Tackling the Singularities at the Endpoints of Time Intervals in Diffusion Models

Pengze Zhang[1*] Hubery Yin[2*] Chen Li[2] Xiaohua Xie[1†]
[1]Sun Yat-sen University [2]Wechat, Tencent Inc.
zhangpz3@mail2.sysu.edu.cn, {hubery, chaselli}@tencent.com, xiexiaoh6@mail.sysu.edu.cn
https://pangzecheung.github.io/SingDiffusion/

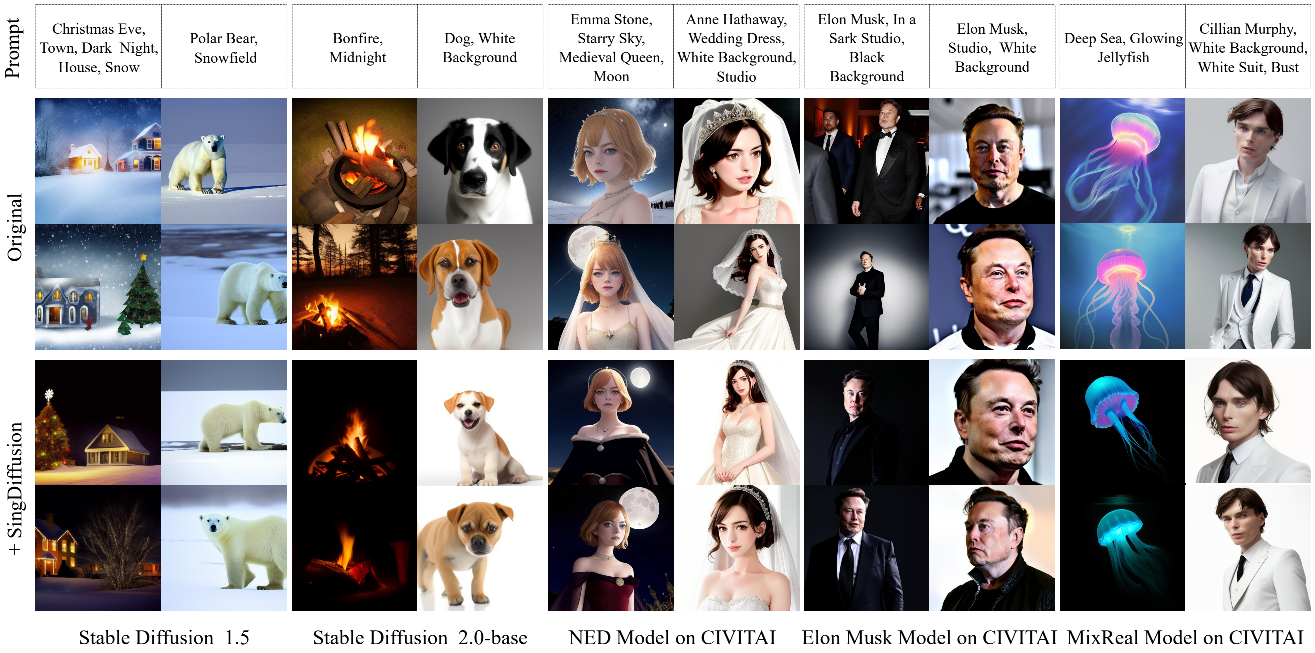
Prompt
Christmas Eve, Town, Dark Night, House, Snow
Polar Bear, Snowfield
Bonfire, Midnight
Dog, White Background
Emma Stone, Starry Sky, Medieval Queen, Moon
Anne Hathaway, Wedding Dress, White Background, Studio
Elon Musk, In a Sark Studio, Black Background
Elon Musk, Studio, White Background
Deep Sea, Glowing Jellyfish
Cillian Murphy, White Background, White Suit, Bust
Original
+ SingDiffusion
Stable Diffusion 1.5
Stable Diffusion 2.0-base
NED Model on CIVITAI
Elon Musk Model on CIVITAI
MixReal Model on CIVITAI

Figure 1. In this paper, we explore singularities theoretically and propose a plug-and-play module SingDiffusion to address the sampling challenge at the initial singular time step. By integrating this module into existing pre-trained models, our approach effectively tackles the difficulties of generating both dark and bright images, and further enhancing overall image quality as confirmed by quantitative analysis.

Abstract

Most diffusion models assume that the reverse process adheres to a Gaussian distribution. However, this approximation has not been rigorously validated, especially at singularities, where $t = 0$ and $t = 1$. Improperly dealing with such singularities leads to an average brightness issue in applications, and limits the generation of images with extreme brightness or darkness. We primarily focus on tackling singularities from both theoretical and practical perspectives. Initially, we establish the error bounds for the reverse process approximation, and showcase its Gaussian characteristics at singularity time steps. Based on this theoretical insight, we confirm the singularity at $t = 1$ is conditionally removable while it at $t = 0$ is an inherent property. Upon these significant conclusions, we propose a novel plug-and-play method ***SingDiffusion*** *to address the initial singular time step sampling, which not only effectively resolves the average brightness issue for a wide range of diffusion models without extra training efforts, but also enhances their generation capability in achieving notable lower FID scores.*

1. Introduction

Diffusion models, generating samples from initial noise by learning a reverse diffusion process, have achieved remark-

[*] Equal contribution. This work was done when Pengze Zhang was an intern at WeChat.
[†] Corresponding Author.

able success in multi-modality content generation, such as image generation [4, 5, 26, 30, 35, 46], audio generation [18, 26, 28], and video generation [14]. These achievements owe much to several fundamental theoretical research, namely Denoising Diffusion Probabilistic Modeling (DDPM) [13, 38], Stochastic Differential Equations (SDE) [40], and Ordinary Differential Equations (ODE) [39, 40]. These approaches are all based on the assumption that the reverse diffusion process shares the same functional form as the forward process. Although an indirect validation of this assumption is given by Song et al. [40], it heavily relies on the existence of solutions to the Kolmogorov forward and backward equations [1], which encounters singularities at the endpoints of time intervals where $t = 0$ and $t = 1$.

The singularity issue is not only a gap in the theoretical formulation of current diffusion models, but also affects the quality of the content they generate. Current applications simply ignore singularity points in their implementation [27, 31, 42] and restrict the time interval to $[\varepsilon_1, 1-\varepsilon_2]$. As a result, the average brightness of the generated images typically hovers around 0 [9, 19] (normalizing brightness to $[-1, 1]$). For example, as shown in Fig. 1, existing pre-trained diffusion models, such as *Stable Diffusion 1.5* (SD-1.5) [31] and *Stable Diffusion 2.0-base* (SD-2.0-base), fail in generating images with pure white or black backgrounds. To address this challenge, Guttenberg et al. [9] add extra offset noise during the training process to allow the network could learn the overall brightness changes of the image. Unfortunately, the offset noise usually disrupts the pre-defined marginal probability distribution and further invalidates the original sampling formula. Lin et al. [19] re-scales the noise schedule to enforce a zero terminal signal-to-noise ratio, and employ the v-prediction technique [36] to circumvent the issue of division by zero at $t = 1$. However, this method only supports models in v-prediction manner and requires substantial training to fine-tune the entire model. Therefore, it is advantageous to devise a plug-and-play method which effectively deals with the singularity issue for any practicable diffusion model without extra training efforts.

In this paper, we begin with a theoretical exploration of the singularity issue at the endpoints of time intervals in diffusion model, and then devise a novel plug-and-play module to address the accompanying average brightness problem in image generation. Through establishing mathematical error bounds rigorously, we first prove the approximate Gaussian characteristics of the reverse diffusion process at all sampling time steps, especially where the singularity issue appears. Following this, we conduct a thorough analysis of this approximation in the vicinity of singularities, and arrive at two significant conclusions: 1) by computing the limit of ‘zero divided by zero’ at $t = 1$, we confirm the corresponding initial singularity is removable; 2) the singularity at $t = 0$ is an inherent property of diffusion models, thus we should adhere to this form rather than simply avoiding the singularity. Following the aforementioned theoretical analysis, we propose the SingDiffusion method, specifically tailored to address the challenge of sampling during the initial singularity. Especially, this method can seamlessly integrate into existing sampling processes in a plug-and-play manner without requiring additional training efforts.

As demonstrated in Fig. 1 and experiments in the appendix, our novel plug-and-play initial sampling step can effectively resolve the average brightness issue. It can be easily applied to a wide range of diffusion models, including the SD-1.5 and SD-2.0-base with ϵ-prediction, SD-2.0 with v-prediction, and various pre-trained models available on the CIVITAI website[1], thanks to its one-time training strategy. Furthermore, our method can generally enhance the generative capabilities of these pre-trained diffusion models in notably improving the FID [11] scores at the same CLIP [29] level on the COCO dataset [20].

2. Related Work

2.1. Reverse process approximation

Denoising Diffusion Probability Models (DDPM) [13, 38] establish a hand-designed forward Markov chain in discrete time, and model the distribution of data in the reverse process. In contrast, Song et al. [40] establish the diffusion model in continuous time, framed as a Stochastic Differential Equation (SDE). Moreover, they reveal an Ordinary Differential Equation (ODE) termed ‘probability flow’, sharing the same single-time marginal distribution as the original SDE. Notably, ODE also has discrete counterparts known as Denoising Diffusion Implicit Models (DDIM) [39].

The assumption that the reversal of the diffusion process has an identical functional form to the forward process is fundamental in the aforementioned methods. Several studies aim to prove this assumption. Song et al. [40] indirectly substantiate this in continuous time by introducing a reverse SDE. Nevertheless, they do not provide error bounds in discrete-time cases, leaving this assumption unverified at discrete-time steps. Additionally, the treatment of singularities at $t = 0$ and $t = 1$ is not addressed in their work. McAllester et al. [24] provide proof in discrete time, showing the density of the reverse process as a mixture of Gaussian distributions. Despite this, it doesn’t qualify as a pure Gaussian distribution. In contrast, to fill this theoretical gap, our approach directly substantiates this assumption by establishing error bounds for the reverse process approximation at both non-singular and singular time steps.

2.2. Singularities in diffusion model

Several studies have focused on investigating the singularity occurring at $t = 0$. Song et al. [40] attempt to bypass

[1] https://civitai.com/

this singularity by initiating their analysis at $t = \varepsilon > 0$ instead of $t = 0$. Therefore, this approach didn't effectively address the core singularity problem. Dockhorn et al. [6] propose a diffusion model on the product space of position and velocity, and avoid the singularity through hybrid score matching. Nevertheless, this approach lacks compatibility with DDPM, SDE and ODE due to the incorporation of the velocity space. Lu et al. [22] employ the x-prediction method to mitigate singularity during training, but did not address the singularity issue during the sampling process.

Therefore, a comprehensive solution to the singularity at $t = 0$ is still pending. Moreover, the singularity at $t = 1$ remains unexplored. To tackle these issues, we thoroughly investigate the sampling process at both $t = 0$ and $t = 1$, and offer theoretical solutions.

2.3. Average brightness issue

Diffusion models have shown significant comprehensive quality and controllability in computer vision, including text-to-image generation [5, 12, 26, 30, 31, 35], image editing [3, 10, 16, 25, 33, 43], image-to-image translation [34, 41, 44], surpassing previous generative models [7, 8, 23, 45]. However, most existing diffusion models ignore the sampling at the initial singular time step, resulting in the inability to generate bright and dark images, i.e., the average brightness issue. Adding offset noise [9] and employing v-prediction [19, 36] are two ways to tackle this problem. However, these methods require fine-tuning for each existing model, consuming a substantial amount of time and limiting their applicability. In contrast, we propose a novel plug-and-play solution targeting the core of the average brightness issue, i.e., the singularity at the initial time step. Our method not only empowers the majority of existing pre-trained models to effectively generate images with the desired brightness level, but also significantly enhances their generative capabilities.

3. Method

To facilitate the clarity of our exploration into singularities from theoretical and practical angles, this section is organized as follows: 1) We start by introducing background and symbols in the Preliminaries. 2) Next, we derive error bounds for the reverse process approximation, confirming its Gaussian characteristics at both regular and singular time steps. 3) We then theoretically analyze and handle the sampling at singular time steps, i.e., $t = 0$ and $t = 1$. 4) Lastly, based on our previous analysis, we propose a plug-and-play method to address initial singularity time step sampling, effectively resolving the average brightness issue.

3.1. Preliminaries

In the realm of generative models, we are consistently provided with a set of training samples denoted as $\{y_i \in \mathbb{R}^d\}_{i=1}^N$, which are inherently characterized by a distribution given by [15]:

$$p(x, t = 0) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - y_i), \tag{1}$$

where $\delta(x)$ denotes the Dirac delta function.

Consider a continuous-time Gaussian diffusion process within the interval $0 \le t \le 1$. Following [17], the distribution of x_t conditioned on x_0 is written by $p(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 I)$, where α_t and σ_t are positive scalar functions of t satisfying $\alpha_t^2 + \sigma_t^2 = 1$, and α_t decreases monotonically from 1 to 0 over time t; the distribution of x_t conditioned on x_s is represented as $p(x_t|x_s) = \mathcal{N}(\alpha_{t|s}x_s, \sigma_{t|s}^2 I)$, where $0 \le s < t \le 1$, $\alpha_{t|s} = \alpha_t/\alpha_s$, $\sigma_{t|s}^2 = 1 - \alpha_{t|s}^2$. Consequently, the forward process is derived as follows:

$$x_t = \alpha_{t|s}x_s + \sqrt{1 - \alpha_{t|s}^2}\, z_s, \tag{2}$$

where the set $\{z_t\}_{t=0}^1$ comprises independent standard Gaussian random variables.

For a discrete-time diffusion process with T steps, the time $i \in \{0, 1, ...T\}$ corresponds to t in the continuous case as $i = t \times T$. Defining $\hat{\beta}_i = 1 - \hat{\alpha}_{i|i-1}^2$, where $\hat{\alpha}_{i|i-1} = \alpha_{i/T|(i-1)/T}$, and '$\hat{}$' denotes the symbol in the discrete-time process, the forward process in Eq. 2 can be rewritten as:

$$\hat{x}_i = \sqrt{1 - \hat{\beta}_i}\,\hat{x}_{i-1} + \sqrt{\hat{\beta}_i}\,\hat{z}_{i-1}, \tag{3}$$

which is equivalent to the forward process outlined in [13].

Taking into account the initial distribution (Eq. 1), the single-time marginal distribution of x_t is given by:

$$p(x_t, t) = \frac{1}{N}\sum_i (2\pi\sigma_t^2)^{-\frac{d}{2}} \exp(-\frac{(x_t - \alpha_t y_i)^2}{2\sigma_t^2}), \tag{4}$$

where d is the dimension of the training samples. As a result, the reverse process can be derived using Bayes' rule:

$$p(x_s|x_t) = p(x_t|x_s)\frac{p(x_s, s)}{p(x_t, t)} = (2\pi\sigma_{s|t})^{-\frac{d}{2}}$$
$$\sum_i \exp(-\frac{1}{2\sigma_{s|t}^2}(x_s - \frac{\alpha_{t|s}\sigma_s^2 x_t}{\sigma_t^2} - \frac{\alpha_s\sigma_{t|s}^2 y_i}{\sigma_t^2})^2)w_i(x_t, t), \tag{5}$$

where $\sigma_{s|t}^2 = \sigma_{t|s}^2\frac{\sigma_s^2}{\sigma_t^2}$, and $w_i(x_t, t) = \frac{\exp(-\frac{(x_t - \alpha_t y_i)^2}{2\sigma_t^2})}{\sum_j \exp(-\frac{(x_t - \alpha_t y_j)^2}{2\sigma_t^2})}$.

3.2. Error bound estimation

Existing diffusion models, such as [13, 38] are based on an assumption that the reverse process in Eq. 5 can be approximated by a Gaussian distribution, when $\hat{\beta}_i$ is small, as given by:

$$\tilde{p}(x_s|x_t) = (2\pi\sigma^2_{s|t})^{-\frac{d}{2}} \exp\Big(-\frac{1}{2\sigma^2_{s|t}}\big(x_s - \frac{\alpha_{t|s}\sigma_s^2 x_t}{\sigma_t^2} - \frac{\alpha_s\sigma^2_{t|s}\bar{y}(x_t,t)}{\sigma_t^2}\big)^2\Big), \tag{6}$$

where $\bar{y}(x_t,t) = \sum_i w_i(x_t,t)y_i$. However, these studies did not furnish the error bounds to support this assumption. To address this theoretical gap, we estimate the error bound as follows.

Proposition 1 (Error Bound Estimated by $\sigma_{s|t}$). $\forall s \in (0,1)$, $\exists \tau \in (s,1)$ *and* $C > 0$, *such that* $\forall t \in (s,\tau]$, $\int_{\mathbb{R}^d} |p(x_s|x_t) - \tilde{p}(x_s|x_t)|dx_s < C\sqrt{\sigma_{s|t}}$.

Proposition 1 demonstrates that when $\sigma_{s|t}$ is small, the forward and reverse processes share the same form, i.e., Gaussian distribution. Since $\hat{\beta}_i = \frac{\sigma_{i/T}}{\sigma_{(i-1)/T}}\sigma_{(i-1)/T|i/T}$ features a term in the $\sigma_{s|t}$ form, also supports the inference that this assumption remains valid when $\hat{\beta}_i$ is small, as highlighted in [13, 38]. It is worth noting that this error is bounded by $\sigma_{s|t}$ instead of $\hat{\beta}_i$ strictly.

However, the error bound estimated by $\sigma_{s|t}$ in Proposition 1 is not sufficient to prove the assumption at the singularity time step $t = 1$. The reason is that when $t = 1$, $\sigma_{s|t}$ approaches 1 as $s \to 1$, which is not a small value. Consequently, the error bound at $t = 1$ in Proposition 1 remains non-negligible.

To tackle this issue, we instead utilize α_s to bound the error at time step $t = 1$, and present a new proposition:

Proposition 2 (Error Bound Estimated by α_s). $\exists \nu \in (0,1)$ *and* $C > 0$, *such that* $\forall \nu \leq s < t \leq 1$, $\int_{\mathbb{R}^d} |p(x_s|x_t) - \tilde{p}(x_s|x_t)|dx_s < C\sqrt{\alpha_s}$.

According to Proposition 2, setting $t = 1$, it has $\alpha_s \to 0$ as $s \to 1$. As a result, the error bound at $t = 1$ assures a small value, affirming the validation of the Gaussian approximation assumption at $t = 1$.

In sum, through Proposition 1 and 2, we have established that the reverse process of the diffusion model can be approximated by Gaussian distribution across all time steps.

3.3. Tackling the singularities

With the theoretical foundation provided in Section 3.2, we can delve into the analysis of singular time steps using Eq. 6. It is worth noting that this section will mainly focus on addressing the singularities present in the discrete-time diffusion model [13], while the treatment of the continuous case [40] will be deferred to the appendix.

3.3.1 The singularities at $t = 1$

Drawing from Eq. 6, a straightforward way is to train a neural network $\bar{y}_\theta(x_t,t)$ for estimating $\bar{y}(x_t,t)$, known as x-prediction. This approach ensures that the approximated reverse process avoids encountering a singularity at $t = 1$. However, for stable training [38], the mainstream choice among current diffusion models is ϵ-prediction. It utilizes a neural network $\epsilon_\theta(x_t,t)$ for estimating $\epsilon(x_t,t) = \frac{x_t - \alpha_t\bar{y}(x_t,t)}{\sigma_t}$, which will encounter singularity. More specifically, substituting $\epsilon(x_t,t)$ for $\bar{y}(x_t,t)$ yields the equation:

$$x_s = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}x_t + \frac{\alpha_s\sigma^2_{t|s}}{\sigma_t^2}\frac{x_t - \sigma_t\epsilon(x_t,t)}{\alpha_t} + \sigma_{s|t}z_t. \tag{7}$$

At $t = 1$, the denominator α_t becomes 0, resulting in a division-by-zero singularity. Theoretically, this singularity is removable because the limit exists:

$$\lim_{t\to 1^-}\frac{x_t - \sigma_t\epsilon(x_t,t)}{\alpha_t} = \lim_{t\to 1^-}\bar{y}(x_t,t) = \frac{1}{N}\sum_i y_i. \tag{8}$$

Regrettably, computing this limit during inference is unfeasible, since $\epsilon(x_1,1) = x_1$ loses all information of the correct sampling direction. Conversely, $\bar{y}(x_t,t)$ retains the correct sampling direction for all $t \in [0,1]$. Particularly at $t = 1$, $\bar{y}(x_1,1) = \frac{1}{N}\sum_i y_i$ encapsulates the information for the initial inference step. Therefore, leveraging x-prediction at the initial time step proves more advantageous than ϵ-prediction.

3.3.2 The singularities at $t = 0$

When $s = 0$ and t is small, the distribution in Eq. 6 degenerates into a singular distribution, i.e., a Gaussian with zero variance, resembling a Dirac delta:

$$\tilde{p}(x_0|x_t) = \delta(x_0 - y_{j_0}), \tag{9}$$

where $j_0 = \arg\min_j |x_t - \alpha_t y_j|$. This singularity directs the sampling process to converge at the correct point $y_{j_0} = \bar{y}(x_0,0)$. Therefore, the singularity at $t = 0$ is an inherent characteristic of diffusion models that do not require avoidance as long we use suitable sampling techniques. For instance, in the final step of the original DDPM sampling method, it has $x_0 = \bar{y}(x_t,t)$. When t is small, $\bar{y}(x_t,t) \approx \bar{y}(x_0,0)$, making this process equivalent to Eq. 9. Thus, there is no need to avoid the singularity.

Besides, we also arrived at this conclusion within the continuous diffusion model, i.e., SDE. More detailed elaboration on this topic can be found in the appendix.

3.4. SingDiffusion

In addition to theoretical issues, the singular sampling issue can also cause average brightness issue in applications, as depicted in Fig. 2. This is mainly because most of the existing method sample images starting at time step $1 - \varepsilon$ using a standard Gaussian distribution, which significantly

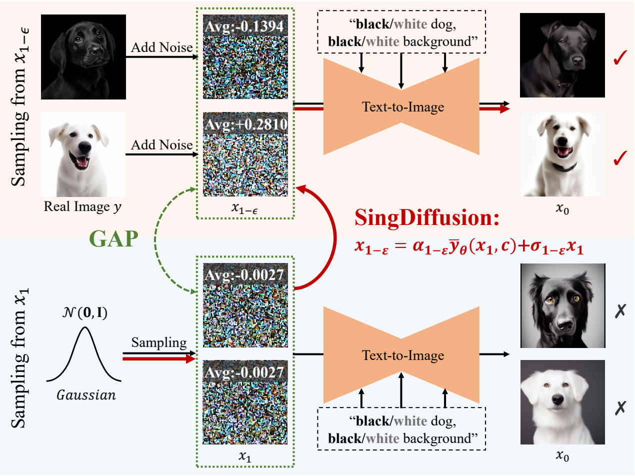
Sampling from $x_{1-\epsilon}$
Add Noise
Add Noise
Real Image y
Avg:-0.1394
Avg:+0.2810
$x_{1-\epsilon}$
"black/white dog, black/white background"
Text-to-Image
✓
✓
x_0
GAP
SingDiffusion:
$x_{1-\varepsilon} = \alpha_{1-\varepsilon}\bar{y}_\theta(x_1, c) + \sigma_{1-\varepsilon}x_1$
Sampling from x_1
$\mathcal{N}(\mathbf{0}, \mathbf{I})$
Gaussian
Sampling
Avg:-0.0027
Avg:-0.0027
x_1
Text-to-Image
"black/white dog, black/white background"
✗
✗
x_0

Figure 2. Illusion of the gap between the sampling from $x_{1-\varepsilon}$ and x_1. Due to the lack of consideration of singular time step sampling in most of the existing methods, they will encounter the average brightness issue. To tackle this, we propose a plug-and-play SingDiffusion method (highlighted in red) to bridge this gap.

Algorithm 1 Training process of SingDiffusion

```
repeat
    $x_0, c \sim p(x_0, c)$, $x_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    Take gradient descent step on $\nabla_\theta \left\|\bar{y}_\theta(x_1, c) - x_0\right\|^2$
until converged
```

Algorithm 2 Sampling process of SingDiffusion

```
$x_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$\varepsilon = 1/T$
$x_{1-\varepsilon} = \alpha_{1-\varepsilon}\bar{y}_\theta(x_1, c) + \sigma_{1-\varepsilon}x_1$
for $t = 1 - \varepsilon, \ldots, \varepsilon$ do
    Calculate $x_{t-\varepsilon}$ using existing sampling algorithms
end for
return $x_0$
```

diverges from the true distribution $p(x_{1-\varepsilon}, 1-\varepsilon)$. To validate this, we select two images (a black dog against a black background and a white dog against a white background). Then we diffuse them for a time period of $1-\varepsilon$ respectively, and calculate the mean value of their latent code. It can be seen that the mean of these two latent codes are -0.14 and +0.28, which are significantly different from the samples obtained from the standard Gaussian distribution. Moreover, as evident from the visualization in Fig. 2, the latent code ($x_{1-\varepsilon}$) of the black dog and white dog are notably darker and brighter compared to x_1 from the Gaussian distribution respectively. Under such distributional differences, according to Proposition 3, employing a Gaussian distribution at $t = 1 - \epsilon$ is equivalent to generating images towards an average brightness of 0 at $t = 1$ (with grayscale ranging from [-1, 1]). Consequently, current methods encounter challenges in generating dark or bright images.

Proposition 3. *Setting $x_{1-\epsilon} \sim \mathcal{N}(0, I)$ is equivalent to sampling the value from standard Gaussian as $\bar{y}(x_1, 1)$ at $t = 1$.*

Based on our analysis of sampling at singular time steps in Section 3.3.1, we propose a novel plug-and-play method SingDiffusion to fill the gap at $t = 1$, thus solving the problem of average brightness issue. Considering a pre-trained model $\epsilon_\theta(x_t, t)$ afflicted by the singularity due to division by zero, we proceed to train a model using x-prediction at $t = 1$. The algorithm of our training and sampling process are shown in Algorithm 1 and Algorithm 2. Firstly, for image-prompt data pairs (x_0, c) in the training process, we use a U-net [32] $\bar{y}_\theta$ to fit $\bar{y}(x_1, t = 1, c)$, where $\bar{y}(x_t, t, c)$ is the extension of $\bar{y}(x_t, t)$ defined in Section 3.2 under the condition c. The loss function can be written as:

$$\mathcal{L} = \mathbb{E}_{x_0, c \sim p(x_0, c), x_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} ||\bar{y}_\theta(x_1, c) - x_0||^2. \quad (10)$$

As our model is only amortized with respect to text embeddings and not the time step, we omit the variable t for $\bar{y}_\theta$, and the U-net does not necessitate a complex architecture.

In the sampling process, we use the new model $\bar{y}_\theta$ in the initial time-step with a DDIM scheduler:

$$x_{1-\varepsilon} = \alpha_{1-\varepsilon}\bar{y}_\theta(x_1, c) + \sigma_{1-\varepsilon}x_1. \quad (11)$$

This equation guarantees $x_{1-\varepsilon}$ adheres to the distribution $p(x_{1-\varepsilon}, 1-\varepsilon)$. Following this, we utilize the existing pre-trained model $\epsilon_\theta(x_t, t, c)/v_\theta(x_t, t, c)$ in ϵ-prediction/v-prediction manner to perform the subsequent sampling steps until it generates x_0. It's worth noting that our method is solely involved in the sampling at $t = 1$, independent of the subsequent sampling process. Consequently, our approach can be once-trained and seamlessly integrated into the majority of diffusion models.

For further improving matching between generated images and input prompts, existing diffusion models typically incorporate classifier-free techniques [26]:

$$o_{guidance} = o_{neg} + w \times (o_{pos} - o_{neg}), \quad (12)$$

where $w \geq 1$ represents the guidance scale, o signifies the output of the diffusion model which can be either $\bar{y}_\theta$, ϵ_θ or v_θ, 'pos' and 'neg' refer to the outputs corresponding to positive and negative prompts respectively. Nevertheless, we notice that when applying this technique at the initial singular time step, the influence of the guidance scale w predominates the results due to the greater directional difference between the negative and positive outputs. To tackle this challenge, we implement a straightforward yet highly effective normalization method:

$$\bar{y}_{\theta_{guidance}} = [\bar{y}_{\theta_{neg}} + w \times (\bar{y}_{\theta_{pos}} - \bar{y}_{\theta_{neg}})]/w. \quad (13)$$

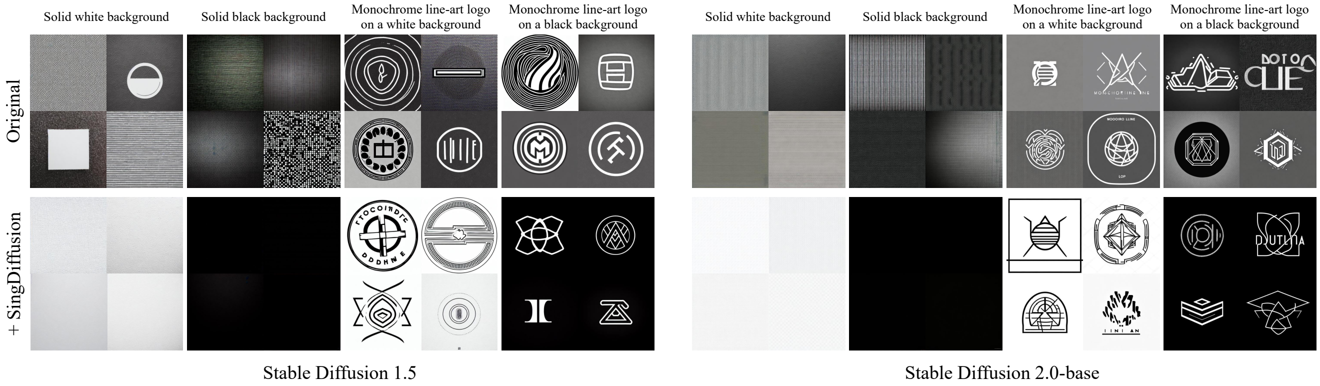
Solid white background
Solid black background
Monochrome line-art logo on a white background
Monochrome line-art logo on a black background
Solid white background
Solid black background
Monochrome line-art logo on a white background
Monochrome line-art logo on a black background
Original
+ SingDiffusion
Stable Diffusion 1.5
Stable Diffusion 2.0-base

Figure 3. Comparison of stable diffusion models and SingDiffusion on average brightness issue.

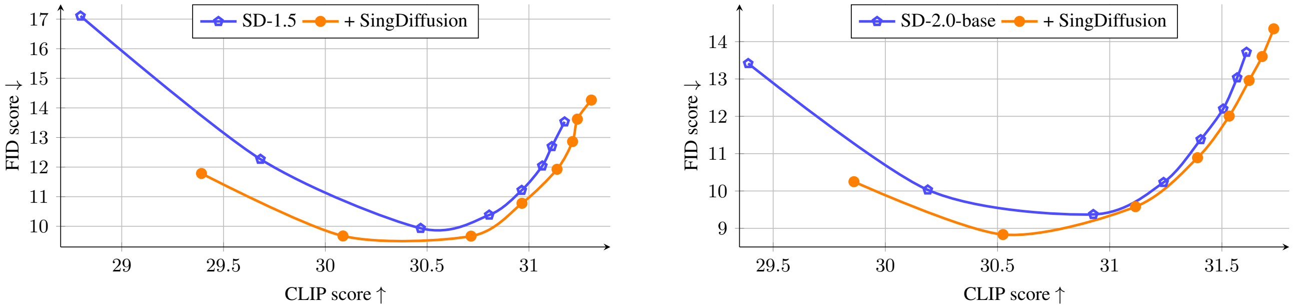

Figure 4. Comparison of Pareto curves between SingDiffusion, SD-1.5, and SD-2.0-base on 30k COCO images, across various guidance scales in [1.5, 2, 3, 4, 5, 6, 7, 8].

4. Experiment

4.1. Implement details

To create a versatile plug-and-play model, our model is trained on the Laion2B-en dataset [37], including 2.32 billion text-image pairs. The images are center-cropped to 512 × 512. Our $\bar{y}_\theta$ follows the U-net structure similar to stable diffusion models but with reduced parameters, totaling 140 million. We utilize the AdamW optimizer [21] with a learning rate of 1e-4 to train our $\bar{y}_\theta$. The training is executed across 64 Nvidia V100 chips with a batch size of 3072, and completes after 180K steps.

During the testing process, all steps, including our initial sampling, are carried out using the default guidance scale of $w = 7.5$. After our initial sampling at the singular time step, all existing pre-trained diffusion models generate images in the DDIM pipeline, executing 50 steps to produce images following a schedule with a total time of $T = 1000$.

4.2. Average brightness issue

To validate the effectiveness of SingDiffusion in addressing the average brightness issue, we select four extreme prompts, including “Solid white/black background”, and “Monochrome line-art logo on a white/black background”.

Table 1. Comparison of average brightness of 100 generated images between stable diffusion models and our SingDiffusion under different prompt conditions. For ’white’/’black’ prompts, higher/lower average brightness is better.

Model	”Solid **white** background”	”Solid **black** background”	”Monochrome line-art logo on a **white** background”	”Monochrome line-art logo on a **black** background”
SD-1.5	141.43	83.09	137.95	113.66
+ Ours	**212.59**	**3.04**	**223.68**	**11.52**
SD-2.0-base	150.52	99.67	136.13	104.45
+ Ours	**227.43**	**0.29**	**228.68**	**10.87**

For each prompt, we generate 100 images using SD-1.5, SD-2.0-base and SingDiffusion methods, and then calculate their average brightness. The results are shown in Table 1. It is remarkably clear that the stable diffusion methods, whether using prompts with ‘black’ or ‘white’ descriptors, tend to generate images with average brightness. However, SingDiffusion, implementing initial singularity sampling, effectively corrects the average brightness of the generated images. For example, under the prompt “solid black background,” our method notably lowers the average brightness from 99.67 to 0.29 for images generated by SD-2.0-base.

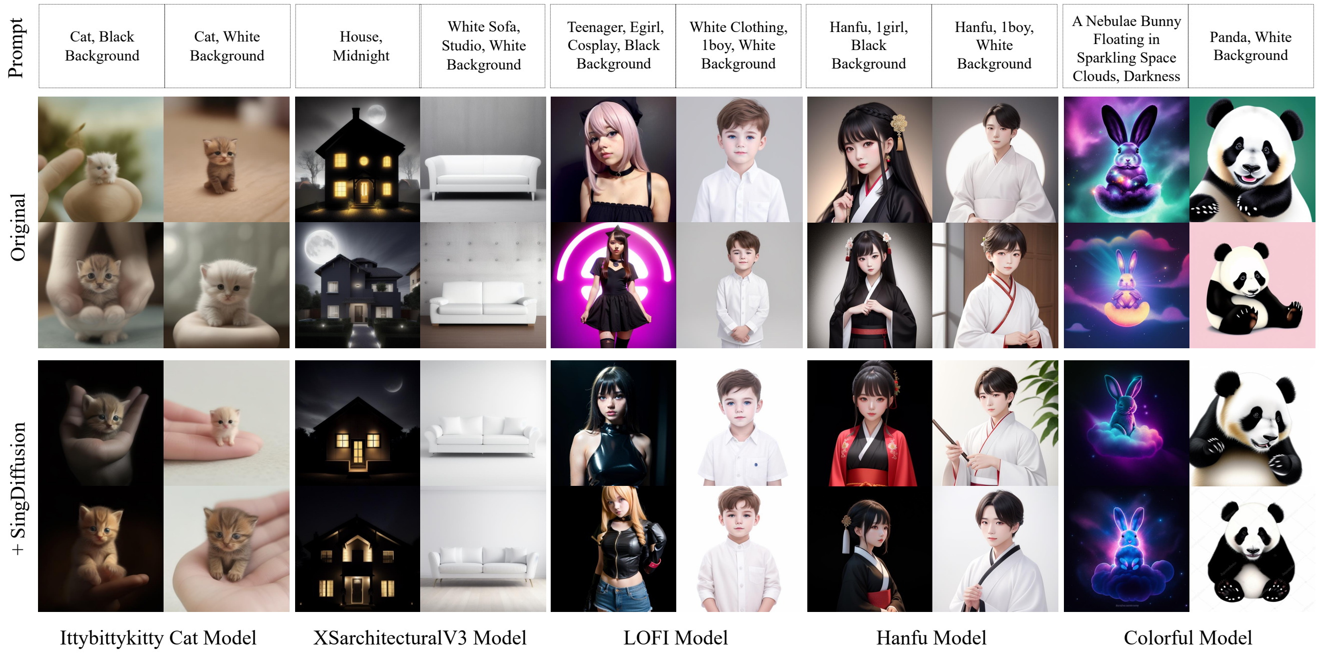
Prompt
Cat, Black Background
Cat, White Background
House, Midnight
White Sofa, Studio, White Background
Teenager, Egirl, Cosplay, Black Background
White Clothing, 1boy, White Background
Hanfu, 1girl, Black Background
Hanfu, 1boy, White Background
A Nebulae Bunny Floating in Sparkling Space Clouds, Darkness
Panda, White Background
Original
+ SingDiffusion
Ittybittykitty Cat Model
XSarchitecturalV3 Model
LOFI Model
Hanfu Model
Colorful Model

Figure 5. Our method can be trained once and seamlessly integrated into the pre-trained models on CIVITAI in a plug-and-play fashion.

Table 2. Comparison of stable diffusion model and SingDiffusion on FID score and CLIP score without classifier guidance.

Model	SD-1.5		SD-2.0-base	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑
Original	31.86	26.70	25.17	27.48
+ SingDiffusion	**21.09**	**27.71**	**18.01**	**28.23**

Moreover, we also provide visualization results for these prompts in Fig. 3. It can be seen that images generated by stable diffusion methods predominantly exhibit a gray tone. In contrast, our SingDiffusion method successfully overcomes this issue, and is capable of generating both dark and bright images.

4.3. Improvement on image quality

To validate our improvements in general image quality, following [35], we randomly select 30k prompts from the COCO dataset [20] as the test set. We employ two metrics for evaluation, including the FID score and CLIP score. Specifically, Fréchet Inception Distance (FID) [11] calculates the Fréchet distance between the real data and the generated data. A lower FID implies more realistic generated data. While the Contrastive Language-Image Pre-training (CLIP) [29] score measures the similarity between the generated images and the given prompts. A higher CLIP score means the generated images better match the input prompts.

First of all, we compare SingDiffusion with SD-1.5 and SD-2.0-base in Table 2 to gauge the model's inherent fitting capability, without using guidance-free techniques. It is evident that SingDiffusion significantly outperforms the existing stable diffusion methods in both FID and CLIP scores. These results highlight the yet-to-be-utilized fitting potential in current stable diffusion models, emphasizing the significance of sampling at the initial singular time step.

Furthermore, inspired by Imagen [35], we plot CLIP v.s. FID Pareto curves by varying guidance values within the range [1.5, 2, 3, 4, 5, 6, 7, 8] in Fig. 4. SingDiffusion exhibits substantial improvements over stable diffusion models, especially noticeable with smaller guidance scales. As the guidance scale increases, SingDiffusion consistently maintains a lower FID compared to stable diffusion for achieving a similar CLIP score. This emphasizes that our approach not only enhances image realism but also ensures better adherence to the input prompts.

4.4. Plug-and-play on other pre-trained models

Since our method is extensively trained on the Laion dataset, it can be easily integrated into the majority of existing diffusion models. To validate this, we download several pre-trained models from the CIVITAI website, each specializing in different image domains like anime, animals, clothing, and so on. To facilitate the application of SingDiffusion to these models, we integrate SingDiffusion into the popular Stable Diffusion Web UI [2] system, and sample images with the Euler ancient method after using our initial singularity sampling process. The outcomes, displayed

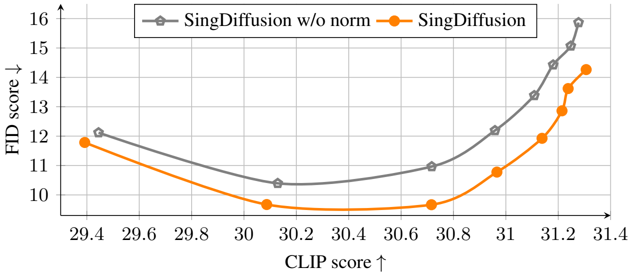

Figure 6. Ablation study of guidance normalization on SD-1.5 across various guidance scales in [1.5, 2, 3, 4, 5, 6, 7, 8].

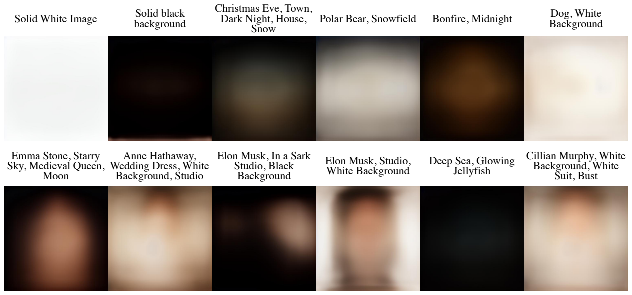
Solid White Image
Solid black background
Christmas Eve, Town, Dark Night, House, Snow
Polar Bear, Snowfield
Bonfire, Midnight
Dog, White Background
Emma Stone, Starry Sky, Medieval Queen, Moon
Anne Hathaway, Wedding Dress, White Background, Studio
Elon Musk, In a Sark Studio, Black Background
Elon Musk, Studio, White Background
Deep Sea, Glowing Jellyfish
Cillian Murphy, White Background, White Suit, Bust

Figure 7. Visualization of $\bar{y}_\theta$ for various prompts.

in Fig. 1 and Fig. 5, showcase SingDiffusion effectively resolving the average brightness issue across all models while preserving their original generative capacities. This demonstrates the adaptability and practicality of our approach.

4.5. Effect on classifier guidance normalization

We conduct an ablation study on the guidance normalization operation (Eq. 13), and represent the CLIP v.s. FID Pareto curves in Fig. 6. It can be seen that the method without normalization exhibits inferior results compared with the full method. As the guidance scale increases, the gap in FID between these two methods gradually widens. This phenomenon primarily arises from significant disparities between $\bar{y}_{\theta_{pos}}$ and $\bar{y}_{\theta_{neg}}$. According to Eq. 13, larger guidance scales may lead to overflow issues. In contrast, normalizing the results with the guidance scale helps keep the outputs within a typical range, thus restoring the FID score.

4.6. Visualization of the $\bar{y}_\theta$

According to Eq. 10, our main goal with $\bar{y}_\theta$ is to model the average image corresponding to each prompt. To confirm this, we employ the prompts presented in Fig. 1 and visualize their corresponding $\bar{y}_\theta$. As demonstrated in Fig. 7, it is clear that $\bar{y}_\theta$ does indeed represent a smoothed average image. For instance, the first four images of the second row resemble average faces, aligning with the input prompt for generating celebrities. Additionally, we notice that $\bar{y}_\theta$ adapts to the prompt's brightness suggestion, appearing brighter or darker accordingly. This highlights our method's ability to effectively capture pertinent lighting conditions, and underscores the significance of the initial singular time step sampling.

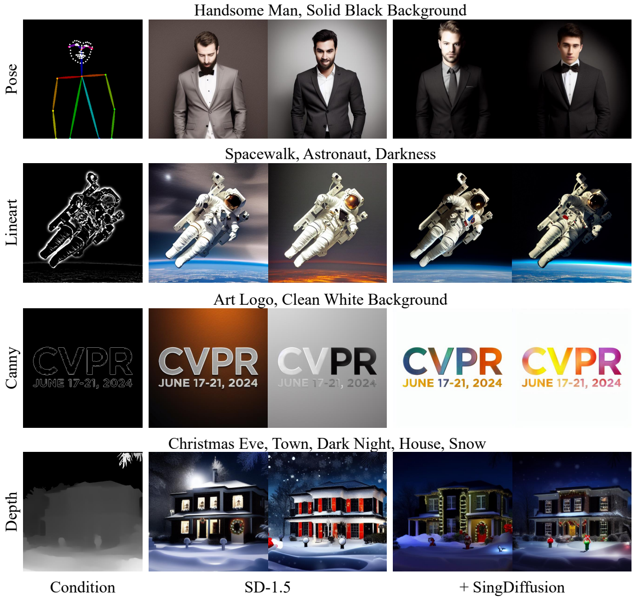
Handsome Man, Solid Black Background
Pose
Spacewalk, Astronaut, Darkness
Lineart
Art Logo, Clean White Background
Canny
CVPR
JUNE 17-21, 2024
Christmas Eve, Town, Dark Night, House, Snow
Depth
Condition
SD-1.5
+ SingDiffusion

Figure 8. SingDiffusion integrates seamlessly with ControlNet.

4.7. Application on ControlNet

Our model can seamlessly integrate with existing diffusion model plugins, such as ControlNet [44]. Since our $\bar{y}_\theta$ is structurally different from ControlNet, ControlNet is adopted after our initial sampling step. The results in Fig. 8 demonstrate that our method is fully compatible with ControlNet and effectively resolves its average brightness issue.

5. Conclusion

In this study, we delve into the singularities of time intervals in diffusion models, exploring both theoretical and practical aspects. Firstly, we demonstrate that at both regular and singular time steps, the reverse process can be approximated by a Gaussian distribution. Leveraging these theoretical insights, we conduct an in-depth analysis of singular time step sampling and propose a theoretical solution. Finally, we introduce a novel plug-and-play method SingDiffusion, addressing the initial time-step sampling challenge. Remarkably, this module substantially mitigates the average brightness issue prevalent in most current diffusion models, and also enhances their generative capabilities.

Acknowledgments. This project is supported by the National Natural Science Foundation of China (62072482, U22A2095)

References

[1] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, pages 313–326, 1982. 2

[2] AUTOMATIC1111. Stable diffusion web ui. https://github.com/AUTOMATIC1111/stable-diffusion-webui, 2022. 7

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 3

[4] Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021. 2, 3

[6] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 3

[9] Nicholas Guttenberg. Diffusion with offset noise. https://www.crosslabs.org/blog/diffusion-with-offset-noise, 2023. 2, 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 7

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

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 2, 3, 4

[14] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. 2

[15] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 26565–26577, 2022. 3

[16] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, 2023. 3

[17] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21696–21707, 2021. 3

[18] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[19] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 2, 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 2, 7

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[22] Yubin Lu, Zhongjian Wang, and Guillaume Bal. Mathematical analysis of singularities in the diffusion model under the submanifold assumption. *arXiv preprint arXiv:2301.07882*, 2023. 3

[23] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision (ECCV)*, page 715–732, 2020. 3

[24] David McAllester. On the mathematics of diffusion models. *arXiv preprint arXiv:2301.11108*, 2023. 2

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[26] Alexander Quinn 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. In *International Conference on Machine Learning (ICML)*, pages 16784–16804, 2022. 2, 3, 5

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[28] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Confer-*

ence on Machine Learning (ICML), pages 8599–8608, 2021. 2

[29] 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 (ICML)*, pages 8748–8763, 2021. 2, 7

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831, 2021. 2, 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 5

[33] 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 (CVPR)*, pages 22500–22510, 2023. 3

[34] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference Proceedings*, 2022. 3

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 36479–36494, 2022. 2, 3, 7

[36] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 25278–25294, 2022. 6

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015. 2, 3, 4

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[40] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 4

[41] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[42] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 2

[43] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18381–18391, 2023. 3

[44] 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 (ICCV)*, pages 3836–3847, 2023. 3, 8

[45] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722, 2022. 3

[46] Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Formulating discrete probability flow through optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2