DiffuseHigh: Training-free Progressive High-Resolution Image Synthesis through Structure Guidance

Younghyun Kim[1*], **Geunmin Hwang**[1*], **Junyu Zhang**[2, 3], **Eunbyung Park**[1,3†]

[1]Department of Artificial Intelligence, Sungkyunkwan University
[2]School of Computer Science and Engineering, Central South University
[3]Department of Electrical and Computer Engineering, Sungkyunkwan University

Abstract

Large-scale generative models, such as text-to-image diffusion models, have garnered widespread attention across diverse domains due to their creative and high-fidelity image generation. Nonetheless, existing large-scale diffusion models are confined to generating images of up to 1K resolution, which is far from meeting the demands of contemporary commercial applications. Directly sampling higher-resolution images often yields results marred by artifacts such as object repetition and distorted shapes. Addressing the aforementioned issues typically necessitates training or fine-tuning models on higher-resolution datasets. However, this poses a formidable challenge due to the difficulty in collecting large-scale high-resolution images and substantial computational resources. While several preceding works have proposed alternatives to bypass the cumbersome training process, they often fail to produce convincing results. In this work, we probe the generative ability of diffusion models at higher resolution beyond their original capability and propose a novel progressive approach that fully utilizes generated low-resolution images to guide the generation of higher-resolution images. Additionally, we integrate an image sharpening operation into our pipeline, further enhancing image quality. Our method obviates the need for additional training or fine-tuning which significantly lowers the burden of computational costs. Extensive experiments and results validate the efficiency and efficacy of our method.

Code — https://github.com/yhyun225/DiffuseHigh
Extended version — https://arxiv.org/abs/2406.18459

Introduction

With the establishment of diffusion models, there have been rapid advancements across various domains, including audio synthesis (Chen et al. 2020; Lam et al. 2021; Liu et al. 2023), image synthesis (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021), video generation (Ho et al. 2022; Wang et al. 2023; Blattmann et al. 2023), and 3D generation (Poole et al. 2022; Wang et al. 2024; Shi et al. 2023; Tang et al. 2023). Notably, text-to-image diffusion models (Rombach et al. 2022a; Saharia et al. 2022b; Ramesh et al. 2022) have attracted considerable attention due to their ability to generate visually captivating images using intuitive, human-friendly natural language descriptions. Stable Diffusion (SD) and Stable Diffusion XL (SDXL), the open-source text-to-image diffusion models trained on a large-scale online dataset (Schuhmann et al. 2022), have emerged as prominent tools for a diverse range of generative tasks. These tasks include but are not limited to image editing (Avrahami, Lischinski, and Fried 2022; Hertz et al. 2022; Tumanyan et al. 2023), inpainting (Rombach et al. 2022a; Saharia et al. 2022a; Lugmayr et al. 2022), super-resolution (Rombach et al. 2022a; Saharia et al. 2022c; Gao et al. 2023), and image-to-image translation (Brooks, Holynski, and Efros 2023; Zhang, Rao, and Agrawala 2023).

Despite the promising performance exhibited by SD and SDXL, they encounter limitations when generating images at higher-resolutions beyond their training resolution. The direct inference of unseen high-resolution samples often reveals repetitive patterns and irregular structures, particularly noticeable in object-centric samples, as discussed in prior works (He et al. 2023; Du et al. 2024). While a straightforward approach might involve training or fine-tuning diffusion models on higher-resolution images, several challenges impede this approach. First, collecting text-image pairs of higher-resolution is not readily feasible. Second, training on large-resolution images demands substantial computational resources due to the increased size of the intermediate features. Furthermore, capturing and learning the features from high-dimensional data often requires a greater model capacity (more model parameters), leading to further computational strain on the training process.

Several tuning-free (Bar-Tal et al. 2023; Lee et al. 2024; He et al. 2023; Du et al. 2024) methods proposed various approaches to adapt pre-trained diffusion models to higher resolutions beyond their original settings. MultiDiffusion (Bar-Tal et al. 2023) and SyncDiffusion (Lee et al. 2024) employ joint diffusion processes with overlapping windows, each corresponding to different region within the generating image. These models can produce images of arbitrary shape, but the resulting image involves object repetition issues since the non-overlapping patches do not correlate to each other, lacking perception of global context during the denoising process. ScaleCrafter (He et al. 2023), on the other hand, extends the receptive field of the diffusion model by

[*]These authors contributed equally.
[†]corresponding author

Figure 1: **Qualitative examples of the proposed DiffuseHigh pipeline.** *DiffuseHigh* enables the pre-trained text-to-image diffusion models (SDXL in this figure) to generate higher-resolution images than the originally trained resolution, e.g., 4×, 16×, without any training or fine-tuning.

dilating the pre-trained convolution weights of the denoising UNet (Ronneberger, Fischer, and Brox 2015). While it effectively addresses repetition issues in certain instances, its success heavily depends on the extensive search of the hyperparameters.

In this work, we investigate the text-to-image diffusion model's capability of generating previously unseen high-resolution images and introduce a novel approach that does not involve any training (or fine-tuning) and additional modules. Moreover, our proposed method does not modify the pre-trained weights or the architecture of the denoising network, which eliminates the labor of searching for the optimal hyperparameters involved in the pipeline, and is more robust to certain hyperparameters. We posit that text-to-image diffusion models trained on internet-scale datasets innately possess the potential to generate images at resolutions higher than their training resolution thanks to its convolutional architecture (Rombach et al. 2022a) and broad data distribution coverage.

We introduce a novel progressive high-resolution image generation pipeline, dubbed *DiffuseHigh*, where relatively low-resolution (training-resolution of the pre-trained diffusion models) images serve as structural guidance for generating higher-resolution images. Inspired by the recent literature (Meng et al. 2021; Podell et al. 2023; Guo et al. 2024), our proposed pipeline involves a noising-denoising loop to synthesize higher-resolution images. First, we generate the low-resolution image using the base diffusion model and upsample it with arbitrary interpolation, e.g., bilinear interpolation. Then, we add sufficient noise to obfuscate the fine details of the interpolated images. Finally, we perform the reverse diffusion process to denoise those images to infuse the high-frequency details to synthesize higher-resolution images and repeat this process until we obtain the desired resolution images. This approach leverages the overall structure from the low-resolution image, effectively addressing repetition issues observed in the prior methods.

However, the 'adding noise to damage the images' approach poses several challenges. If we add too much noise, then we lose most of the structure in the low-resolution images, resulting in repetitive outcomes similar to those we generate from scratch. On the other hand, if we introduce a minimal amount of noise, the generated higher-resolution images do not show notable differences from the interpolated images, losing the opportunity to synthesize high-frequency details. In addition, finding adequate noise relies on both the content of the image and the pre-trained models, which makes it challenging to offer precise suggestions to users.

To resolve the issues above, we propose a principled way of preserving the overall structure from the low-resolution image for the suggested progressive pipeline. We employ a frequency-domain representation to extract the global structure as well as detailed contents from the low-resolution images. More specifically, we adopt the Discrete Wavelet Transform (DWT) to obtain essential contents, e.g., the approximation coefficient, which we then incorporate into the denoising procedure to ensure that the resulting image remains consistent and does not deviate excessively. Fig. 2 provides an overview of the overall pipeline of our method.

The contributions of our work are summarized as follows:

- We suggest a novel training-free progressive high-resolution image synthesis pipeline called *DiffuseHigh*, in which a lower-resolution image acts as a guide for generating higher-resolution images.
- Our proposed method involves Discrete Wavelet Transform (DWT)-based structure guidance during the denoising process, which enhances both the structural properties and fine details of the generated samples.
- We conduct comprehensive experiments and ablation studies on high-resolution image synthesis, demonstrating the superiority and versatility of our method.

Related Works

Text-to-Image Generation Recently, diffusion models (DMs) have gained popularity for their ability to produce high-quality images (Peebles and Xie 2023), showcasing great potential in text-to-image generation (Nichol et al. 2021; Ho et al. 2022; Ramesh et al. 2022). Especially the pioneering work, Stable Diffusion (Rombach et al. 2022a) and Stable Diffusion XL (Podell et al. 2023) have garnered broad attention due to their astonishing image quality and computational efficiency. Moreover, thanks to their large-scale training, they have been applied to various text-to-image tasks (Li et al. 2024; Nichol et al. 2021) by fine-tuning (Ruiz et al. 2023) or using training-free (Ramesh et al. 2021) methods.

High-resolution Image Synthesis Despite advancements in diffusion-based image synthesis methods, achieving high-resolution image generation remains elusive. Direct inference of SD and SDXL produces samples with repetitive patterns and irregular structures (He et al. 2023). Previous studies have tackled these challenges through training from scratch or fine-tuning (Xie et al. 2023; Zheng et al. 2023; Guo et al. 2024). However, these methods often necessitate substantial computational resources and considerable amount of high-resolution text-paired training dataset. Consequently, there is a growing trend towards training-free methods for generating high-resolution images.

ScaleCrafter (He et al. 2023) employs dilated convolution to modify the receptive field of convolutions in denoising UNet, enabling high-resolution image generation without the need for training. FouriScale (Huang et al. 2024) further introduces low pass filtering on each input feature map to remove aliasing effect, caused by spatial down-sampling of features, thus constructing consistency in frequency spectrum across different scales of the feature maps. HiDiffusion (Zhang et al. 2023) identifies that the object repetition problem primarily originates from the deep blocks in the denoising UNet and proposes alternative UNet which dynamically adjust the feature map size during the denoising process. Additionally, they successively reduce the computational burden by modifying the self-attention blocks of the UNet. However, we argue that modifying the weights or the architecture of the pre-trained diffusion model has risk of degrading the model performance, often resulting in undesirable deformations in images (See Fig.4). DemoFusion (Du et al. 2024) leverages skip residual connections and dilated sampling to generate higher-resolution images in a progressive manner. Despite their efforts, it suffers from the irregular patterns and repetition of small objects in localized areas of the result images, and also from the slow generation speed. AccDiffusion (Lin et al. 2024) addresses these issues with patch-wise prompt and improved dilated sampling, but still suffers from extremely slow inference speed.

Concurrently, ResMaster (Shi et al. 2024) also proposed an algorithm that leverages the low-frequency information of the latent of the guidance image, in order to provide desirable global semantics during the denoising process. Different from theirs, we explicitly obtain structural guidance from the reconstructed image by utilizing DWT.

Method

Our work aims to generate higher-resolution images over training size given textual prompts with a text-to-image diffusion models in a training-free manner. In this work, we mainly utilize SDXL (Podell et al. 2023) as our base model. We provide preliminaries related to our work in the appendix of the extended version (Kim et al. 2024).

Problem Formulation

Given a text description y and SDXL $D_\phi(\cdot)$ trained on fixed-size images $x_0^L \in \mathbb{R}^{h\times w\times 3}$, our objective is to generate higher resolution images $x_0^H \in \mathbb{R}^{H\times W\times 3}$ without training or modifying ϕ, where $h \ll H, w \ll W$.

Progressive High-Resolution Image Generation

We first present the progressive high-resolution image generation strategy of *DiffuseHigh*, equipped with a pretrained SDXL model. Initially, given text prompt y, our method starts with a clean image $x_0^L \in \mathbb{R}^{h\times w\times 3}$, either generated by the SDXL or provided by the user. Assuming alignment between the generated image and the provided text, we employ arbitrary interpolation, e.g., bilinear interpolation, to upscale the image:

$$\tilde{x}_0^H = \texttt{INTERP}(x_0^L) \in \mathbb{R}^{H\times W\times 3} \tag{1}$$

Note that the details of $\tilde{x}_0^H$ lack clarity due to the nature of the interpolation, which entails averaging neighboring pixel values to compose newly introduced pixels.

In order to infuse the appropriate details into $\tilde{x}_0^H$, we first add noise corresponding to the diffusion timestep $\tau < T$ to its latent code $\tilde{z}_0^H = \mathcal{E}(\tilde{x}_0^H)$:

$$\tilde{z}_\tau^H = \tilde{z}_0^H + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\tau^2 I), \tag{2}$$

where σ_τ^2 is the variance of the Gaussian noise at timestep τ. We selected the noising diffusion timestep τ where the

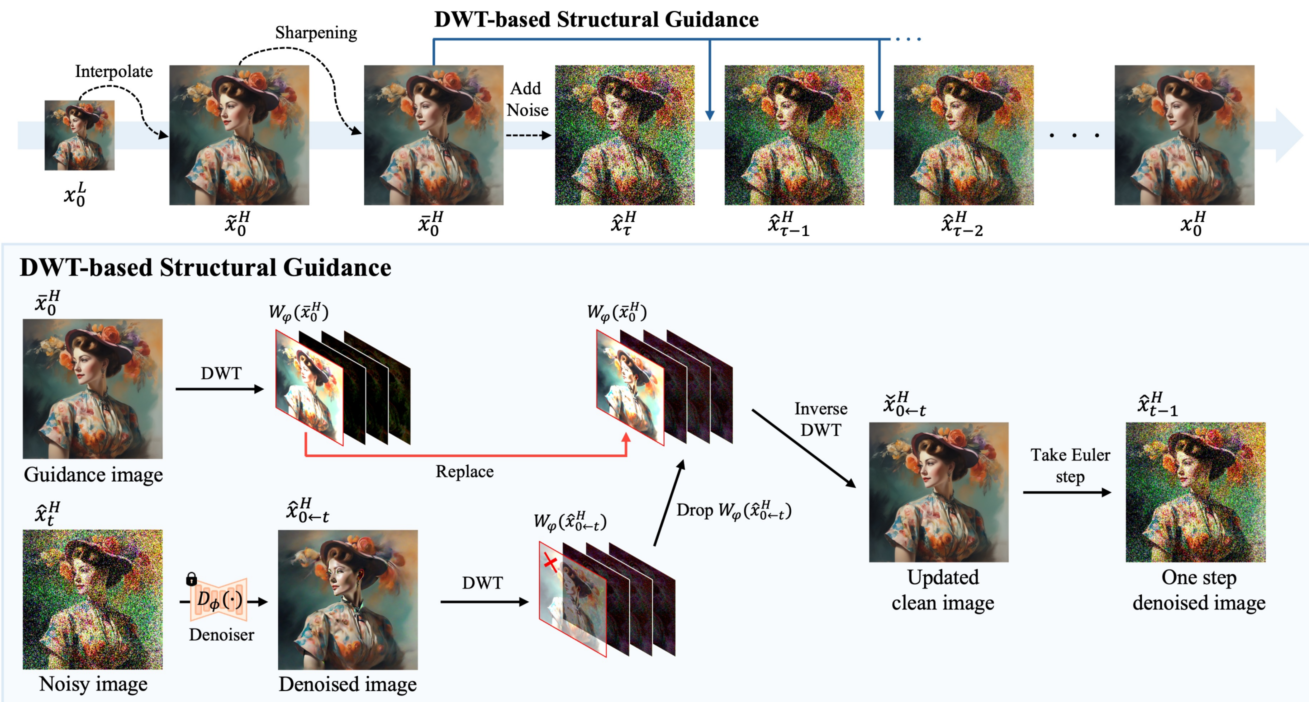

Figure 2: **Progressive High-Resolution *DiffuseHigh* Pipeline.** Overall pipeline of our proposed *DiffuseHigh*. For simplicity, we did not depict transformation between latent space and pixel space.

noisy image reconstructed from the latent decoder $\mathcal{D}(\cdot)$, $\hat{x}_\tau^H = \mathcal{D}(\hat{z}_\tau^H)$, preserves the global structures. Then the denoising network $D_\phi(\cdot)$ performs the iterative reverse process on the noisy latent representation $\hat{z}_\tau^H$ to recover the clean latent z_0^H. Finally, we obtain the desired high-resolution image $x_0^H \in \mathbb{R}^{H\times W\times 3}$ by employing the latent decoder, i.e., $x_0^H = \mathcal{D}(z_0^H)$. We repeat this process iteratively until we obtain the desired higher-resolution image.

The noising-denoising technique adopted in our work gradually projects the sample onto the manifold of natural, highly detailed images that the diffusion model has learned. As shown in Make-a-Cheap-Scaling (Guo et al. 2024), this process enables the injection of high-frequency details into the interpolated high-resolution image. Nonetheless, we observed numerous instances where solely applying this simple approach degraded the image quality, typically suffering from repeated small objects or deformed local details in the image. This lead us to develop a more principled way to uphold the overall structure and maintain the quality of the generated higher-resolution images.

Structural Guidance through DWT

To remedy the aforementioned issues, we hereby introduce a **structural guidance** by incorporating a Discrete Wavelet Transform (DWT). This method aims to enhance the fidelity of generated images by encouraging the preservation of crucial features from the low-resolution input.

Let φ be the two-dimensional scaling function, and ψ^H, ψ^V, ψ^D the two-dimensional wavelets, each corresponding to the horizontal (H), vertical (V), and diagonal directions (D), respectively. Then, the single level 2D-DWT decomposition of the image x can be written as follows:

$$\mathtt{DWT}(x) := \{W_\varphi(x)\} \cup \{W_{\psi^i}(x)\}_{i\in\{H,V,D\}}, \tag{3}$$

where $W_\varphi(x)$ the approximation coefficient, and $W_{\psi^i}(x)$ the detail coefficients along the direction $i \in \{H, V, D\}$.

Considering that $W_\varphi(x)$ contains the global features of the image x, given an interpolated image $\tilde{x}_0^H \in \mathbb{R}^{H\times W\times 3}$ obtained from Eq. (1), we extract its low-frequency component $W_\varphi(\tilde{x}_0^H)$ utilizing the DWT, which encapsulates the overall structure and coarse details of the image. Then, during the progressive denoising process, we replace the low-frequency component of the estimated clean image $\hat{x}_{0\leftarrow t}^H = \mathcal{D}(\hat{z}_{0\leftarrow t}^H)$, with the extracted low-frequency component at timestep 't' as follows:

$$\check{x}_{0\leftarrow t}^H = \mathtt{iDWT}(\{W_\varphi(\tilde{x}_0^H)\} \cup \{W_{\psi^i}(\hat{x}_{0\leftarrow t}^H)\}_{i\in\{H,V,D\}}) \tag{4}$$

where $\hat{z}_{0\leftarrow t}^H = D_\phi(\hat{z}_t^H; \sigma_t)$ and $\mathtt{iDWT}$ denotes inverse DWT. Then, the updated estimated clean image $\check{x}_{0\leftarrow t}^H$ is encoded back into the latent space to sample the next noisy latent.

Previous studies (Ho, Jain, and Abbeel 2020; Rombach et al. 2022a) present that the reverse process over each timestep performs denoising on different **levels** of the image, from semantic to perceptual, or low-frequency to high frequency details. Since the global structures and low frequency details are fixed and barely changed at the latter part of the denoising process, we found it beneficial to apply our

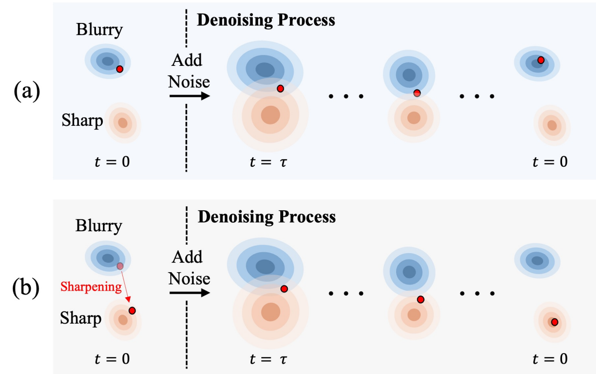

Figure 3: **Data sample toward sharp data distribution mode with sharpening.** (a) Without sharpening, (b) With sharpening. Red dot represents the data point. We encourage data point to move toward the sharp data distribution mode during denoising process by sharpening the blurry image.

structural guidance only at the early stages of our denoising process. Furthermore, this strategy significantly lowers the computational burden of our pipeline since our pipeline acquires low-frequency guidance from the reconstructed image, which requires frequent transition between the latent space and the pixel space. Empirically, we found that applying structural guidance steps $\delta = 5$ out of $\tau = 15$ steps yields the best results.

Boosting the Image Quality with Sharpening

Our proposed structural guidance effectively transfers the correct global context from a low-resolution image to a high-resolution image, maintaining the global coherence of the image. However, the generated image often appears blurry with smooth textures. We hypothesize that this blurriness arises from the interpolation used in our pipeline for the following reasons: (1) It is apparent that the diffusion models trained on large-scale datasets have the prior of blurry samples. Adding noise to the interpolated image, which lies near the blurry data distribution mode, is more likely to result in a blurry image after the denoising process (Fig. 3 (a)). (2) Interpolating a low-resolution image involves averaging neighboring pixel values, thus creating smooth transitions between pixels. These low intensity changes in object boundaries and edges are easily incorporated into the low-frequency information and subsequently transferred to the target image through our DWT-based structural guidance.

To address the blurriness issue, we apply the sharpening operation to the interpolated image $\tilde{x}_0^H$:

$$\bar{x}_0^H = (\alpha + 1)\tilde{x}_0^H - \alpha \mathcal{S}(\tilde{x}_0^H) \tag{5}$$

where $\mathcal{S}$ is an arbitrary smoothing operation and α is the sharpness factor that controls the magnitude of the sharpness. This behavior slightly moves the sample point closer to the sharp data distribution mode, resulting in a sharp and clear sample after the denoising process (Fig. 3 (b)), and also causes meaningful intensity changes at edges and boundaries of the interpolated image. Surprisingly, we found that simply sharpening the image significantly alleviates the aforementioned issues. We further provide extensive analysis on this phenomenon in the appendix. The overall pipeline of *DiffuseHigh* is illustrated in Fig. 2.

Experiments

In this section, we report the qualitative and quantitative results of our proposed *DiffuseHigh*. We also provide extensive ablation studies to validate the efficacy of our method thoroughly.

Implementation Details

We mainly conducted our experiments with SDXL (Podell et al. 2023), which is capable of generating 1K resolution images. We validate our method by generating images at different resolutions, 2048×2048, 2048×4096, and 4096×4096. We used 50 EDM scheduler (Karras et al. 2022) steps to generate images. We fixed our hyperparameters to $\tau = 15$ and $\delta = 5$. We utilized Gaussian blur and sharpness factor $\alpha = 1.0$ for our sharpening operation. Hyperparameters are set equally in every experiment.

Baselines

We compare our method against two groups of baselines; training-free methods and super-resolution (SR) methods. For training-free methods, we selected (1) **ScaleCrafter** (He et al. 2023), (2) **FouriScale** (Huang et al. 2024), (3) **HiDiffusion** (Zhang et al. 2023), (4) **DemoFusion** (Du et al. 2024), and (5) **AccDiffusion** (Lin et al. 2024). Each of these baselines are capable of generating higher-resolution images over the trained resolution with SDXL in a training-free manner. For SR methods, we compare our method to two popular SR models, namely (1) **SDXL+SD-Upscaler** (Rombach et al. 2022b) and (2) **SDXL+BSRGAN** (Zhang et al. 2021), since it is intuitive to first generate an image and then apply super-resolution models to obtain higher-resolution images. In the main text, we mainly compare our method against training-free methods. We leave comparison to super-resolution methods in the appendix.

Evaluation

We utilized the LAION-5B (Schuhmann et al. 2022) dataset as a benchmark for the image generation experiments. Following previous works (Du et al. 2024), we randomly sampled 1K captions and generated images corresponding to each caption. We selected Frechet Inception Distance (FID_r) (Heusel et al. 2017), Kernel Inception Distance (KID_r) (Bińkowski et al. 2018), and CLIP Score (Radford et al. 2021) as our evaluation metrics. Note that FID_r and KID_r require resizing the images to a resolution of 299^2, which is undesirable for assessing the high-frequency details of the image. To further provide the concrete evaluation, we also adopted patch FID (FID_p) (Chai et al. 2022) and patch KID (KID_p) as our evaluation metrics. In detail, we randomly cropped 1K patches from each generated image and measured the performance with randomly sampled 10K images from the LAION-5B dataset.

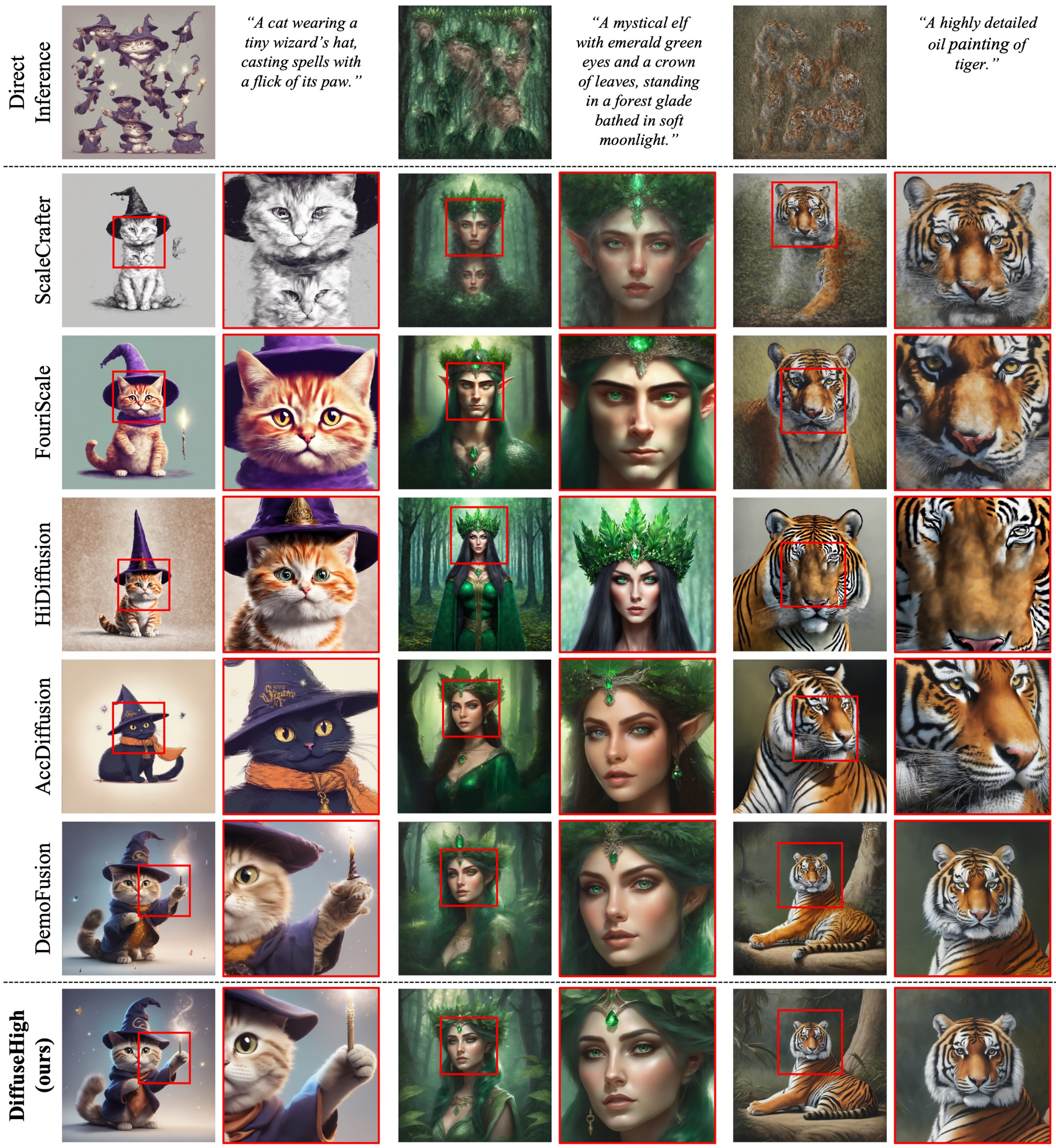

Figure 4: **Qualitative comparison to baselines in 4096 $\times$ 4096 resolution experiment.** Please **ZOOM-IN** the figure in order to see the details of each image.

	2048 × 2048						2048 × 4096						4096 × 4096					
Methods	FID_r (↓)	KID_r (↓)	FID_p (↓)	KID_p (↓)	CLIP (↑)	Time	FID_r (↓)	KID_r (↓)	FID_p (↓)	KID_p (↓)	CLIP (↑)	Time	FID_r	KID_r (↓)	FID_p (↓)	KID_p (↓)	CLIP (↑)	Time
Direct Inference	94.11	0.021	66.72	0.020	29.96	27	139.44	0.052	79.67	0.026	27.82	79	173.70	0.067	92.42	0.032	25.12	261
ScaleCrafter	79.95	0.015	59.32	0.016	29.56	31	132.43	0.044	114.94	0.055	25.42	145	112.81	0.031	108.63	0.047	27.07	465
FouriScale	67.05	0.010	51.13	0.013	30.85	55	117.52	0.036	114.93	0.055	26.49	154	97.40	0.023	104.62	0.041	28.37	460
HiDiffusion	75.59	0.011	54.77	0.013	29.17	16	102.79	0.026	81.68	0.027	26.68	43	127.55	0.044	153.98	0.081	24.44	87
AccDiffusion	57.64	<u>0.008</u>	38.71	0.013	<u>31.38</u>	78	<u>58.61</u>	0.025	68.84	<u>0.014</u>	28.02	271	64.12	0.012	57.56	0.022	30.90	638
DemoFusion	<u>57.16</u>	**0.007**	<u>35.54</u>	<u>0.010</u>	**31.95**	63	63.56	<u>0.010</u>	<u>49.08</u>	0.017	<u>29.13</u>	216	<u>62.82</u>	<u>0.010</u>	<u>48.97</u>	<u>0.017</u>	**31.59**	487
DiffuseHigh	**56.90**	<u>0.008</u>	**34.14**	**0.009**	31.26	32	**53.83**	**0.007**	**33.73**	**0.009**	**30.50**	58	**56.09**	**0.007**	**38.93**	**0.010**	<u>31.32</u>	147

Table 1: **Quantitative results of higher-resolution image generation experiments**. Hereinafter, we represent the best results with **bold** and second best with <u>underline</u>. We measured the inference time (sec) of each method by averaging the time generating 10 images in a single NVIDIA H100 gpu. We ran the official code of each baseline and obtained the results.

Results

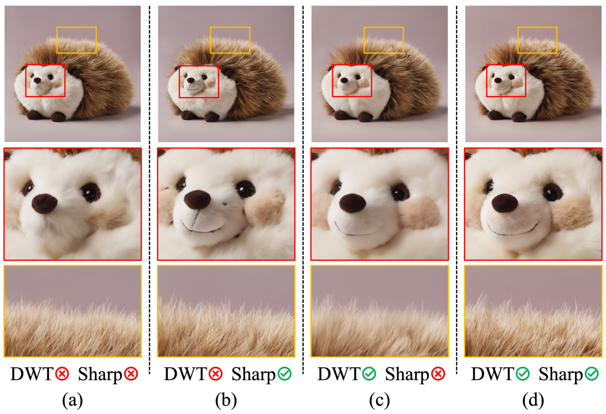

Figure 5: **Ablating each component of *DiffuseHigh*.** 'DWT' denotes the DWT-based structural guidance and 'Sharp' denotes the sharpening operation. Each sample has 4K resolution, generated from the same 1K image.

Qualitative Comparison We compare our method to baselines qualitatively in Fig. 4. In terms of training-free methods, while ScaleCrafter, FouriScale, and HiDiffusion partially alleviate the object repetition problem, they often fail to capture correct global semantics, particularly at higher resolution. We argue that since these methods alter the pre-trained weights or the architecture of the UNet, they might have the risk of ruining the powerful generation ability of the diffusion model at higher resolution. DemoFusion preserves the overall structure of the image well. However, their approach frequently introduces small repeated objects in the local area of the result image, and also require considerable inference time, due to their MultiDiffusion (Bar-Tal et al. 2023)-style generation scheme. Leveraging structural properties of the low-resolution images, *DiffuseHigh* exhibits correct global structures, while also showcasing favorable textures and high-frequency details.

Quantitative Comparison We report the quantitative evaluation results in Tab. 1. As observed, our method surpassed nearly every training-free baseline method in every resolution experiment, in terms of FID_r, KID_r, FID_p, and KID_p. These results demonstrate that our proposed *DiffuseHigh* not only synthesizes visually approving results but also favorable textures and patterns corresponding to the higher-resolution images. One notable observation is that our metric scores does not differ a lot along the different resolutions compared to others, which proves the efficacy of our proposed method to transfer correct structures. Also, our method showcased superior performance on CLIP score, which highlights the ability of our pipeline to generate semantically correct images given text prompts. DemoFusion showed a better CLIP score compared to ours in the 2K and 4K experiments, but the difference is negligible. Moreover, our method achieved superb inference time thanks to our partial denoising process, which starts denoising process from the intermediate diffusion timestep.

Ablation Studies

Structural Guidance and Sharpening We validate the role of each component involved in our pipeline. As illustrated in Fig. 5, our structural guidance enables the generated image to preserve essential structures. By forcing the denoising process to maintain the low-frequency details of the sample, which is obtained from well-structured low-resolution images, samples with our DWT-based structural guidance present desirable structures and shapes. However, samples without structural guidance tend to have deformed shapes (mouth of the hedgehog in Fig. 5 (a)) or artifacts (dots around the face in Fig. 5 (b)). Also, we observed that the sharpening operation involved in our pipeline further enhances the quality of the image, particularly on blurred object boundaries or smoothed textures of the image (Fig. 5 (c) and (d)). We also leave quantitative results in the appendix.

δ	$\text{FID}_r(\downarrow)$	$\text{KID}_r(\downarrow)$	$\text{FID}_p(\downarrow)$	$\text{KID}_p(\downarrow)$	CLIP ($\uparrow$)	Time
3	56.35	0.007	**38.67**	0.010	31.31	223 sec
5	**56.09**	0.007	38.93	0.010	31.32	258 sec
7	56.18	0.007	38.95	0.010	**31.33**	293 sec
10	56.49	0.007	39.31	0.011	31.32	346 sec
15	56.54	0.007	39.50	0.011	31.26	435 sec

Table 2: **Evaluation with varying** δ. We generated 10K images with randomly sampled captions from the LAION-5B dataset. We generated 4K images starting from the same 1K images generated by SDXL to ensure the fair comparison. The inference time is measured with NVIDIA A100 gpu.

DWT-based Structural Guidance Steps We conduct the experiment with varying δ to assess the validity of our proposed structural guidance. As shown in Tab. 2, FID_r decreases as δ approaches 5, and then increases as δ gets large. This observation suggests that our proposed structural guidance effectively facilitates the preservation of the desired structures, while an excessive guidance steps inhibit the generation of rich high-frequency details. Additionally, in terms of FID_p, $\delta = 3$ yielded the highest score and $\delta = 5$ the second highest score, but the difference is negligible. Nevertheless, we observed that small δ often fail to guide the correct global semantics. Therefore, we selected $\delta = 5$ as the optimal hyperparameter throughout this paper.

Conclusion

We present a training-free progressive high-resolution image synthesis pipeline using a pre-trained diffusion model. Our proposal involves leveraging generated low-resolution images as a guiding mechanism to effectively preserve the overall structure and intricate details of the contents. We also propose a novel principled way of incorporating structure information into the denoising process through frequency domain representation, which allows us to retain the essential information presented in low-resolution images. The extensive experiments with the pre-trained SDXL have shown that the proposed *DiffuseHigh* generates higher-resolution images without commonly reported issues in the existing approaches, such as repetitive patterns and irregular structures.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, Artificial Intelligence Graduate School Program(Sungkyunkwan University)) and the National Research Foundation (NRF) grant (RS-2024-00337548). This work was also supported by the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469), and Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT2401-01.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *International Conference on Machine Learning*, 1737–1752. PMLR.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chai, L.; Gharbi, M.; Shechtman, E.; Isola, P.; and Zhang, R. 2022. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision*, 170–188. Springer.

Chen, N.; Zhang, Y.; Zen, H.; Weiss, R. J.; Norouzi, M.; and Chan, W. 2020. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Du, R.; Chang, D.; Hospedales, T.; Song, Y.-Z.; and Ma, Z. 2024. Demofusion: Democratising high-resolution image generation with no costs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6159–6168.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10021–10030.

Guo, L.; He, Y.; Chen, H.; Xia, M.; Cun, X.; Wang, Y.; Huang, S.; Zhang, Y.; Wang, X.; Chen, Q.; et al. 2024. Make a Cheap Scaling: A Self-Cascade Diffusion Model for Higher-Resolution Adaptation. *arXiv preprint arXiv:2402.10491*.

He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Huang, L.; Fang, R.; Zhang, A.; Song, G.; Liu, S.; Liu, Y.; and Li, H. 2024. FouriScale: A Frequency Perspective on Training-Free High-Resolution Image Synthesis. *arXiv preprint arXiv:2403.12963*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 35: 26565–26577.

Kim, Y.; Hwang, G.; Zhang, J.; and Park, E. 2024. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. *arXiv preprint arXiv:2406.18459*.

Lam, M. W.; Wang, J.; Huang, R.; Su, D.; and Yu, D. 2021. Bilateral denoising diffusion models. *arXiv preprint arXiv:2108.11514*.

Lee, Y.; Kim, K.; Kim, H.; and Sung, M. 2024. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36.

Li, Y.; Wang, H.; Jin, Q.; Hu, J.; Chemerys, P.; Fu, Y.; Wang, Y.; Tulyakov, S.; and Ren, J. 2024. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *NeurIPS*, 36.

Lin, Z.; Lin, M.; Zhao, M.; and Ji, R. 2024. AccDiffusion: An Accurate Method for Higher-Resolution Image Generation. *arXiv preprint arXiv:2407.10738*.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35: 36479–36494.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022c. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shi, S.; Li, W.; Zhang, Y.; He, J.; Gong, B.; and Zheng, Y. 2024. ResMaster: Mastering High-Resolution Image Generation via Structural and Fine-Grained Guidance. *arXiv preprint arXiv:2406.16476*.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Xie, E.; Yao, L.; Shi, H.; Liu, Z.; Zhou, D.; Liu, Z.; Li, J.; and Li, Z. 2023. DiffFit: Unlocking Transferability of Large Diffusion Models via Simple Parameter-Efficient Fine-Tuning. *arXiv preprint arXiv:2304.06648*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *IEEE International Conference on Computer Vision*, 4791–4800.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, S.; Chen, Z.; Zhao, Z.; Chen, Z.; Tang, Y.; Chen, Y.; Cao, W.; and Liang, J. 2023. HiDiffusion: Unlocking High-Resolution Creativity and Efficiency in Low-Resolution Trained Diffusion Models. *arXiv preprint arXiv:2311.17528*.

Zheng, Q.; Guo, Y.; Deng, J.; Han, J.; Li, Y.; Xu, S.; and Xu, H. 2023. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. *arXiv preprint arXiv:2308.16582*.