HiDiffusion: Unlocking Higher-Resolution Creativity and Efficiency in Pretrained Diffusion Models

Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhao Chen, Yao Tang, and Jiajun Liang[†]

MEGVII Technology
{zhangshen1915, chaowechan, yhao.chen0617}@gmail.com
{zhaozhenyu, tangyao02, liangjiajun}@megvii.com

Project Page: https://hidiffusion.github.io/

Abstract. Diffusion models have become a mainstream approach for high-resolution image synthesis. However, directly generating **higher-resolution** images from pretrained diffusion models will encounter unreasonable object duplication and exponentially increase the generation time. In this paper, we discover that object duplication arises from feature duplication in the deep blocks of the U-Net. Concurrently, We pinpoint the extended generation times to self-attention redundancy in U-Net's top blocks. To address these issues, we propose a tuning-free higher-resolution framework named HiDiffusion. Specifically, HiDiffusion contains Resolution-Aware U-Net (RAU-Net) that dynamically adjusts the feature map size to resolve object duplication and engages Modified Shifted Window Multi-head Self-Attention (MSW-MSA) that utilizes optimized window attention to reduce computations. we can integrate HiDiffusion into various pretrained diffusion models to scale image generation resolutions even to 4096×4096 at 1.5-6× the inference speed of previous methods. Extensive experiments demonstrate that our approach can address object duplication and heavy computation issues, achieving state-of-the-art performance on higher-resolution image synthesis tasks.

Keywords: Higher-Resolution Image Synthesis · High-Efficiency Diffusion

1 Introduction

Generative model has witnessed an explosion of diffusion models of growing capability and applications [13, 32, 38–40]. Being trained on a large volume of images (Laion 5B [37]), Stable Diffusion (SD) [30, 32] can generate fixed-size (e.g. 512×512 for SD 1.5 [32]) high-quality images given text or other kinds of prompts. However, it is limited to synthesizing images with higher resolutions (e.g. 2048×2048). The limitation has two perspectives: **(i) Feasibility**.

[†]Corresponding author

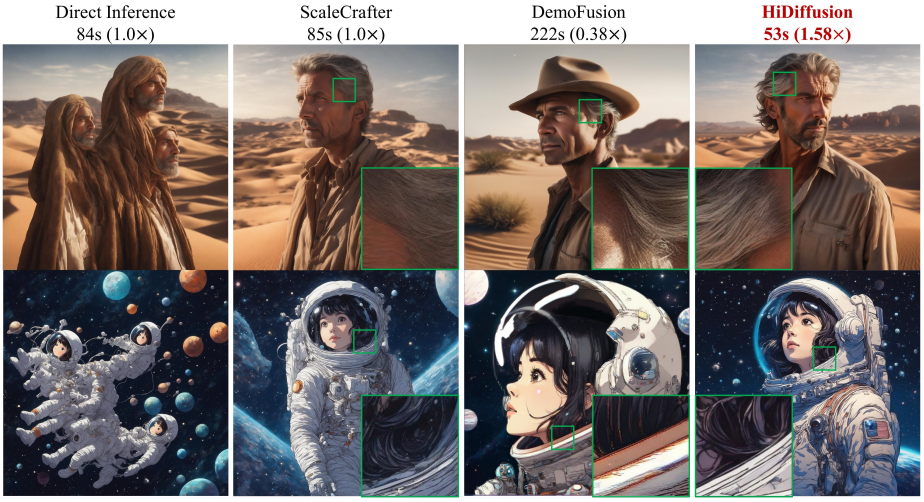


Fig. 1: 2048×2048 resolution images based on SDXL [30]. The first line of the text indicates the generation methods, while the second line indicates the cost time and inference speed relative to direct inference. Our Hidiffusion can generate reasonable and realistic high-resolution images with high efficiency. Compared to previous methods, ours exhibits richer fine-grained details and is 1.58× faster than Scalerafter [11], 4.18× faster than DemoFusion [8]. Best viewed when zoomed in.

Diffusion models lack scalability in higher-resolution image generation. As illustrated in the Direct Inference column of Fig. 1, when directly inferencing to generate 2048×2048 resolution images for SDXL [30] that being trained on 1024×1024 resolution, the generated images exhibit unreasonable object duplication and inexplicable object overlaps. **(ii) Efficiency.** As resolution increases, the time cost becomes more and more unacceptable. For example, SD 1.5 can generate a 512×512 resolution image in only 3s, whereas it takes 165s to generate a 2048×2048 image on an NVIDIA V100 with 50 DDIM steps. The low efficiency of diffusion models in higher-resolution synthesis makes it impractical for real-world applications. We ask: **Can Stable Diffusion efficiently synthesize images with resolution beyond the training image sizes?**

Existing methods answer feasibility question from three perspectives: (i) Collecting enough higher-resolution images to retrain a diffusion model for higher-resolution synthesis [30]. (ii) Leveraging additional super-resolution models [32, 44] to upscale low-resolution images. (iii) Modifying the operation [16] or architecture [11] of U-Net, or creating a new synthesis schedule [8] for higher-resolution synthesis in a tuning-free way. While those perspectives can mitigate object duplication, the first two require large-scale high-resolution datasets and the training process is costly. The tuning-free methods can leverage the power of the pretrained diffusion model, but they suffer from insufficient image details or low inference efficiency, as shown in Fig. 1.

In this paper, we aim to resolve object duplication and generate higher-resolution images with fine details in a tuning-free way. Different from previous

methods, we explore a new perspective by investigating the feature map in the U-Net. Our observation reveals that the generated image is highly correlated with the feature map of deep blocks in structures and feature duplication happens in the deep blocks. The highly duplicated features guide the synthesis direction, resulting in object duplication. We propose a simple yet effective method called Resolution-Aware U-Net (RAU-Net). RAU-Net involves Resolution-Aware Downsampler (RAD) and Resolution-Aware Upsampler (RAU) to align the feature map size with the deep block of U-Net. In contrast to ScaleCrafter [11], which attributes object duplication to the limited receptive field of convolutions and requires determining the parameters of *each* convolution in U-Net, we discover feature duplication and propose a more concise solution RAU-Net that modifies the parameters of only *two* convolutions. Therefore, RAU-Net can be more readily integrated into various diffusion models. The slight modifications can also better preserve the capabilities of the pretrained model, and consequently retain more fine-grained image details. To further improve the higher-resolution image quality, we propose a Switching Threshold to boost the fine details of generated higher-resolution images. Our proposed approach requires no further fine-tuning and can be seamlessly integrated into diffusion models.

Besides higher-resolution feasibility, efficiency is another important concern. Numerous works focus on reducing the sampling step [20,23,24,27,35,38] and the acceleration of diffusion U-Net [2, 20, 26]. These acceleration methods enhance the inference efficiency but also compromise the generated image quality. In this paper, we unearth that the dominant time-consuming global self-attention in the top blocks exhibits surprising locality. Inspired by this observation, we propose Modified Shifted Window Multi-head Self-Attention (MSW-MSA) and replace the global self-attention with it in higher-resolution synthesis. This substitution needs no further fine-tuning. Compared to the previous local attention method [22], our method uses large window attention and shifts windows across timestep to accommodate diffusion models. Empirically, MSW-MSA achieves significant acceleration without compromising image quality.

We combine RAU-Net and MSW-MSA into a unified tuning-free framework for higher-resolution image generation, dubbed HiDiffusion. We conduct qualitative and quantitative experiments to validate the effectiveness of our method. Specifically, HiDiffusion can scale the resolution of SD 1.5 [32] and SD 2.1 [32] from 512×512 to 2048×2048, scale SDXL Turbo [36] from 512×512 to 1024×1024, and scale SDXL [30] from 1024×1024 to 4096×4096. Moreover, HiDiffusion is 1.5-2.7× faster than vanilla SD and is 1.5-6× faster than previous methods in higher-resolution image generation. We hope this work can provide valuable guidance for future research on the scalability of diffusion models.

2 Related Work

High-Resolution Image Synthesis. The application of diffusion models in high-resolution image generation poses a significant challenge. Existing methods have primarily concentrated on diffusion in lower-dimensional spaces (la-

tent diffusion) [32], or divided the generative process into multiple sub-problems [14, 15, 41, 42]. Nevertheless, these solutions render the framework is highly intricate. Recently, there has been a growing interest in exploring tuning-free approaches for variable-sized adaptation. [16] propose an attention scaling factor for variable-sized image synthesis. MultiDiffusion [1] and SyncDiffusion [17] manipulated the generation process by binding together multiple diffusion generation processes. Despite their advancements, these approaches still exhibit object duplication. Recently, ScaleCrafter [11] mitigates object duplication through redilation that can dynamically adjust the convolutional receptive field during inference. ScaleCrafter can effectively address object duplication but somewhat degrades the image quality. DemoFusion [8] proposed a novel progressive generation schedule with skip residual and dilated sampling. It can generate high-quality high-resolution images, but the long generation time limits its practicality. Different from the previous method, we turn our attention to investigating the properties of the feature map of U-Net. We discover that feature duplication leads to object duplication and propose RAU-Net to effectively resolve it.

Diffusion Model Acceleration. As diffusion model training and inference is time-consuming, particularly in the context of high-resolution images, various methods [4, 19, 29] have been extensively investigated to accelerate the training and inference of diffusion models. Unlike fast sampling approaches [23, 35, 38, 40] that improve sampling schemes to reduce the sampling step. ToMeSD [2], SnapFusion [20] and DeepCache [26] speeded up an off-the-shelf diffusion model without training by exploiting natural redundancy in diffusion models. However, they all somewhat compromise image quality. In this paper, through the analysis of the locality of global self-attention in the top blocks, we develop a simple yet effective method MSW-MSA that significantly accelerates the generation of higher-resolution images without the need for fine-tuning and does not compromise image quality. Compared to previous local attention implementations [9, 22], MSW-MSA leverages locality observation, allowing it to better adapt to diffusion models and achieve a superior trade-off between speed and quality.

3 Method

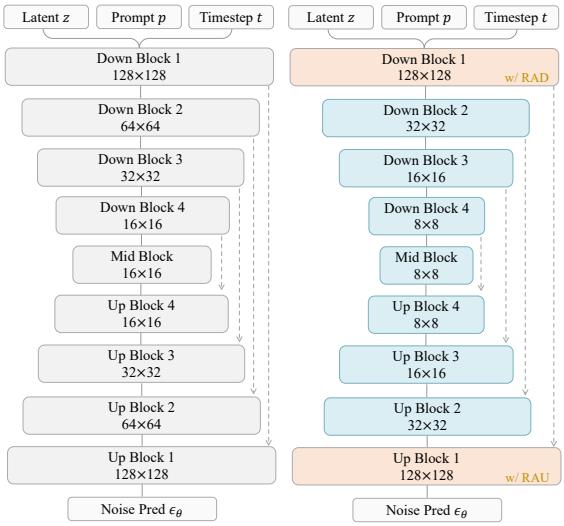
3.1 Preliminaries

U-Net architecture. The neural backbone of Stable Diffusion is implemented as a U-Net [6, 32, 33], which contains several Down Blocks, Up Blocks, and a Mid Block, as shown in Fig. 2a. The Mid Block remains unchanged in our method. Consequently, we omit it for the sake of simplicity. Each Down Block and Up Block can be written respectively as:

$$y = \mathcal{D}(\mathcal{F}(x, t, p)), \tag{1}$$

$$y = \mathcal{U}(\mathcal{F}(x, t, p)), \tag{2}$$

where x is the latent feature, t is the timestep, p is the prompt, $\mathcal{F}$ incorporates ResNet [10] layers and Vision Transformer [7] layers, which maintain the dimen-



(a) Vanilla U-Net **(b)** HiDiffusion RAU-Net

Fig. 2: Comparison between vanilla Stable Diffusion's U-Net architecture and our proposed HiDiffusion RAU-Net architecture on 1024×1024 resolution with SD 1.5 [32]. Parameters in all blocks are frozen. The main difference lies in the blue Blocks (differ in the dimensions of feature map) and orange Blocks (Our proposed RAD and RAU modules are incorporated into Block 1.).

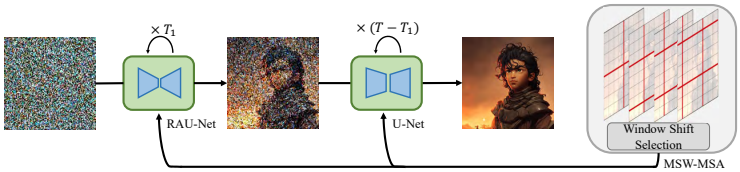


Fig. 3: The framework of HiDiffusion.

sions of the feature map. $\mathcal{D}(*)$ represents the downsampler and $\mathcal{U}(*)$ represents the upsampler. $\mathcal{D}(*)$ and $\mathcal{U}(*)$ in vanilla U-Net are computed as:

$$\mathcal{D}(x) = \mathcal{C}_{3,1,2,1}(x), \tag{3}$$

$$\mathcal{U}(x) = \mathcal{C}_{3,1,1,1}(interp(x, 2)), \tag{4}$$

where $\mathcal{C}_{k,p,s,d}$ means convolution filter with kernel size as k, padding size as p, stride as s, dilation rate as d. $interp(x, \beta)$ denotes an interpolation function that upsample the resolution by a factor of β.

Content generation over timestep. The diffusion model progressively performs the denoising process according to the noise schedule. Recent research [5, 13, 25, 32, 43] has found that the diffusion model displays varying denoising behavior over timestep. Diffusion models denoise from structures to details. They generate the low-frequency component in the early denoising stage and the high-frequency component in the late denoising stage.

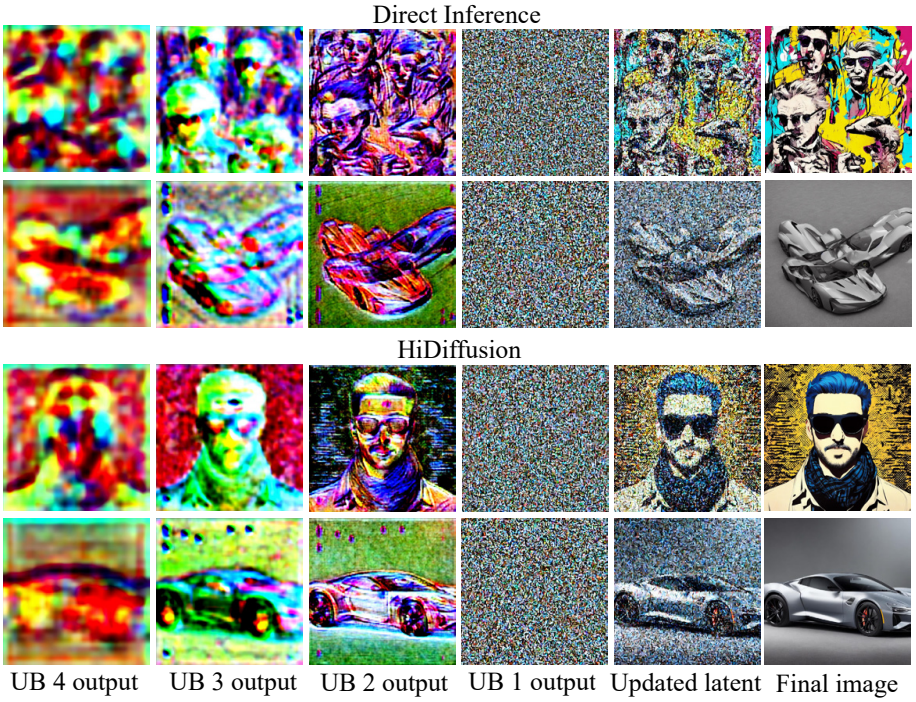


Fig. 4: 1024×1024 resolution images based on SD 1.5 [32]. We visualize the output feature map of U-Net blocks of the 30th step (50 DDIM steps). UB: Up Block. The object structure of the generated images (the last column) is highly correlated with the feature of the deep blocks (UB4, UB3, UB2) in the U-Net. Feature duplication happens when directly generating higher-resolution images and the duplicated features guide the generation direction to object duplication. HiDiffusion can mitigate feature duplication, enabling the generation of reasonable high-resolution images.

3.2 HiDiffusion

The HiDiffusion framework comprises two components: Resolution-Aware U-Net (RAU-Net) and Modified Shifted Window Multi-head Self-Attention (MSW-MSA). The RAU-Net is designed to overcome object duplication when scaling to higher resolution. MSW-MSA is introduced to improve the inference efficiency of diffusion for higher-resolution image synthesis. The overall framework of HiDiffusion is present in Fig. 3. For each method, we initially present the motivation experiments and then introduce our methods. This section is based on the 1024×1024 resolution image generation with SD 1.5 [32]. For other models and extreme resolution, please refer to the appendix for details.

Resolution-Aware U-Net. In this work, we investigate the feature map in the U-Net, aiming to uncover the cause of object duplication and resolve it.
Obeservation. *The generated image is highly correlated with the feature map of deep blocks in structures and feature duplication happens in the deep blocks.*

We provide empirical evidence to demonstrate this observation in Fig. 4. We discover that the the structure of the generated image follows the structural information of the feature maps of the deep blocks. The top block of the U-Net only maps the feature map of deep blocks to noise estimation. We also discover that feature duplication happens in the deep blocks of the U-Net, meaning that the features contain repeated structural information. The highly duplicated features guide the generation direction, resulting in object duplication in the image.

Based on the observation, we aim to reduce the feature duplication in the deep blocks to generate higher-resolution images. As the higher-resolution feature size of deep blocks is larger than the corresponding size in training, these blocks may fail to incorporate feature information globally to generate a reasonable structure. We contend that if the size of the higher-resolution features of deep blocks is reduced to the corresponding size in training, these blocks can generate reasonable structural information and alleviate feature duplication.

Inspired by this motivation, we propose Resolution-aware U-Net (RAU-Net), a simple yet effective method to dynamically resize the features to match the deep blocks. An illustrative comparison of the vanilla SD 1.5 U-Net and RAU-Net in generating 1024×1024 resolution images is presented in Fig. 2. We incorporate our Resolution-Aware Downsapler (RAD) in Down Block 1 as a substitute for the original downsampler, and likewise, we replace the original upsampler with the Resolution-Aware Upsampler (RAU) in Up Block 1. RAD downsamples the feature map to guarantee the dimensions of the resulting feature map align with those of the corresponding training images, thereby matching with the deep blocks. On the other hand, RAU simply upsamples the feature size to the desired resolution. Specifically, the RAD and RAU can be written as follows:

$$\mathcal{RAD}(x, \alpha) = \mathcal{R}(\mathcal{C}_{3,1,2,1}(x), \alpha), \tag{5}$$

$$\mathcal{RAU}(x, \beta) = \mathcal{C}_{3,1,1,1}(interp(x, \beta)), \tag{6}$$

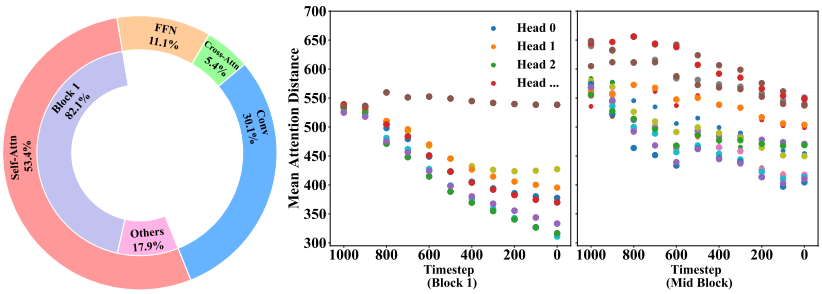
where α is the downsampling factor, β is the upsampling factor. $\mathcal{R}(*)$ can be achieved by adjusting the conventional downsampler parameters:

$$\mathcal{R}(\mathcal{C}_{3,1,2,1}(x), \alpha) = \mathcal{C}_{3,p,\alpha,d}(x), \tag{7}$$

or achieved by using adaptive pooling:

$$\mathcal{R}(\mathcal{C}_{3,1,2,1}(x), \alpha) = ada\_pool(\mathcal{C}_{3,1,2,1}(x), \frac{\alpha}{2}). \tag{8}$$

We mainly choose the first variant in this paper. We conduct ablations about RAD in the appendix. RAU can be simply achieved with an upscale interpolation such as bilinear interpolation. For 1024×1024 image generation, we need to downsample the feature map by a factor of 4 to match the deep blocks, i.e., $\alpha = 4$. This downsampling factor is twice the downsampling factor of the conventional downsampler. We set $d = 2$, $p = 2$. For RAU, we only need to adjust the interpolation factor to 4. Compared to the original samplers, both RAD and RAU do not introduce additional trainable parameters. Therefore, RAD and RAU can be integrated into vanilla U-Net without further fine-tuning.



(a) Operation consumption. **(b)** Mean attention distance.

Fig. 5: Analysis of the time consumption and mean attention distance. (a) The self-attention operation of Block 1 significantly dominates the time consumption. (b) A pronounced locality is evident in the self-attention mechanism of the top blocks.

Upon incorporating RAU-Net into SD 1.5 [32], we address the object duplication problem. But we also bring blurry images with bad details (please refer to the appendix). As mentioned in 3.1, diffusion models denoise from structures to details. Our RAU-Net introduces additional downsampling and upsampling operations, leading to a certain degree of information loss. In the early stages of denoising, RAU-Net can generate reasonable structures with minimal impact from information loss. However, in the later stages of denoising when generating fine details, the information loss in RAU-Net results in the loss of image details and a degradation in quality. Consequently, we establish a **Switching Threshold** T_1, such that when the denoising steps $t < T_1$, RAU-Net is employed, conversely, when the denoising steps $t \geq T_1$, vanilla U-Net is utilized. This simple adjustment can effectively counteract the information loss brought about by RAU-Net, significantly improving the image quality. Moreover, we observed that the parameter T_1 is not sensitive, with settings between 10 and 40 for 1024×1024 generation yielding notably superior performance with 50 DDIM steps. Please refer to the appendix for details.

Modified Shifted Window Attention Stable Diffusion combined with RAU-Net is capable of generating higher-resolution images with high quality. However, it still faces an efficiency challenge: unaffordable slow speed in generating higher-resolution images. In this section, we revisit the consumption and properties of operations in U-Net, trying to accelerate the diffusion model.

Obeservation. *The self-attention of the top blocks takes the dominant consumption. However, it demonstrates locality.*

Given a latent feature map with 128×128 (corresponding to 1024×1024 resolution in pixel space), Fig. 5a shows the time consumption of each operation in SD 1.5. It can be observed that self-attention, especially in Block 1 (i.e. Down Block 1 and Up Block 1), takes the dominant consumption. Driven by previous local self-attention works in vision [22, 28], we visualize the mean attention distance for each head across different timesteps of Block 1 (top block) and Mid Block (deep block), as shown in Fig. 5b. We surprisingly find that the self-attention mechanism in the top blocks demonstrates a pronounced locality.

Certain heads are observed to attend to approximately half of the image, while others focus on even more confined regions close to the query location.

According to this observation, it is suggested to propose local self-attention for efficient computation. We delve into how to design local attention to ensure acceleration while maintaining image quality. Based on window attention [22], we propose Modified Shifted Window Attention (MSW-MSA), a simple yet effective approach for lossless acceleration to replace the original global attention. Specifically, MSW-MSA has two modifications: **(i) Large window attention**. Fig. 5b indicates that the mean attention distance of the top blocks is local but not very small. This suggests that the small window attention utilized in vision field [22] may not be suitable for the diffusion model. To achieve a balance between acceleration and image quality, we opt for a larger window attention. The experimentally validated window size is $(H/2, W/2)$, where H and W respectively represent the height and width of the input feature (please refer to the appendix for the results from small to large window sizes). **(ii) Shift window operation across timestep**. Window shift operation is needed to introduce cross-window connections. However, Stable Diffusion transformer block has only one self-attention module that is unable to process window attention and shift window attention successively. To apply shift window operation, we propose to shift different strides based on the timesteps. Specifically, we adopt a random selection strategy, where at each timestamp, we randomly select a stride parameter from a fixed set of shift strides. This approach enables the integration of information from diverse windows. Our MSW-MSA can be written as:

$$y = \text{MSW-MSA}(x, w, s(t)) + x, \tag{9}$$

where w is the window size, $s(t)$ is the shifted stride function dependant on the timestep t.

We substitute the global self-attention in Block 1 with MSW-MSA. It is worth noting that while other blocks can integrate MSW-MSA, the resulting efficiency gains are not substantial. Experiments demonstrate that our MSW-MSA approach can significantly reduce time consumption **without compromising image quality** in higher-resolution image synthesis.

4 Experiments

4.1 Experiment Settings

In this work, we evaluate the performance of our HiDiffusion on SD 1.5 [32], SD 2.1 [32], SDXL Turbo [36] and SDXL [30]. We apply our approach to text-guided image synthesis on higher resolution ranging from 4× to even 16× times the training image resolution. For quantitative evaluation, we use Frechet Inception Distance (FID) [12] to measure the realism of the output distribution. FID downsamples all images to a common size of 299, which ignores high-resolution details. We further use patches FID (pFID) [3] to evaluate image details. CLIP Score [31] is proposed to evaluate the alignment between image and text. We

Method	Resolution	Latency (s) ↓	ImageNet			COCO		
			FID ↓	pFID ↓	CLIP ↑	FID ↓	pFID ↓	CLIP ↑
SD 1.5	1024 × 1024	16.23	25.55	36.36	0.295	38.21	49.20	0.309
SD 1.5 + HiDiffusion		**8.26(1.96×)**	**21.81**	**30.86**	**0.307**	**21.36**	**31.59**	**0.323**
SD 2.1		12.99	24.63	36.15	0.299	31.33	37.43	0.314
SD 2.1 + HiDiffusion		**7.33(1.77×)**	**22.34**	**32.71**	**0.309**	**20.77**	**31.51**	**0.326**
SDXL Turbo		5.72	74.23	76.08	0.300	23.45	35.10	0.325
SDXL Turbo+ HiDiffusion		**4.65(1.23×)**	**27.76**	**32.63**	**0.317**	**20.89**	**32.90**	**0.330**
SD 1.5	2048 × 2048	165.76	53.03	35.96	0.284	78.53	42.82	0.286
SD 1.5 + HiDiffusion		**58.38(2.83×)**	**27.33**	**33.42**	**0.307**	**28.93**	**34.70**	**0.321**
SD 2.1		118.32	60.60	41.64	0.281	82.74	47.62	0.289
SD 2.1 + HiDiffusion		**45.33(2.61×)**	**30.67**	**37.14**	**0.305**	**32.87**	**35.76**	**0.320**
SDXL		84.24	27.48	30.67	0.300	28.71	32.44	0.318
SDXL + HiDiffusion		**53.29(1.58×)**	**22.22**	**28.27**	**0.314**	**20.89**	**29.21**	**0.332**
SDXL†	4096 × 4096	769.65	118.30	89.96	0.277	-	-	-
SDXL + HiDiffusion†		**286.97(2.68×)**	**64.12**	**74.91**	**0.299**	-	-	-

Table 1: Comparison of vanilla Stable Diffusion and our Hidiffusion in zero-shot text-guided image synthesis on ImageNet and COCO dataset. † means we generate 1K images for quantitative evaluation due to the heavy computational burden.

compare our HiDiffusion with other methods on ImageNet [34] and COCO [21] datasets. Without further elaboration, we generate 10K (10 per class) images to compute metrics for ImageNet evaluation and generate 40,504 (1 caption per image) images from COCO 2014 validation captions to compute metrics for COCO evaluation. We use xFormers [18] by default. The model latency is measured on a single NVIDIA V100 with a batch size of 1.

We introduce the parameter setting of SD 1.5, please refer to the appendix for the parameter setting of other models. For 1024×1024 generation, we incorporate RAD and RAU in Block 1 and set $\alpha = \beta = 4$. We set the window size as $(64, 64)$. The predefined set of shift strides is $\{(0, 0), (16, 16), (32, 32), (48, 48)\}$. All experiments are conducted with 50 DDIM steps. The classifier-free guidance scale is 7.5. The switching threshold T_1 is set as 20. When extended to 2048×2048, we can simply set $\alpha = 8$. However, a sharp change in resolution caused by interpolation may bring blurriness, hence we adopt a progressive approach by incorporating RAU and RAD with $\alpha = \beta = 4$ into Block 1 and Block 2, respectively, This allows the feature map to gradually match the deep blocks. Please refer to the appendix for more details.

4.2 Main results

In this section, We incorporate our method into SD 1.5 [32], SD 2.1 [32], SDXL Turbo [36] and SDXL [30] to evaluate the effectiveness of our method. SD 1.5 and SD 2.1 are capable of generating images with 512×512 resolution. We integrate HiDiffusion into them to scale the resolution to 1024×1024 and 2048×2048. We use HiDiffusion to scale the generation resolution of SDXL Turbo to 1024×1024. For SDXL, which is trained for generating 1024×1024 images, we incorporate our

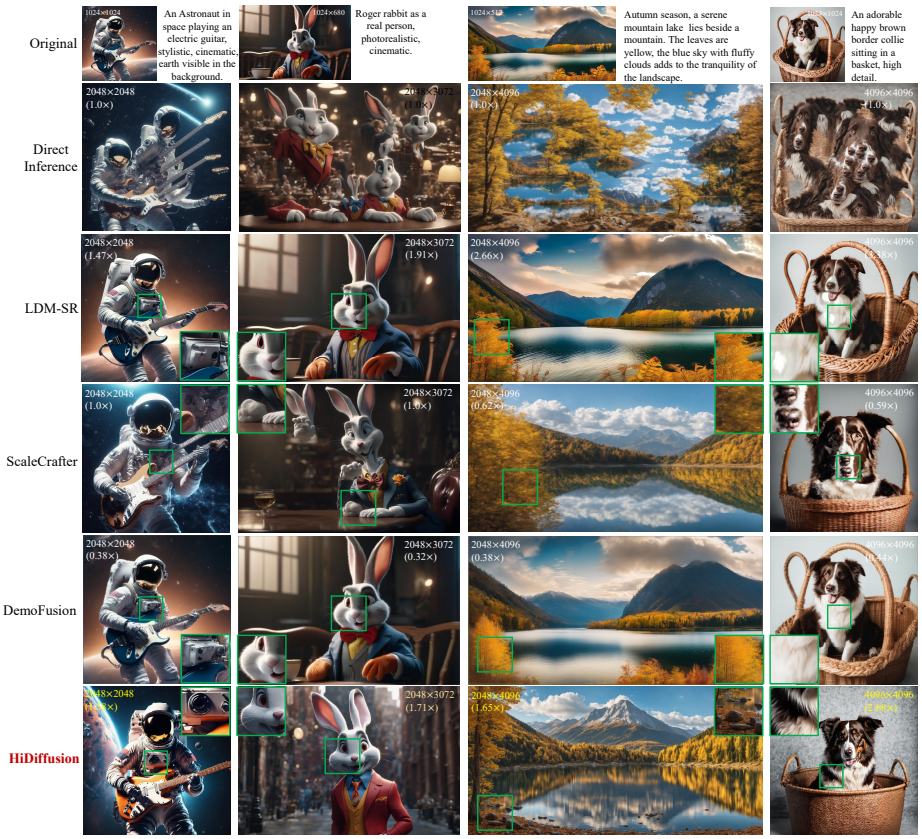


Fig. 6: Qualitative comparison with other methods based on SDXL [30]. The input prompt is located to the right of the original image. The first line of text in the image indicates the image resolution, while the second line indicates the inference speed relative to direct inference. Best viewed when zoomed in.

method to scale the resolution to 2048×2048 and 4096×4096. Besides fixed aspect ratios, we also generate images with various aspect ratios, such as 512×2048, 1280×1024 and 2048×4096, and so on, please refer to the appendix. We compare our method with vanilla SD and the higher-resolution synthesis methods ScaleCrafter [11] and DemoFusion [8]. For the acceleration of diffusion model, we compare the diffusion acceleration method Token Merge for Stable Diffusion (ToMeSD) [2] and DeepCache [26] with our proposed MSW-MSA. Moreover, we compare our method with super-resolution method for a thorough evaluation, even though the latter requires a large number of high-resolution images and extra training efforts to train a super-resolution model.

Comparision with vanilla SD. In Fig. 6, we show qualitative comparison between the vanilla SD (direct inference) and our method. It can be easily seen the vanilla SD suffers from object duplication and degradation in visual quality as well. In contrast, our HiDiffusion mitigates the duplication problem

Backbone	Method	Resolution	Latency (s) ↓	ImageNet			COCO		
				FID ↓	pFID ↓	CLIP ↑	FID ↓	pFID ↓	CLIP ↑
SD 1.5	ScaleCrafter [11]	1024 × 1024	17.94	54.90	74.39	0.302	81.68	87.32	0.318
	HiDiffusion (ours)		**8.26**	**50.14**	**65.80**	**0.307**	**80.73**	**84.19**	**0.322**
SD 2.1	ScaleCrafter [11]		14.58	57.87	75.39	0.303	80.02	84.69	0.321
	HiDiffusion (ours)		**7.33**	**50.20**	**68.69**	**0.309**	**79.19**	**82.47**	**0.325**
SD 1.5	ScaleCrafter [11]	2048 × 2048	287.89	65.04	86.83	0.299	98.30	104.49	0.309
	HiDiffusion (ours)		**58.38**	**53.35**	**65.95**	**0.307**	**86.15**	**86.82**	**0.319**
SD 2.1	ScaleCrafter [11]		216.48	78.12	102.14	0.295	110.75	122.49	0.299
	HiDiffusion (ours)		**45.33**	**57.71**	**72.09**	**0.306**	**88.56**	**87.07**	**0.317**
SDXL	ScaleCrafter [11]		85.83	49.97	72.64	0.310	**82.14**	90.45	0.329
	DemoFusion [8]		222.79	48.36	66.41	0.311	85.92	85.59	0.331
	HiDiffusion (ours)		**53.29**	**47.01**	**62.29**	**0.315**	84.66	**84.58**	**0.333**
SDXL	ScaleCrafter [11]	4096 × 4096	1298.39	78.90	102.63	0.305	-	-	-
	DemoFusion [8]		1735.58	**58.93**	76.53	**0.311**	-	-	-
	HiDiffusion (ours)		**286.97**	64.12	**74.91**	0.307	-	-	-

Table 2: Comparison of high-resolution generation method and our HiDiffusion in zero-shot text-guided image synthesis on ImageNet and COCO dataset.

and holds more realistic image structures simultaneously. The quantitative results are shown in Tab. 1. Our approach outperforms vanilla SD in both quality and image-text alignment. We achieve much better metric scores across all experiment settings, especially for the images with much higher resolution (a significant FID improvement from 78.53 to 28.93 for SD 1.5 on COCO dataset with 2048×2048 resolution). Furthermore, HiDiffusion significantly accelerates diffusion inference. For instance, when incoporating HiDiffusion, SDXL is 2.68× faster than the vanilla model when generating images with 4096×4096 resolution.

Comparison with higher-resolution synthesis methods. We present a qualitative comparison between ScaleCrafter [11], DemoFusion [8] and our method in Fig. 6. We observe that all three methods can generate reasonable structures. But our method can generate much richer details than ScaleCrafter and DemoFusion. Tab. 2 shows quantitative results. We generate 1K images for both ImageNet and COCO evaluations due to the heavy computational burden. Our method outperforms ScaleCrafter across almost all models and achieves comparable or better performance than DemoFusion. It is worth noting that we significantly surpass ScaleCrafter and DemoFusion in efficiency: HiDiifusion is 1.5-5× faster than ScaleCrafter and is 4-6× faster than DemoFusion.

Comparison with diffusion super-resolution models. Instead of directly generating higher-resolution images using a single diffusion model, a more commonly used approach in the community is to generate images with original resolution using Stable Diffusion and scale them to higher resolution using an extra super-resolution model. Although this approach requires additional high-resolution training datasets and extensive training efforts to train a large super-resolution model, we compare it for a thorough comparison, despite the inherent unfairness to our one-stage and training-free method. We compare our method with a pretrained Stable Diffusion super-resolution model LDM-SR [32]. Tab. 3 shows the quantitative results. In terms of generation efficiency, our method out-

Method	Resolution	Latency (s) ↓	FID ↓	pFID ↓	CLIP ↑
SD 1.5 + LDM-SR*	1024 × 1024	18.61	**17.56**	36.39	**0.307**
SD 1.5 + HiDiffusion		**8.26(2.25×)**	21.81	**30.86**	**0.307**
SD 2.1 + LDM-SR*		18.48	**18.54**	37.99	0.308
SD 2.1 + HiDiffusion		**7.33(2.52×)**	22.34	**32.71**	**0.309**
SD 1.5 + LDM-SR*	2048 × 2048	73.03	**17.32**	35.57	**0.308**
SD 1.5 + HiDiffusion		**58.28(1.25×)**	27.33	**33.42**	0.307
SD 2.1 + LDM-SR*		72.90	**18.16**	41.87	**0.308**
SD 2.1 + HiDiffusion		**45.33(1.61×)**	30.67	**37.14**	0.304
SDXL + LDM-SR*		57.28	**46.73**	68.98	**0.315**
SDXL + HiDiffusion		**53.29(1.07×)**	47.01	**62.29**	**0.315**
SDXL + LDM-SR*	4096 × 4096	**227.57**	**57.04**	79.66	**0.315**
SDXL + HiDiffusion		286.97	64.12	**74.91**	0.307

Table 3: Comparison of diffusion super-resolution and our method in zero-shot text-guided image synthesis on ImageNet dataset. * is a two-stage method, requiring extra high-resolution datasets and training efforts to train a large super-resolution model. Our approach is one-stage and can generate high-resolution images without any extra high-resolution data collection and training costs.

performs LDM-SR significantly when based on SD 1.5 and SD 2.1. On SDXL, our speed is roughly on par with that of LDM-SR. Both LDM-SR and our method are capable of generating plausible structures. However, our method exhibits lower pFID. This indicates that the images generated by our method are more detailed. We visualize the synthesized samples in Fig. 6. Compared with LDM-SR, a distinction can be observed in terms of visual image detail quality. Our method directly generates content on a 2048×2048 or 4096×4096 canvas, resulting in higher richness, sharper characteristics, and fine-grained details.

Comparison with diffusion acceleration method. We compare our method with ToMeSD [2] and DeepCache [26]. The comparison is implemented at the resolutions of 1024×1024 and 2048×2048 based on SD 1.5 with RAU-Net. Tab. 4 shows the quantitative results on ImageNet [34]. Our proposed MSW-MSA outperforms ToMeSD and DeepCache across almost all metrics. Although ToMeSD and DeepCache have achieved significant acceleration, they have also compromised image quality. In contrast, our MSW-MSA can achieve or even surpass the acceleration effects of ToMeSD and DeepCache without compromising image quality. Please refer to the appendix for the visual sample comparison.

4.3 Ablation study

We ablate the components of HiDiffusion on 1024×1024 resolution image generation based on SD 1.5. Tab. 5 shows quantitative results of all possible combinations. It can be seen that both RAU-Net and MSW-MSA bring improvements in performance and speed. RAU-Net uses RAD to adjust the feature size to the training dimensions, resulting in speed benefits. Meanwhile, the window attention in MSW-MSA prevents an overwhelming number of tokens from global interaction, thereby reducing token homogenization and enhancing performance.

Method	Resolution	Latency (s)	FID ↓	pFID ↓	CLIP ↑
Baseline	1024×1024	14.31	22.93	32.80	**0.307**
ToMeSD [2]		8.73	22.76	35.23	0.305
DeepCache (interval=3) [26]		**6.62**	25.31	35.61	0.306
DeepCachee (interval=2) [26]		8.49	24.07	33.89	0.306
MSW-MSA (ours)		8.26	**21.80**	**30.86**	**0.307**
Baseline	2048×2048	151.58	28.21	33.79	0.306
ToMeSD [2]		67.02	27.49	34.38	0.305
DeepCache (interval=3) [26]		70.70	34.49	42.13	0.305
DeepCache (interval=2) [26]		90.20	32.72	41.44	**0.307**
MSW-MSA (ours)		**58.38**	**27.33**	**33.42**	**0.307**

Table 4: Quantitative evaluation of diffusion acceleration methods and our proposed MSW-MSA in zero-shot text-guided image synthesis on ImageNet. Baseline indicates SD 1.5 with RAU-Net.

RAU-Net	MSW-MSA	Latency (s)	FID ↓	pFID ↓	CLIP ↑
		16.23	25.55	36.36	0.295
✓		14.31	22.93	32.80	0.307
	✓	10.15	23.28	33.84	0.297
✓	✓	**8.26**	**21.81**	**30.86**	**0.307**

Table 5: Ablation about the components of HiDiffusion with SD 1.5 in zero-shot text-guided image synthesis on ImageNet.

This perspective is discussed in [16]. When both are used simultaneously, the optimal result can be achieved. We provide visual samples in the appendix.

5 Conclusion

In this paper, we propose a tuning-free framework named HiDiffusion for higher-resolution image generation. HiDiffusion includes Resolution-Aware U-Net (RAU-Net) that makes higher-resolution generation possible and Modified Shifted Window Multi-head Self-Attention (MSW-MSA) that makes higher-resolution generation efficient. Empirically, HiDiffusion can be incorporated into SD 1.5 [32], 2.1 [32], XL [30], and XL Turbo [36], and scale them to generate 1024×1024, 2048×2048, or even 4096×4096 resolution images, while significantly reducing inference time. Compared to previous higher-resolution generation methods, we can generate images with richer details in less inference time. We hope our work can bring insight to future works about the scalability of diffusion models.

Limitations and future work: Our approach involves directly harnessing the intrinsic potential of Stable Diffusion without any additional training or fine-tuning, hence some inherent issues posed by Stable Diffusion persist, such as the requirement for prompt engineering to obtain more promising images. Furthermore, we can explore better ways to integrate with super-resolution models to achieve more amazing generation outcomes.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: ICML (2023) 4
2. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: CVPRW. pp. 4598–4602 (2023) 3, 4, 11, 13, 14
3. Chai, L., Gharbi, M., Shechtman, E., Isola, P., Zhang, R.: Any-resolution training for high-resolution image synthesis. In: ECCV. pp. 170–188. Springer (2022) 9
4. Chen, Y.H., Sarokin, R., Lee, J., Tang, J., Chang, C.L., Kulik, A., Grundmann, M.: Speed is all you need: On-device acceleration of large diffusion models via gpu-aware optimizations. In: CVPR. pp. 4650–4654 (2023) 4
5. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: CVPR. pp. 11472–11481 (2022) 5
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. vol. 34, pp. 8780–8794 (2021) 4
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 4
8. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: Demofusion: Democratising high-resolution image generation with no $$$. In: CVPR (2024) 2, 4, 11, 12
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR. pp. 12873–12883 (2021) 4
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 4
11. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In: ICLR (2024) 2, 3, 4, 11, 12
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 9
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020) 1, 5
14. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. arXiv preprint arXiv:2301.11093 (2023) 4
15. Jiménez, Á.B.: Mixture of diffusers for scene composition and high resolution image generation. arXiv preprint arXiv:2302.02412 (2023) 4
16. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free diffusion model adaptation for variable-sized text-to-image synthesis. arXiv preprint arXiv:2306.08645 (2023) 2, 4, 14
17. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. In: NeurIPS (2023) 4
18. Lefaudeux, B., Massa, F., Liskovich, D., Xiong, W., Caggiano, V., Naren, S., Xu, M., Hu, J., Tintore, M., Zhang, S., Labatut, P., Haziza, D.: xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers (2022) 10
19. Li, L., Li, H., Zheng, X., Wu, J., Xiao, X., Wang, R., Zheng, M., Pan, X., Chao, F., Ji, R.: Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. pp. 7105–7114 (2023) 4
20. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In: NeurIPS (2024) 3, 4

21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014) 10
22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021) 3, 4, 8, 9
23. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: NeurIPS. pp. 5775–5787 (2022) 3, 4
24. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022) 3
25. Ma, H., Zhang, L., Zhu, X., Feng, J.: Accelerating score-based generative models with preconditioned diffusion sampling. In: ECCV. Springer (2022) 5
26. Ma, X., Fang, G., Wang, X.: Deepcache: Accelerating diffusion models for free. In: CVPR (2024) 3, 4, 11, 13, 14
27. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: CVPR. pp. 14297–14306 (2023) 3
28. Pan, X., Ye, T., Xia, Z., Song, S., Huang, G.: Slide-transformer: Hierarchical vision transformer with local self-attention. In: CVPR. pp. 2082–2091 (2023) 8
29. Pan, Z., Gherardi, R., Xie, X., Huang, S.: Effective real image editing with accelerated iterative diffusion inversion. In: ICCV. pp. 15912–15921 (2023) 4
30. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 1, 2, 3, 9, 10, 11, 14
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 9
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 14
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015) 4
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**, 211–252 (2015) 10, 13
35. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022) 3, 4
36. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 3, 9, 10, 14
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS. pp. 25278–25294 (2022) 1
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 1, 3, 4
39. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS. pp. 11895–11907 (2019) 1
40. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021) 1, 4

41. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint arXiv:2309.03350 (2023) 4
42. Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., Li, Z.: Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. arXiv preprint arXiv:2304.06648 (2023) 4
43. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: CVPR. pp. 22552–22562 (2023) 5
44. Zheng, Q., Guo, Y., Deng, J., Han, J., Li, Y., Xu, S., Xu, H.: Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. arXiv preprint arXiv:2308.16582 (2023) 2