TCFG: Truncated Classifier-Free Guidance for Efficient and Scalable Text-to-Image Acceleration

Xiaomeng Fu[12*], Jia Li[3†*]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]The University of Hong Kong
fuxiaomeng@iie.ac.cn lijia@eee.hku.hk

Abstract

Diffusion models have achieved remarkable success in image and video generation due to their powerful generative capabilities. However, they suffer from slow inference speed and high computational costs. Existing acceleration methods for diffusion models may compromise model performance and struggle to generalize across diverse diffusion model architectures and downstream tasks. To address these issues, we propose a model-agnostic and highly scalable acceleration strategy for text-controlled image generation. Specifically, we dynamically modulate the text guidance coefficience and truncate redundant text-related computations during the denoising process. Experimental results demonstrate that our approach achieves significant model acceleration while preserving precise text-image alignment, showcasing the potential for a wide range of diffusion models and downstream applications. We will release the code upon acceptance.

1. Introduction

Nowadays, Diffusion Models (DMs) [17, 48, 49] have gained significant popularity in both academia and industry, and have dominated diverse applications [2, 3, 13, 18, 28, 37, 39, 44, 58] due to their superior capabilities to generate high-quality images. Especially, text-to-image generation emerges as a particularly prominent application, significantly advancing the development of Artificial Intelligence Generated Content (AIGC). By leveraging billions of parameters and extensive training datasets, these large-scale pretrained text-to-image diffusion models demonstrate robust generative priors, thereby facilitating numerous downstream vision tasks [10, 11, 36, 42, 53, 56].

Despite the remarkable capabilities of text-to-image diffusion models, their relatively slow inference speed remains a significant obstacle to widespread implementation. Two primary factors hinder the inference speed of these models. Firstly, in contrast to other one-step generative models such as Variational Autoencoders (VAEs) [24] and Generative Adversarial Networks (GANs) [12], diffusion models necessitate a series of iterative steps to reconstruct initial noise into a coherent image. Secondly, to achieve enhanced generative capabilities, existing diffusion models are increasing in size, resulting in slower speed for each denoising step. To address these speed and efficiency challenges, various optimization strategies have been proposed, which can be broadly classified into two categories: inference sampling optimization and model structural optimization.

The former [20, 32, 33, 45, 46, 50, 57] seek to develop more efficient samplers to minimize the number of denoising steps, based on advanced stochastic differential equation (SDE) solvers or various knowledge distillation methods. However, these techniques encounter performance limitations when attempting to compress the process into very few steps, as the sharp reduction in sampling steps can alter the convergence dynamics of the denoising process.

The latter methods [5, 6, 9, 23, 34, 54, 55] focus on more efficient denoising networks, with techniques including quantization, pruning, feature caching, etc. However, these methods usually modify the internal structure and computation flow of a given model, making its specific strategy not directly applicable across various diffusion architectures and limiting the adaptability for downstream tasks that rely on certain features or layers.

Given the limitations of existing acceleration methods and the widespread adoption of Classifier Free Guidance (CFG) in modern diffusion models, we focus on exploring the potential optimization within CFG to develop a universal acceleration strategy for diffusion models.

Our method begins with a comprehensive analysis of the role of text input within CFG. We demonstrate that the influence of text input diminishes once image semantics are

*Equal contribution, †Corresponding author

well-established. This occurs because images with rich semantics provide more precise conditioning signals compared to textual descriptions, resulting in model outputs independent of the input text. Based on this observation, we conclude that text-based guidance becomes redundant after semantic convergence. Consequently, we propose Truncated CFG (TCFG), which estimates semantic convergence and accordingly terminates text-related computations.

To further accelerate the inference, we expedite semantic convergence to enable earlier truncation of text-related computations. To achieve this, we reformulate CFG as a gradient descent optimizer with the guidance scale serving as the learning rate. Inspired by learning rate decay and warmup strategies commonly used in gradient-based optimization, we adapt these techniques to the CFG framework to dynamically adjust the guidance scale. This promotes both faster semantic convergence and more plausible generated semantics.

We evaluate our approach through extensive experiments on pretrained text-to-image diffusion models and their downstream applications. Experimental results demonstrate superior performance compared to existing acceleration methods. Additionally, our approach exhibits seamless compatibility with diffusion-based downstream tasks while achieving notable improvements in text-image alignment fidelity. In summary, the primary contributions of this work are as follows:

- We conduct a comprehensive investigation into the impact of text input along the denoising process and demonstrate that the control of text inputs becomes negligible once the image semantics stabilize. Based on this insight, we propose TCFG, a straightforward yet efficient method for accelerating diffusion models.
- We reinterpret CFG as a gradient descent optimizer, providing a novel insight into the underlying mechanisms of CFG. Furthermore, we introduce an adaptive guidance scale strategy incorporating techniques of decay and warmup, which brings improvements in both image generation quality and text-image alignment.
- We perform comprehensive experiments on pretrained text-to-image diffusion models and their downstream applications, demonstrating the effectiveness of our proposed method. Notably, our approach is model-agnostic and highly scalable, which supports seamless integration in various downstream tasks, exhibiting superior applicability and versatility.

2. Related Works

2.1. Efficient Diffusion Models

Due to space constraints, we provide a concise overview of contemporary research on efficient diffusion models. Current approaches fall into two primary categories: **inference sampling optimization** and **model structural optimization**.

Inference sampling optimization focuses on minimizing the number of iterations required during the inference stage. These strategies can be further divided into training-free and training-based methods. Training-free methods [20, 33, 50, 57, 59] analyze the diffusion models from the lens of stochastic differential equations (SDEs) or Probability Flow Ordinary Differential Equations (PF-ODEs), accelerating the sampling process by faster equation solvers. Training based methods [8, 32, 45, 46] employ knowledge distillation techniques, transferring learned behavior from multi-step teacher models to few-step student architectures. Despite achieving notable improvements, inference sampling optimization encounters limitation when attempting to sample with a minimal number of steps. Reducing sampling steps will disrupt the convergence dynamics and potentially leading to corrupted images.

Model structural optimization seeks to develop computationally efficient denoising networks via architectural modifications. These methods can also be divided into training-based and training-free methods. Training-based methods [6, 9, 23, 55] directly reduce the parameters within the denoising net. Training-free methods [5, 34, 54] reduces computational redundancy via dynamic inference mechanisms like token merging or feature reuse. However, model structural optimization is not conducive to downstream tasks, as they manipulate the inner structure of the denoising network, which may conflict with the modification required by the downstream tasks.

2.2. Classifier Free Guidance

Several studies [21, 29, 47] have discussed the properties of the classifier free guidance process. Lin et al. [29] observe that the images generated with a large guidance scale exhibit excessive magnitude, leading to over-exposure. To mitigate this issue, they propose rescaling the image in each iteration of the denoising process. Shen et al. [47] identify spatial inconsistency in the CFG process and propose assigning different guidance scales to distinct objects. Karras et al. [21] interpret CFG as an extrapolation between conditional and unconditional scores and reveal that its underlying mechanism involves using a suboptimal score to direct the generation process towards more plausible outputs. To the best of our knowledge, we are the first work introducing a novel perspective of Classifier-Free Guidance through gradient descent theory. Building on this insight, we develop a decay and warmup strategy that significantly improves CFG's convergence properties.

3. Preliminary

To introduce TCFG, it is necessary to understand the foundational principles of Diffusion Models and CFG first. Due

to space limit, we provide a brief introduction to diffusion models and CFG here.

Diffusion Models. Diffusion models [17, 48, 51] learn to invert a predefined forward noising process. For Denoising Diffusion Probabilistic Models (DDPM), the forward process gradually adds Gaussian noise to data x_0 over T steps until $q(x_T) \sim \mathcal{N}(0, \mathbb{I})$:

$$x_{t+1} = \sqrt{1-\beta_t}x_t + \sqrt{\beta_t}\epsilon \quad \epsilon \in \mathcal{N}(0, \mathbb{I}) \tag{1}$$

Sampling iteratively (by a denoising network) reverses this process via $p_t(x_{t-1}|x_t;\theta) \sim \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \sigma_t^2)$, where:

$$\mu_\theta = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t)) \tag{2}$$

Classifier Free Guidance. Classifier Free Guidance [16] combines conditional (c) and unconditional ($\varnothing$) diffusion models during inference for text-aligned generation:

$$\hat{\epsilon}_\theta^{\text{CFG}} = \epsilon_\theta(x_t, \varnothing) + \lambda \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing)) \tag{3}$$

where λ balances image quality and text alignment.

4. Method

Our research aims to develop a model-agnostic acceleration strategy while maintaining generation quality. The core idea is to identify and eliminate text-related computations that negligibly impact the final quality of generated images. Through analysis of the denoising process, particularly the temporal influence of text inputs, we observe that text conditioning becomes largely redundant once visual semantics are well-established. Based on this observation, we propose Truncated CFG (TCFG), which divides the denoising process into several time phrases based on semantic establishment status and drops the unnecessary text computations in CFG. To further accelerate the process, we reframe CFG as a gradient descent optimizer, incorporating learning rate decay and warmup techniques to facilitate faster semantic convergence.

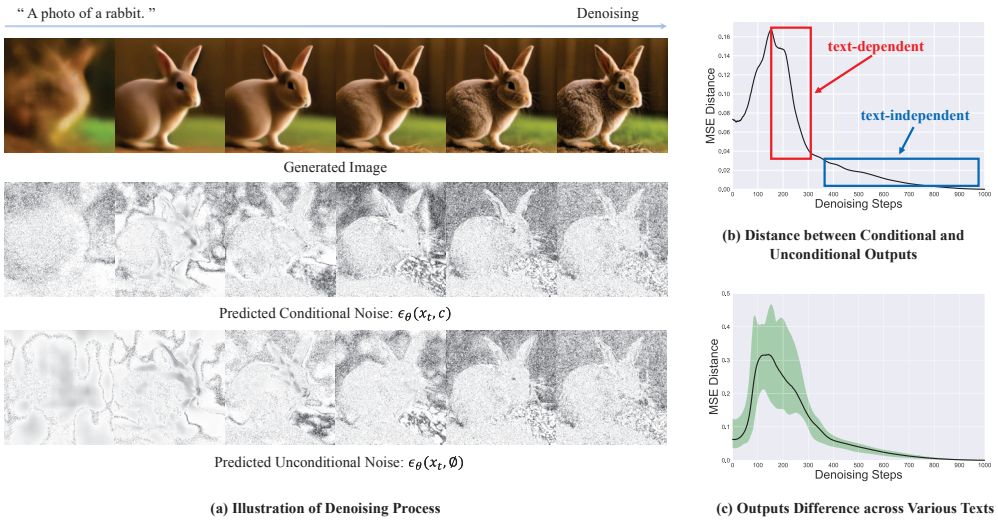
4.1. Analysis on Denoising Process

Recent studies [7, 26, 35, 38] have revealed that the denoising process of the diffusion models comprises two distinct stages: the **semantics planning stage**, which involves searching the latent space to establish spatial information representing semantic structure, and the **fidelity improving stage**, which then enhances to regional details of the elements to improve the overall visual fidelity. Fig. 1(a) illustrates the denoising process for the prompt "a photo of a rabbit". It can be observed that: 1) Diffusion models initially generate the geometric contour of the rabbit, followed by the addition of finer details, which aligns with the findings of previous research. 2) The predicted noise, with ($\epsilon_\theta(x_t, c)$, shown in the 2nd row) and without ($\epsilon_\theta(x_t, \varnothing)$, shown in the 3rd row) the input text, exhibits significant variation at the beginning of the denoising process but converges to similar patterns thereafter. Based on the observation, we hypothesize that **the text input primarily influences the semantics planning stage and have limited effectiveness in the fidelity improving stage.**

To validate the hypothesis, we conducted quantitative analysis experiments on outputs under different text controls. As shown in Fig. 1(b), we calculated the distance between outputs with and without text guidance. The differences swiftly diminish with increasing timesteps, indicating a significant reduction in the correlation between outputs and the textual inputs. Additionally, as demonstrated in Fig. 1(c), we investigated and visualized the output differences under various text input conditions. The green area represents output variance when using semantically similar inputs(e.g. 'a photo of a gray rabbit' vs 'a photo of a white rabbit') or dissimilar text inputs (e.g. 'a photo of a gray rabbit' vs 'a girl on Mars'). It can be observed that the impact of text on the output also significantly decreases during the denoising process. These results indicate that the distance converges during the fidelity improvement stage, suggesting that outputs are relatively insensitive to input texts, regardless of the content. This observation is intuitive: accurate conditions generally prevails over ambiguous conditions when conditions conflict, as models tend to favor outputs with higher probability. In this context, accurate conditions provide more precise guidance and correspond to outputs with higher probability. Specifically, the denoising U-Net considers both the visual input x_t and the text input c as conditions. In the fidelity improving stage, the semantics of the visual input are well-established, providing more precise guidance compared to the potentially ambiguous text input. When text inputs conflict with visual information, the U-Net preferentially follows the more accurate visual guidance, which correlates with higher-probability outputs, effectively superseding the text input's influence.

4.2. Truncated CFG (TCFG)

Based on the above observations, we explicitly divide the denoising process into two time range: the classical CFG time range $[t', T]$, which shape the semantics according to the input text, and the truncated time range $[0, t')$, where text-unrelated finer details are added based on the well-established semantics. Within the truncated time range, the outputs of the U-Net, with or without the text input exhibit minimal variance, leading to significant computational redundancy. Consequently, we drop the prediction for $\epsilon_\theta(x_t, c)$ in truncated time range due to its diminished guid-



(a) Illustration of Denoising Process

(b) Distance between Conditional and Unconditional Outputs

(c) Outputs Difference across Various Texts

Figure 1. (a) Illusion of the denoising process (1st row). We also represent the conditional noise and unconditional noise across the denoising process (2nd and 3rd row), respectively. (b) The distance between conditional and unconditional outputs throughout the denoising process. (c) The distance between outputs with various textual inputs.

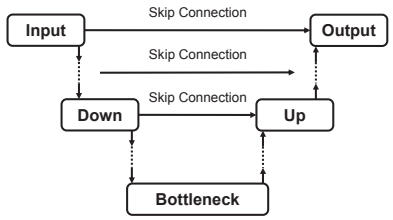
ance accuracy compared to $\epsilon_\theta(x_t, \varnothing)$ and propose TCFG:

$$\hat{\epsilon}_\theta^{\text{TCFG}} = \begin{cases} \hat{\epsilon}_\theta^{\text{CFG}}, & t \in [t', T] \\ \epsilon_\theta(x_t, \varnothing), & t \in [0, t') \end{cases} \tag{4}$$

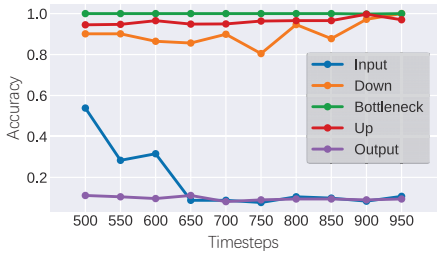
where t' is the dividing time. Note that an increased t' results in an earlier truncated time range, making a greater number of predictions for $\epsilon_\theta(x_t, c)$ to be discarded, thereby enhancing the efficiency of the sampling process. Conversely, late t' corresponds to short truncated time range, leading to more redundant sampling process.

The dividing time t'. For TCFG, the dividing time is crucial as it delineates the truncated time range and subsequently impacts the ultimate outputs. However, choosing an appropriate dividing time is challenging as the optimal timing varies with each sampling process and corresponds to the complexity of the generated content. Complex semantics requires more time to establish and vice versa. An ideal dividing time should exhibit the following attributes: Firstly, the geometry of the corresponding visual condition $x_{t'}$ should be sufficiently well-defined to overwhelm the text condition. Secondly, it should be as early as possible to reduce text-related computations. Therefore, the fundamental objective in estimating the dividing time is to determine the precise timestep when the semantics are fully established. A straightforward approach would be to directly utilize the distance between $\epsilon_\theta(x_t, c)$ and $\epsilon_\theta(x_t, \varnothing)$ as an indicator for semantics (similar to that in Fig. 1(a)). However, this method often results in a late t'. Specifically, due to the predefined noise schedule inherent in diffusion models, the visual input x_t is initially dominated by the injected noise at the early stage of the denoising process, making the final outputs fragile and unstable. Consequently, even when the semantics have converged, a substantial distance caused by the dominated injected noise may persist. The distance only converges after the noise injected in adjacent steps neutralizes, thereby leading to a late t'.

To explore which feature can effectively represent the semantics and serves as a more reliable indicator, we conduct a probing experiment. We extract internal representations from various positions within the denoising U-Net backbone (including Input, Output, Downsample, Upsample and Bottleneck blocks) and employ these representations for classification tasks. The results are shown in Figure 2(b). It can be observed that features of the input and output within the U-Net show very limited discriminative ability in the early stage of the denoising process, achieving accuracy comparable to the nearly random level. This phenomenon can be attributed to severe input perturbation by noise, which is directly propagated to the output through skip connections. The presence of noise in x_t significantly reduces the discriminative capability of both input and output features. The bottleneck features, however, achieve optimal performance due to their maximal distance from the input and absence of skip connections. Based on this experimental evidence, we employ the bottleneck features as the semantic indicator and measure the distance between bottleneck features with and without text input to capture the



(a) Architecture of the Denoising U-net



(b) Results of the Probing Experiment

Figure 2. (a) We extract features from the denoising U-Net for classification tasks to determine which features can best represent the semantics of the generated images. (b) The accuracy of different features, a higher accuracy means corresponding features contain more semantics that benefit the classification task.

convergence of semantics:

$$d = ||h(c) - h(\varnothing)||_2^2 \tag{5}$$

where $h(c)$ and $h(\varnothing)$ are bottleneck features with/without text input, respectively. For simplicity we omit the timestep subscript t. To achieve more precise measurement of semantic convergence, we further compute the variation in distance between adjacent timesteps, which measures the rate of semantic convergence. Initially, the convergence speed is rapid; however, it decelerates as the semantics are established. We identified the deceleration timestep as the temporal division using an off-the-shelf online changepoint detection method [1]:

$$\Delta d_t = d_t - d_{t-1} \tag{6}$$

$$\mathcal{M}(\Delta d_{1:t}) = \mathbf{1}[p(r_t|\Delta d_{1:t-1}, \Delta d_t) > \tau] \tag{7}$$

where $\mathcal{M}$ is the online changepoint detection model, r_t is the possibility that current timestep t is a changepoint. Eq. 7 indicates that if the probability exceeds the threshold τ, we consider the current time t as the changepoint and assign it as the dividing time.

4.3. Rescheduling the CFG Process

We propose TCFG, which eliminates the prediction for $\epsilon_\theta(x_t, c)$ once visual semantics are adequately established.

In this section, we focus on accelerating semantic establishment by rescheduling the guidance scale λ, aiming to achieve a more efficient sampling process. An intuitive idea to accelerate the semantic establishment is to employ a larger guidance scale factor λ. However, directly scaling typically results in oversaturated outputs with degraded global structure [43]. To address this challenge, we conduct an in-depth analysis of the CFG process. We reformulate the CFG process from the lens of gradient descent, where the guidance scale λ can be interpreted as the learning rate. Subsequently, we reschedule the guidance scale by incorporating learning rate decay and learning rate warmup techniques that are widely utilized in gradient-based model training.

CFG as a Gradient Descent Optimizer. In CFG, each denoising step ($x_t \rightarrow x_{t-1}$) comprises two subprocesses: the first subprocess estimates the clean image using the predicted CFG noise (Eq. 3), while the second subprocess introduces noise to the estimated clean image to ensure alignment with the noise level at the current timestep. Notably, the second subprocess is a predefined non-parametric operation that does not impact the final output. Consequently, the estimated clean image is the primary focus of our attention. This estimated clean image can be expressed as follows:

$$\begin{aligned} \hat{x}_0 &= x_0(\varnothing) + \lambda \cdot (x_0(c) - x_0(\varnothing)) \\ x_0(\varnothing) &= \mathbb{E}[x_0|x_t, \varnothing] \quad x_0(c) = \mathbb{E}[x_0|x_t, c] \end{aligned} \tag{8}$$

where $x_0(c)$ and $x_0(\varnothing)$ represent predicted clean images with/without the text input, respectively. For brevity, we omit the subscript t. This equation can be further reformulated into a gradient descent process:

$$\begin{aligned} \hat{x}_0 &= x_0(\varnothing) - \eta \nabla_{x_0(\varnothing)} \mathcal{L}(x_0(\varnothing)) \\ \mathcal{L}(x_0(\varnothing)) &= ||x_0(\varnothing) - x_0(c)||_2^2 \quad \eta = \lambda/2 \end{aligned} \tag{9}$$

This reformulation reveals that during the sampling process, the classical CFG functions as a gradient descent optimizer, aiming to align the unconditional predicted clean image $x_0(\varnothing)$ with the conditional clean image $x_0(c)$. The guidance scale λ, served as the learning rate, is constant for the classical CFG. In the following, we propose to integrate the learning rate decay and learning rate warmup strategies to accelerate the convergence of the optimization process. By doing so, we aim to accelerate the convergence of the semantics and thereby achieve a more efficient sampling process.

Learning Rate Decay. Initially, we adopt a large learning rate to facilitate rapid exploration of the semantic space at the beginning of the denoising process. As the denoising process advances and improved text alignment is achieved, we transition to a smaller learning rate to refine this alignment and ensure stability. Besides, compared to the classical CFG, we increase the initial guidance scale to

$C\lambda_0 (C > 1)$ and apply a decay factor γ every k steps. The learning rate decay can be expressed as:

$$\lambda(t) = \max(C\lambda_0 \gamma^{\lfloor \frac{t}{k} \rfloor}, \lambda_1) \tag{10}$$

where λ_1 serves as a lower bound to ensure the efficacy of the guidance throughout the process.

Learning Rate Warmup. Warmup represents a strategy where a smaller learning rate is utilized during the initial phase of training. During the initial period, the model's weights and parameters have not yet fully adapted to the underlying data distribution. The application of a smaller learning rate during the warmup phase facilitates a smooth transition, enabling the model to gradually adjust and stablize before transitioning to higher learning rates [19, 31]. Similarly, we incorporate a warmup period within the denoising process, which offers several notable advantages: Firstly, it enables the predicted clean image to diverge from the initial standard Gaussian noise and converge towards the image manifold, thereby establishing a robust foundation for subsequent text alignment process. Secondly, it supports the use of the higher learning rates without causing the estimated clean image to deviate from the image manifold, which is a primary cause of the corrupted results. Thirdly, it accelerates the establishment of semantics, thereby advancing the truncated time range. In practice, we set the guidance scale $\lambda = 1$ during the warmup period, as it further omits the predictions of the $x_0(\varnothing)$.

4.4. Rescheduled TCFG

In summary, we introduce the following modifications to TCFG to further accelerate the establishment of semantics. First, we incorporate a warmup period during the initial denoising steps, which facilitates the use of a higher learning rate in subsequent denoising process. Second, we increase the initial learning rate and implement a gradual decay as the denoising process progresses. Both these two adjustments contribute to advancing the truncated time range, thereby further reducing the number of required queries. The proposed TCFG can be implemented by adjusting the guidance scale λ:

$$\lambda(t) = \begin{cases} 1, & t \in [t_w, T] \quad \%\text{Warmup} \\ \text{Eq. 10}, & t \in [t', t_w) \quad \%\text{CFG with decay} \\ 0, & t \in [0, t') \quad \%\text{Truncate} \end{cases} \tag{11}$$

where t_w is the dividing time for the warmup period.

5. Experiments

5.1. Experimental Setups

Models, Datasets and Baselines. We evaluate our method on the widely utilized SD 1.5 and SD 2 [41]. Additionally, we assess our method on models specifically designed for downstream tasks, namely DreamBooth [42] for customization, ScaleCrafter [14] for high-resolution image generation, and ControlNet [56] for image-to-image translation. Following prior work, we utilize the MS-COCO [30] validation set as the text inputs to evaluate text-to-image generation. We also evaluate our method on Pick-a-Pic [25] and HPD v2 [52] to further demonstrate the effectiveness of our method. For baselines, we consider two categories: more efficient samplers, including DDIM [49] with fewer steps and DPM-Solver [33], and more efficient networks, such as DeepCache [34] and TokenMerge [5].

Evaluation Metrics. In our study, we evaluate our method across four distinct yet crucial dimensions: 1) Computational Efficiency, measured by the Number of Function Evaluations (NFE), latency and overall time speedup. 2) Set-Level Image Quality, utilizing the Frechet Inception Distance (FID) [15] and Kernel Inception Distance (KID) [4] to quantify the proximity of the generated images to the real image distribution. 3) Sample-Level Image Quality, employing no-reference Image Quality Assessment methods, namely MUSIQ [22] and LAION-Aes [27]. Besides, we also leverage CLIPScore [40] to evaluate the textual alignment. 4) Human Preference, leveraging HPSv2 [52] to evaluate the alignment of the generated images with human visual perception.

Implementation Details. All experiments are conducted using a single NVIDIA V100 GPU. For all baseline methods, the classifier-free guidance scale λ_0 is set to 7.0. Our method accelerates the image generation process through the rescheduling of the classifier-free guidance scale, involving hyperparameters t_w for the warmup period and $\{C, \gamma, k, \lambda_1\}$ for the learning rate decay. Specifically, t_w is set to 900 with C, γ, k, λ_1 values of 1.5, 0.8, 5 and 4.0 separately. This configuration implies that the first 100 timesteps are reserved for warmup and an initial guidance scale enlargement by 1.5 times, followed by a decay of 0.8 for every 5 denoising steps, ensuring that the guidance scale remains no less than 4.0.

5.2. Performance on Text-to-Image Models

Quantitative Results. The NFE of our method varies across individual samples due to the truncation of the CFG based on the varying convergence of semantic establishment among different samples. Here, we report the average NFE across all images here (100 NFEs per sample decreased to 68.76 NFEs per sample). To ensure a fair comparison, we select efficient samplers with comparable NFEs (70 NFEs v.s. our 68.76 NFEs). The performance of our method is presented in Table 1. Notably, our method achieves a 1.7× acceleration without any degradation in set-level image quality, as evidenced by FID score from 65.18 to 66.43 (a negligible change) and the KID metric remain-

Table 1. The performance on Text-To-Image Models (SD 1.5) on MS-COCO.

	NFE	Latency(s)↓	Speedup↑	FID↓	KID↓	MUSIQ↑	LAION-Aes↑	HPSv2↑	CLIPScore↑
DDIM-50† [49]	100	3.58	1×	65.18	3.02	5.17	5.25	28.31	30.54
DDIM-35 [49]	70	2.70	1.33×	77.86	3.51	5.19	5.22	26.31	30.23
DPM-Solver-35 [33]	70	4.71	0.76×	70.92	3.38	4.91	5.18	27.19	30.31
DeepCache [34]	100	**2.10**	**1.70**×	82.16	3.52	4.85	5.22	27.22	30.22
TokenMerge [5]	100	2.27	1.58×	101.52	3.77	4.48	5.04	24.63	28.87
TCFG (ours)	100→68.76	2.11	**1.70**×	**66.43**	**3.02**	**5.53**	**5.42**	**28.97**	**30.49**

†We leverage DDIM-50 as the sampler for our method and structure optimization baselines (DeepCache and TokenMerge).

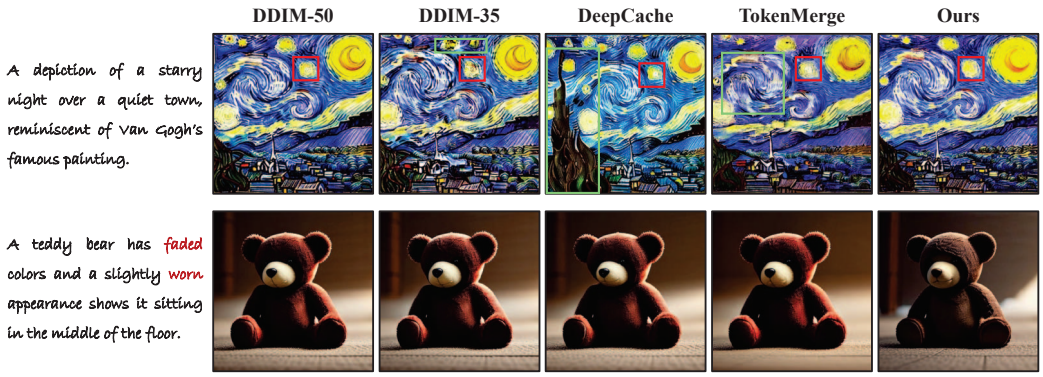


Figure 3. The qualitative results on pretrained text-to-image models (SD v1.5). Our method can generate high-fidelity images compared to the original DDIM-50 while further improve the semantics and image-text alignment of the generated image.

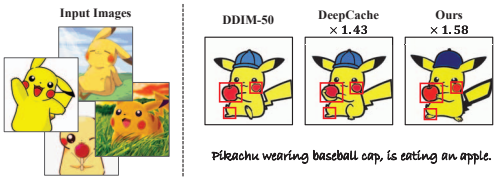


Figure 4. The qualitative results on DreamBooth which is tailored for personal subject generation. Our method can achieve simultaneous acceleration and semantic refinement.

ing constant. In contrast, DeepCache and TokenMerge suffer from significant degradation in both FID and KID. Furthermore, we observe a significant improvement in aesthetic score and human preference, which underscores the superiority of our adaptive guidance scale approach.

Qualitative Results. We present qualitative results in comparision with the baseline models in Figure 3, where DeepCache, TokenMerge and our method leverage the DDIM-50 sampler. Regions demarcated by red rectangles reveal spatial distortions in existing acceleration methods, while green-rectangle annotations highlight distinct flaws (e.g. extra element in DeepCache, local blurring in TokenMerge, artifacts in DDIM-35). Notably, DeepCache exhibits significant color style shift (from purple to blue tones). In contrast, our proposed method demonstrate negligible variance when compared to images generated using the DDIM-50 sampler. Additionally, our method can produce images that are more closely aligned with the input texts due to the rescheduling of the guidance scale. In case of the teddy bear, our method generates the image with a better background with higher fidelity and richer details, which fully reflects the word of "floor" mentioned in the text input. Furthermore, the generated image exhibits a darker hue, which is more consistent with the input text's description of "faded color" and "worn appearance".

5.3. Performance on Downstream Applications

DreamBooth. DreamBooth [42] is intended to extend the language-vision dictionary of the diffusion foundation models, enabling it to associate new vocabulary with specific subjects that users wish to generate, offering a highly personalized and adaptable approach to image generation. We present the results of our method accelerating DreamBooth in Figure 4. Notably, our method demonstrates a substantial acceleration rate while exhibiting the capability to modify semantics to better align with the specific subject (the corrected round blush and more polished feet marked by red rectangles). In contrast, DeepCache, suffering from spatial inconsistency, exhibits disturbed semantics, as evidenced by

Figure 5. The qualitative results on ScaleCrafter which is tailored for high-resolution image synthesis. Our method not only facilitates the acceleration of the DDIM-50 process but also corrects distortions, demonstrating superior semantic convergence.

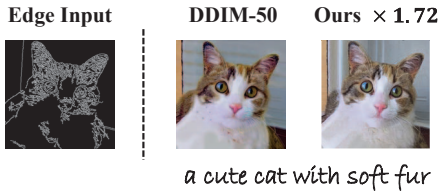
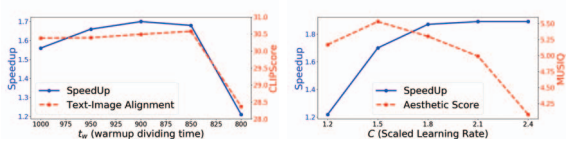


Figure 6. The qualitative results on ControlNet which is designed to incorporate fine-grained control into pretrained text-to-image models.

the apple highlighted by red rectangle.

ScaleCrafter. ScaleCrafter [14] aims to address the challenge of generating high-resolution images using pretrained diffusion models. By broadening the perception field of convolutional kernels within the U-Net, ScaleCrafter achieves high-fidelity high resolution image synthesis in a training-free manner. However, this technique also obstructs the application of current acceleration methods, as both ScaleCrafter and acceleration methods manipulate the internal connection of the U-Net, leading to conflicts in the computational workflow. In contrast, our method seamlessly accelerates ScaleCrafter while achieving superior text-image alignment. Notably, as illustrated in Figure 5, compared to the image generated by DDIM-50, our method further corrects distortions and errors present in the bear's face, such as distorted and poorly defined eyes.

ControlNet. ControlNet [56] is devised to incorporate fine-grained spatial conditional control (such as the edge map) into pre-trained diffusion models. Specifically, this incorporation requires an externally pretrained network for the spatial condition, thereby also limiting the direct applicability of current acceleration methods. In contrast, our method can seamlessly accelerate ControlNet as demonstrated in Figure 6. Besides, we terminate the computation for the extra control net branch at the same time we truncate the CFG as the semantics are already established and the current visual input x_t dominates both the additional spatial control and the text control. Given that the semantics are mainly provided by the spatial input, the semantic development is fast, resulting in an even higher speedup (x1.72).



(a) ablation on the warmup stage. (b) ablation on the scaled learning rate.

Figure 7. Ablation study on (a) the warmup stage t_w and (b) the scaled learning rate C.

5.4. Ablation Study

The warmup stage. t_w is the temporal boundary demarcating the warmup period ($[t_w, T]$). A lower t_w corresponds to an extended warmup period. To evaluate the influence of t_w, we conducted an ablation study. The results are shown in Figure 7(a), which reveals that incorporating a warmup stage facilitates the convergence of the dynamics, compared to no warmup (i.e. $t_w = 1000$). However, over warmup hinders the semantic planning efficacy, inducing text-image misalignment.

The scaled learning rate C. We examined the influence of the scaled learning rate C, which modulates the initial learning rate (i.e. the guidance scale in CFG). As demonstrated in Figure 7(b), elevated C enhances semantic convergence rates, facilitating the semantic construction. Nevertheless, excessively high C induce oversaturation in generated images, thereby degrading image quality (lower MUSIQ).

6. Conclusion

In this paper, we introduce TCFG, a model-agnostic and highly scalable acceleration strategy applicable to both pretrained diffusion models and their downstream applications. We first analyze the influence of text input on the denoising process, observing that its impact diminishes once image semantics stabilize. Based on this observation, we propose TCFG, which truncates text-related computations upon semantic convergence. To maximize acceleration through earlier truncation, we reframe CFG as a gradient descent optimizer, incorporating learning rate decay and warmup strategies to enhance convergence dynamics. Comprehensive experiments across pretrained diffusion models and downstream tasks demonstrate TCFG's dual benefits: accelerated inference and improved alignment between input prompts and generated images.

References

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007. 5

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*. 1

[3] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 1

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 6

[5] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4599–4603, 2023. 1, 2, 6, 7

[6] Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2024. 1, 2

[7] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 3

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2

[9] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 16716–16728, 2023. 1, 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[11] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023. 1

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[13] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 1

[14] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 6, 8

[15] 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*, 30, 2017. 6

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

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[19] Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? underlying mechanisms and improvements. *arXiv preprint arXiv:2406.09405*, 2024. 6

[20] Tero Karras, Miika Aittala, Samuli Laine, and Timo Aila. Elucidating the design space of diffusion-based generative models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 26565–26577, 2022. 1, 2

[21] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024. 2

[22] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6

[23] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: Architecturally compressed stable diffusion for efficient text-to-image generation. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023. 1, 2

[24] Diederik P Kingma and Max Welling. Auto-encoding variational {Bayes}. In *The Second International Conference on Learning Representations*, 2014. 1

[25] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 6

[26] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[27] LAION-AI. aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor, 2022. 6

[28] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1

[29] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 2

[30] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[31] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020. 6

[32] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 1, 2, 6, 7

[34] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024. 1, 2, 6, 7

[35] 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*, 2022. 3

[36] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[37] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 1

[38] Ji-Hoon Park, Yeong-Joon Ju, and Seong-Whan Lee. Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications*, 248:123231, 2024. 3

[39] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[41] 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. 6

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 6, 7

[43] 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. 5

[44] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 1

[45] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 1, 2

[46] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025. 1, 2

[47] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379, 2024. 2

[48] 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, 3

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 6, 7

[50] 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*, 2020. 1, 2

[51] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021. 3

[52] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score

v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 6

[53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

[54] Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattn: Attention compression for diffusion transformer models. *arXiv preprint arXiv:2406.08552*, 2024. 1, 2

[55] Dingkun Zhang, Sijia Li, Chen Chen, Qingsong Xie, and Haonan Lu. Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models. *arXiv preprint arXiv:2404.11098*, 2024. 1, 2

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 6, 8

[57] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2024. 1, 2

[58] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022. 1

[59] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2