. As a result, learning effective policies becomes challenging.
- ***Sacrificing diversity for better alignment***. To achieve higher alignment score, the model may learn unnecessary policies. For example, with prompts like "*a bear washing dishes*", cartoon-like images are more likely to get higher rewards than realistic photographs because the prompts are often depicted in a cartoon style in pre-training data. With sparse rewards, model fine-tuned via naive RL algorithms (*e.g.*, DDPO [8]) may learn these unnecessary policies about styles, resulting in generating only cartoon-like images, as shown in Figure 1. This shows a trade-off between alignment and diversity, where alignment is improved at the expense of diversity [54, 71].

The challenge of sparse reward has attracted widespread attention in traditional RL [20, 62]. The classic solutions are constructing additional rewards by various techniques, such as reward shaping [40, 50], to achieve dense reward functions [13, 19, 27, 43]. Unfortunately, these solutions are not suitable for diffusion models because it is hard to evaluate the noisy images in the denoising process. This motivates us to ask: *How can we mitigate the negative effects of sparse rewards when using RL to train diffusion models*?

In this paper, we introduce a novel RL-based fine-tuning framework for diffusion-based text-to-image generation to address the challenge of sparse reward, which we refer to as **B^2-DiffuRL** [2]. Our framework employs two strategies. The first one is ***backward progressive training*** (BPT), applied to the training stage. Initially, we focus training on only the final timesteps of the image generation process, as shown in Figure 2 (c). As training rounds increase, we gradually extend the training interval backward to cover all timesteps, and achieve training on the entire denoising process in the end. The second strategy is ***branch-based sampling*** (BS), applied to the sampling stage. For each training interval in denosing process, we perform branch sampling to get multiple samples under each branch, as shown in Figure 2 (d). Within each branch, we only select the best and worst samples to form a contrastive sample pair for RL training.

[2] B^2-DiffuRL is short for **B**ackward progressive training and **B**ranch-based sampling for **R**einforcement **L**earning in **Diffu**sion models.

Our framework has the following three capabilities: (1) ***Better prompt-image alignment***. With small training interval, BPT strategy enables the models to easily and quickly learn the policies for the later timesteps of generation. As the model becomes proficient in these later timesteps, it progressively learns to manage the earlier timesteps of the denoising process. By mitigating the complexity of dealing with the entire process from the outset, BPT reduces the learning difficulty associated with sparse rewards. Moreover, with BS strategy, the contrastive samples within the same branch share identical states and actions up to the start of the training interval. By comparing the contrastive samples, the models can accurately identify how much the denosing policies of the current training interval contribute to the final image during training. (2) ***Maintaining diversity when improving alignment***. Denoised from the same intermediate state, the contrastive samples share similar coarse-grained visual information (*e.g.*, image styles) but receive different rewards. It prevents the models from learning unnecessary policies (*e.g.*, about image styles) as shortcuts to achieve high rewards, thus helping maintain diversity. (3) ***Compatibility***. Although we mainly compare with the current state-of-the-art RL-based fine-tuning algorithm called DDPO [8] in this paper, our framework is compatible with any previous optimization algorithm such as policy gradient [57], DPO [48, 63] and DPOK [17]. Experiments show that applying B^2-DiffuRL can improve effectiveness of different algorithms in terms of both alignment and diversity.

Our contributions can be summarized as: (1) We investigate the problem of RL-based diffusion models fine-tuning for improving prompt-image alignment, and for the first time highlight the challenge of sparse reward. (2) We propose a compatible RL-based fine-tuning framework named B^2-DiffuRL, employing backward progressive training and branch-based sampling strategies, to address the above challenge. (3) Extensive experimental results on Stable Diffusion [52] show the effectiveness of B^2-DiffuRL in terms of both alignment and diversity when compatible with different RL algorithms, without increasing computational cost.

2. Related Work

2.1. Text-to-Image Diffusion Models

Diffusion models have gained substantial attention for their ability to generate high-quality samples [23, 59, 60, 67]. One of the primary applications of diffusion models is image generation [5, 24]. These models have been shown to produce images that are both high in fidelity and diversity, rivaling the outputs of Generative Adversarial Networks (GANs) [15, 18]. The extension of diffusion models to text-to-image generation has opened up possibilities for creating images from textual descriptions [70]. Works like DALL-E [49] and Imagen [55] have demonstrated that diffusion models can be effectively conditioned on textual input to produce corresponding images. Despite their success, text-to-image diffusion models often suffer from the issue of prompt-image misalignment [31, 44].

2.2. Reinforcement Learning with Sparse Reward

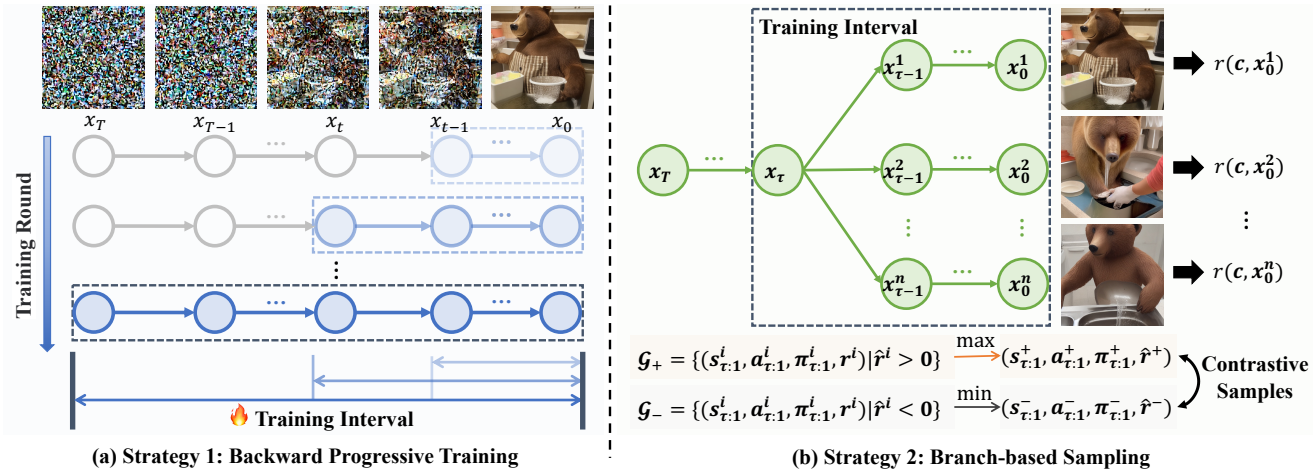
Reinforcement Learning (RL) is a learning paradigm in which an agent learns to make decisions by interacting with an environment to maximize cumulative rewards [29, 45]. Applications of RL span various domains, including gaming, robotics, finance, and healthcare [11, 39]. Recently, RL has played an important role in alignment. For example, RL has been leveraged to fine-tune large language models (LLMs), ensuring that the generated outputs align with human values and intentions [9]. One of the significant challenges in RL is dealing with sparse rewards, where feedback signals are infrequent and the agent must explore extensively to discover rewarding states [50, 62]. Traditional RL algorithms struggle in such settings due to the inefficiency in learning from limited feedback [20, 40]. Various techniques have been proposed to address this challenge [3, 42], such as reward shaping [13, 19, 27, 43], where additional heuristic rewards are provided to guide the agent. However, these classic RL strategies can not be applied to our problem directly, since it is difficult to evaluate the noisy images during denoising process.

2.3. Improving Alignment of Diffusion Models

Early diffusion models focused primarily on the quality and fidelity of the generated images [15, 23, 59]. However, as the demand for a more interactive and user-driven generation grew, improving alignment between prompts and generated images is crucial for enhancing the usability and reliability of these models in practical applications [16, 34, 53, 72]. The initial approaches to conditioning diffusion models on text prompts employ a variety of techniques, including both classifier guidance [15] and classifier-free guidance [22]. With the advent of LDMs [51], subsequent researches focus on fine-tuning pre-trained models to enhance alignment [26, 33]. Recently, RL has been employed to fine-tune the text-to-image diffusion models [8, 10, 17, 32, 46, 63, 65, 66]. However, the issue of sparse rewards limits the performance of such methods in prompt-image alignment, and even sacrifices a lot of diversity in order to improve controllability. In this paper, by mitigating the negative effects of sparse rewards, we further develop the application of RL in training diffusion models.

3. Method

In this section, we first introduce how to train diffusion models with RL. Then we highlight the challenge of sparse reward in this context. Finally, we introduce



(a) Strategy 1: Backward Progressive Training

(b) Strategy 2: Branch-based Sampling

Figure 3. (**Method**) We propose the framework B^2-DiffuRL, employing two strategies to address the challenge of sparse rewards. (a) Backward progressive training strategy: We focus initially on the final timesteps of the denoising process and gradually extend the training interval to earlier timesteps, easing the learning difficulty associated with sparse rewards. (b) Branch-based sampling strategy: We perform branch-based sampling at the beginning of each training interval. Comparisons between samples within the same branch provide a clear indication of whether the policies of the current training interval positively contribute to the final images.

B^2-DiffuRL, employing two strategies to address this challenge. B^2-DiffuRL can be compatible with different RL algorithms, such as DDPO [8], DPO [63] and DPOK [17].

3.1. Problem and Challenge

Text-to-Image Diffusion Models. Text-to-image diffusion models iteratively refine random noise into a coherent image that matches the given prompt [52]. The process of diffusion models consists of two phases: the forward process and the reverse process [23]. In the forward process, an image $\mathbf{x}_0$ is gradually corrupted into pure noise $\mathbf{x}_T$ through T steps, where Gaussian noise is added at each step. The reverse process aims to generate an image from pure noise conditioned on a textual description $\mathbf{c}$ by denoising iteratively [23, 58]:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathbf{c}), \sigma_t \mathbf{I}^2), \quad (1)$$

where μ_θ is predicted by a diffusion model parameterized by θ, and σ_t is the fixed timestep-dependent variance.

Training Diffusion Models with RL. The denoising process of diffusion models can be formulated as a *sequential decision-making problem*. Therefore, this process can be viewed through the lens of RL, where each step in the denoising process is considered as a decision made by an agent (the diffusion model). Following this formulation, the *state* s_t at each timestep is represented by $(\mathbf{c}, t, \mathbf{x}_t)$, *i.e.*, the text prompt, the current timestep, and the noisy image at the current timestep. The sequence of states represents the gradual refinement from noise to the final image. The *action* a_t at each timestep involves denoising by sampling the next noisy image $\mathbf{x}_{t-1}$. The *policy* π_θ, parameterized by θ, defines the action selection strategy. In this context, the policy is defined as $\pi_\theta(a_t \mid s_t) = p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$. The *reward* can be defined as a prompt-image alignment score $r(\mathbf{c}, \mathbf{x}_0) \in \mathbb{R}$, which is given by human preferences or model evaluations. A larger reward means a better prompt-image alignment. To improve the prompt-image alignment of diffusion models, we can execute RL-based training by maximizing the following objective:

$$\mathcal{J}_{\mathrm{RL}}(\theta) = \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}), \mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{c})} \left[r(\mathbf{x}_0, \mathbf{c}) \right], \quad (2)$$

where $p(\mathbf{c})$ follows a uniform distribution, meaning that we randomly sample prompts from a candidate set of prompts. To construct the training data for RL, we first collect denoising trajectories via sampling based on diffusion models. Then we can update parameters θ via gradient descent [41].

Challenge of Sparse Reward. However, the reward $r(\mathbf{x}_0, \mathbf{c})$ is only available at the end of the image generation process. This sparsity of reward makes it challenging for the diffusion model to identify which actions during the denoising process positively impact the final alignment and reward them appropriately. As a result, the diffusion model struggles to learn effective strategies and may even adopt unnecessary or incorrect ones. The classic RL strategies, such as constructing additional rewards, are not suitable here because it is difficult to evaluate the noisy images during the denoising process. This motivates us to develop new RL strategies for training diffusion models to mitigate the negative effects of sparse rewards. For a comprehensive discussion on the challenge of sparse reward, we refer the readers to Appendix C.

3.2. Strategy 1: Backward Progressive Training

The conventional training methods involve training the model across all timesteps of the denoising process from the beginning. However, due to the complexity and large noise present in the early timesteps, the training process can be unstable and inefficient, especially with sparse rewards. We hypothesize that focusing on the final timesteps, where the generated images are more coherent and less noisy, could provide a more stable foundation for the RL training. By mastering these final timesteps first, the model can incrementally handle the earlier, noisier stages more effectively, leading to overall better performance and control. We call this strategy as backward progressive training (BPT). Formally, let T represent the total number of timesteps in the denoising process. Initially, we train the model on the last τ timesteps, where $\tau < T$. Therefore, each trajectory sampled for training consists of τ timesteps:

$$\{s_t, a_t, \pi_\theta(a_t \mid s_t) | t = \tau, \tau-1, ..., 1\} \text{ with reward } r(\mathbf{x}_0, \mathbf{c}), \tag{3}$$

which can be abbreviated as $(s_{\tau:1}, a_{\tau:1}, \pi_{\tau:1}, r)$ without ambiguity. As training progresses, the training interval is extended backward by incorporating more timesteps, ultimately covering the entire range from T to 1. The training objective during each phase remains consistent with Eq. (2). Following DDPO, we use policy gradient estimation [30, 57] and the gradient is:

$$\nabla_\theta \mathcal{J}_{\text{BPT}} = -\mathbb{E}\left[\sum_{t=1}^{\tau} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})}{p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})\, \hat{r}(\mathbf{x}_0, \mathbf{c})\right], \tag{4}$$

where θ_{old} is the parameters of diffusion model prior to update and $\hat{r}$ is the normalized value of reward r (see Appendix D.1 for details). The expectation is taken over sampled denoising trajectories.

Previous works fine-tune diffusion models along the entire denoising process from x_T to x_0, with the sparse reward r_0. With such sparse reward, it is difficult for the model to directly learn effective network parameters for the entire denoising process. We propose BPT to make the model learn the denoising process from x_τ to x_0 first. As training progresses, τ is gradually increased to T, and the model learns to manage the earlier timesteps after becoming proficient in later timesteps. This is easier than directly learning the entire denoising process. By applying BPT, the model can more effectively learn how to denoise when only x_0, state at the last timestep, has a reward. We refer the readers to Appendix C for a comprehensive discussion.

3.3. Strategy 2: Branch-based Sampling

The sparse rewards make it difficult to tell whether actions on certain timesteps during denoising have a positive or negative effect on the final alignment. To further mitigate this issue, we introduce the strategy of branch-based sampling (BS). When constructing training data for RL, we perform branch sampling at the beginning of training interval $[\tau, 1]$, as shown in Figure 3 (b). Within each branch, we divide the sampled denoising trajectories (distinguished by the superscript i) into two groups:

$$\begin{aligned} \mathcal{G}_+ &= \left\{ \left(s^i_{\tau:1}, a^i_{\tau:1}, \pi^i_{\tau:1}, \hat{r}^i\right) \middle| \hat{r}^i := \hat{r}(\mathbf{x}^i_0, \mathbf{c}) > 0 \right\}, \\ \mathcal{G}_- &= \left\{ \left(s^i_{\tau:1}, a^i_{\tau:1}, \pi^i_{\tau:1}, \hat{r}^i\right) \middle| \hat{r}^i := \hat{r}(\mathbf{x}^i_0, \mathbf{c}) < 0 \right\}, \end{aligned} \tag{5}$$

where group $\mathcal{G}_+$ consists of trajectories with positive rewards (if available), and group $\mathcal{G}_-$ consists of trajectories with negative rewards (if available). We then select the trajectory $(s^+_{\tau:1}, a^+_{\tau:1}, \pi^+_{\tau:1}, \hat{r}^+)$ with the best reward from the positive group and the trajectory $(s^-_{\tau:1}, a^-_{\tau:1}, \pi^-_{\tau:1}, \hat{r}^-)$ with the worst reward from the negative group to form a contrastive sample pair for RL. The gradient of the contrastive sample pair is:

$$\begin{aligned} \nabla_\theta \mathcal{J}_{\text{BS}} = -\mathbb{E}\Bigg(\sum_{t=1}^{\tau} \Bigg[& \frac{p_\theta(\mathbf{x}^+_{t-1} \mid \mathbf{x}^+_t, \mathbf{c})}{p_{\theta_{\text{old}}}(\mathbf{x}^+_{t-1} \mid \mathbf{x}^+_t, \mathbf{c})} \nabla_\theta \log p_\theta(\mathbf{x}^+_{t-1} \mid \mathbf{x}^+_t, \mathbf{c}) \hat{r}^+ \\ & + \frac{p_\theta(\mathbf{x}^-_{t-1} \mid \mathbf{x}^-_t, \mathbf{c})}{p_{\theta_{\text{old}}}(\mathbf{x}^-_{t-1} \mid \mathbf{x}^-_t, \mathbf{c})} \nabla_\theta \log p_\theta(\mathbf{x}^-_{t-1} \mid \mathbf{x}^-_t, \mathbf{c}) \hat{r}^- \Bigg] \Bigg). \end{aligned} \tag{6}$$

By isolating the impact of actions outside the training interval on the final images, the comparison between the contrastive samples directly reflects how much the actions within the training interval contribute to the reward. Branch-based sampling strategy provides clear signals to the model, allowing the model to focus on actions that truly drive positive outcomes. Therefore, it further mitigates the impact of reward sparsity and facilitates more efficient learning of effective policies. Moreover, by avoiding learning unnecessary policies (*e.g.*, image styles), our approach can also maintain the diversity of generated images, which will be demonstrated and discussed in the following section. We emphasize that B^2-DiffuRL does not increase computational cost of RL algorithms, as discussed in Appendix D.3.

4. Experiments

In this section, we evaluate the effectiveness of B^2-DiffuRL in terms of improving prompt-image alignment and maintaining diversity. We first compare our method with existing state-of-the-art method DDPO [8]. Then, we focus on ablation studies on the proposed two strategies, as well as the compatibility and generalization ability. For simplicity, we refer to B^2-DiffuRL as ours in this section.

4.1. Experimental Setup

Diffusion Models. Following the previous work [8], we use Stable Diffusion (SD) v1.4 as the backbone diffusion model, which has been widely used in academia and industry. We apply LoRA to UNet for efficient fine-tuning [26]. We employ DDIM [58] algorithm for sampling. Following

Figure 4. **(Samples)** Examples of images generated by different methods on three templates. For each set of images, we use the same random seed. Our method achieves better prompt-image alignment compared to vanilla Stable Diffusion and DDPO.

the previous work [8], we set the total denoising timesteps $T = 20$. The weight of noise is set to 1.0, which decides the degree of randomness of each denoising in DDIM. Each experiment is conducted with three different seeds.

Prompt Templates. In the sampling phase, we construct the prompts based on three different templates. The three prompt templates consider the behavior of the object, the attribute of the object, and the positional relationship between the objects in turn, which we believe can cover a wide range of commonly used prompts in image generation. (1) Template 1:"*a(n) [animal] [activity]*". We use this template designed by DDPO. The animal is chosen from the list of 45 common animals, and randomly matched with an activity from the list:"*riding a bike*", "*playing chess*" and "*washing dishes*". (2) Template 2: "*[color] [fruit/vegetable]*". This template focuses on object attributes. To construct a list of color-fruit/vegetable combinations, we query GPT-4 [1] about fruits/vegetables' names and their common colors. We require each item to have at least 3 colors, and we end up building 40 prompts for this template. (3) Template 3: "*[object 1] [predicate] [object 2]*". The predicates refer to positional relationship. We construct the prompts based on the annotations of Visual Relation Dataset [38]. We choose four predicates: "*on*", "*under*", "*on the left of*", and "*on the right of*", and end up with 40 prompts for this template. The prompts mentioned above are only used for training. In order to evaluate the generalization ability, we further construct prompts that will not be used in training. The full prompt lists are shown in the Appendix H.

Rewards. We score the prompt-image alignment by BERTScore and CLIPScore, and use them as reward functions: (1) BERTScore is introduced by DDPO [8], in which one uses the visual language model, such as LLaVA [36], to generate a description of the image, and then uses BERT's recall metric [14] to measure the semantic similarity between the prompt and the description. (2) CLIPScore is simply the similarity between text embedding and image embedding measured by CLIP model [7, 47]. We recommend using CLIPScore as reward function due to the instability of BERTScore, as shown in Appendix F.1. For implementations, we use 7b half-precision LLaVA v1.5 model [35], DeBERTa xlarge model [21] (a variant of BERT model), and ViT-H-14 CLIP model [47], respectively. To improve the stability of training, we normalize the rewards, as described in detail in Appendix D.1.

Evaluation Metrics. In this paper, we focus on both prompt-image alignment and image diversity. For alignment, we use BERTScore [8] and CLIPScore [7, 47] as metrics, the same as reward functions. A higher BERTScore or CLIPScore represents better prompt-image alignment. For diversity, following previous works [2, 4, 7, 73], we use inception score (IS) as the metric. A higher inception score represents better image diversity.

4.2. Qualitative Evaluation

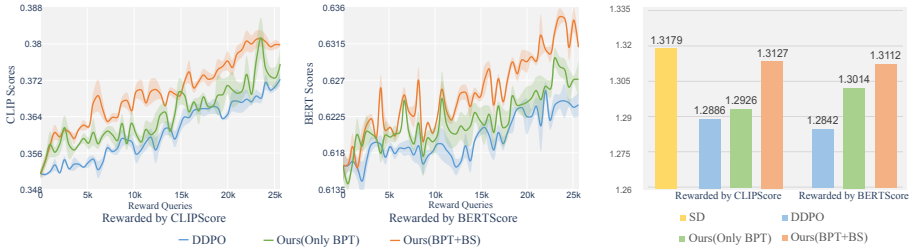
We first evaluate the performance of our method and DDPO on the three prompt templates rewarded by CLIPScore. We use our method and DDPO respectively to fine-tune the diffusion model. After the same round of training, we sample some images from original model and fine-tuned models, as shown in Figure 4. The results qualitatively show that our method performs better than DDPO in improving the prompt-image alignment. We also conduct human preference test over 80 independent human raters (from undergrad to Ph.D.), who are asked to pick the best fit to prompt among three images generated by different models. As shown in Figure 7, the images generated by our method get higher preference rates than original SD and DDPO on all the three

(a) Template 1: a(n) [animal] [activity] (b) Template 2: [color] [fruit/vegetable] (c) Template 3: [object1] [predicate] [object2]

Figure 5. **(Alignment)** Alignment curves of our method and DDPO on three prompt templates.

Methods	Temp. 1	Temp. 2	Temp. 3
SD	1.3179	1.4133	1.3582
DDPO	1.2886	1.3323	1.3273
Ours	**1.3127**	**1.3579**	**1.3348**

Table 1. **(Diversity)** IS ↑ of images generated by the SD [52], DDPO [8], and ours on three templates. There is a trade-off between alignment and diversity, while our method helps maintain diversity.



(a) Alignment (measured by CLIPScore ↑ and BERTScore ↑) (b) Diversity (measured by IS ↑)

Figure 6. **(Ablation Study)** We separately evaluate the impact of each proposed strategy on prompt-image alignment and image diversity with template 1. (a) Both BPT and BS strategies help improve prompt-image alignment. (b) BS strategy also helps to maintain image diversity.

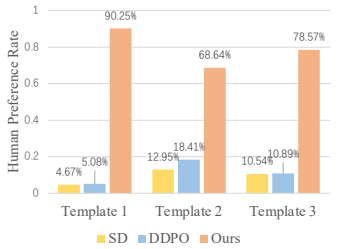


Figure 7. **(Human Evaluation)** Human preference rates for prompt-image alignment of images generated by SD, DDPO and our method.

prompt templates. Also, images by our method are more diverse than those by DDPO. For example, on template 1, all images by DDPO adopt a cartoon style, while those by ours keep original styles of SD; on templates 2 and 3, backgrounds of the images by DDPO tend to reduce to a single color, while those by ours do not. This can be seen more clearly from Appendix G.

4.3. Quantitative Evaluation

We compare our method with DDPO quantitatively in terms of prompt-image alignment and diversity.

Prompt-Image Alignment. Figure 5 shows the curve of CLIPScore when fine-tuning the diffusion models using our method and DDPO as the amount of reward queries increases. We can observe that our method almost always achieves higher CLIPScore during fine-tuning on all the three prompts. This shows that our approach can improve prompt-image alignment better with the same number of reward queries compared to DDPO, which is due to our proposed two strategies.

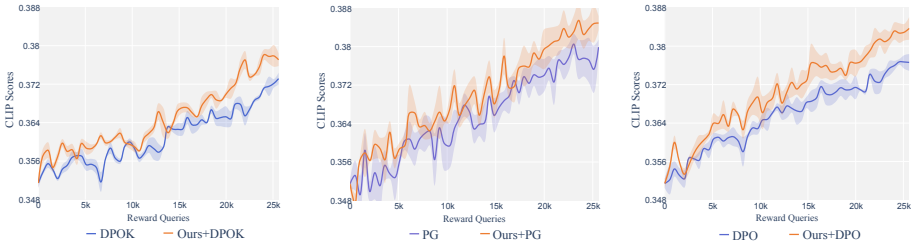
Image Diversity. We evaluate the diversity of the images generated by original SD and the models fine-tuned by our method and DDPO. The results are shown in Table 1. After 25.6k reward queries during fine-tuning, both the models trained by ours and DDPO exhibit a reduction in diversity, since there is an inherent trade-off between alignment and diversity [71]. However, we find that the models trained by our method have a smaller reduction in diversity on all templates. For example, on template 1, the diversity of the model trained by our method decreases much less than that of DDPO, and is basically the same as the original model. Overall, our method can mitigate the reduction in image diversity during RL-based diffusion model fine-tuning.

4.4. Ablation Study

We separately evaluate the impact of each proposed strategy on alignment and diversity respectively.

Ablation Study on BPT Strategy. To evaluate the effectiveness of BPT, we fine-tune Stable Diffusion with only BPT strategy, rewarded by CLIPScore and BERTScore respectively. As shown in Figure 6 (a), regardless of the reward function, our proposed BPT strategy outperforms DDPO in terms of alignment. As we previously analyzed, BPT simplifies learning by training in stages, alleviating the negative effects of sparse rewards, and thus improving alignment. Moreover, since we only train models on timesteps of current training interval instead of all the denoising process, the computation costs of our method are less than DDPO for each queried reward.

Ablation Study on BS Strategy. The effectiveness of BS strategy on prompt-image alignment is shown in Figure 6 (a). We can observe that, based on BPT, the BS strategy further improves alignment in terms of both BERTScore and CLIPScore. By comparing contrastive

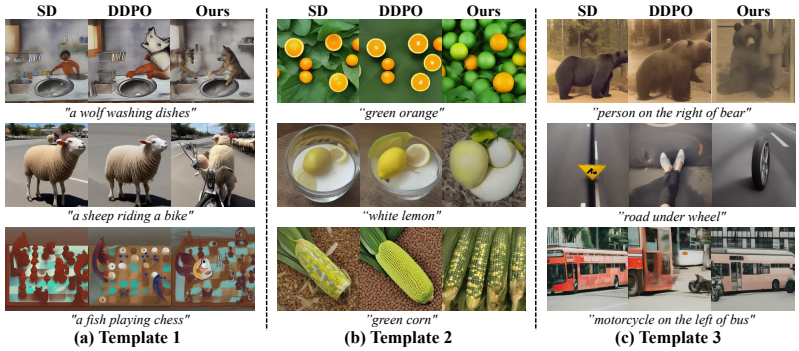


(a) DPOK and our method (b) PG and our method (c) DPO and our method

Figure 8. **(Compatibility: Alignment)** Alignment curves of our method on template 1 when compatible with difference RL algorithms.

Methods	Vanilla	Ours+Vanilla
SD	1.3179	-
DPOK	1.2785	**1.3005**
PG	1.2462	**1.2896**
DPO	1.2895	**1.3051**

Table 2. **(Compatibility: Diversity)** IS ↑ of images on template 1 generated by our method when compatible with different RL algorithms.



(a) Template 1 **(b) Template 2** **(c) Template 3**

Figure 9. **(Generalization)** Examples of images generated by SD, DDPO and ours on three templates. The prompts are not used in training. We use the same seed for sampling.

Methods	Temp. 1	Temp. 2	Temp. 3
SD	0.3515	0.3168	0.2977
DDPO	0.3698	0.3175	0.3134
Ours	**0.3748**	**0.3252**	**0.3183**

Table 3. **(Generalization)** Prompt-image alignment (measured by CLIPScore ↑) of the generated images by SD, DDPO and our method on the prompts based on three templates. The prompts are not used during the training process.

samples, BS provides a clear indication of how much the policies of current training interval contribute to final images. This helps the model to learn effective policies. Additionally, as shown in Figure 6 (b), diversity of generated images always suffers from reduction since there is a trade-off between alignment and diversity. Fortunately, the BS strategy helps models avoid learning unnecessary policies, thereby contributing to maintaining image diversity. With BS strategy, diversity of the fine-tuned model decreases less, and even achieves similar diversity as the original SD.

4.5. Compatibility

Our framework B^2-DiffuRL is compatible with various RL algorithms, not limited to DDPO. We further apply B^2-DiffuRL to some widely used RL algorithms in diffusion model fine-tuning, including DPOK [17], policy gradient (PG) [57] and direct preference optimization (DPO) [48, 63]. The implementation details are shown in Appendix D.1. On the one hand, as we can see from Figure 8, when compatible with different RL algorithms, our method can help each of them to improve alignment to a greater extent. On the other hand, as shown in Table 2, while all algorithms reduce the diversity of generated images, our method can help mitigate the reduction. These experimental results further illustrate the effectiveness of our method in terms of both prompt-image alignment and diversity when applied to various RL algorithms.

4.6. Generalization Ability

Models fine-tuned by our method show generalization capabilities. We generate 1,600 images on the prompts based on the corresponding templates but not belong to the training lists, and test the prompt-image alignment on CLIPScore. As shown in Table 3, compared with DDPO, the models fine-tuned with our method also perform better on these prompts not used for training. Figure 9 shows examples of images generated on these prompts, qualitatively illustrating the good generalization ability of the models fine-tuned with our method. More samples can be seen in Appendix G.

5. Conclusions

In this work, we mitigated the issues of prompt-image misalignment in text-to-image diffusion models by reinforcement learning (RL). We highlight the challenge of sparse reward when training diffusion models with RL. By introducing a compatible RL-based fine-tuning framework B^2-DiffuRL that leverages backward progressive training and branch-based sampling strategies, we effectively mitigated the negative effects of sparse reward. Using Stable Diffusion as backbone, we performed extensive experiments with various kinds of text prompts. Both qualitative and quantitative experimental results demonstrate that, compared with naive RL-based diffusion model training method, the proposed framework achieves better prompt-image alignment while sacrificing less image diversity.

Acknowledgement

This work was supported by the National Key Research & Development Project of China (2024YFB3312900), the National Natural Science Foundation of China (62376243, 62441605, 62037001), Zhejiang Provincial Natural Science Foundation of China (LD25F020001), Key R&D Program of Zhejiang (2025C01128), and the Starry Night Science Fund at Shanghai Institute for Advanced Study (Zhejiang University). All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer, 2019. 6, 17

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017. 3

[4] Shane Barratt and Rishi Sharma. A note on the inception score, 2018. 6, 17

[5] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 3

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[7] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. Renaissance: A survey into ai text-to-image generation in the era of large model, 2023. 6, 17

[8] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. 2023. 1, 2, 3, 4, 5, 6, 7, 17

[9] 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. 3

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2024. 3

[11] Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 17

[13] Rati Devidze, Parameswaran Kamalaruban, and Adish Singla. Exploration-guided reward shaping for reinforcement learning under sparse rewards. *Advances in Neural Information Processing Systems*, 35:5829–5842, 2022. 2, 3

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[15] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 1, 3

[16] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[17] 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, 8, 14

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3, 17

[19] Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. *Advances in Neural Information Processing Systems*, 35:15281–15295, 2022. 2, 3

[20] Joshua Hare. Dealing with sparse rewards in reinforcement learning. *arXiv preprint arXiv:1910.09281*, 2019. 2, 3

[21] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. 6

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 1, 3, 4

[24] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 3

[25] Steven C. H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey, 2018. 2

[26] 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, 2021. 3, 5

[27] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33:15931–15941, 2020. 2, 3

[28] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024. 1

[29] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996. 1, 3

[30] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002. 5

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[32] 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. 1, 3

[33] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility, 2024. 3

[34] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 3

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 6

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 6

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 16

[38] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 6

[39] Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE communications surveys & tutorials*, 21(4):3133–3174, 2019. 3

[40] Farzan Memarian, Wonjoon Goo, Rudolf Lioutikov, Scott Niekum, and Ufuk Topcu. Self-supervised online reward shaping in sparse-reward environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2369–2375. IEEE, 2021. 2, 3

[41] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *The Journal of Machine Learning Research*, 21(1): 5183–5244, 2020. 4

[42] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018. 3

[43] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, pages 278–287, 1999. 2, 3

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021. 3

[45] Athanasios S Polydoros and Lazaros Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2): 153–173, 2017. 3

[46] 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

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 6

[48] 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. 3, 8

[49] Aditya Ramesh, Mikhail Pavlov, Scott Gray Gabriel Goh, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 1, 3

[50] Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pages 4344–4353. PMLR, 2018. 2, 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 4, 7

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[54] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradely, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023. 2

[55] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed

Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[56] 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

[57] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3, 5, 8, 12

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 4, 5

[59] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in neural information processing systems*, 2020. 1, 3

[60] Fareena Sultan, John U Farley, and Donald R Lehmann. A meta-analysis of applications of diffusion models. *Journal of marketing research*, 27(1):70–77, 1990. 3

[61] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 17

[62] Alexander Trott, Stephan Zheng, Caiming Xiong, and Richard Socher. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[63] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. 1, 3, 4, 8

[64] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[65] 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

[66] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8941–8951, 2024. 3

[67] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 3

[68] Zhongqi Yue, Jiankun Wang, Qianru Sun, Lei Ji, Eric I Chang, Hanwang Zhang, et al. Exploring diffusion time-steps for unsupervised representation learning. *arXiv preprint arXiv:2401.11430*, 2024. 2

[69] Zhongqi Yue, Pan Zhou, Richang Hong, Hanwang Zhang, and Qianru Sun. Few-shot learner parameterization by diffusion time-steps. *arXiv preprint arXiv:2403.02649*, 2024. 2

[70] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 3

[71] Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmark. *arXiv preprint arXiv:2405.01719*, 2024. 2, 7

[72] 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

[73] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 6, 17