ResMaster: Mastering High-Resolution Image Generation via Structural and Fine-Grained Guidance

Shuwei Shi[1], **Wenbo Li**[2], **Yuechen Zhang**[2], **Jingwen He**[2], **Biao Gong**[3], **Yinqiang Zheng**[1*]
[1]The University of Tokyo
[2]The Chinese University of Hong Kong
[3]Ant Group
{shishuwei666, fenglinglwb}@gmail.com, yqzheng@ai.u-tokyo.ac.jp

Abstract

Diffusion models excel at producing high-quality images; however, scaling to higher resolutions, such as 4K, often results in structural distortions, and repetitive patterns. To this end, we introduce ResMaster, a novel, training-free method that empowers resolution-limited diffusion models to generate high-quality images beyond resolution restrictions. Specifically, ResMaster leverages a low-resolution reference image created by a pre-trained diffusion model to provide structural and fine-grained guidance for crafting high-resolution images on a patch-by-patch basis. To ensure a coherent structure, ResMaster meticulously aligns the low-frequency components of high-resolution patches with the low-resolution reference at each denoising step. For fine-grained guidance, tailored image prompts based on the low-resolution reference and enriched textual prompts produced by a vision-language model are incorporated. This approach could significantly mitigate local pattern distortions and improve detail refinement. Extensive experiments validate that ResMaster sets a new benchmark for high-resolution image generation.

1 Introduction

Recently, diffusion models (Chefer et al. 2023; Epstein et al. 2023; Rombach et al. 2022; Chen, Laina, and Vedaldi 2024; Peebles and Xie 2023; Brooks, Holynski, and Efros 2023; Niu et al. 2025; Shi et al. 2024) have significantly advanced the fields of image generation, drawing considerable attention from the research community. Although representative models such as Stable Diffusion XL (Podell et al. 2023) (SDXL), DALL·E 3 (OpenAI 2023) and Midjourney (Midjourney 2023) can generate high-quality images, they perform well only within resolutions of 1024×1024. This limitation hinders their applications that require generated images with higher resolutions.

Several methods (Bar-Tal et al. 2023; He et al. 2023; Du et al. 2024; Zhang et al. 2023; Huang et al. 2024a) achieve higher-resolution image generation by adapting pre-trained diffusion models (*e.g.*SDXL) without additional re-training. The pioneering work, Multi-Diffusion (Bar-Tal et al. 2023), generates higher-resolution images by stitching generated overlapping patches. However, this approach cannot guarantee global consistency, since there is no explicit structural guidance during generation. Moreover, it often results in repetitive patterns that correspond to the provided prompt due to patch-wise generation. To alleviate the above issues, some methods (He et al. 2023; Zhang et al. 2023; Huang et al. 2024a) adopt whole image generation instead of the patch-wise counterpart. Specifically, they either employ dilated convolutions or reduce the image feature resolution at specific network locations to allow the receptive field of the pre-trained model to accommodate the higher-resolution image generation process. However, applying dilated convolutions hampers the generative capabilities inherited from pre-trained diffusion models. Meanwhile, downsampling and upsampling image features result in information loss. These operations lead to obvious structural distortions and repetitive patterns, as demonstrated in Figure 1. On the other hand, DemoFusion (Du et al. 2024) introduces dilated sampling to improve global consistency. Nonetheless, due to the lack of fine-grained guidance, repetitive patterns continue to plague this approach, as shown in Figure 1. In summary, existing methods struggle to achieve structure consistency and reasonable local detail generation.

Figure 1: **Comparisons of 4096×4096 image generation based on SDXL (Podell et al. 2023).** Our ResMaster can generate more faithful texture details and structured contents (*e.g.*, head) compared to state-of-the-art methods.

*Corresponding author.

Figure 2: **Selected multiple aspect-ratio images generated by ResMaster versus SDXL (Podell et al. 2023).** SDXL can synthesize high-quality 1024×1024 images. ResMaster can further generate 4096×4096 resolution results or more without retraining the text-to-image diffusion model, maintaining high quality. **Best viewed ZOOMED-IN.**

To this end, we propose **ResMaster**, a novel higher-resolution image generation method that employs *Structural and Fine-Grained Guidance* to ensure structural integrity and local detail generation. Specifically, ResMaster implements low-frequency component swapping using the low-resolution image generated at each sampling step to maintain structural coherence in higher-resolution outputs. Additionally, to mitigate repetitive patterns and enhance the generation of fine details, we employ localized fine-grained guidance using tailored image prompts and enriched textual prompts. The image prompts, derived from the generated low-resolution counterparts, contain critical semantic and structural information. Simultaneously, the enriched textual prompts produced by a pre-trained vision-language model (VLM) contribute to image generation on more complex and faithful patterns.

With these techniques, ResMaster can generate high-resolution images with better-structured compositions, as well as more faithful and richer local patterns compared to the state-of-the-art methods (see Figure 1). Moreover, it can generate high-quality and higher-resolution images at different aspect ratios, as demonstrated in Figure 2. Extensive qualitative and quantitative experiments have demonstrated that ResMaster achieves the state-of-the-art performance in higher-resolution image generation.

2 Related Work

Text-to-Image Diffusion Models. Text-to-image diffusion models (Dhariwal and Nichol 2021; Feng et al. 2024; Ho et al. 2022; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020a; Nichol and Dhariwal 2021; Ramesh et al. 2022; stability.ai 2022) represent a cutting-edge advancement in generative technology, leveraging the power of diffusion probabilistic models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020b) to synthesize high-quality images from textual descriptions. These models operate by gradual denoising a noisy input through iterative refinement steps. Furthermore, Latent Diffusion Models (LDMs) (Rombach et al. 2022) perform a similar process within a compact latent space, improving both the efficiency and scalability of the model, maintaining high fidelity in the generated images. This approach has been further refined in SDXL (Podell et al. 2023) and other models (Zheng et al. 2024b; Teng et al. 2023; Huang et al. 2024b; Lu et al. 2024; Zhu et al. 2024),

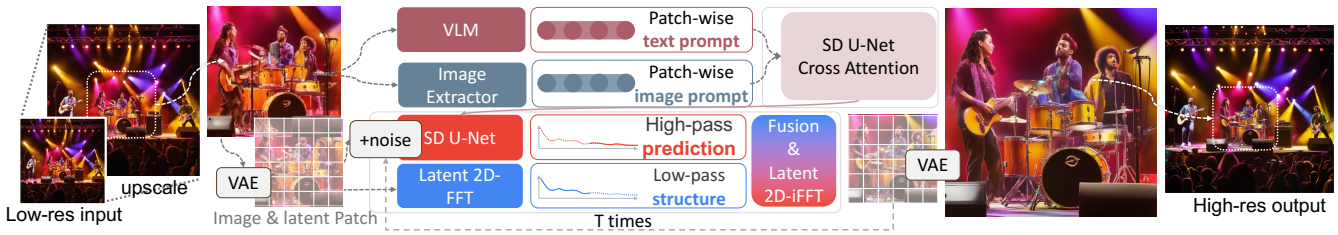

Figure 3: **The overall framework of ResMaster.** ResMaster is a patch-based denoising diffusion model that includes structural and fine-grained guidance. Fine-grained guidance utilizes an Image Extractor and a Large Vision-Language Model to extract region-aware image prompts and re-caption text prompts, respectively. These conditions are then used together via Cross Attention to guide the denoising process of the current patch. Furthermore, structural guidance ensures the structure of the generated image through low-frequency swapping with latent 2D-FFT.

which are applied in creative fields. However, despite these advancements, challenges remain, particularly in achieving higher resolution outputs (e.g., 4K) without compromising the generative quality.

High-Resolution Image Synthesis. Previous studies can be categorized into training-based methods (Zheng et al. 2024a; Teng et al. 2023; Chen et al. 2023a; Guo et al. 2024; Yang et al. 2024) and training-free methods (Bar-Tal et al. 2023; Jin et al. 2024; He et al. 2023; Lee et al. 2023; Haji-Ali, Balakrishnan, and Ordonez 2023; Lin et al. 2024a,b; Zhang et al. 2023). The training-based methods are trained on specific sets of images within the range of target resolution and aspect ratio. For example, Any-size-Diffusion (Zheng et al. 2024a) leverages a selected set of images with a restricted range of ratios to optimize the diffusion model. PixArt-Σ (Chen et al. 2024) uses a weak-to-strong training strategy to train their model on the higher-quality training data. However, these methods are still limited by the resolution of their training images and cannot generate high-quality images beyond a specific range. In contrast, training-free strategies aim to utilize pre-trained diffusion models without additional training phases. For instance, MultiDiffusion (Bar-Tal et al. 2023) addresses high-resolution synthesis by altering fusing multiple denoising paths. However, it suffers from local repetitions and distortion of object structures. To alleviate the aforementioned issues, some methods (He et al. 2023; Huang et al. 2024a) employ dilated convolutions to increase the receptive field of the convolution kernel. These methods can alleviate some of the issues with structural distortions and repetitive patterns. However, when applied to higher-resolution image generation tasks, issues such as structural distortion and a decline in the quality of local detail generation begin to emerge. DemoFusion (Du et al. 2024) improves the accuracy of target structures and reduces the appearance of repeated objects through skip residual and dilated sampling. It shows further improvement in generating high-quality images, but it is still affected by repetitive objects and chaotic local details. In this paper, we propose a method with Structural and Fine-Grained Guidance, which ensures structural accuracy while enhancing the details of high-resolution images.

3 Methodology

3.1 Preliminaries

Latent Diffusion Model. Our methods are built on the forefront text-to-image diffusion model, SDXL (Podell et al. 2023), which belongs to the series of LDMs. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, the encoder $\mathcal{E}$ in LDM first encodes it into latent representation $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{H/8\times W/8\times C}$. Then, forward diffusion and denoising are conducted in the latent space. In the forward process, the noise is gradually added to the latent z within T steps, represented as

$$q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}\right) = N\left(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t \boldsymbol{I}\right), \quad (1)$$

where β_t is the variance schedule, and $t \in \{1, ..., T\}$. On the other hand, in the backward process, a Unet ϵ_θ is used to predict the noise iteratively, eventually yielding results under the guidance of the text prompt y. The object of this stage can be formulated as:

$$L = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right], \quad (2)$$

where τ_θ is the text encoder of CLIP (Radford et al. 2021).

Patch-based Diffusion Model. Multi-Diffusion (Bar-Tal et al. 2023) initially generates higher-resolution images through a denoising process that utilizes overlapping patches. This approach is widely adopted in subsequent works (Lee et al. 2023; Du et al. 2024) due to its flexibility and convenience. Specifically, given a latent representation $z_t \in \mathbb{R}^{H/8\times W/8\times C}$, it is first partitioned into patches $z_{n,t} \in \mathbb{R}^{h\times w\times C}$ with a specified window size $[h, w]$ and stride $[d_h, d_w]$, resulting in a total of $N = \left(\frac{(H/8-h)}{d_h}+1\right) \times \left(\frac{(W/8-w)}{d_w}+1\right)$ patches. Each patch is individually denoised, with overlapping areas averaged at each step.

3.2 Model Framework

Our ResMaster generates high-resolution images guided by their low-resolution counterparts. As shown in Figure 3, we first use SDXL to create a low-resolution image $\mathbf{I}^L$ based on the prompt p. This image is then upsampled to the target resolution using bicubic interpolation, resulting in $\mathbf{B}^L$, which is divided into N equal-sized overlapping patches. We then

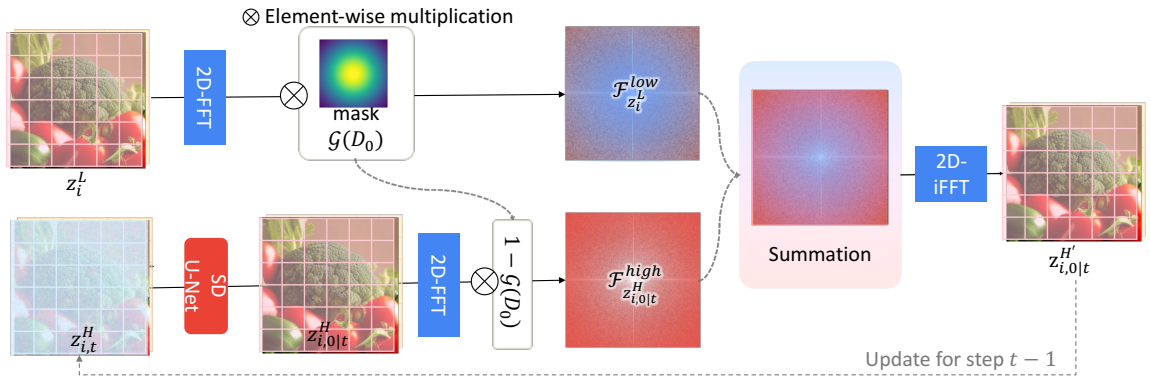

Figure 4: **The overall pipeline of Structural Guidance.** We use 2D Fast Fourier Transform (2D-FFT) to convert the image patch latent to frequency domain and apply a Gaussian low-pass filter to extract low-frequency information for exchange. This low-frequency information is then fused with the generated high-frequency information and converted back to spatial domain.

apply structural guidance and fine-grained guidance. a) **Structural Guidance**: To improve structural coherence, we use the VAE encoder to transform the i-th low-resolution patch of $\mathbf{B}^L$ into $\mathbf{z}_i^L$, serving as a guidance map. The i-th high-resolution noise patch at time t, dubbed $\mathbf{z}_{i,t}^H$, is mapped to a preliminary estimate $\mathbf{z}_{i,0|t}^H$. We align the low-frequency components of $\mathbf{z}_{i,0|t}^H$ with $\mathbf{z}_i^L$ for structural guidance. b) **Fine-grained Guidance**: Each low-resolution patch of $\mathbf{B}^L$ is processed by Image Condition Extractor and Large Vision Language Model to yield fine-grained image and textual representations. These representations are injected into the generative network via cross-attention to guide the noise prediction more accurately.

3.3 Structural Guidance

Due to the distribution disparity between the training data and target high-resolution images, previous patch-based diffusion models often exhibit structural distortions and repetitive patterns, impairing visual quality. To enhance structural rationality, we use generated low-resolution images for structural guidance. A common approach, as noted in (Dhariwal and Nichol 2021), updates $\mathbf{z}_{0|t}^H$ to align with $\mathbf{z}^L$ via gradient decay. However, this method increases time and memory consumption and introduces blurriness from the upsampled low-resolution image into the generated result. To mitigate these issues, we propose low-frequency (*i.e.*, structure (Choi et al. 2021)) swapping, as illustrated in Figure 4, which is efficient and resource-friendly. We are the first to innovatively introduce this approach to high-resolution image generation to address structural distortion issues, and we apply this concept to image latents (Ren et al. 2024; Si et al. 2024). We perform this operation on each patch, and for simplicity, we omit the subscript i when describing the low-frequency swapping process. Specifically, at time step t, we predict $\mathbf{z}_{0|t}^H$ from $\mathbf{z}_t^H$ and replace its low-frequency components with low-resolution guidance image $\mathbf{z}^L$, resulting in $\mathbf{z}_{0|t}^{H'}$. This ensures proper structural guidance, formulated as follows:

$$\mathcal{F}_{\mathbf{z}^L}^{low} = \texttt{FFT\_2D}(\mathbf{z}^L) \odot \mathcal{G}(D_0)\,, \tag{3}$$

$$\mathcal{F}_{\mathbf{z}_{0|t}^H}^{high} = \texttt{FFT\_2D}(\mathbf{z}_{0|t}^H) \odot (1 - \mathcal{G}(D_0))\,, \tag{4}$$

$$\mathbf{z}_{0|t}^{H'} = \texttt{IFFT\_2D}(\mathcal{F}_{\mathbf{z}^L}^{low} + \mathcal{F}_{\mathbf{z}_{0|t}^H}^{high})\,, \tag{5}$$

where FFT_2D is the 2D Fast Fourier Transform, and IFFT_2D is its inverse. $\mathcal{G}$ represents the Gaussian low-pass filter, and D_0 is the normalized cutoff frequency. The term $\mathbf{z}_{0|t}^H$ is calculated by:

$$\mathbf{z}_{0|t}^H \approx \left(\mathbf{z}_t^H - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t^H\right)\right)/\sqrt{\bar{\alpha}_t}\,, \tag{6}$$

where $\boldsymbol{\epsilon}_\theta$ is the denoising Unet, α_t is the prescribed variance schedule and $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. Then, the final result z_{t-1}^H is derived as:

$$q\left(\mathbf{z}_{t-1}^H \mid \mathbf{z}_t^H, \mathbf{z}_{0|t}^{H'}\right) = \mathcal{N}\left(\mathbf{z}_{t-1}^H; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{z}_t^H, \mathbf{z}_{0|t}^{H'}\right), \tilde{\beta}_t\mathbf{I}\right)\,. \tag{7}$$

To further enhance the model's generative capability, we introduce a scaled cosine decay coefficient c_t that reduces the value of D_0 as the denoising timestep t progresses. This process can be formulated as $D_0^t = D_0 * c_t$, where $c_t = ((1+\cos\left(\frac{T-t}{T}\times\pi\right))/2)$. This ensures that the structure is well-formed in the early stages, while reducing structural guidance in the later stages, allowing the model to focus more on generating local details. However, structural guidance alone may not suffice to generate locally faithful and rich details because of the low-resolution guidance image, necessitating fine-grained guidance.

3.4 Fine-Grained Guidance

Tailored image prompts. To further enhance local structure consistency and mitigate repeated patterns resulting from the identical text prompt across different patches (Bar-Tal et al. 2023; Du et al. 2024), we use Image Condition Extractor which consists of a CLIP image encoder and a linear projection network to customize multiple image prompts, inspired by previous successful methodologies (Ye et al. 2023). These

methods have been proven to generate images consistent with specific attributes based on the information from the input images. First, a CLIP image encoder extracts the class token from the low-resolution image, mapping it to representative image features. After that, we use a small linear projection network to project the image embedding into a sequence of features. Decoupled cross-attention mechanisms are then employed to integrate these image and text features into the pre-trained text-to-image diffusion model, formulated as:

$$\mathbf{X}' = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{\mathrm{T}}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\mathrm{T}} + \lambda\,\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{\mathrm{I}}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\mathrm{I}}\,, \tag{8}$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}_q$, $\mathbf{K}_{\mathrm{T}} = \boldsymbol{c}_{\mathrm{T}}\mathbf{W}_k$, $\mathbf{V}_{\mathrm{T}} = \boldsymbol{c}_{\mathrm{T}}\mathbf{W}_v$, $\mathbf{K}_{\mathrm{I}} = \boldsymbol{c}_{\mathrm{I}}\mathbf{W}'_k$, and $\mathbf{V}_{\mathrm{I}} = \boldsymbol{c}_{\mathrm{I}}\mathbf{W}'_v$ represent the query, key, and values of text and image features respectively, and λ is the weighting factor.

This design, with patch-wise image prompts, significantly alleviates the issue of repetitive patterns, resulting in improved local structures. However, due to the lack of details in low-resolution images, the extracted image prompts only provide structural guidance and cannot fully activate the network's generative capabilities. Thus, more informative text prompts are needed. Existing methods using global text face two issues. First, the patch-based method does not align well with global text. Second, global text often lacks detailed local descriptions. Therefore, incorporating more detailed descriptions is essential, as evidenced by existing text-to-image methods (Chen et al. 2024; Betker et al. 2023).

Enriched textual prompts. To obtain more detailed descriptions for image patches, we introduce a pre-trained large Vision-Language Model (VLM), Share-Captioner (Chen et al. 2023b), to re-caption low-resolution image patches. This model has proven effective in better aligning textual and visual information and reducing hallucinations (Chen et al. 2024). Our experiments indicate that providing the full image's description as context to the VLM does not yield additional benefits and may lead to hallucinations. Therefore, the instruction we give to the VLM is "*Describe the following image patch in detail.*" Following this instruction, the VLM generates a detailed prompt for each patch. These enriched prompts enable our method to produce richer local details, resulting in higher visual quality.

4 Experiments

In this section, we report the qualitative and quantitative results and ablation studies. We validate the performance of ResMaster based on the SDXL (Podell et al. 2023).

4.1 Implementation Details

This subsection introduces some hyperparameter settings of our ResMaster. In the structural guidance module, we set the initial normalized cutoff frequency D_0 to 1.0 to maintain control over object structures during the early stages of generation. Within the fine-grained guidance module, we utilize IP-Adapter (Ye et al. 2023) to inject image prompts into SDXL (Podell et al. 2023) as the condition. IP-Adapter (Ye et al. 2023) is capable of generating images corresponding to the content of the image prompts. Herein, we set the weight of the image prompt λ to 0.8. Our framework is built on the patch-based diffusion model. We follow (Du et al. 2024) to partition the entire noise into patches with a specified size [1024, 1024] and stride [64,64].

4.2 Comparison

We compare our method with the following representative generative approaches: (i) SDXL (Podell et al. 2023) Direct Inference, which uses pre-trained SDXL to directly infer the target resolution images. (ii) SDXL+BSRGAN. We use the classic super-resolution method BSRGAN (Zhang et al. 2021) to conduct image super-resolution. This is a traditional approach to increase image resolution which is proven to lack the local details in (Lin et al. 2024b; Du et al. 2024). (iii) SCALECRAFTER (He et al. 2023), a method that generates high-resolution images directly using dilated convolutions or large convolutional kernels. (iv) DemoFusion (Du et al. 2024), a high-resolution image generation method based on MultiDiffusion (Bar-Tal et al. 2023), using dilated sampling to ensure global structural consistency. (v) FouriScale (Huang et al. 2024a), a method employing dilated convolutions coupled with a low-pass operation and a padding-then-crop strategy. (vi) HiDiffusion (Zhang et al. 2023), an efficient high-resolution image generation method utilizing a Resolution-Aware operation to align the feature map size with the deep block of U-Net.

4.3 Qualitative Results

Figure 5 shows a visual comparison of different models, each producing 4096×4096 resolution results. We select complex real-world scenes and examples prone to local pattern repetition and structural distortion to demonstrate the superiority of our method. Firstly, in the case of the first complex scene, the image generated by SDXL has a lower resolution and lacks fine-grained generation guidance for local content. As a result, the faces of people and the complex background structures are not clear and complete. BSRGAN performs super-resolution based on SDXL, partially eliminating the blurriness in upscaled results of SDXL. However, it is evident that BSRGAN merely sharpens the low-resolution results, making it unsuitable for high-resolution image generation tasks. High-resolution image generation requires more local details and, in some scenarios, the ability to correct and complete the inherent issues in low-resolution images. From the faces and objects in the results, it can be seen that ScaleCrafter, HiDiffusion, Fouriscale, and DemoFusion all introduce varying degrees of structural distortion and local pattern repetition. This is mainly due to the fact that ScaleCrafter and Fouriscale both use dilated convolutions, which weaken their ability to maintain structure and handle high-frequency details in high-resolution image generation. The use of upsampling and downsampling operations at certain positions in HiDiffusion may cause information loss in image features, impacting the final denoising output. DemoFusion lacks fine-grained guidance for local content generation, leading to repetitive patterns and chaotic structures in the objects. In contrast, ResMaster effectively restores facial features and improves the structure of complex objects, making them clearer,

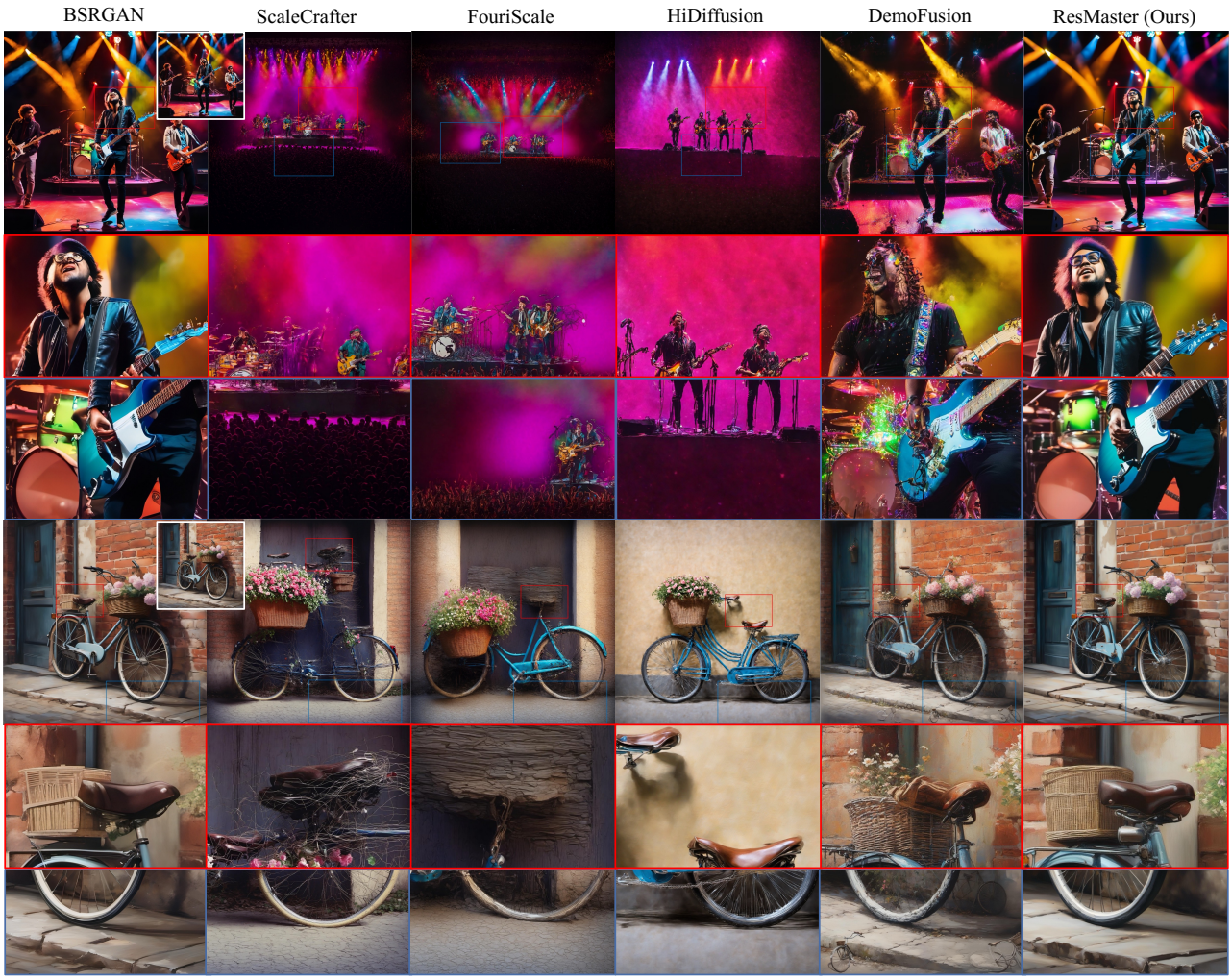

Figure 5: **Qualitative comparisons with other methods.** All results are presented at a resolution of 4096×4096. Some areas have been zoomed in.

Figure 6: **The Ablation study of three components used in ResMaster**: Structural Guidance (SG), Tailored Image Prompts (TIP) and Enriched Textual Prompts (ETP). All results are presented at a resolution of 4096×4096.

more complete, and aesthetically pleasing. Similarly, the results in the bicycle case confirm our observations. BSRGAN still fails to provide more details. ScaleCrafter, HiDiffusion, and Fouriscale have weak structure preservation and chaotic details. DemoFusion produces repeated bicycles, alters the seat structure, and exhibits pattern confusion between different objects (flowers in the rear basket). In conclusion, our proposed ResMaster can improve more faithful details while ensuring structural accuracy, owing to our proposed structural and fine-grained guidance. This capability is particularly crucial in high-resolution image generation, avoiding the occurrence of repetitive patterns and structural distortions.

4.4 Quantitative Results

For the fair evaluation of model performance, we conduct quantitative experiments on the dataset of Laion-5B (Schuhmann et al. 2022) with a large amount of image-caption pairs. We randomly sample $1K$ captions as the text prompts for the high-resolution image generation. Additionally, we randomly sample $10K$ images from Laion-5B as a real image set. We adopt 3 metrics following prior works (He et al. 2023; Du et al. 2024): Frechet Inception Distance(FID) (Heusel et al. 2017), Inception Score(IS) (Salimans et al. 2016) and CLIP Score (Radford et al. 2021) to evaluate both image quality and semantic similarity between image features and text prompts. Among them, FID_r and IS_r require resizing the test

Resolution	Method	$FID_r \downarrow$	$IS_r \uparrow$	$FID_c \downarrow$	$IS_c \uparrow$	CLIP ↑	Time
2048×2048	SDXL Direct Inference (Podell et al. 2023)	90.33	13.13	63.47	21.74	29.18	1min
	SDXL + BSRGAN (Zhang et al. 2021)	**67.00**	17.10	<u>42.79</u>	22.36	<u>31.64</u>	1min
	SCALECRAFTER (He et al. 2023)	78.95	16.23	58.86	21.71	30.23	1min
	FouriScale (Huang et al. 2024a)	72.49	17.26	49.82	23.75	29.53	1min
	HiDiffusion (Zhang et al. 2023)	73.24	16.06	53.02	24.36	28.99	1min
	DemoFusion (Du et al. 2024)	69.81	<u>17.95</u>	45.42	<u>24.53</u>	31.45	2min
	ResMaster (Ours)	<u>68.16</u>	**18.01**	**40.22**	**24.62**	**31.79**	1min
4096×4096	SDXL Direct Inference (Podell et al. 2023)	173.42	7.50	89.46	16.42	24.54	5min
	SDXL + BSRGAN (Zhang et al. 2021)	<u>67.08</u>	17.14	**50.89**	15.30	30.61	1min
	SCALECRAFTER (He et al. 2023)	107.46	11.06	107.88	10.94	29.77	9min
	FouriScale (Huang et al. 2024a)	93.30	13.58	87.99	12.10	26.29	8min
	HiDiffusion (Zhang et al. 2023)	103.79	12.36	91.64	11.49	28.06	2min
	DemoFusion (Du et al. 2024)	76.03	<u>17.85</u>	56.75	<u>16.48</u>	<u>30.83</u>	11min
	ResMaster (Ours)	**65.43**	**18.44**	<u>55.09</u>	**16.51**	**30.95**	6min

Table 1: **Quantitative comparison results**. The best results are marked in **bold**, and the second best results are marked by <u>underline</u>.

images to 299^2, which may influence the evaluation results for high-resolution images. For more reasonable evaluation, we follow (Zheng et al. 2024a) to crop and resize some local patches at $1K$ resolution to compute FID_c and IS_c. We report quantitative results at two different resolutions. The inference time is performed on a single NVIDIA Tesla 40G-A100 GPU. As shown in Table 1, ResMaster has achieved state-of-the-art performance across multiple metrics. Due to our fine-grained guidance strategy, ResMaster achieves better detail accuracy and semantic alignment, as reflected in the improved CLIP score. In comparisons of FID metrics at various resolutions, ResMaster consistently ranks in the top two. SDXL+BSRGAN, which strictly adheres to low-resolution inputs, has been shown to suffer from blurriness in high-resolution image generation and lacks the ability to produce rich details (Du et al. 2024). Meanwhile, ResMaster ensures higher generation quality while achieving faster inference speeds compared to other representative methods. Generating a $4K$ image is 5 minutes faster than DemoFusion.

4.5 Ablation Study

Quanlitative Results of Ablation Study. ResMaster primarily consists of two components: structural guidance and localized fine-grained guidance. The localized fine-grained guidance includes Tailored Image Prompts and Enriched Textual Prompts. To visually present the contribution of each module, we progressively display the effects brought by the introduction of each module, as shown in Figure 6. The images we present are generated at a resolution of 4096×4096. ResMaster is built on patch-based multi-diffusion. Therefore, when all modules are removed, the model degrades to Multi-Diffusion (Bar-Tal et al. 2023). We present the results of the base model, Multi-Diffusion. Without the intervention of guidance strategies, Multi-Diffusion exhibits structural distortions and repeated patterns. With the introduction of the proposed guidance strategies, structural distortions and repeated patterns disappear. Specifically, when we introduce the structural guidance strategy, the issue of structural distortions is resolved. However, due to the low resolution of the guidance images, local pattern details exhibit misalignment. We further introduce Tailored Image Prompts which can significantly alleviate the issue, making the structures and details clearer. Nonetheless, Tailored Image Prompts do not introduce additional information to supplement local details. Further incorporating Enriched Textual Prompts enhances the richness of details in high-resolution images. In conclusion, when combined, they leverage their respective strengths and functionalities, resulting in impressive generative outcomes.

Quantitative Results of Ablation Study. The quantitative results of the ablation study are shown in Table 2. All the results are performed in the resolution of 4096×4096. Table 2 shows that compared to the baseline model Multi-Diffusion, the use of Structural Guidance(SG) significantly improves quantitative results. Furthermore, incorporating Tailored Image Prompts (TIP) and Enriched Textual Prompts (ETP) further enhances the model's performance. This demonstrates the effectiveness of each module we proposed.

Method	$FID_r \downarrow$	$IS_r \uparrow$	$FID_c \downarrow$	$IS_c \uparrow$	CLIP ↑
Base (Multi-Diffusion)	92.80	9.33	68.8	13.63	27.19
Base+SG	79.43	14.59	64.24	15.79	28.39
Base+SG+TIP	66.36	18.14	56.86	16.33	30.28
Base+SG+TIP+ETP (ResMaster)	**65.43**	**18.44**	**55.09**	**16.51**	**30.95**

Table 2: **Quantitative comparison results of the ablation study.** The best results are marked in **bold**. Base refers to the model that does not include the method proposed in this paper, namely Multi-Diffusion.

5 Conclusion

In this paper, we introduce ResMaster, a training-free, patch-based diffusion model for high-resolution image generation. ResMaster employs structural and fine-grained guidance strategies, ensuring the overall structural integrity of high-resolution images while also achieving reasonable and rich local details. ResMaster can generate images of any scale and aspect ratio. Extensive experiments have demonstrated the superior capabilities of ResMaster.

Acknowledgments

This research was supported in part by JSPS KAKENHI Grant Numbers 24K22318, 22H00529, 20H05951, JST-Mirai Program JPMJMI23G1.

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

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.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024. PixArt-σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. *arXiv preprint arXiv:2403.04692*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023a. PIXART-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5343–5353.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14347–14356.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Du, R.; Chang, D.; Hospedales, T.; Song, Y.-Z.; and Ma, Z. 2024. DemoFusion: Democratising High-Resolution Image Generation With No $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36: 16222–16239.

Feng, Y.; Gong, B.; Chen, D.; Shen, Y.; Liu, Y.; and Zhou, J. 2024. Ranni: Taming Text-to-Image Diffusion for Accurate Instruction Following. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

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*.

Haji-Ali, M.; Balakrishnan, G.; and Ordonez, V. 2023. ElasticDiffusion: Training-free Arbitrary Size Image Generation. arXiv:2311.18822.

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*.

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.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47): 1–33.

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

Huang, S.; Gong, B.; Feng, Y.; Chen, X.; Fu, Y.; Liu, Y.; and Wang, D. 2024b. Learning Disentangled Identifiers for Action-Customized Text-to-Image Generation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2024. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36.

Lee, Y.; Kim, K.; Kim, H.; and Sung, M. 2023. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36: 50648–50660.

Lin, M.; Lin, Z.; Zhan, W.; Cao, L.; and Ji, R. 2024a. CutDiffusion: A Simple, Fast, Cheap, and Strong Diffusion Extrapolation Method. *arXiv preprint arXiv:2404.15141*.

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

Lu, Z.; Wang, Z.; Huang, D.; Wu, C.; Liu, X.; Ouyang, W.; and Bai, L. 2024. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*.

Midjourney. 2023. Midjourney.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. 8162–8171. PMLR.

Niu, M.; Cun, X.; Wang, X.; Zhang, Y.; Shan, Y.; and Zheng, Y. 2025. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, 111–128. Springer.

OpenAI. 2023. Dalle-3.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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*.

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*, abs/2204.06125.

Ren, W.; Yang, H.; Zhang, G.; Wei, C.; Du, X.; Huang, S.; and Chen, W. 2024. ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. *arXiv preprint arXiv:2402.04324*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

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.; Gong, B.; Chen, X.; Zheng, D.; Tan, S.; Yang, Z.; Li, Y.; He, J.; Zheng, K.; Chen, J.; et al. 2024. MotionStone: Decoupled Motion Intensity Modulation with Diffusion Transformer for Image-to-Video Generation. *arXiv preprint arXiv:2412.05848*.

Si, C.; Huang, Z.; Jiang, Y.; and Liu, Z. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4733–4743.

Song, J.; Meng, C.; and Ermon, S. 2020a. Denoising Diffusion Implicit Models. *ArXiv*, abs/2010.02502.

Song, J.; Meng, C.; and Ermon, S. 2020b. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

stability.ai. 2022. Stable Diffusion 2.0 Release.

Teng, J.; Zheng, W.; Ding, M.; Hong, W.; Wangni, J.; Yang, Z.; and Tang, J. 2023. Relay Diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*.

Yang, Z.; Jiang, H.; Hong, W.; Teng, J.; Zheng, W.; Dong, Y.; Ding, M.; and Tang, J. 2024. Inf-DiT: Upsampling Any-Resolution Image with Memory-Efficient Diffusion Transformer. *arXiv preprint arXiv:2405.04312*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.

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. 2024a. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7571–7578.

Zheng, W.; Teng, J.; Yang, Z.; Wang, W.; Chen, J.; Gu, X.; Dong, Y.; Ding, M.; and Tang, J. 2024b. CogView3: Finer and Faster Text-to-Image Generation via Relay Diffusion. *arXiv preprint arXiv:2403.05121*.

Zhu, R.; Pan, Y.; Li, Y.; Yao, T.; Sun, Z.; Mei, T.; and Chen, C. W. 2024. SD-DiT: Unleashing the Power of Self-supervised Discrimination in Diffusion Transformer. *arXiv preprint arXiv:2403.17004*.