Training Diffusion Models Towards Diverse Image Generation with Reinforcement Learning

Zichen Miao[1], Jiang Wang[2], Ze Wang[1], Zhengyuan Yang[2], Lijuan Wang[2], Qiang Qiu[1], Zicheng Liu[3]
Purdue University[1] Microsoft Corporation[2] Advanced Micro Devices, Inc.[3]
{miaoz, zewang, qqiu}@purdue.edu {jiangwang, zhengyang, lijuanw}@microsoft.com zicheliu@amd.com

Abstract

Diffusion models have demonstrated unprecedented capabilities in image generation. Yet, they incorporate and amplify the data bias (e.g., gender, age) from the original training set, limiting the diversity of generated images. In this paper, we propose a diversity-oriented fine-tuning method using reinforcement learning (RL) for diffusion models under the guidance of an image-set-based reward function. Specifically, the proposed reward function, denoted as Diversity Reward, utilizes a set of generated images to evaluate the coverage of the current generative distribution w.r.t. the reference distribution, represented by a set of unbiased images. Built on top of the probabilistic method of distribution discrepancy estimation, Diversity Reward can measure the relative distribution gap with a small set of images efficiently. We further formulate the diffusion process as a multi-step decision-making problem (MDP) and apply policy gradient methods to fine-tune diffusion models by maximizing the Diversity Reward. The proposed rewards are validated on a post-sampling selection task, where a subset of the most diverse images are selected based on Diversity Reward values. We also show the effectiveness of our RL fine-tuning framework on enhancing the diversity of image generation with different types of diffusion models, including class-conditional models and text-conditional models, e.g., StableDiffusion.

1. Introduction

Generative modeling on different modalities has been an active research area [5, 8, 34]. Empowered by large-scale models and datasets, the diffusion model develops remarkable capabilities in modeling high-dimensional, complex distributions, which enables photo-realistic conditional image generation under various types of guidance, e.g., classifier guidance [8] and textual guidance [29].

However, diffusion models inherit and amplify data bias from the large-scale, uncurated training data [3], showing undesired behaviors under specific conditions. For instance, text-to-image diffusion models show significant gender bias for particular occupations. Given the text prompt, 'a photo of a programmer', StableDiffusion [29] generates almost all males, whereas female programmers are dramatically underrepresented, which is shown in Figure 1. This 'conditional mode collapse' also happens in class-conditional diffusion models [8], where generated images show high uniformity under high guidance scales.

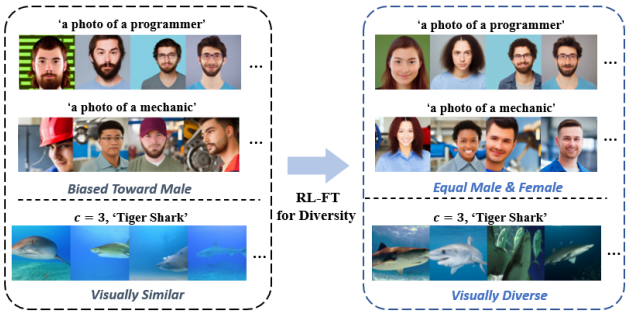


Figure 1. Illustration of non-diverse & biased images generated with both class-conditional and text-conditional diffusion models. Our diversity-oriented RL fine-tuning framework can tune both models for diverse & unbiased image generation.

To advance diverse class-conditional image generation, [27] retrains diffusion models with a re-weighted loss function, while [33] proposes a classifier-based diversity guidance. Meanwhile, both text prompt engineering [9, 26] and text prompt tuning [40] methods have been proposed for unbiased text-to-image generation. However, they all adopt task-specific designs that lack the generalizability to other types of diffusion models.

In this paper, we propose a general fine-tuning framework to further enhance the generation diversity of diffusion models under the guidance of a small set of unbiased reference images. Specifically, we design a class of reward functions, denoted as *Diversity Reward*, that estimate the coverage of the generative distribution w.r.t. the reference distribution, where each population is represented by a small set of images. To fine-tune diffusion models with *Diversity*

Reward, we apply reinforcement learning (RL) as a versatile optimization framework. Note that our RL fine-tuning framework does not depend on specific types of diffusion models and thus is generalizable.

Reward function design poses as the core of RL. Unlike ImageReward [39] adopted by [5, 10], where the reward is predicted by a reward model based on a single image, our *Diversity Reward* measures the overlap between the generative and the reference populations based on a set of samples. We explore several designs of *Diversity Reward*. First, we adopt the principal distribution distance measurement, Maximum Mean Discrepancy [13] (MMD), as one instantiation of *Diversity Reward*. It estimates the distribution distance with a few samples from each population, and we propose to use the negative MMD value to indicate the coverage of the current generative distribution. Moreover, formulating the groups of images as a Gaussian Process [38] in the feature space, we apply mutual information to measure the dependency between the generated and reference image random variables.

To provide more dense supervision to RL fine-tuning, we further estimate the contribution of each generated image among the group with the marginal utility [23]. In this way, we assign each image a distinct reward value. By maximizing the proposed *Diversity Reward*, RL fine-tuning expands the model coverage and mitigates the discrepancy between the generative and reference distributions.

We first validate the effectiveness of *Diversity Reward* with a post-sampling selection task, where we try to select the most diverse subsets with the proposed reward function. We then evaluate our diversity-oriented RL fine-tuning framework on both class-conditional and text-conditional diffusion models.

We summarize our contribution as follows:

- We propose theoretically supported *Diversity Reward* for estimating the diversity of diffusion models with small sets of generated and reference images.
- We propose to use the marginal utility to estimate the reward for each image from *Diversity Reward*.
- We propose a general RL fine-tuning framework equipped with *Diversity Reward* to further enhance the generative diversity of diffusion models.

2. Related Works

2.1. Diffusion Models

Diffusion models [14, 34] gradually transform multi-mode distributions into the standard Gaussian. By learning each step and constructing reverse processes, they show state-of-the-art performances in image synthesis [8, 14, 29]. Among them, conditional diffusion models [8, 29] have shown the ability to generate photo-realistic images with high correspondence to given conditions. Specifically, [8] equips diffusion models with an additional classifier, leading to high-fidelity class-conditional image generation. While text-conditional diffusion models generate images based on text prompts [2, 24, 30, 41]. For instance, StableDiffusion, built on top of latent diffusion models(LDM) [29], augments the diffusion models with the text encoder of CLIP [28], enabling generating high-quality images aligned with the given text prompts. However, both show a lack of diversity in conditional image generation, especially under high guidance scales. We propose a diversity-oriented RL fine-tuning method to mitigate this issue.

2.2. Unbiased Image Generation

Mitigating bias in deep models has been explored intensively in discriminative models [36, 37]. Yet, there are limited works on diverse generative modeling [7, 19], where most of them focus on GAN-based models. Several recent works have been proposed for learning debiased diffusion models. [33] propose hardness score guidance to sample images from low-density regions with class-conditional diffusion models, while [27] adopts a re-weighted loss to train diffusion models that generate data from long-tailed classes. For text-conditional diffusion models, [3] proposed to diversify model outputs by ethical intervention, and [9] proposed to directly add attribute words to the prompt. Yet, these prompt engineering methods have limitations such as being ambiguous, and not always generating diverse images reliably. [40] apply prompt tuning to learn inclusive prompts with a small set of unbiased images, yet it needs to manually specify the diversity ratio during generation. Different from these works, our RL fine-tuning framework works with both class and text conditional diffusion models that automatically fit the diversity ratio in the reference images.

2.3. RL Fine-tuning Diffusion Models

Inspired by tuning large language models with reinforcement learning (RL) [25], several works have been proposed to formulate the diffusion model as a multi-step decision-making problem and apply RL methods for various objectives. [10] introduces a policy gradient method for training diffusion models for a more efficient sampler. [5, 11] build upon [10, 21] to better align text-to-image diffusion models to human preferences using a policy gradient algorithm, where they both adopt a single-image-based reward function [39]. Different from these works, we first propose an image-set-based reward, i.e., *Diversity Reward*, that measures the gap between generative and reference distribution efficiently. It further enables the diversity-oriented RL fine-tuning for diffusion models.

3. Preliminary

3.1. Diffusion Models

As proposed in [14, 34], diffusion models diffuse the original data distribution on x_0 with a sequential forward

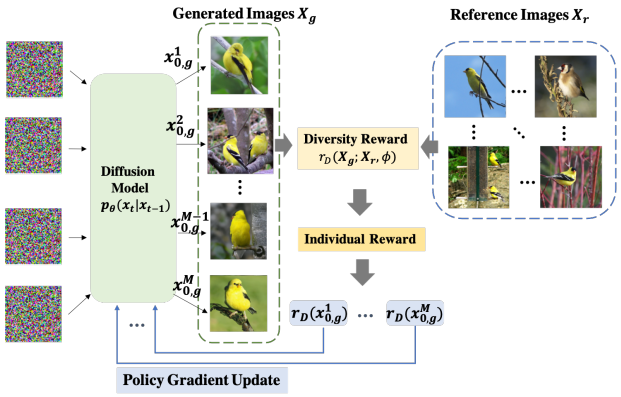


Figure 2. Illustration of the proposed general diversity-oriented RL fine-tuning. The proposed *Diversity Reward* uses a pair of sets of generated images and diverse reference images to assess the diversity of the current model. After estimating the individual reward for each image, we apply RL to maximize reward values. Our method has no model-specific design and is generalizable to a broad range of diffusion models.

Markov process in T total steps. By using a transition distribution $q(x_t|x_{t-1})$ in each step t, which iteratively adds noise to the data, diffusion models gradually destruct the original data distribution $p(x_0|c)$ (c denotes the condition) to standard Gaussian $p(x_T|c) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. To sample from random noise x_T, [14] parameterizes a reverse Markovian process with neural networks,

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

where θ is the parameters of the neural network. Starting from random noise x_T, diffusion models create a denoising trajectory $\{x_T, x_{T-1}, ..., x_0\}$, and obtains sample x_0 in the end.

The optimization of θ is performed by minimizing the KL divergence between $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$, which can be represented by the following objective,

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

where the posterior mean of the forward diffusion process $\hat{\mu}(x_0, t)$ is a linear combination of x_0 and x_t.

3.2. Fine-tuning Diffusion Model with Reinforcement Learning

Fine-tuning a pre-trained diffusion model with reinforcement learning (RL) can be achieved by optimizing the objective below,

$$L = \mathbb{E}_{c, x_0 \sim p_\theta(x_0|c)}[r(x_0, c)], \quad (3)$$

where $p_\theta(x_0|c)$ denotes the sampling distribution, and $r(x_0, c)$ denotes the reward defined over image x_0 and the condition c. [5, 10] formulate the reverse diffusion process as the following Markov decision process (MDP),

$$\begin{aligned} &s_t \triangleq \{x_t, c, t\}, \quad a_t \triangleq x_{t-1} \\ &\pi(a_t|s_t) \triangleq p_\theta(x_{t-1}|x_t, c), \\ &P(s_{t+1}|s_t, a_t) \triangleq \{\delta(x_{t-1}), \delta(c), \delta(t-1)\} \\ &R(s_t, a_t) \triangleq \mathbb{1}\{t=0\} \cdot r(x_0, c), \end{aligned} \quad (4)$$

where s_t, a_t, denote state, action, π is the policy, and $\delta(x)$ denotes the Dirac distribution where all probability mass concentrates on x. In this formulation, the diffusion model serves as the policy network and parameterizes the Markov transition kernel $P(s_{t+1}|s_t, a_t)$. The trajectory consists of T timesteps, ending with a sampled image x_0.

This formulation further enables the estimation of policy gradients. With access to both likelihood and gradients of likelihood, we follow the formulation in [5, 10], which makes direct Monte Carlo estimates of $\nabla_\theta L$, by sampling, and then perform parameter update. Specifically, we adopt the importance sampling estimator [16], and substitute the per-step return with the final reward $r(x_0, c)$,

$$\nabla_\theta L = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(x_{t-1}|x_t, c)}{p_{\theta'}(x_{t-1}|x_t, c)} \nabla_\theta \log p_\theta(x_{t-1}|x_t, c) r(x_0, c)\right] \quad (5)$$

in which the expectation is taken over the trajectories sampled with $p_{\theta'}$, i.e., the previous sampler. The estimator becomes less accurate if $p_{\theta'}$ deviates too much from p_θ. We adopt the trust region [31] for regularizing the change of θ w.r.t. θ', which in practice we adopt the clipping proposed in proximal policy optimization [32].

4. Method

In this section, we first describe the proposed *Diversity Reward*, which estimates the discrepancy between the generative and the reference distributions defined by a small set of diverse images. We then show how the proposed reward can be seamlessly incorporated into the diffusion model fine-tuning with reinforcement learning (RL) towards diverse image generation.

4.1. Diversity Reward

We first present our formulation and show several design choices for *Diversity Reward*. Current rewards for fine-tuning diffusion models are mostly single-image reward. For instance, ImageReward [39], proposed for enhancing text-image alignment, evaluates behaviors of the diffusion model with a single image. However, it is intractable to design a reward function based on a single image to evaluate the diversity. We propose to use a set of generated images to characterize the distribution and propose a reward function that measures the overlap between the current generative

distribution and the reference distribution using two sets of sampled images.

Problem Formulation. Given the generative distribution of the diffusion model $p_\theta(x_0|c)$ and the reference distribution $p_r(x_r)$, we propose the *Diversity Reward* to measure the overlap, which can also be represented as negative distance between p_θ and p_r, $-\mathcal{D}(p_\theta, p_r)$. MMD [13] is a principal method in probabilistic theory to estimate $\mathcal{D}(p_\theta, p_r)$ given a few samples from each population. Formally, we denote M generated images sampled from the diffusion model under condition c as $\mathbf{X_g} = \{x^1_{0,g}, x^2_{0,g}, ..., x^M_{0,g}\}$, where $x^m_{0,g} \overset{\text{i.i.d.}}{\sim} p_\theta(x_0|c)$, and N reference images as $\mathbf{X_r} = \{x^1_r, x^2_r, ..., x^N_r\}$ sampled from the reference distribution, $x^n_r \overset{\text{i.i.d.}}{\sim} p_r(x_r)$. MMD then estimates $\mathcal{D}(p_\theta, p_r)$ in the feature spaces of network ϕ with a measure d,

$$r_D(\mathbf{X_g}; \mathbf{X_r}, \phi) = -d(\mathbf{Z_g}, \mathbf{Z_r}) \approx -\mathcal{D}(p_\theta, p_r), \quad (6)$$

where $\mathbf{Z_g}, \mathbf{Z_r}$ denotes the transformed deep features of $\mathbf{X_g}, \mathbf{X_r}$ with model $\phi(\cdot)$, and $d(\cdot, \cdot)$ is a distance measure in the feature space. Specifically, $\mathbf{Z_g} = \{Z^1_g, Z^2_g, ..., Z^M_g\}$, $Z^m_g = \phi(x^m_{0,g})$, and $\mathbf{Z_r} = \{Z^1_r, Z^2_r, ..., Z^N_r\}$, $Z^n_r = \phi(x^n_r)$.

Alternatively, we can also view the *Diversity Reward* with a measure of the dependency between random variables, $x_{0,g}$andx_r, where we can naturally adopt mutual information as a proxy measurememt. As the relationship between image samples are better characterized in the feature space of ϕ, we compute the mutual information of $\mathbf{Z_g}, \mathbf{Z_r}$, and our *Diversity Reward* is defined as,

$$r_D(\mathbf{X_g}; \mathbf{X_r}, \phi) = I(\mathbf{Z_g}, \mathbf{Z_r}) \quad (7)$$

We also explore another non-differentiable design of the *Diversity Reward* based on the Recall metric [20], detailed in Appendix A.

4.1.1 Choices of ϕ

We propose two choices of $\phi(\cdot)$; one is the pre-trained InceptionV3 [35] for class-conditional diffusion models, and the other is the pre-trained visual encoder of CLIP [28] for text-to-image diffusion models.

4.1.2 Selection of Reference Images

We select $\mathbf{X_r}$ with a small set of diverse images in a task-specific manner. For instance, we randomly select around 50 images per class from the training set for class-conditional generation, and select a set of images with balanced attributes, e.g., 25 photos for people w/ & w/o eyeglasses.

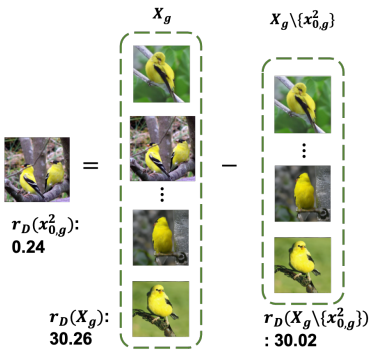


Figure 3. Illustration of the proposed individual reward, where the reward for a single image is estimated based on the *Diversity Reward* of two image sets.

4.1.3 Specific Designs of Diversity Reward

MMD. Maximum mean discrepancy (MMD) can be written as

$$\mathcal{D}_\mathcal{F}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y), \quad (8)$$

in which $\mathcal{F}$ denotes the witness function. We propose to exploit MMD as a *Diversity Reward* that characterizes the coverage of p_θ w.r.t. p_r. Specifically, with kernel function K_ϕ in the feature space of ϕ, we represent the MMD reward as,

$$\mathcal{D}_\mathcal{F}(p_\theta, p_r) = ||\mathbb{E}_{x^m_{0,g} \sim p_\theta} f^*(x^m_{0,g}) - \mathbb{E}_{x^n_r \sim p_r} f^*(x^n_r)||^2, \quad (9)$$

where the witness function is defined as,

$$f^*(x) = \mathbb{E}_{x^m_{0,g} \sim p_\theta} K_\phi(x^m_{0,g}, x) - \mathbb{E}_{x^n_r \sim p_r} K_\phi(x^n_r, x), \quad (10)$$

and $K_\phi(x, y) = K(\phi(x), \phi(y)) = K(Z_x, Z_y)$.

Note that with a small batch of generated and real samples, i.e., $\mathbf{X_g}, \mathbf{X_r}$, we adopt an unbiased estimation of $\mathcal{D}_\mathcal{F}(P_g, P_r)$ with $d(\mathbf{Z_g}, \mathbf{Z_r})$, which is the L^2 norm between the mean of kernel matrices. As the reward function is maximized in RL, we take the negative MMD distance as our MMD diversity reward,

$$r^{\text{MMD}}_D(\mathbf{X_g}; \mathbf{X_r}, \phi) = -\frac{1}{M(M-1)} \sum_{i \neq j}^{M} K_\phi(x^i_{0,g}, x^j_{0,g}) - \frac{1}{N(N-1)} \sum_{i \neq j}^{N} K_\phi(x^i_r, x^j_r) + \frac{2}{MN} \sum_{i=1}^{M} \sum_{i=1}^{N} K_\phi(x^i_{0,g}, x^j_r). \quad (11)$$

We use the polynomial kernel as the kernel function K.

Mutual Information. Mutual information of $\mathbf{Z_g}, \mathbf{Z_r}$ can be written as,

$$I(\mathbf{Z_g}, \mathbf{Z_r}) = H(\mathbf{Z_g}) + H(\mathbf{Z_r}) - H(\mathbf{Z_g}, \mathbf{Z_r}), \quad (12)$$

where $H(X)$ denotes the entropy of X. As $P(\mathbf{Z_g}), P(\mathbf{Z_r})$ are intractable, an approximation to $I(\mathbf{Z_g}, \mathbf{Z_r})$ with only a few samples is required. [4] have proposed an effective method to estimate mutual information with samples, yet an additional neural network is needed to be trained, which can induce extra computational and time costs. Instead, we propose to model $\mathbf{Z_g}, \mathbf{Z_r}$ with *Gaussian Process* (GP). By modeling the covariance (Cov) of $\mathbf{Z_g}, \mathbf{Z_r}$ using the kernel function K,

$$\text{Cov}(Z_g^m, Z_r^n) = \text{Cov}(m, n) = K(Z_g^m, Z_r^n), \tag{13}$$

we can represent the entropy of $H(Z)$ with,

$$H(Z) = \frac{1}{2}\ln|\text{Cov}(Z, Z)| + \frac{D}{2}\ln(1 + \frac{\pi}{2}), \tag{14}$$

where $|A|$ is the determinant of matrix A.

By removing constants, Equation 12 can be represented as

$$d(\mathbf{Z_g}, \mathbf{Z_r}) = \frac{1}{2}\ln|Cov(\mathbf{Z_g}, \mathbf{Z_g})| - \frac{1}{2}\ln|Cov(\mathbf{Z_{g,r}}, \mathbf{Z_{g,r}})|, \tag{15}$$

where $\mathbf{Z_{g,r}}$ is the concatenation of $\mathbf{Z_g}, \mathbf{Z_r}$. In terms of the kernel function K, we select the RBF kernel [38]. We denote this design of r_D as,

$$r_D^{\text{GP-MI}}(\mathbf{X_g}; \mathbf{X_r}, \phi) = \ln\frac{|Cov(\mathbf{Z_g}, \mathbf{Z_g})|}{|Cov(\mathbf{Z_{g,r}}, \mathbf{Z_{g,r}})|}. \tag{16}$$

4.2. Diversity-Oriented RL Fine-tuning

4.2.1 Individual Reward

Note that the diversity rewards we propose in the previous section are based on a set of generated images $\mathbf{X_g} = \{x_{0,g}^1, x_{0,g}^2, ..., x_{0,g}^M\}$. In other words, for all M chains of states and actions, the diffusion model, i.e., policy network, only receives a single reward. This sparse reward issue can cause unstable RL learning and sub-optimal results [1, 12, 32].

To generate dense reward values, i.e., assigning each member in $\mathbf{X_g}$ with a reward, we propose to adopt the *marginal utility* in submodular set cover problem [23] to estimate the individual contribution of $x_{0,g}^m$ to the overall diversity reward. Specifically, the marginal utility of e w.r.t. a subset S is expressed as,

$$F(e|S) = F(S \cup \{e\}) - F(S), \tag{17}$$

where F is a function defined over sets. It is required that F is monotone, i.e., $F(e|S) \geq 0, \forall e \in V \setminus S$ and $S \subset V$, where V is the whole set. As r_D^{MMD} and r_D^{MI} satisfy this condition, we propose individual rewards $\hat{r}_D$ as follows,

$$\begin{aligned}
&\hat{r}_D^{\text{GP-MI}}(x_{0,g}^m; \mathbf{X_r}, \phi) \\
&= r_D^{\text{GP-MI}}(\mathbf{X_g}; \mathbf{X_r}, \phi) - r_D^{\text{GP-MI}}(\mathbf{X_g} \setminus \{x_{0,g}^m\}; \mathbf{X_r}, \phi) \\
&\hat{r}_D^{\text{MMD}}(x_{0,g}^m; \mathbf{X_r}, \phi) \\
&= r_D^{\text{MMD}}(\mathbf{X_g}; \mathbf{X_r}, \phi) - r_D^{\text{MMD}}(\mathbf{X_g} \setminus \{x_{0,g}^m\}; \mathbf{X_r}, \phi)
\end{aligned} \tag{18}$$

4.2.2 RL fine-tuning with Diversity Rewards

Applying our proposed diversity rewards to RL fine-tuning introduced in Section 3.2, we now can tune diffusion models towards diverse conditional image generation. Specifically, the objective function used in RL fine-tuning for diversity (RLD) now becomes,

$$L_{RLD} = \mathbb{E}_{c, x_{0,g}^1, ..., x_{0,g}^M \sim p(x|c)}[\sum_m r^D(x_{0,g}^m)]. \tag{19}$$

Note that we compute the individual reward and have $r^D(x_{0,g}^m) = \hat{r}_D^{\text{GP-MI}}(x_{0,g}^m; \mathbf{X_r}, \phi)$ or $r^D(x_{0,g}^m) = \hat{r}_D^{\text{MMD}}(x_{0,g}^m; \mathbf{X_r}, \phi)$.

To optimize Equation 19, we adopt the policy gradient in Equation 5 which supports multiple steps of parameter updates,

$$\nabla_\theta L_{RLD} = \mathbb{E}\left[\sum_{m=1}^{M}\sum_{t=0}^{T}\frac{p_\theta(x_{t-1}^m|x_t^m, c)}{p_{\theta'}(x_{t-1}^m|x_t^m, c)}\nabla_\theta \log p_\theta(x_{t-1}^m|x_t^m, c) r(x_{g,0}^m, c)\right]. \tag{20}$$

We summarize the proposed diversity-oriented RL fine-tuning in Algorithm 1.

5. Experiments

In this section, we present experiments to validate each component of the proposed method. We first validate the effectiveness of the proposed reward functions in Section 5.1.

Algorithm 1 Diversity-Oriented Diffusion Models with RL Fine-tuning.

Require: pre-trained diffusion model $p_{\theta_{pre}}$, reference images with each condition c: $\{\mathbf{X_r}|c\}$, diversity reward function $r_D(\mathbf{X_g}; \mathbf{X_r}, \phi)$.
Initialize $p_\theta = p_{\theta_{pre}}$
while θ not converged **do**
  For each condition $c \sim p(c)$, sample M images $\mathbf{X_g}|c = \{x_{0,g}^1, ..., x_{0,g}^M\}, x_{0,g}^m \sim p_\theta(x_0|c)$, together with their intermediate states $x_{1:T,g}^m$.
  Compute diversity rewards with the selected reward function for each condition c, $r_D(\mathbf{X_g}; \mathbf{X_r}, \phi, c)$.
  Assign reward to $x_{0,g}^m$ with individual reward (18).
  Take E_{in} rounds of policy gradient steps with (20).
end while
return Fine-tuned diffusion model p_θ.

(a) Baseline Model (b) RL Fine-tuned Model w/ r_D^{MMD} (c) Reference images

Figure 4. Comparisons with baseline model on ImageNet-128x128 generation (high guidance scale $s = 4.0$). Reference images are randomly selected from the training set of ImageNet. Our method fine-tunes the class-conditional model to generate images with higher visual diversity, e.g., diverse backgrounds, guided by the diverse reference images.

In Section 5.2 and Section 5.3, we demonstrate that, equipping with the proposed reward functions, our reinforcement learning (RL) framework can fine-tune both class-conditional and text-conditional diffusion models to generate diversified images guided by small reference sets.

5.1. Diversity Reward Evaluation

Post-Sampling Selection. Given a set of O generated images under condition c, we aim at selecting a subset of M images that have the maximum diversity. Specifically, on the 128x128 ImageNet class-conditional generation, we first sample 100K images from the pre-trained diffusion model [8]. We view the sampled images as 10K image sets, where each set of $O = 10$ images is sampled from the same (class) condition c. As for subset selection, we evaluate the reward values for all $\binom{O}{M}$ possible subsets and select the one with the largest reward. As for the $\mathbf{X_r}$ in computing rewards, we randomly select 50 images of the same class c from the ImageNet training set. We then evaluate the FID, Precision, and Recall of the selected 50K images.

Reward Evaluation Results. We first present the quantitative results of post-sampling selection in Table 1. Compared with the baseline random selection, all of the proposed rewards show their effectiveness in selecting the diverse subset in the post-sampling selection task. Specifically, subsets selected with max-reward criteria have higher Recall than the baseline random selection, which in turn improves the overall FID by a large margin. On the other hand, subsets selected with the min-reward criteria show low diversity compared with baseline selection, where both Recall and FID drop significantly. The results above show that proposed rewards are effective indicators in measuring the overlap between $\mathbf{X_g}$ and $\mathbf{X_r}$. We provide more results including visualizations in Appendix B.

Table 1. Reward Evaluation on ImageNet 128x128 post-sampling selection task. Compared with random selection, subsets selected with our maximum *Diversity Reward* achieve higher diversity (higher Recall and lower FID), while the ones selected with minimum reward have lower diversity (lower Recall and higher FID).

Select Criterion	Recall	Precision	FID
Baseline (Random)	40.22	85.77	8.10
w/ max MMD	**47.66**	83.26	**6.22**
w/ max GP-MI	45.31	84.14	6.81
w/ min MMD	27.69	92.39	14.94
w/ min GP-MI	31.24	88.26	10.32

Table 2. Results on ImageNet-128x128 class-conditional generation. Our method improves the diversity (Recall) of baseline models under various guidance scales.

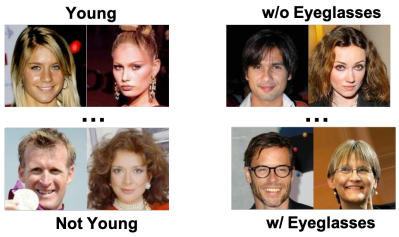
Method	Recall	Precision	FID
Baseline ($s = 4.0$)	36.15	82.96	24.49
Ours (GP-MI)	45.31	81.72	23.81
Ours (MMD)	**47.66**	83.26	**23.42**
Baseline $s = 3.0$	40.35	82.10	23.00
Ours (GP-MI)	46.31	81.74	22.31
Ours (MMD)	**49.31**	81.26	**22.08**
Baseline $s = 2.0$	45.95	79.20	21.19
Ours (GP-MI)	47.83	79.05	20.95
Ours (MMD)	**49.30**	78.82	**20.48**
Baseline $s = 1.0$	55.10	74.16	19.86
Ours (GP-MI)	56.13	73.52	19.49
Ours (MMD)	**59.63**	**74.84**	**19.19**

5.2. Diverse Class-Conditional Image Generation

In this section, we validate our method with class-conditional diffusion models. On both ImageNet [18] and CIFAR-10/100 [17], we show that our RL fine-tuning framework with the proposed rewards can adapt diffusion models to generate images with enhanced diversity.

Diverse Fine-tuning on ImageNet. We fine-tune the pre-trained ImageNet 128x128 conditional diffusion model and guided classifier proposed in [8] with RL and the diversity rewards. Note that we only fine-tune the first 100 classes

(a) Baseline generation, all 'Young' people without 'Eyeglasses'



(b) Unbiased reference images for attributes 'Young' and 'Eyeglasses'

(c) Unbiased generation for attribute 'Eyeglasses'

(d) Unbiased generation for attribute 'Young'

Figure 5. RL Fine-tuning on StableDiffusion [29] for unbiased face generations. The text prompt used is, 'a headshot of a person'. Reference Images are selected from CelebA [22], which have equal numbers of images with positive and negative attributes, e.g., 25 images w/ glasses, 25 images w/o glasses. Baseline generations are biased as they are young people without glasses. Our method tunes the SD to generate unbiased facial images under attributes 'Eyeglasses' and 'Young' respectively.

of ImageNet for efficiency consideration. For $\mathbf{X_r}$, we randomly select $N = 50$ images per class from the training set as in Section 5.1. As for evaluation, we select 20 non-overlapped images from each class as the real-image set with 2K images, and generate 5K images from the fine-tuned model. We then evaluate the Precision, Recall [20], and FID to measure both the diversity and quality of generated images. In RL fine-tuning, we adopt the LoRA [15] to the attention layers of both the diffusion model and guided classifier. We set $M = 5$, learning rate to $1e-4$, and fine-tune for 100 epochs. More implementation details are provided in Appendix C.

We present the quantitative results of RL fine-tuning on multiple guidance scales, $s = \{1.0, 2.0, 3.0, 4.0\}$. As shown in Table 2, RL fine-tuned models with all the proposed rewards show superior performances in all guidance scales. Specifically, compared with baseline models, fine-tuned models show higher Recall, which in turn leads to improvement in FID. We also provide visualizations in Figure 4, where images from the fine-tuned model, following the guidance of the reference set, are more diverse compared to the baseline generation results.

Diverse Fine-tuning on CIFAR. To compare with other benchmarks, we provide RL fine-tuning results on CIFAR [17]. Following [27], we first obtain pre-trained DDPM models on both CIFAR-10 and CIFAR-100. Compared with [27], where diverse diffusion models are re-trained from scratch, our method only fine-tunes a small portion of the parameters in a more efficient way. As for the $\mathbf{X_r}$, we also randomly select 50 images per class from the training set. More implementation details are provided in Appendix C. As shown in Table 3, our method achieves comparable results with [27], despite being guided by only a small reference set and with much fewer parameters.

5.3. Unbiased Text-Conditional Image Generation

In this section, we show how our method can improve diversity in text-to-image (T2I) generation.

5.3.1 Experimental Settings.

We adopt the settings in [40], and conduct experiments on tuning StableDiffusion [29] (SD) toward diverse face generation under the guidance of a small set of reference images $\mathbf{X_r}$. Specifically, we leverage our method to fine-tune T2I diffusion models to generate face images with a uniform distribution in a set of attributes, such as gender, skin tone, and age. For example, we tune SD to generate equal

Table 3. Results on CIFAR-10/100. Our method achieves comparable performances with task-specific retraining method [27].

Dataset	Method	Recall	F_s	IS	FID
CIFAR100	DDPM	0.65	0.97	13.65	3.11
	CBDM [27]	0.67	0.97	14.03	2.72
	Ours (GP-MI)	0.67	0.97	13.97	2.83
	Ours (MMD)	**0.69**	0.97	13.87	**2.70**
CIFAR10	DDPM	0.64	0.99	9.80	3.16
	CBDM [27]	0.65	0.99	9.63	3.03
	Ours (GP-MI)	0.65	0.99	9.71	3.04
	Ours (MMD)	**0.67**	0.99	9.68	**2.98**

Method	$\mathbb{D}_{\text{KL}}^{male}$	$\mathbb{D}_{\text{KL}}^{pale\ skin}$	$\mathbb{D}_{\text{KL}}^{young}$	$\mathbb{D}_{\text{KL}}^{eyeglass}$	$\mathbb{D}_{\text{KL}}^{mustache}$	$\mathbb{D}_{\text{KL}}^{smile}$
SD [29]	0.343	0.308	0.578	0.375	0.111	0.134
EI [3]	0.143	0.644	0.423	0.531	0.693	0.189
HPS [9]	1e-5	2.8e-3	0.027	0.371	0.241	4.4e-3
PD [7]	0.322	0.165	0.131	0.272	0.063	0.146
CD [19]	0.309	0.074	0.284	0.301	0.246	0.469
ITI-GEN [40]	1×10^{-6}	0	1×10^{-4}	2×10^{-4}	4.5×10^{-4}	1.0×10^{-3}
Ours (MMD)	1×10^{-3}	1×10^{-3}	2×10^{-4}	6×10^{-4}	1×10^{-3}	9×10^{-4}
Ours (GP-MI)	2×10^{-4}	5×10^{-4}	$\mathbf{2 \times 10^{-6}}$	$\mathbf{1.5 \times 10^{-4}}$	$\mathbf{3 \times 10^{-4}}$	$\mathbf{8 \times 10^{-4}}$

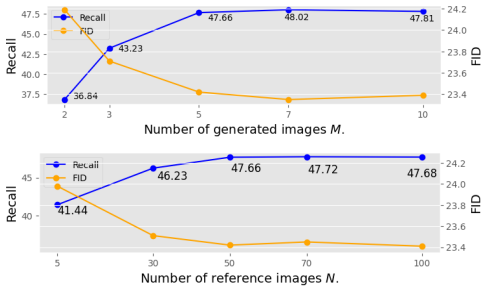


Figure 6. (Table) Comparison with baselines on unbiased face generation with StableDiffusion. Our method achieves competitive results with baseline text prompt engineering & tuning methods while being a generalizable framework. (Figure) Ablation experiments on the number of generated and reference images used in evaluating *Diversity Reward*.

numbers of faces with and without eyeglasses. We adopt the proposed two diversity rewards with RL fine-tuning, and apply LoRA [15] adaptation to SD. The text prompt for tuning is set as 'a headshot of a person'. More implementation details are provided in Appendix D.

Reference Images. We select $\mathbf{X_r}$ from CelebA [22], which contains facial images annotated with 40 attributes. Specifically, we randomly select 50 images for each binary attribute, with 25 from the positive attribute set and the rest of 25 from the negative attribute set. For instance, for attribute 'Eyeglasses', we select 25 reference images of faces with eyeglasses, and 25 reference images of faces without glasses, as shown in Figure 5b.

Baselines and Evaluation Metrics. We compare our method with the following methods. 1)SD [29] 2) Ethical Intervention (EI) [3] 3) Hard Prompt Searching (HPS) [9] 4) Prompts Debiasing (PD) [7] 5) Custom Diffusion (CD) [19] 6) ITI-GEN [40]. As for the evaluation, we sample 200 images from the fine-tuned SD and use the CLIP [28] to predict the attribute for each generated image. For attribute a, this gives an attribute distribution of generated images P_{gen}^a. Following [6, 7, 40], we adopt KL divergence to measure the discrepancy between P_{gen}^a and the targeted uniform attribute distribution P_{uni}^a, i.e., $\mathbb{D}_{\text{KL}}[P_{gen}^a || P_{uni}^a]$. We provide more details on evaluation in Appendix D.

5.3.2 Results.

As shown in Figure 6 (Table), our RL fine-tuning framework achieves better performance than most of the baselines. As for the comparison with ITI-GEN [40], a diverse image generation method specifically designed for StableDiffusion, our method achieves comparable performances while having the property of being generalizable to other diffusion models. We provide more results with other attributes in Appendix D. Moreover, we present visualizations in Figure 5. As shown in Figure 5c & 5d, SD fine-tuned with our method can generate images with uniform distributions for attributes 'Eyeglasses' and 'Young', following the guidance of $\mathbf{X_r}$, which greatly enhances the diversity over the baseline generation as in Figure 5a.

5.4. Ablation Experiments

In this section, we study the effects of the size of generated images M and reference images N used to evaluate the reward. Specifically, we adopt the ImageNet experiment as in Section 5.2. More details are provided in Appendix E

5.4.1 Choice of M

We test $M = 2, 3, 5, 7, 10$, and set $N = 50$. As shown in Figure 6 (Top Figure), using small numbers of generated images ($M = 2, 3$) leads to degraded results, as it can be hard to represent the distribution with 2,3 samples. On the other hand, we find that $M = 5$ can achieve reasonable results while increasing M further can not improve the results even with more cost induced.

5.4.2 Choice of N

We test $N = 5, 30, 50, 70, 100$, and set $M = 5$. More experimental details are provided in Appendix E. As shown in Figure 6 (Bottom Figure), $N \geq 30$ reference images lead to reasonable results, while $N = 5$ images can cause performance degradation.

6. Conclusion

We present a general diversity-oriented fine-tuning framework for diffusion models with reinforcement learning (RL), under the guidance of a small set of diverse reference images. We propose several designs of *Diversity Reward* with theoretical justification, and design an RL fine-tuning framework that adapts the generative distribution of diffusion models to align with the diverse reference distribution. The effectiveness of the proposed reward functions and the overall RL fine-tuning framework is validated on multiple datasets with various diffusion models.

References

[1] 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. 5

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[3] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022. 1, 2, 8

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 5

[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, 2, 3

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2(4):5, 2022. 8

[7] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 2, 8

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 6

[9] 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, 2, 8

[10] Ying Fan and Kangwook Lee. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023. 2, 3

[11] 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. 2

[12] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018. 5

[13] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel method for the two-sample-problem. *Journal of Machine Learning Research*, 13:723–773, 2012. 2, 4, 1

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[15] 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. 7, 8, 1, 2

[16] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002. 3

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 7

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 6

[19] 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. 2, 8

[20] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 4, 7, 1

[21] 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. 2

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 7, 8

[23] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020. 2, 5

[24] 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. 2

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 2

[26] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022. 1

[27] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18434–18443, 2023. 1, 2, 7

[28] 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. 2, 4, 8

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 7, 8

[30] 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. 2

[31] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015. 3

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

[33] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2022. 1, 2

[34] 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. 1, 2

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4, 1, 2

[36] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5310–5319, 2019. 2

[37] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 2

[38] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006. 2, 5, 1

[39] 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. 2, 3

[40] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 1, 2, 7, 8

[41] 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. 2