

# SwiftBrush : One-Step Text-to-Image Diffusion Model with Variational Score Distillation

Thuan Hoang Nguyen Anh Tran

VinAI Research, Hanoi, Vietnam

{v.thuannh5, v.anhtt152}@vinai.io



Figure 1. Samples generated in a **single step** by SwiftBrush. Our distillation method can produce images of high-fidelity roughly **20 times** faster compared to Stable Diffusion.

## Abstract

*Despite their ability to generate high-resolution and diverse images from text prompts, text-to-image diffusion models often suffer from slow iterative sampling processes. Model distillation is one of the most effective directions to accelerate these models. However, previous distillation*

*methods fail to retain the generation quality while requiring a significant amount of images for training, either from real data or synthetically generated by the teacher model. In response to this limitation, we present a novel image-free distillation scheme named **SwiftBrush**. Drawing inspiration from text-to-3D synthesis, in which a 3D neural radiance field that aligns with the input prompt can be obtained*

*from a 2D text-to-image diffusion prior via a specialized loss without the use of any 3D data ground-truth, our approach re-purposes that same loss for distilling a pretrained multi-step text-to-image model to a student network that can generate high-fidelity images with just a single inference step. In spite of its simplicity, our model stands as one of the first one-step text-to-image generators that can produce images of comparable quality to Stable Diffusion without reliance on any training image data. Remarkably, SwiftBrush achieves an FID score of **16.67** and a CLIP score of **0.29** on the COCO-30K benchmark, achieving competitive results or even substantially surpassing existing state-of-the-art distillation techniques.*

## 1. Introduction

Diffusion models are gaining significant attention from the research community as they have achieved remarkable results in various generative tasks, including image generation [9, 24, 38, 39], video synthesis [10], 3D modeling [3, 26, 44], audio generation [15], and text creation [6, 16]. Especially, text-to-image diffusion models, which combine the power of language models and high-quality diffusion models [30, 34, 41], have revolutionized the way that people create visual content based on textual descriptions. Synthetic images indistinguishable from real photos are now possible with only a few clicks. Additionally, with techniques like ControlNet [47] or DreamBooth [33], extra control to the generation outputs can be made to further boost the expressiveness of the text-to-image models and bring them closer to mass users. However, despite their impressive success, these models exhibit a noteworthy drawback of slow inference speed. The primary reason for this lies in the iterative nature of their sampling, presenting significant challenges for deployment on consumer devices.

Several approaches have been proposed to improve the inference speed of diffusion-based text-to-image generation. However, the most effective direction is time-step distillation, which aims to reduce the number of sampling steps of the diffusion model with the network backbone unchanged and minimum reduction in generation quality. Starting from the pioneering work [35], more advanced and effective distillation methods have been proposed, and a few recent ones could train efficient student networks that only require one forward step to synthesize images from input captions. In this work, we are interested in the distillation-based approach for one-step text-to-image generation. Particularly, Meng et al. [23] employ a two-stage distillation, where the student tries to first match its output with the classifier-free guided output of the teacher and then bootstrap itself to reduce the inference steps gradually. Such a method is complex and time-consuming, requiring at least 108 A100 GPU days [19]. Based on the work of Song et

al. [40], Latent Consistency Model (LCM) [22] enforces the consistency of all points on the ODE trajectory in the latent space of Stable Diffusion [41] and unlike [23], they directly combine the two distillation stages into one. Although both Guided Distillation and LCM significantly reduce the number of inference steps to 2 or 4, their 1-step inference yields blurry and unsatisfying results. More sophisticated, Instaflow [19] proposes a technique called “reflow” to straighten the sampling trajectory of the teacher before distilling. While their proposed techniques proven to be superior to the original diffusion framework in one-step text-to-image distillation, InstaFlow requires an expensive 4-stage training schedule. Also, it is noteworthy that the effectiveness of these methods relies heavily on a large number of text-image pairs, which is not always the case due to the limited accessibility of such data.

In this work, we aim to develop a novel distillation method for one-step text-to-image generation with higher quality and a more approachable training process characterized by an image-free and straightforward mechanism. Building upon the recent advancements in text-to-3D techniques, our work takes inspiration from their ability to generate high-quality 3D Neural Radiance Fields (NeRF) without utilizing 3D ground truth data. Such a remarkable achievement is facilitated by employing a powerful pre-trained 2D text-to-image model to assess whether the image rendered from NeRF is realistic or not, akin to the discriminator in GANs. This observation underpins the motivation for our SwiftBrush approach, as we recognize that these principles can be innovatively adapted for text-to-image generation making image supervision unnecessary. SwiftBrush, therefore, emerges as a conceptually novel one-step text-to-image distillation process, uniquely capable of operating without 2D image supervision. This strategic adaptation not only opens up a change of view in the distillation of text-to-image diffusion models into a single-step generators but also points out the potential of blending principles from text-to-3D to text-to-image synthesis.

As a result, our method has successfully enabled the generation of high-quality images with remarkable details. Quantitatively, our one-step model achieves promising zero-shot results, 16.67 for the FID score and 0.29 for the CLIP score on the MS COCO 2014 dataset (30K images). Notably, this is the first time a distilled one-step model outperforms the previous approaches without a single image used for training. A demonstration of generated images from our distilled model is shown in Fig. 1.

## 2. Related Work

**Text-to-Image Generation.** Previously, text-to-image models relied on GANs and focused on only small-scale object-centric domains such as flowers and birds [31, 46].

Recently, these types of systems have advanced rapidly with the availability of web-scale datasets of text-image pairs such as LAION5B [37], large language models such as T5 [28], and large vision-language models such as CLIP [27]. DALL-E [29] was the first autoregressive model to demonstrate a remarkable zero-shot ability in creating images from text captions by simply scaling up both the network and dataset size. Following that, a plethora of methods have been used for text-to-image generation, including masked generative transformers [4], GAN-based, and diffusion models. Based on Latent Diffusion Model [32], Stable Diffusion is an open-source diffusion-based generator that has gained widespread popularity among artists as well as researchers. Recently, text-to-image GANs are receiving attention again with notable work like StyleGAN-T [36], or GigaGAN [13]. While they are extremely fast as no iterative sampling is required, they lack behind diffusion models in terms of scalability and realism due to the fact that they suffer from instability and the infamous issue “mode collapse”. To the best of our knowledge, so far only two above-mentioned work succeeded in applying GANs for text-to-image using tons of intricate techniques and auxiliary modules. Conversely, diffusion models can be easily scale up to generate high-quality and sharp samples, yet they require lengthy sampling process. Here, our objective is to create a large-scale generative model that combines the speed of GANs with the quality of diffusion models.

**Text-to-3D Generation.** Instead of learning a generative model directly from 3D supervision, many research works have utilized the rich knowledge of 2D prior for 3D generation tasks. For instance, one of the earliest works, Dream Fields [12], leverages CLIP [27] to guide the generated images so that rendered images from many camera views highly align with the text caption. On the other hand, pioneer works like DreamFusion [26] and Score Jacobian Chaining [42] (SJC) propose two different but equivalent frameworks where a 2D text-to-image generative teacher is used to generate 3D objects from textual descriptions. Subsequent works seek to improve these text-to-3D methods in various ways. Among them, Magic3D [17] and ProlificDreamer [43] are two remarkable works that effectively enhance the generation quality of 3D assets. Magic3D consists of two stages where they first obtain a low-resolution 3D representation via a DreamFusion-style optimization and then refine its texture to get the final high-fidelity 3D mesh. Meanwhile, ProlificDreamer proposes variational score distillation involving a secondary teacher to bridge the gap between the teacher and the 3D NeRF. Influenced by the latter work, our method adapts the same technique to develop a one-step text-to-image generator.

**Knowledge Distillation** introduced by Hinton et al. [8] is one of the methods falling under the umbrella of transfer

learning. This algorithm draws its main inspiration from the human learning process, where knowledge is conveyed from a more knowledgeable teacher to a less knowledgeable student. In the context of diffusion models, works have been proposed for faster learning of student models from a pre-trained diffusion teacher. The most straightforward approach involves direct distillation [21], wherein a student model is supervised from the teacher’s output after sampling for 50-1000 steps, which can be prohibitive. Recent work [35, 40] avoids the lengthy sampling during distillation by various bootstrapping methods. Afterward, a series of methods [19, 22, 23] adapt these methods into text-to-image settings and inherit their limitation of being image-dependent in training. In contrast, a concurrent work [7] learns a time-conditioned model to predict the outcome of the teacher at any given time step, all without the need for image supervision. Likewise, our approach is similarly image-free yet achieves significantly better results and offers a much simpler distillation design.

### 3. Proposed Method

#### 3.1. Preliminary

**Diffusion Models** consist of two processes: a forward process that gradually adds noise and another reverse process to predict the distribution of gradually denoised data points. The forward process incrementally introduces noise to an initial data point  $x_0$ , drawn from the distribution  $q_0(x_0)$ , resulting in  $q_t(x_t|x_0) = \mathcal{N}(\alpha_t x_0, \sigma_t^2 I)$ , where  $\{(\alpha_t, \sigma_t)\}_{t=1}^T$  is the noise schedule. The data point  $x_t$  at each time-step  $t$  is drawn from  $q_t(x_t|x_0)$  using a noise  $\epsilon \sim \mathcal{N}(0, I)$ :

$$x_t = \alpha_t x_0 + \sigma_t \epsilon. \quad (1)$$

After  $T$  time steps, the final data point is expected to be pure Gaussian noise  $x_T \sim \mathcal{N}(0, I)$ .

Conversely, the backward process starts from  $x_T \sim \mathcal{N}(0, I)$  and iteratively denoises for  $T$  time steps with each step  $t$  converts from noisy variable  $x_{t+1}$  to less-noisy variable  $x_t$ , guided by a model  $\epsilon_\psi$  that predicts the noise  $\epsilon$  in Eq. (1). The network weights  $\psi$  are trained by minimizing:

$$\mathcal{L}_{uncond}(\psi) = \mathbb{E}_{t, \epsilon \in \mathcal{N}(0, 1)} \|\epsilon_\psi(x_t, t) - \epsilon\|_2^2 \quad (2)$$

where  $t$  is uniformly samples within  $\{1, \dots, T\}$ .

While common diffusion models operate in the pixel space, Latent Diffusion Models (LDMs)[32] model the diffusion process in the latent space generated by a pre-trained regularized autoencoder, which typically features representations of smaller resolution, thus enhancing both training and inference efficiency.

**Text-to-Image Diffusion Models.** In contrast to unconditional diffusion models that generate outputs freely, text-conditioned diffusion models steer the sampling process using an extra prompt  $y$ . This guides the model to produce

outputs that are not only photorealistic but also closely adhere to the provided text description. The objective to train such models is formulated as follows:

$$\mathcal{L}_{diff}(\psi) = \mathbb{E}_{t,y,\epsilon \in \mathcal{N}(0,1)} \|\epsilon_\psi(x_t, t, y) - \epsilon\|_2^2 \quad (3)$$

which is slightly different from the unconditional diffusion loss Eq. (2). Due to the prompt conditioning, the model can deliver more controllable generation compared to its unconditional counterparts. Nevertheless, the implementations of many leading methods [1, 30, 34] remain inaccessible to the public. Stable Diffusion, which primarily utilizes the LDM framework, stands out as the first openly available large-scale model, significantly propelling text-to-image synthesis's widespread adoption and versatility.

**Score Distillation Sampling (SDS)** is a distillation technique for pre-trained diffusion models, effectively applied in generating 3D assets [17, 26, 42]. It utilizes a pre-trained text-to-image diffusion model, which predicts diffusion noise from text condition  $y$ , denoted as  $\epsilon_\psi(x_t, t, y)$ . The method optimizes a single 3D NeRF, parameterized by  $\theta$ , to align with a given text prompt. Given camera parameters  $c$ , a differentiable rendering function  $g(\cdot, c)$  is used to render an image at the camera view  $c$  from the 3D NeRF. Here, the rendered image  $g(\theta, c)$  is utilized to optimize the weight  $\theta$  through a loss function whose gradient can be approximated by:

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon,c} \left[ w(t)(\epsilon_\psi(x_t, t, y) - \epsilon) \frac{\partial g(\theta, c)}{\partial \theta} \right] \quad (4)$$

where  $t \sim \mathcal{U}(0.02T, 0.98T)$ ,  $T$  is the maximum timesteps of the diffusion model,  $\epsilon \sim \mathcal{N}(0, I)$ ,  $x_t = \alpha_t g(\theta, c) + \sigma_t \epsilon$ ,  $y$  is the input text, and  $w(t)$  is a weighting function

Despite the advancements in text-to-3D synthesis, empirical studies [26, 42] indicate that SDS often encounters issues like over-saturation, over-smoothing, and reduced diversity. The same degradations can be observed if we naively apply SDS in our framework, as shown in Sec. 4.3.

**Variational Score Distillation (VSD)** is introduced in ProlificDreamer [43] to address previously mentioned issues of SDS by modifying the loss slightly:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{VSD} = & \mathbb{E}_{t,\epsilon,c} [w(t)(\epsilon_\psi(x_t, t, y) \\ & - \epsilon_\phi(x_t, t, y, c)) \frac{\partial g(\theta, c)}{\partial \theta}] \end{aligned} \quad (5)$$

Again,  $x_t = \alpha_t g(\theta, c) + \sigma_t \epsilon$  is the noisy observation of the rendered image at camera view  $c$ . VSD sets itself apart from SDS by introducing an additional score function tailored to the images rendered from the 3D NeRF at the camera pose  $c$ . This score is derived by fine-tuning a diffusion model  $\epsilon_\phi(x_t, t, y, c)$  with the diffusion loss described below:

$$\min_{\epsilon_\phi} \mathbb{E}_{t,c,\epsilon} \|\epsilon_\phi(x_t, t, y, c) - \epsilon\|_2^2 \quad (6)$$

As proposed in ProlificDreamer,  $\epsilon_\phi$  is parameterized by Low-Rank Adaption [11] (LoRA) and initialized from the pre-trained diffusion model  $\epsilon_\psi$  with some added layers for conditioning the camera view  $c$ . Note that at each iteration  $i$  of the optimization process,  $\epsilon_\phi$  needs to adapt with the current distribution of  $\theta$ . Therefore, ProlificDreamer interleaves between finetuning  $\epsilon_\phi$  and optimizing  $\theta$ . With these algorithmic enhancements, ProlificDreamer has significantly advanced its capabilities, enabling the generation of NeRFs and the creation of exceptional textured meshes. Such improvement directly inspires us to adapt VSD to the task of one-step text-to-image diffusion-based distillation.

### 3.2. SwiftBrush

**Motivation.** While SDS and VSD is designed explicitly for the text-to-3D generation task, they loosely connect to that goal via the rendered image  $g(\theta, c)$  of the 3D NeRF. As a matter of fact, instead of using NeRF rendering, we can replace it with any function that output 2D images to suit our needs. Inspired by this motivation, we propose to substitute the NeRF rendering  $g(\theta, c)$  with a text-to-image generator that can directly synthesize a text-guided image in one step, effectively converting the text-to-3D generation training into one-step diffusion model distillation.

**Design Space.** We adopt the same approach as [43], with modifications to the design space to suit our task better. To begin with, we make use of two teacher models: a pre-trained text-to-image teacher  $\epsilon_\psi$  and one additional LoRA teacher  $\epsilon_\phi$ . Also, we remove the conditioning on camera view  $c$  from the LoRA teacher as it is unnecessary in our case, and we use classifier-free guidance for both teachers. Then, we replace the NeRF, which is overfitted for a specific user-provided prompt in a text-to-3D setting, with a generalized one-step text-to-image student model  $f_\theta(z, y)$ . Our student model  $f_\theta$  takes a random Gaussian noise  $z$  and a text prompt  $y$  as inputs. Both the LoRA teacher and student model are initialized with the weight of the text-to-image teacher. Next, we train both the student model and the LoRA teacher alternately using Eq. (5) and Eq. (6) while freezing the text-to-image teacher. A pseudo-code and system figure can be viewed in Alg. 1 and Fig. 2.

**Student Parameterization.** Given a pretrained text-to-image diffusion model  $\epsilon_\theta$ , it is possible to directly use its output for the student model, i.e.,  $f_\theta(z, y) = \epsilon_\theta(z, T, y)$ , where  $T$  is the maximum timestep of the pretrained model. However, in our case, the pretrained model of choice is Stable Diffusion, which is inherently designed to predict the added noise  $\epsilon$ . In contrast, our objective is to refine the student model so that it predicts a clean, noise-free image  $x_0$ . Therefore, there is a large domain gap between what we want the student to learn and the output of the student. For the sake of easy training, we empirically re-parameterize

---

**Algorithm 1** SwiftBrush Distillation

---

```
1: Require: a pretrained text-to-image teacher  $\epsilon_\psi$ , a  
   LoRA teacher  $\epsilon_\phi$ , a student model  $f_\theta$ , two learning rates  
    $\eta_1$  and  $\eta_2$ , a weighting function  $\omega$ , a prompts dataset  
    $Y$ , the maximum number of time steps  $T$  and the noise  
   schedule  $\{(\alpha_t, \sigma_t)\}_{t=1}^T$  of the teacher model  
2: Initialize:  $\phi \leftarrow \psi, \theta \leftarrow \psi$   
3: while not converged do  
4:   Sample input noise  $z \sim \mathcal{N}(0, I)$   
5:   Sample text caption input  $y \sim Y$   
6:   Compute student output  $\hat{x}_0 = f_\theta(z, y)$   
7:   Sample timestep  $t \sim \mathcal{U}(0.02T, 0.98T)$   
8:   Sample added noise  $\epsilon \sim \mathcal{N}(0, I)$   
9:   Compute noisy sample  $\hat{x}_t = \alpha_t \hat{x}_0 + \sigma_t \epsilon$   
10:   $\theta \leftarrow \theta - \eta_1 [\omega(t) (\epsilon_\psi(\hat{x}_t, t, y) - \epsilon_\phi(\hat{x}_t, t, y)) \frac{\partial \hat{x}_0}{\partial \theta}]$   
11:  Sample timestep  $t' \sim \mathcal{U}(0, T)$   
12:  Sample added noise  $\epsilon' \sim \mathcal{N}(0, I)$   
13:  Compute noisy sample  $\hat{x}_{t'} = \alpha_{t'} \hat{x}_0 + \sigma_{t'} \epsilon'$   
14:   $\phi \leftarrow \phi - \eta_2 \nabla_\phi \|\epsilon_\phi(\hat{x}_{t'}, t', y) - \epsilon'\|^2$   
15: end while  
16: return trained student model  $f_\theta$ 
```

---

the student output as follows:

$$f_\theta(z, y) = \frac{z - \sigma_T \epsilon_\theta(z, T, y)}{\alpha_T}$$

which is a realization of Eq. (1) if we set  $t = T$ ,  $x_t = z$ ,  $\epsilon \approx \epsilon_\theta(z, T, y)$  and  $x_0 \approx f_\theta(z, y)$ . So basically, this re-parameterization transforms the noise-prediction output of the pretrained model into the “predicted  $x_0$ ” form, which is empirically demonstrated in Sec. 4.3 to be helpful for the student model to learn.

## 4. Experiments

### 4.1. Experimental Setup

**Evaluation Metrics.** We conduct our evaluation process by assessing the performance of SwiftBrush against several well-established methods, such as Guided Distillation [23], Instaflo [19], LCM [22], and BOOT [7], across two zero-shot text-to-image benchmarks: COCO 2014 [18] and Human Preference Score v2 [45] (HPSv2). On the COCO 2014, we follow the conventional evaluation protocol [13, 32, 34, 36], computing FID and CLIP score on a 30K subset of its validation set as our primary metrics. Here, we use ViT-G/14 as the backbone for evaluating CLIP scores. In the case of HPSv2, we adopt their evaluation procedure to assess text-image alignment across four different aspects: anime, concept-art, paintings, and photos.

**Experiment Protocols.** In these experiments, we rely on Stable Diffusion v2.1 [41] as our teacher model, and when-

ever possible, we directly utilize the reported values from other research. As for InstaFlow and LCM, we use the their provided pretrained models to generate images for the quantitative evaluation. Conversely, we are compelled to re-implement BOOT since they did not provide any metrics for their text-to-image models, and their codebase remains entirely closed.

**Implementation Details.** Similar to prior research in the literature, we use student models with the same architectures as those of the teachers and initialize all of the student parameters with the teacher model. In the training process, specific configuration parameters are used. For the LoRA teacher model, we apply a learning rate of 1e-3, a LoRA rank of 64, and an alpha value of 108. Meanwhile, the student model’s learning rate is set at 1e-6 with exponential moving average (EMA) enabled. Adam optimizer [14] and a batch size of 64 are used to train both the student and LoRA teacher for around 65K iterations. Thanks to the image-free nature of SwiftBrush, we do not require access to any image training set, and only the prompts are needed to distill the generator. Here, we utilize 4M captions from the large-scale text-image dataset JourneyDB [25]. Regarding BOOT, we closely follow their implementation guidelines to the best of our abilities and train the student model until its convergence, except that we use Stable Diffusion v2.1 instead of DeepFloyd IF [5] as the teacher.

## 4.2. Results

**Quantitative Results.** We first evaluate the proposed method on standard text-to-image generation benchmarks. The quantitative comparison with the previous distillation, both image-dependent and image-free, and the original Stable Difusion teacher is shown in Tab. 1. Despite being trained with only text captions, SwiftBrush significantly improves the quality of 1-step inference and achieves better photo-realism as well as text-image alignment against other methods, with exceptions of InstaFlow and Stable Diffusion (25 sampling steps) in FID score. However, it is worth noting that Instaflo is so complex, requiring a 4-stage distillation process, while ours is end-to-end and straightforward. We further conduct a quantitative evaluation with the HPSv2 score, a metric trained to capture human preference. Again, we obtain the best score in all aspects among all mentioned works, which is only a slight degradation compared to the original Stable Diffusion teacher (Tab. 2). Moreover, according to the anonymous survey carried out by us (Fig. 3), our method consistently outperforms Instaflo in terms of image fidelity and text-to-image accuracy.

**Qualitative Results.** We report the visual comparison between our method and the mentioned baseline methods for the one-step text-to-image generation task. As seen in Fig. 4, LCM outputs blurry and over-smoothed images in

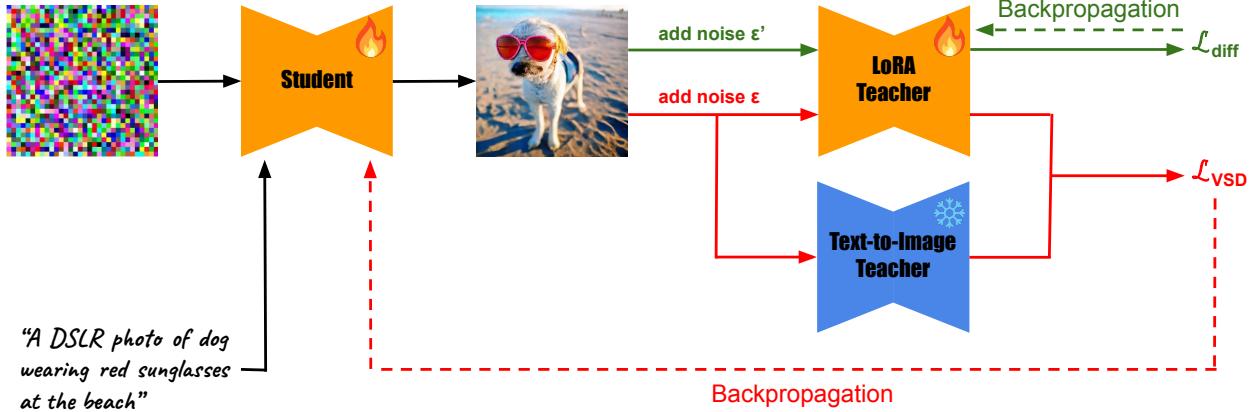


Figure 2. **SwiftBrush overview.** Our system allows training a one-step text-to-image student network from a frozen pretrained teacher with the help of an additional trainable LoRA teacher. The student network inputs a text prompt and random noise to synthesize an image. The image is then added noise and sent to two teachers, alongside the prompt and a randomly drawn timestep, to compute the gradient of VSD loss, which backpropagates back to the student. Apart from that, the LoRA teacher is also updated using diffusion loss. Similar to ProlificDreamer [43], we update the LoRA teacher (green flow) and the student (red flow) alternatively .

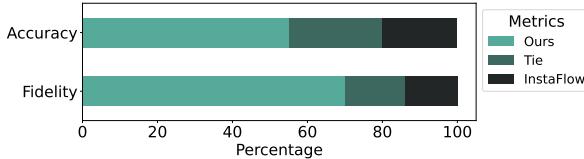


Figure 3. **User Study.** We ask 400 users to compare the generated images of InstaFlow and our method using 30 random text prompts. We report the rate at which each method is selected as better for image quality and text-to-image accuracy.

the 1-step regime. BOOT faithfully produces images that match well with the text caption, but its results still fall short in terms of fidelity. In contrast, a much better quality can be observed in the image produced by Instaflow and SwiftBrush. InstaFlow, nevertheless, cannot handle complex or unusual input prompts properly, which can be noticed in the third and fifth images at the third column of Fig. 4. On top of that, it occasionally produces non-realistic results even for simple prompts like in Fig. 4 at the fourth image of the third column. When, in fact, the outputs generated by SwiftBrush are not only of exceptional quality but also exhibit a remarkable alignment between text descriptions and the corresponding images. Moreover, our results are comparable with those sampled from Stable Diffusion with 25 steps, whereas it is understandable that with only one step, Stable Diffusion cannot create outputs of good quality.

### 4.3. Analysis

**The importance of LoRA teacher.** In Fig. 6, the importance of including the LoRA teacher in the training process is clearly highlighted. We compare the student outputs sampled with different input noises after 10K iteration when using the same text prompts. With the guidance of the LoRA teacher, the student progressively learns to create realistic

Method	Steps	FID-30K ↓	CLIP-30K ↑
Guided Distillation <sup>†</sup>	1	37.3	0.27
LCM <sup>†</sup>	1	35.56	0.24
Instaflow	1	<b>13.10<sup>†</sup></b>	<u>0.28<sup>§</sup></u>
BOOT <sup>‡</sup>	1	48.20	0.26
Ours	1	<u>16.67</u>	<b>0.29</b>
SD 2.1*	25	13.45	0.23
SD 2.1*	1	202.14	0.06

Table 1. Comparison of our method against other works based on FID metric and CLIP score on the COCO 2014 dataset. <sup>†</sup> means that we obtain the score from the corresponding papers. <sup>§</sup> means that we obtain the score using the provided pretrained models of the corresponding papers. <sup>‡</sup> means that we re-implement the work and report the score by ourselves. \* means that we report the score using the pretrained models. We use DPM-Solver [20] to sample Stable Diffusion. **Bold** and underlined numbers are best and second best, respectively.

Models	Human Preference Score v2 ↑			
	Anime	Photo	Concept Art	Paintings
LCM <sup>†</sup>	22.61	22.71	22.74	22.91
InstaFlow <sup>†</sup>	<b>25.98</b>	<u>26.32</u>	<u>25.79</u>	<b>25.93</b>
BOOT <sup>‡</sup>	25.29	25.16	24.40	24.61
Ours	<b>26.91</b>	<b>27.21</b>	<b>26.32</b>	<b>26.37</b>
SD 2.1*	27.48	26.89	26.86	27.46

Table 2. Comparison of our method against other works based on HPSv2 score in 1-step regime. <sup>†</sup> means that we obtain the score using the provided pretrained models of the corresponding papers. <sup>‡</sup> means that we re-implement the work and report the score by ourselves. \* means that we obtain the score from [45]. **Bold** and underlined numbers are best and second best, respectively.

images (1<sup>st</sup> column). In contrast, in the absence of an effective LoRA teacher, the student’s early training outputs, despite being highly aligned with the text prompt, are overly

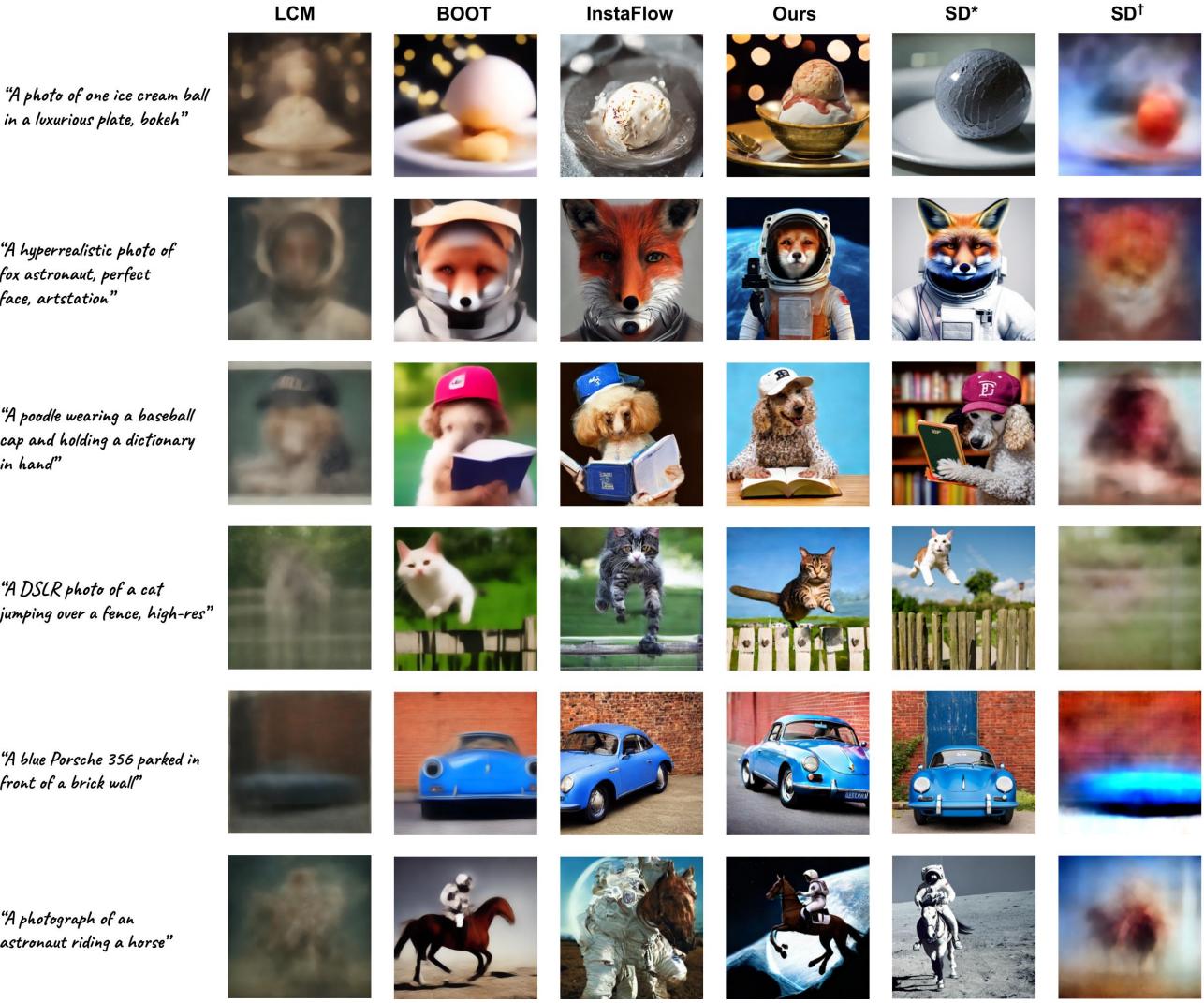


Figure 4. Sample images generated by LCM, BOOT, Instaflow, Ours, SD with 25 sampling steps ( $SD^*$ ) and SD with 1 sampling step ( $SD^\dagger$ ). Images at the same row are sampled from the same text prompt, while images at the same column are from the same model. We use DPM-Solver [20] to sample Stable Diffusion

saturated and unrealistic ( $2^{nd}$  column). These results indicate a fundamental breakdown in the student model’s ability to mimic the teacher model’s distribution, as can be further demonstrated based on the flat red curve in both FID and CLIP score in Fig. 5 (a) and (b).

In addition, even when employing the LoRA teacher, it is crucial to use a sufficiently large rank in the LoRA weight for stable distillation of the student model. Without this, the model tends to undergo severe mode collapse despite initially producing reasonably good results. The green curves in Fig. 5 clearly illustrate that, while the FID score of a student model with a 4-rank LoRA weight initially decreases, it deteriorates starting from 8K steps, resulting in lousy images in Fig. 6 ( $3^{rd}$  column). Additionally, the corresponding CLIP score behaves similarly where there is a slight rise before it worsens in the end. Besides, We also provide extra

visuals for this ablative study in Sec. 7.

**The importance of student parameterization.** Apart from the LoRA teacher, the re-parameterization of the student (Sec. 3) also plays an important role in the SwiftBrush training. Without the it, the student is unable to to create realistic images and it only produces noisy results, as can be noticed from the  $4^{th}$  column of Fig. 6. This can be further backed up by the orange curve in both FID and CLIP score in Fig. 5 (a) and (b), where student without re-parameterization perform consistently worst among all models.

**Interpolating Latent.** We conduct a controlled experiment where the noise input  $z$  remains the same while varying the input prompt  $y$  (Fig. 7). Interestingly, the SwiftBrush-distilled student seemingly features enhanced control and editability compared to the original diffusion model due to

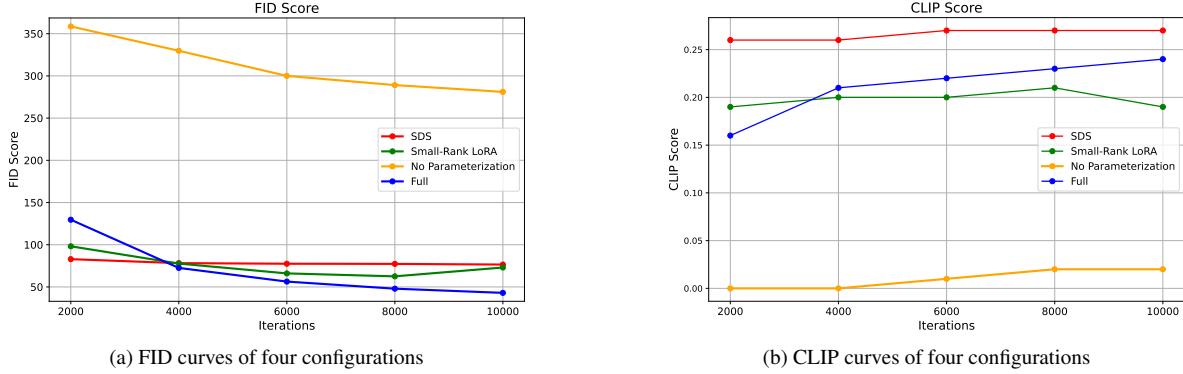


Figure 5. Ablation study on the impact of the LoRA teacher as well as the student parameterization. We train all four configurations with same hyperparameters described in Sec. 4 but only with 10K iterations to observe their training dynamics at the early phase. Then, we evaluate FID and CLIP score on COCO 2014 but with only a 5K subset of images. Here, Full means all of our proposed techniques are used. No Parameterization means we exclude the re-parameterization trick. Small-Rank LoRA means we use only 4-rank LoRA teacher. SDS means we remove the LoRA teacher and use only SDS loss.

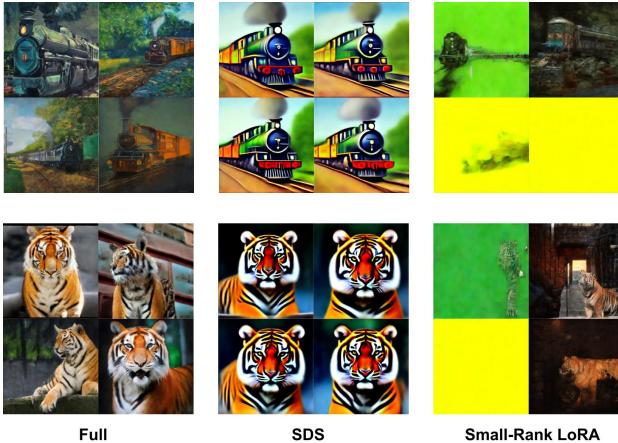


Figure 6. Visual results of the ablative study, where the first two rows are generated with the input prompt “An oil painting of a train”, whereas the remaining two are sampled from “An DSLR photo of a tiger in the city”.

its one-step nature. Additionally, we present further results of latent space interpolation in Sec. 7. In this process, the student model, distilled from the pretrained Stable Diffusion teacher, demonstrates a seamless transition in its image generation.

## 5. Conclusion and Future Work

In this work, we have introduced a new image-free distillation scheme, **SwiftBrush**, which leverages insights from text-to-3D synthesis to speed up text-to-image generation. Empirically, the results presented in this paper demonstrate the effectiveness of our approach in making text-to-image generation more efficient and accessible. We hope our work will encourage broader use of text-to-image generation systems by substantially lowering the inference cost.

SwiftBrush inevitably produce lower quality samples



Figure 7. Results of interpolating the input prompts. The prompts used here are selected from a standard template: ‘A DSLR photo of a {animal} {action}’. Here, ‘animal’ is one of following: owl, raccoon, cat, or panda, while ‘action’ is one of following: reading a book, drinking a latte or eating a pizza.

compared to the teacher which utilizes multi-step sampling for inference. Another drawback is that the current design only focuses on single-step student model and cannot support few-step generation as did in Luo et al. [22] to improve image fidelity. In future work, we aim to investigate the feasibility of extending SwiftBrush to a few-step generation so that we can trade computation for quality. Furthermore, we also find it compelling to explore SwiftBrush-like training where only one teacher is required. Such an exploration could provide tremendous benefits in scenarios where the computational resource is limited. Finally, it is intriguing to see if techniques such as DreamBooth [33], ControlNet [47] or InstructPix2Pix[2] can be integrated with SwiftBrush, allowing instant generation of various applications.

## References

- [1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 4
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. *IEEE Computer Vision and Pattern Recognition Conference*, 2023. 8
- [3] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models. *International Conference on Computer Vision*, 2023. 2
- [4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv preprint arXiv:2301.00704*, 2023. 3
- [5] DeepFloyd. DeepFloyd IF. <https://www.deepfloyd.ai/deepfloyd-if>, 2023. 5
- [6] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. *International Conference on Learning Representations*, 2023. 2
- [7] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Josh Susskind. BOOT: Data-free Distillation of Denoising Diffusion Models with Bootstrapping. *arXiv preprint arXiv:2306.05544*, 2023. 3, 5, 1
- [8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NIPS Deep Learning Workshop*, 2014. 3
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 2020. 2
- [10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video Diffusion Models. *Advances in Neural Information Processing Systems*, 2022. 2
- [11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*, 2022. 4
- [12] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-Shot Text-Guided Object Generation with Dream Fields. *IEEE Computer Vision and Pattern Recognition Conference*, 2022. 3
- [13] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for Text-to-Image Synthesis. *IEEE Computer Vision and Pattern Recognition Conference*, 2023. 3, 5
- [14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 2015. 5
- [15] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *International Conference on Learning Representations*, 2021. 2
- [16] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM: Improves Controllable Text Generation. *Advances in Neural Information Processing Systems*, 2022. 2
- [17] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-Resolution Text-to-3D Content Creation. *IEEE Computer Vision and Pattern Recognition Conference*, 2023. 3, 4
- [18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision*, 2014. 5
- [19] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One Step is Enough for High-Quality Diffusion-Based Text-to-Image Generation. *arXiv preprint arXiv:2309.06380*, 2023. 2, 3, 5, 1
- [20] 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*, 2022. 6, 7
- [21] Eric Luhman and Troy Luhman. Knowledge Distillation in Iterative Generative Models for Improved Sampling Speed. *arXiv preprint arXiv:2101.02388*, 2021. 3
- [22] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2, 3, 5, 8, 1
- [23] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On Distillation of Guided Diffusion Models. *IEEE Computer Vision and Pattern Recognition Conference*, 2023. 2, 3, 5, 1
- [24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. *International Conference on Machine Learning*, 2021. 2
- [25] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. JourneyDB: A Benchmark for Generative Image Understanding. *Advances in Neural Information Processing Systems*, 2023. 5
- [26] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. *International Conference on Learning Representations*, 2023. 2, 3, 4
- [27] 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. *International Conference on Machine Learning*, 2021. 3
- [28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3
- [29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. *International Conference on Machine Learning*, 2021. 3
- [30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 4
- [31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. *International Conference on Machine Learning*, 2016. 2
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *IEEE Computer Vision and Pattern Recognition Conference*, 2022. 3, 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. *IEEE Computer Vision and Pattern Recognition Conference*, 2023. 2, 8
- [34] 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. *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5
- [35] Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. *International Conference on Learning Representations*, 2022. 2, 3
- [36] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *International Conference on Machine Learning*, 2023. 3, 5
- [37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022. 3
- [38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *International Conference on Machine Learning*, 2015. 2
- [39] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations*, 2021. 2
- [40] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. *International Conference on Machine Learning*, 2023. 2, 3
- [41] StabilityAI. Stable Diffusion. <https://github.com/Stability-AI/stablediffusion>, 2021. 2, 5
- [42] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. *IEEE Computer Vision and Pattern Recognition Conference*, 2023. 3, 4
- [43] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *Advances in Neural Information Processing Systems*, 2023. 3, 4, 6
- [44] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel View Synthesis with Diffusion Models. *International Conference on Learning Representations*, 2023. 2
- [45] 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. 5, 6
- [46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. *International Conference on Computer Vision*, 2017. 2
- [47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *International Conference on Computer Vision*, 2023. 2, 8

# SwiftBrush : One-Step Text-to-Image Diffusion Model with Variational Score Distillation

## Supplementary Material

### 6. Training Cost and Inference Speed

**Training Cost.** When training our model on a single A100 40GB GPU, we can use a batch size of up to 16 and approximately 15,825 training iterations per day. Specifically, we train the model with 65,000 gradient updates. Therefore, distilling Stable Diffusion using SwiftBrush requires  $65,000/15,825 \approx 4.11$  A100 GPU days. Additionally, even though the original BOOT requires 500,000 iterations to be fully trained as reported in [7], our re-implementation only needs about 100,000 to converge, which takes approximately  $100,000/18,000 \approx 5.56$  A100 GPU day with a batch size of 16 per GPU.

**Inference Speed.** Next, we measure the inference time of our method using a batch size of 1 on an A100 40GB GPU. The inference time of one-step UNet to generate the latent from random noise is 25 ms, while those of text encoder and VAE are 10 ms and 5 ms, respectively. Consequently, the total inference time of SwiftBrush is  $25 + 10 + 5 = 40$  ms. Furthermore, the speed of the text encoder and VAE in BOOT are the same as ours; however, there is a small overhead in the UNet (30ms) as additional layers are used according to their implementation guidelines. Therefore, the total inference time for BOOT is 45ms.

**Compare with Others.** We also sum up inference speed and training time of existing methods and ours in Tab. 3. Since Guided Distillation [23] does not provide details on the inference time or access to the source code, we have decided not to include its speed in our report. Also, we use the provided pretrained model to measure the inference time of LCM [22] which is slightly faster than ours. In terms of training costs, Guided Distillation and InstaFlow [19] require a considerable amount of time, limiting their practical use. Conversely, LCM converges quickly but produces poor-quality images with one-step inference. Our method, however, offers both training efficiency and high-quality output.

### 7. Additional Visual Results

**Uncurated Samples.** We display uncurated images conditioned on 25 random text prompts. The images generated by various models are presented in Fig. 8, 9, 10 and 11, allowing for better visual comparison between our model and other methods.

**Additional Visuals of Analysis.** To further back up our

Method	Inference Time (ms) ↓	Training Time (days) ↓
Guided Distillation	-	108.8 <sup>†</sup>
LCM	37 <sup>§</sup>	1.3 <sup>*</sup>
Instaflow	90 <sup>†</sup>	199.2 <sup>†</sup>
BOOT	45 <sup>‡</sup>	5.56 <sup>‡</sup>
Ours	40	4.11

Table 3. Comparison of our method against other works based on inference and training time. <sup>†</sup> means that we obtain the numbers from [19]. <sup>§</sup> means that we obtain the number using the provided pretrained models. <sup>\*</sup> means that we obtain the numbers from [22]. <sup>‡</sup> means that we re-implement the work and report the numbers by ourselves. The units for inference speed and training time are milliseconds and A100 GPU days, respectively.

claim in Sec. 4.3, we provide some uncurated samples using the text prompt “An DSLR photo of a tiger in the city”. The images generated by four configurations at every 2000 iterations are presented in Fig. 14, allowing for a better visual comparison of each component’s importance in our proposed method across different training stages.

**Additional Visuals of Interpolation.** Apart from investigating the role of each component in SwiftBrush, we also explore its properties when interpolating text prompt condition  $y$  (Fig. 12) or noise input  $z$  (Fig. 13). It is evident that when the input text prompt  $y$  is interpolated, our model generates a seamless transition in the output, showcasing excellent editability and controllability. On the other hand, it is noteworthy that the quality is considerably lower when interpolating the noise input  $z$ , leading to a poor transition between two endpoints.

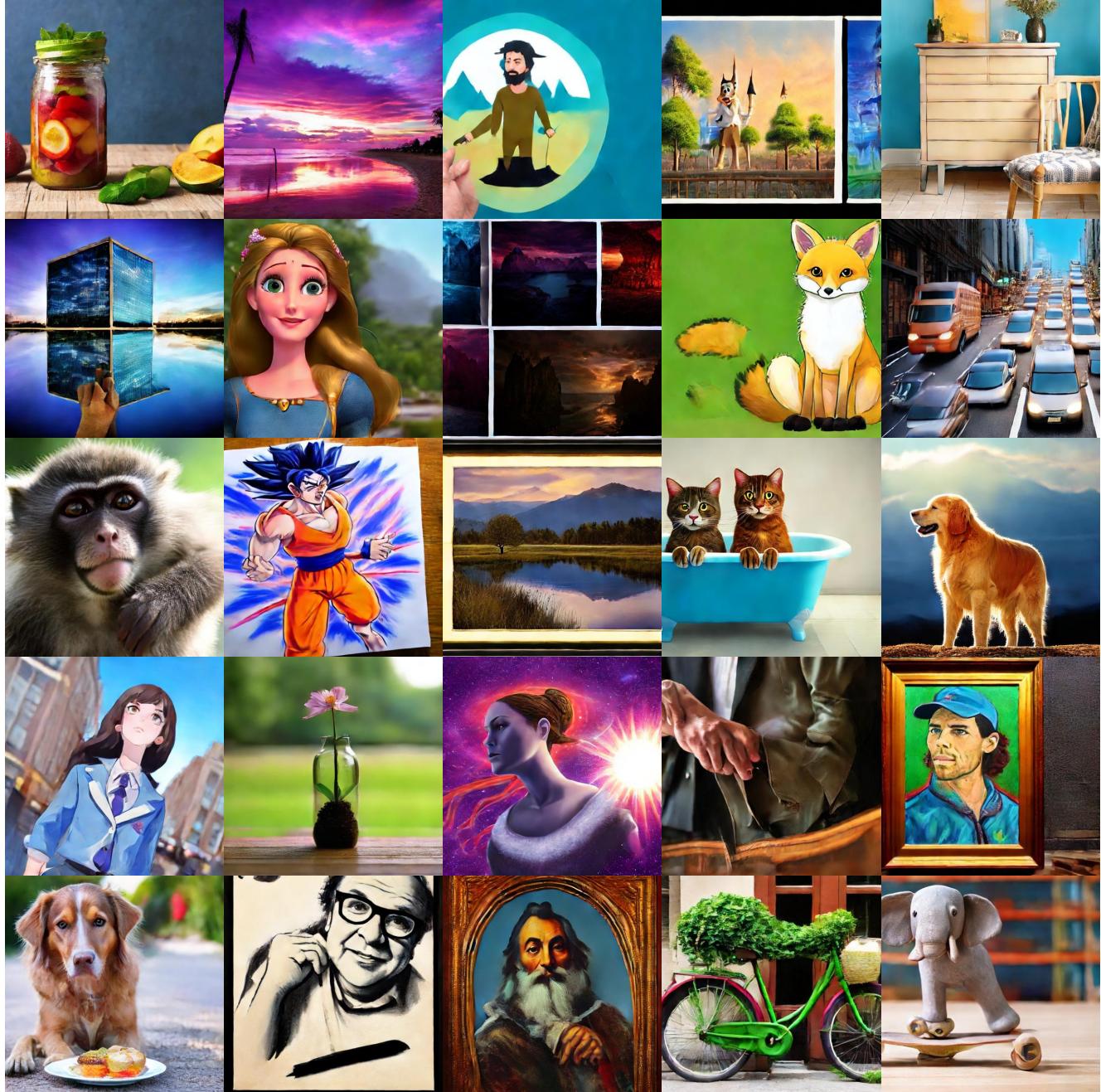


Figure 8. Uncurated samples from one-step **SwiftBrush**.

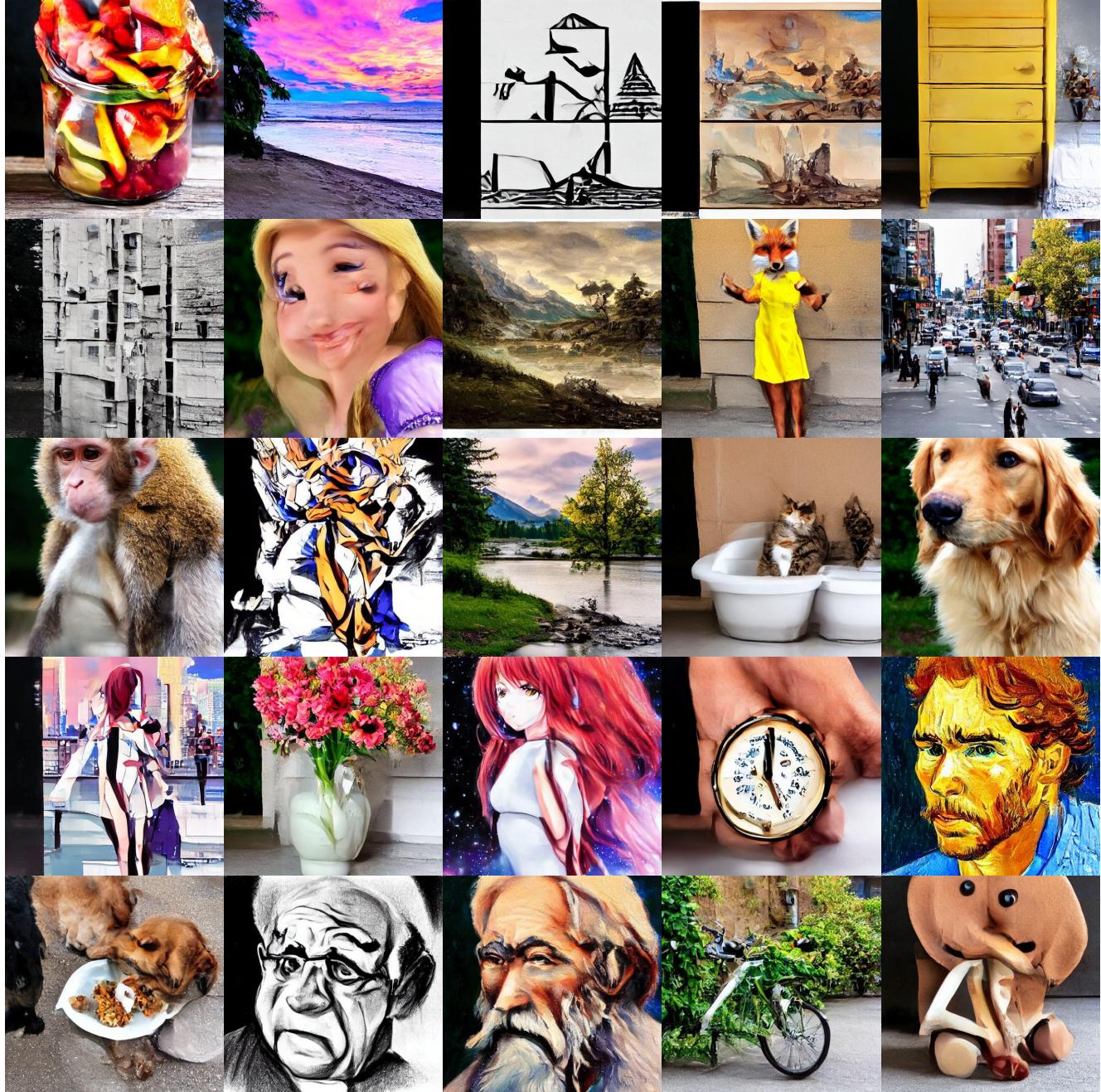


Figure 9. Uncurated samples from one-step **InstaFlow**. We use their provided pretrained model to generate.

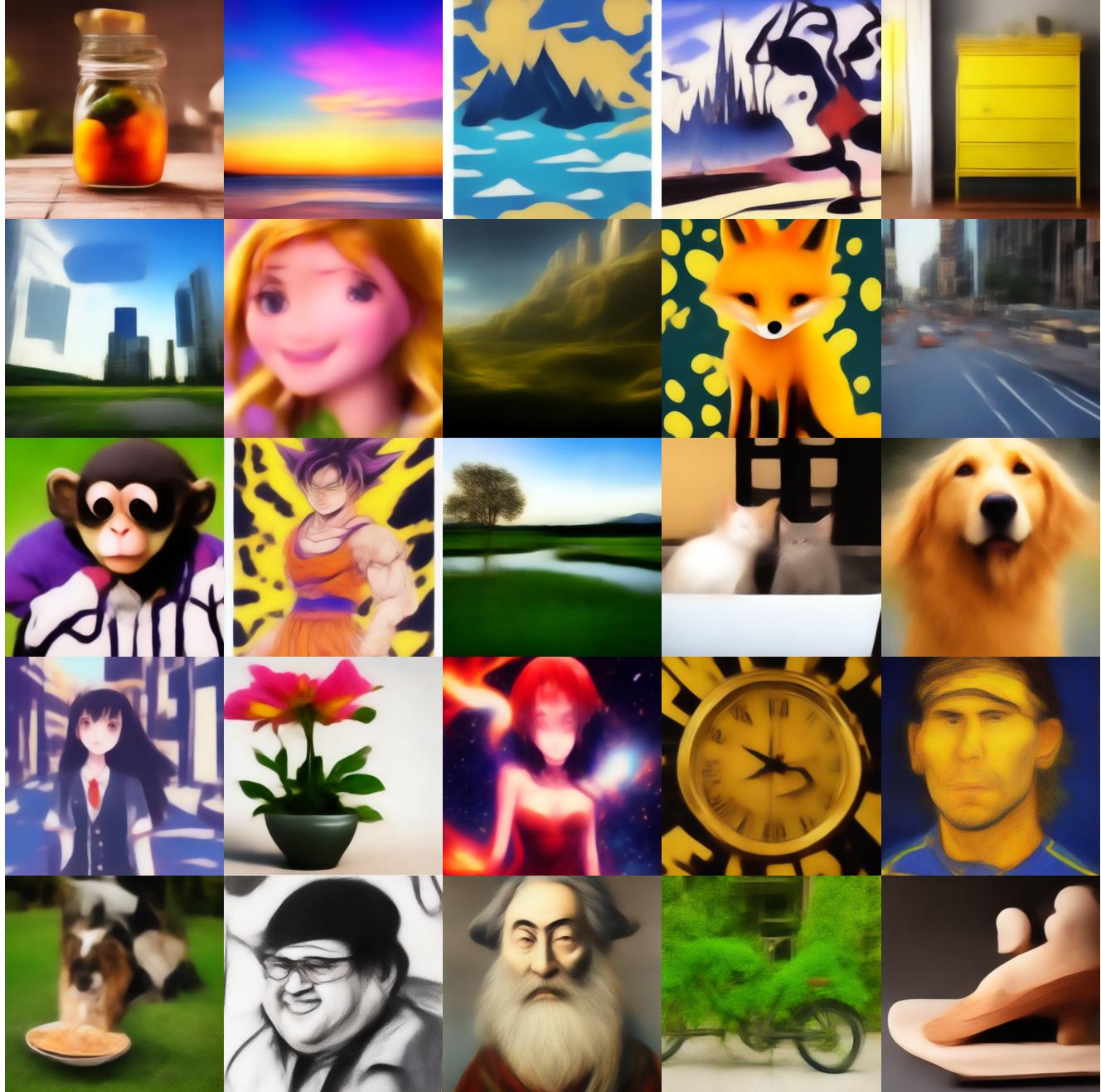


Figure 10. Uncurated samples from one-step **BOOT**. We re-implement and generate the images by ourselves.

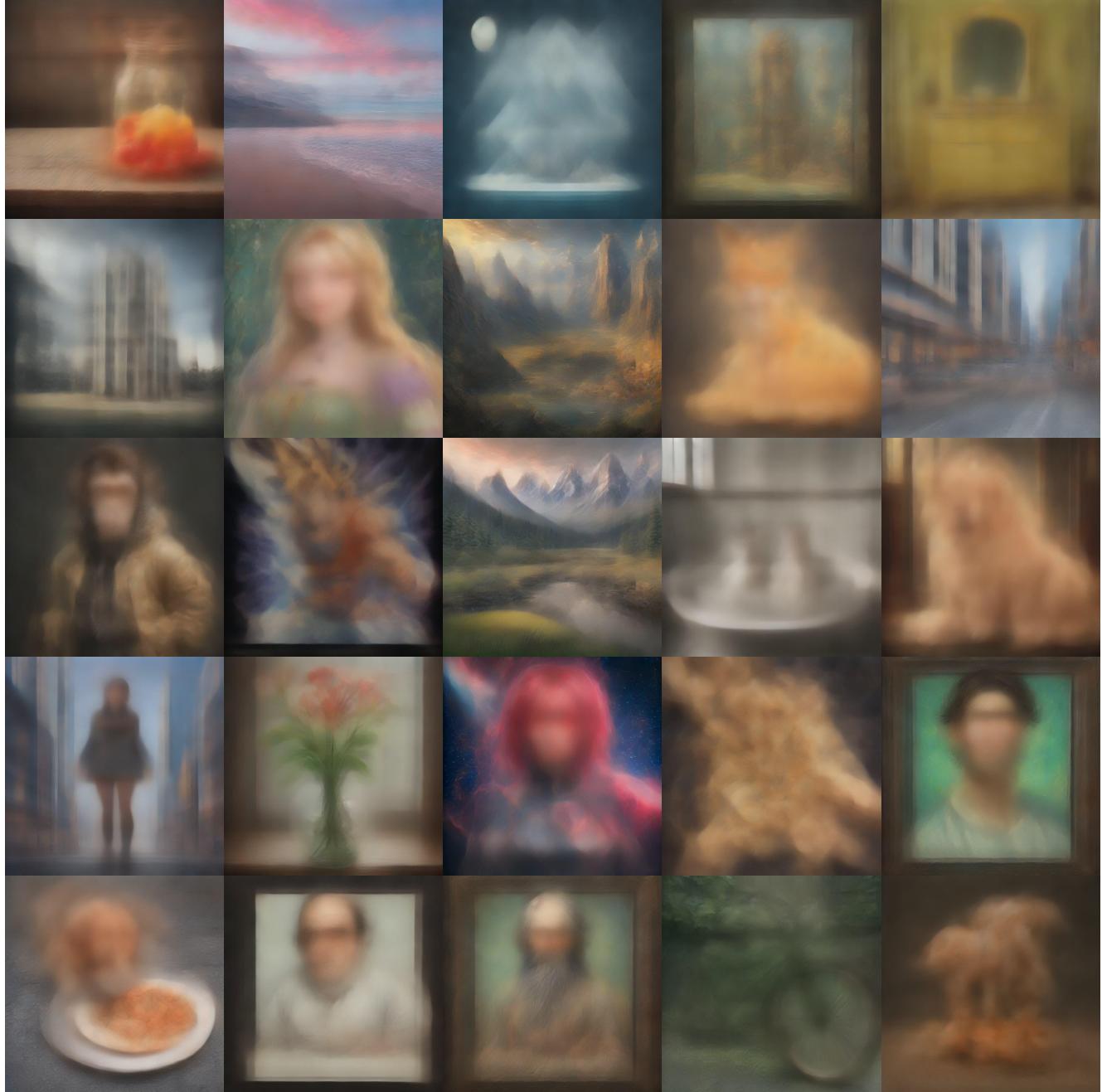


Figure 11. Uncurated samples from one-step **LCM**. We use their provided pretrained model to generate.

Dog



Cat



Figure 12. Results of interpolating the input prompt. The prompts used here are selected from a standard template: “A DSLR photo of a {animal} reading a book”. Here, ‘animal’ is dog or cat and we interpolate the text embedding using alpha-blending. Same noise input  $z$  is used for images at each row.

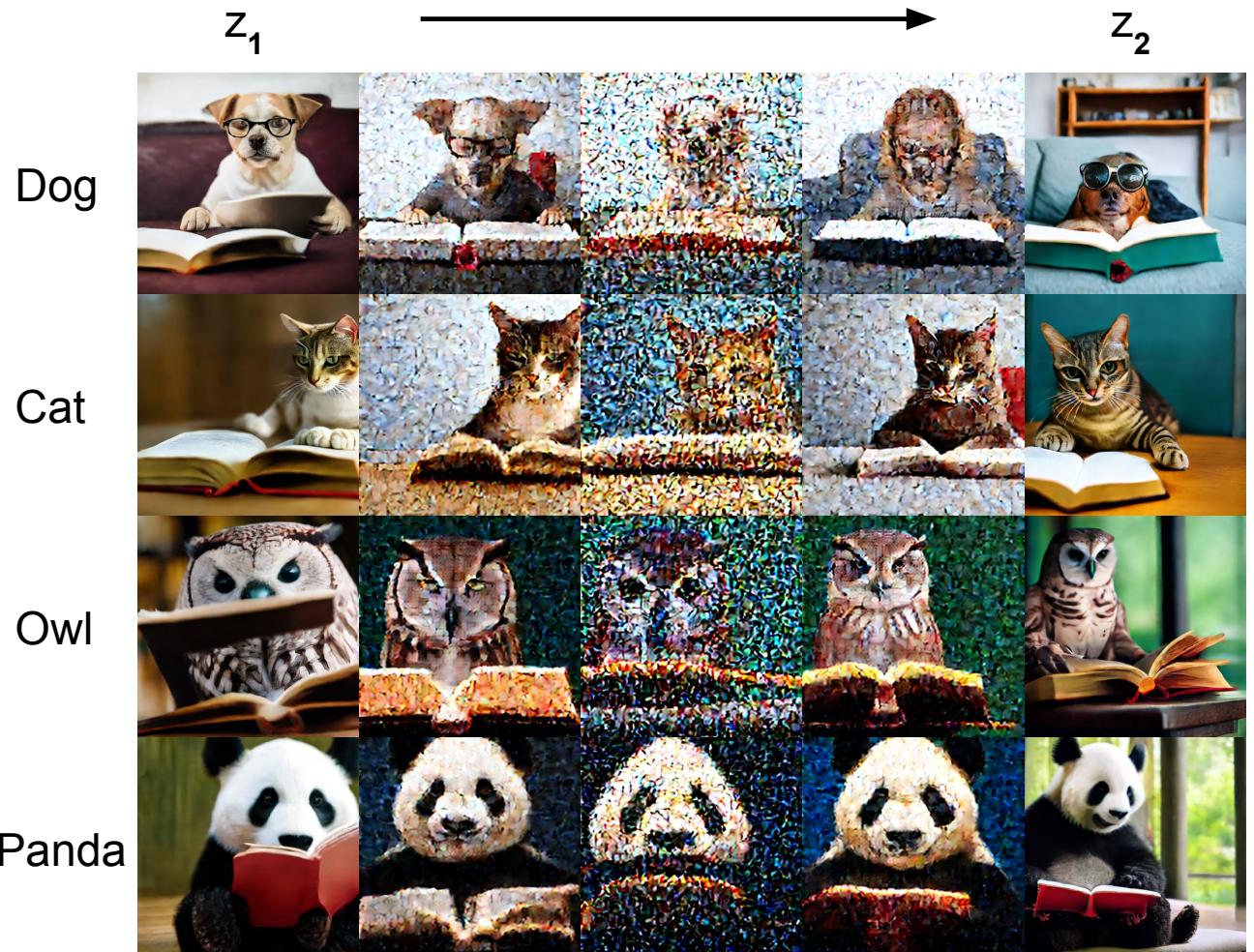


Figure 13. Results of interpolating the noise input. The prompts used here are selected from a standard template: “A DSLR photo of a {animal} reading a book”. Here, ‘animal’ is dog, cat, owl or panda and we interpolate the noise input using alpha-blending. Same text input  $y$  is used for images at each row.

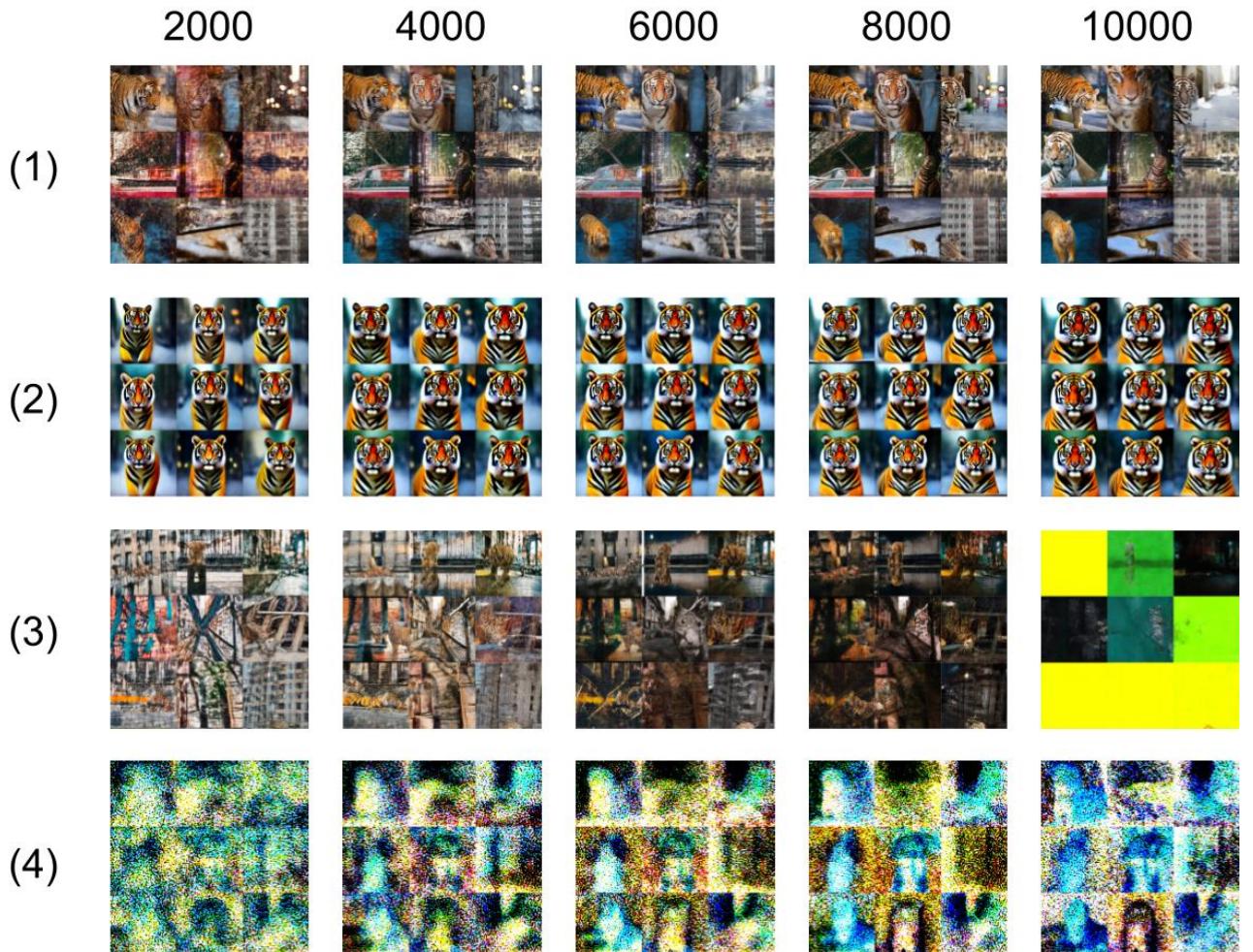


Figure 14. Visual results of the ablative study, where all images are generated with the input prompt “An DSLR photo of a tiger in the city”. Images at same column are generated at the same iteration of training, whereas those at same row are generated using the same configuration. Here, (1) means “Full”, (2) means “SDS”, (3) means “Small-Rank LoRA” and (4) means “No Parameterization”.