

# ResDiT: Evoking the Intrinsic Resolution Scalability in Diffusion Transformers

Yiyang Ma<sup>1,\*</sup> Feng Zhou<sup>1,\*</sup> Xuedan Yin<sup>2</sup> Pu Cao<sup>1</sup> Yonghao Dang<sup>1</sup> Jianqin Yin<sup>1,†</sup>  
<sup>1</sup>Beijing University of Posts and Telecommunications <sup>2</sup>Tsinghua University  
{yym2024, zhoufeng, caopu, dyh2018, jqyin}@bupt.edu.cn, yxd23@mails.tsinghua.edu.cn



Figure 1. **Qualitative examples of the proposed ResDiT**, which enables the pre-trained T2I models to generate high-resolution images than the originally trained resolution, without any training or fine-tuning. **Best view ZOOM-IN.**

## Abstract

Leveraging pre-trained Diffusion Transformers (DiTs) for high-resolution (HR) image synthesis often leads to spatial layout collapse and degraded texture fidelity. Prior work mitigates these issues with complex pipelines that first perform a base-resolution (i.e., training-resolution) denoising process to guide HR generation. We instead explore the intrinsic generative mechanisms of DiTs and propose ResDiT, a training-free method that scales resolution efficiently. We identify the core factor governing spatial layout, position embeddings (PEs), and show that the original

PEs encode incorrect positional information when extrapolated to HR, which triggers layout collapse. To address this, we introduce a PE scaling technique that rectifies positional encoding under resolution changes. To further remedy low-fidelity details, we develop a local-enhancement mechanism grounded in base-resolution local attention. We design a patch-level fusion module that aggregates global and local cues, together with a Gaussian-weighted splicing strategy that eliminates grid artifacts. Comprehensive evaluations demonstrate that ResDiT consistently delivers high-fidelity, high-resolution image synthesis and integrates seamlessly with downstream tasks, including spatially controlled generation.

\*Equal contribution.

†Corresponding author.

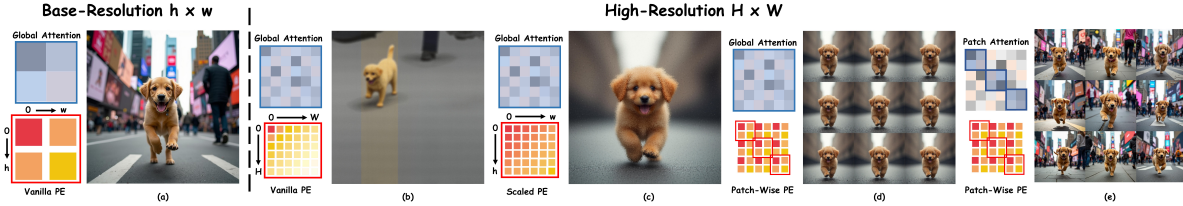


Figure 2. **Disentangling PE and attention range in high-resolution DiT synthesis.** Systematic interventions on positional embeddings (PEs) and attention range across resolutions. (a) At base resolution, a DiT with global attention and vanilla PE produces coherent layouts and fine details. (b) When directly applied to high resolution, layout collapse occurs as the subject becomes shrunken and misplaced due to a mismatch between PE and the attention field. (c) Using a scaled PE restores spatial arrangement but yields blurred details. (d) Applying patch-wise base-resolution PEs ensures correct local structure within each patch, yet details remain degraded. (e) Introducing patch-level local attention further enhances fine details. These results show that positional embeddings determine spatial arrangement, while the attention receptive-field scale governs detail fidelity in DiTs.

## 1. Introduction

Text-to-image generation models have garnered widespread attention due to their impressive capabilities and wide range of applications, such as customized content creation and individualized media synthesis [6, 29]. Recently, a new generation of state-of-the-art models, such as the Flux [19] and SD3 [10], has established a powerful paradigm for high-fidelity image synthesis by introducing transformer architectures [27] that capture long-range dependencies and enable scalable global modeling to support high-quality image generation. However, these models often struggle to generate images beyond their training resolution, with high-resolution inference leading to noticeable degradation or even complete failure [4]. This limitation constrains their applicability in tasks that require high-resolution outputs.

A straightforward solution is to train or fine-tune models directly at high resolutions [12, 15, 22, 31, 39]. This approach demands high-quality, high-resolution datasets and incurs significant computational costs. As a result, various training-free methods have been proposed as more efficient alternatives [1, 5, 8, 13, 16–18, 20, 38, 40, 43, 44]. Some methods are tailored to specific model architectures (U-Net) and do not readily generalize to DiT-based models [13, 16, 44]. Other methods adopt a two-stage paradigm, where a base-resolution image is first generated and then used to guide high-resolution synthesis [4, 8, 9, 18]. While effective, it heavily depends on the base-resolution denoising image and introduces unnecessary complexity. More importantly, they fundamentally treat high-resolution generation as a super-resolution task, relying on external guidance rather than unlocking the model’s intrinsic capability to generate high-resolution content.

In this work, we investigate how to address two major challenges in high-resolution synthesis, layout collapse and degraded details, from an intrinsic, mechanics-based perspective. Since attention serves as the key spatial mechanism enabling token interactions in DiTs, we begin by an-

alyzing two crucial spatial factors within attention: positional embeddings (PEs) [36] and the attention range field. As illustrated in Fig. 2, we systematically intervene on these components under different resolutions. In the base-resolution setting (Fig. 2(a)), a DiT equipped with global attention and vanilla PE at resolution  $h \times w$  generates images with both coherent global layouts and high-quality details. When we directly apply the same global attention and vanilla PE to a high-resolution latent  $H \times W$  at test time (Fig. 2(b)), the model suffers from layout collapse: the main subject becomes shrunken and misplaced, revealing a clear mismatch between the extrapolated PE and the expanded attention field. To decouple these factors, we keep the global attention range at high resolution but replace the HR latent’s PE with its base-resolution counterpart (Fig. 2(c)). This “scaled PE” [7] operation largely restores correct spatial arrangement, yet the generated details remain blurry and low-fidelity. We then introduce a patch-wise base-resolution PE that is tiled across the HR canvas (Fig. 2(d)). Experiments show that PE primarily governs the spatial arrangement of objects, but a mismatch between the expanded attention field and the one used during training still leads to detail loss. This observation is further confirmed in Fig. 2(e): once we adopt patch-level local attention [24] (effectively generating multiple regions independently), the image details improve substantially. These findings lead to an important mechanistic insight about DiTs: positional embeddings dictate spatial arrangement, while the attention receptive-field scale critically determines the model’s ability to generate high-quality details.

Building on these observations, we introduce ResDiT, a training-free framework for high-resolution image synthesis. To improve spatial arrangement at high resolutions, we employ a scaled PE [7] mechanism that maps HR positional embeddings back to the training resolution, ensuring that spatial relationships are generated correctly. Meanwhile, to enhance local detail quality, we propose an overlapping patch partitioning and splicing strategy. This design

enables local attention [24] computation alongside the original global attention, and uses Gaussian-weighted splicing in overlapping regions to suppress grid artifacts during synthesis. To further integrate the strengths of the two attention receptive fields, we introduce a patch-wise spectral fusion technique that preserves high-frequency components from patch attention—which carry fine details—while retaining the low-frequency components from global attention, which encode coherent spatial structure.

In summary, our contributions are as follows:

- We provide a mechanistic analysis of DiTs under high-resolution inference. Through controlled interventions on positional embeddings and attention range, we reveal that PE determines spatial arrangement, while attention receptive-field mismatch causes detail degradation. This explains the root causes of layout collapse and low-fidelity details in HR synthesis.
- We propose ResDiT, a fully training-free method for high-resolution synthesis in DiTs. ResDiT corrects spatial arrangement via scaled PE and restores details through overlapping patch attention with Gaussian splicing. A patch-wise spectral fusion module further combines global structure with local detail fidelity.
- We perform extensive experiments and ablation studies on high-resolution image synthesis, showcasing the capability of our approach. We also integrate ResDiT with an off-the-shelf spatial-control method, showing the compatibility and practical utility.

## 2. Related Work

### 2.1. Text-to-Image Synthesis

Diffusion models [14] have become the dominant paradigm for text-to-image (T2I) synthesis, largely replacing GAN-based [11] approaches due to their stability and ability to generate high-fidelity images from complex prompts. Early T2I diffusion systems primarily relied on U-Net backbones [33], and latent diffusion [32] models such as Stable Diffusion [28] further improved efficiency by performing denoising in a learned latent space, enabling high-resolution generation at manageable computational cost. More recent models move beyond the U-Net architecture and adopt Diffusion Transformers (DiTs) [27], which treat images as sequences of latent tokens and scale more effectively. SD3 [10] exemplifies this shift by fully transitioning to a DiT-based design, while FLUX [19] and its variants extend this direction with large multimodal Transformer [10] blocks and flow-based [21, 23] training objectives. These advances yield stronger prompt fidelity, richer fine-grained detail, and more flexible resolution control. Our work targets this family of Transformer-based T2I models and enables training-free resolution scaling for SD3 and FLUX.

### 2.2. Training-free High-Resolution Image Synthesis

Despite considerable advances in image synthesis, generating high-resolution images (e.g., 3K and above) remains challenging. Training-free methods [1, 5, 8, 13, 16–18, 20, 37, 38, 40, 43, 44] have emerged as a promising direction, as they leverage powerful pre-trained diffusion models and avoid the substantial data and computational costs of re-training at higher resolutions. In the U-Net–based diffusion regime, several training-free approaches address high-resolution generation within a single denoising process: ScaleCrafter [13] enlarges the receptive field via dilated convolutions, while PBC [44] introduces virtual zero-padded boundaries to provide correct spatial context, yielding more coherent global structures. However, these techniques are closely tied to convolutional U-Net architectures and do not directly transfer to DiT-based models. With the shift toward Diffusion Transformers, training-free high-resolution methods have largely adopted two-stage pipelines: they first perform a base-resolution generation and then use this trajectory to guide sampling at higher resolutions. For example, I-Max [9] models base-resolution flows as projections of high-resolution flows and exploits the linear interpolation property of rectified flows to construct dynamic guidance that steers the high-resolution trajectory. HiFlow [4], similarly, derives a virtual reference flow from base-resolution information to regularize high-resolution sampling and fuse coarse structure with fine details. While effective, these methods inherently rely on base-resolution guidance and introduce additional complexity into the denoising process.

## 3. ResDiT

Here, we propose ResDiT, a training-free method that scales resolutions of pre-trained Diffusion Transformers (DiTs) by exploiting their intrinsic generative properties. As illustrated in Fig. 2, ResDiT restructures the vanilla attention operation into two targeted branches. The first branch performs global attention with scaled position embedding to rectify the overall layout, while the second applies patch-level attention to recover fine-grained local details Sec. 3.2. By employing the proposed Minimum-Overlap Partitioning and Gaussian Weighting Splicing strategies Sec. 3.3, feature patches can be divided and seamlessly reunited without introducing grid artifacts. Furthermore, ResDiT incorporates a Patch-Wise Spectral Fusion module to merge the two branches in the frequency domain, enabling a clean separation between layout-level and detail-level components and their effective integration Sec. 3.4.

### 3.1. Preliminaries

**Positional Embedding (PE).** Positional embedding provides spatial priors for Transformer architectures by en-

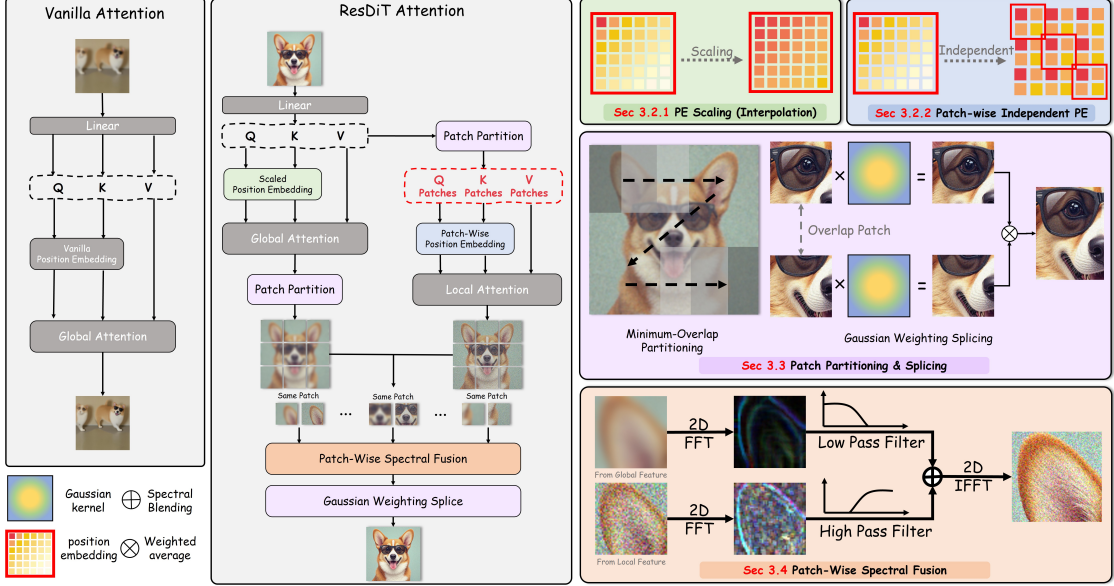


Figure 3. **Overview of ResDiT.** ResDiT restructures the vanilla attention mechanism in Diffusion Transformers (DiTs) into two complementary branches to enable training-free resolution scaling. Specifically, the global branch performs global attention with scaled positional embeddings to preserve the overall spatial layout, while the local branch applies patch-level attention to enhance fine-grained details. To maintain continuity across patches, we propose a Minimum-Overlap Partitioning strategy that ensures contextual consistency at patch boundaries and a Gaussian Weighting Splicing scheme that smoothly fuses overlapping regions without introducing grid artifacts. Finally, a Patch-Wise Spectral Fusion module combines the outputs of both branches in the frequency domain, merging low-frequency structural information from the global branch with high-frequency detail components from the local branch, resulting in high-fidelity and high-resolution generation.

coding coordinate information into feature representations. Among various designs, Rotary Position Embedding (RoPE) [36] is a widely used scheme that encodes relative positions through rotation in the embedding space, and it has been adopted in recent T2I models such as FLUX. Since our method operates directly on positional indices, it is compatible with both RoPE and other positional encoding variants.

**Diffusion Transformer (DiT).** DiT replaces the U-Net backbone with a pure Transformer architecture that processes image features as token sequences. Given image tokens  $\mathbf{x} \in \mathbb{R}^{s \times c}$ , the core computation in each Transformer block is the self-attention operation, defined as:

$$\mathbf{Q} = \mathbf{x}W_Q, \quad \mathbf{K} = \mathbf{x}W_K, \quad \mathbf{V} = \mathbf{x}W_V,$$

$$\text{Attn} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \quad (1)$$

This standard attention formulation serves as the basis for our two-branch attention restructuring in ResDiT.

### 3.2. Position Embedding Rectification

The original position embedding (PE) is not well-suited for high-resolution (HR) generation. As we discussed in Fig. 2,

we present two specified PE rectification patterns for each attention branch.

**PE Scaling (Interpolation).** For the first branch, to retain a consistent global layout, we interpolate the position embedding from the base resolution to the target high resolution. This constrains the position information of the entire feature map to the range that the pre-trained model is familiar with, helping it preserve the global structural skeleton of the generated image. Formally, let  $H, W$  be the height and width of a high-resolution feature map, the corresponding 2D position indices  $p_h$  and  $p_w$  are defined as:

$$(p_h, p_w) \in \{0, 1, \dots, H-1\} \times \{0, 1, \dots, W-1\}, \quad (2)$$

We then scale the position indices to the model’s training range:

$$(p_h, p_w) \in \left\{\frac{0}{s_h}, \frac{1}{s_h}, \dots, \frac{H-1}{s_h}\right\} \times \left\{\frac{0}{s_w}, \frac{1}{s_w}, \dots, \frac{W-1}{s_w}\right\}, \quad (3)$$

where  $s_h = H/h$ ,  $s_w = W/w$  and  $h, w$  represent the height and width at the training resolution. The scaled indices are then used to compute positional embeddings.

**Patch-wise Independent PE.** In the second branch, to align with patch-wise attention, each feature patch is assigned an independent positional embedding to maintain strong detail generation and reinforce local fidelity.

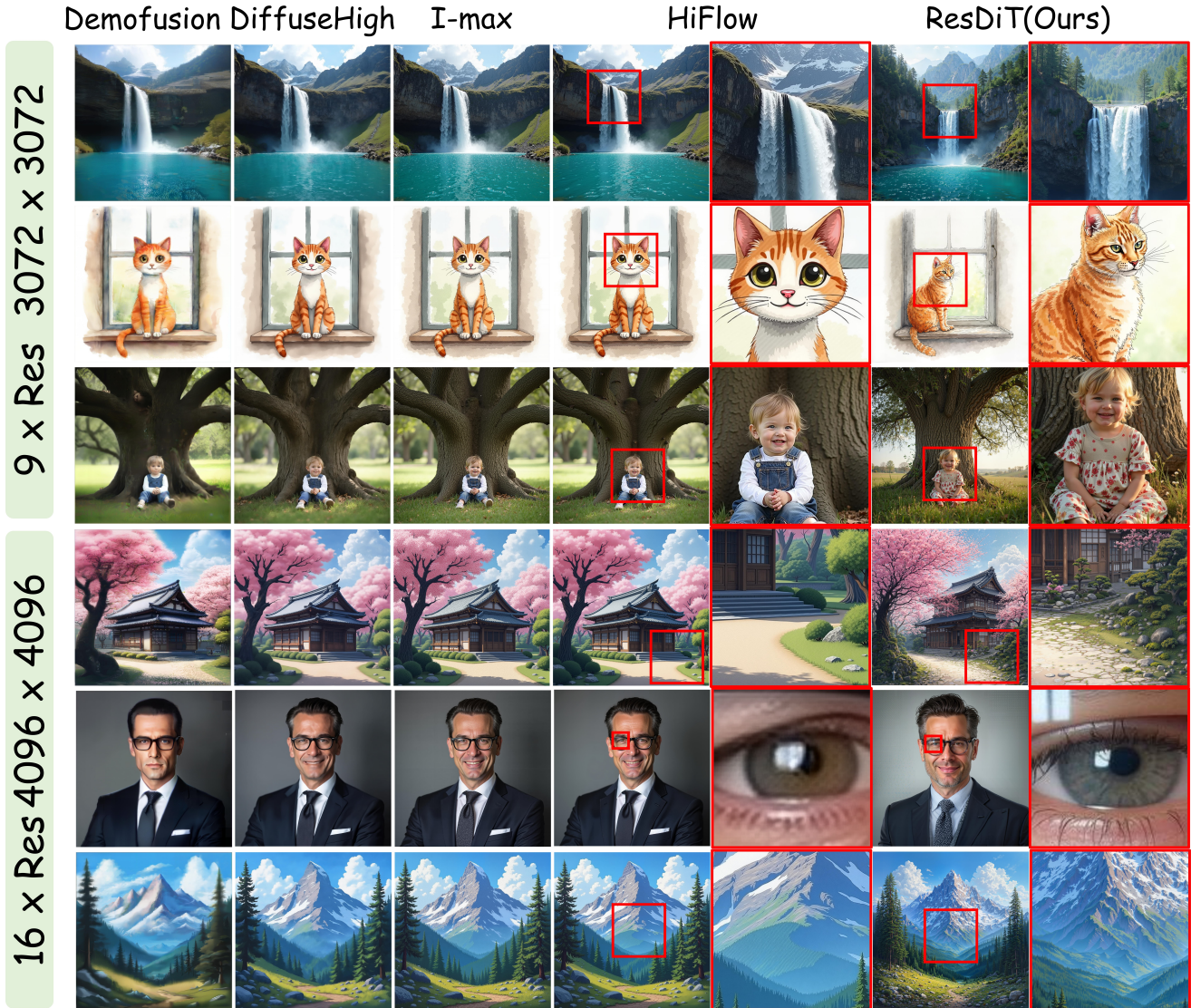


Figure 4. **Qualitative comparison with baselines.** ResDiT achieves a coherent global structure without relying on base resolution image information, while simultaneously delivering richer and more delicate local details in high-resolution outputs compared to existing methods. We further compare ResDiT with sota methods in terms of the capacity to generate fine-grained local details. **Best View ZOOM-IN.**

### 3.3. Patch Partitioning & Splicing

As discussed in Fig. 2, global attention over a high-resolution feature map often causes blurred textures and loss of fine details because the model is forced far beyond the spatial scale it was trained on. A natural solution is to confine attention to patch-level regions matching the training resolution. However, naïvely partitioning the feature map into a rigid grid introduces visible seams and grid artifacts along patch boundaries. To overcome this, we introduce partitioning and splicing strategies that enable patch-level attention while maintaining smooth, artifact-free continuity across patches.

**Minimum-Overlap Partitioning.** Non-overlapping feature patches can cause undesired and non-negligible image discontinuities, as features near patch boundaries lack contextual information from adjacent regions. To address this issue, we adopt a minimum-overlap partitioning strategy, where neighboring patches slightly overlap and thus share boundary context, smoothing feature transitions and reducing visible artifacts.

Concretely, along a single spatial axis of length  $H$  with patch size  $h$ , we choose an integer  $N$  such that  $N > H/h$ , and place the  $k$ -th patch ( $k = 1, \dots, N$ ) with starting index

$$t_k = \frac{(k-1)(H-h)}{N-1}. \quad (4)$$

Resolution	Method	KID↓	KID <sub>p</sub> ↓	IS↑	IS <sub>p</sub> ↑	CLIP↑	User Study↑
3072 × 3072	Demofusion [8]	0.0211	0.0342	12.20	10.21	31.92	3.1
	DiffuseHigh [18]	0.0195	0.0213	12.61	10.13	32.74	4.2
	I-Max [9]	0.0192	0.0207	<b>12.96</b>	10.48	32.73	4.2
	HiFlow [4]	<u>0.0190</u>	<b>0.0194</b>	12.87	<u>10.67</u>	<u>32.76</u>	<u>4.6</u>
	ResDiT(Ours)	<b>0.0189</b>	<u>0.0199</u>	<u>12.91</u>	<b>10.87</b>	<b>32.85</b>	<b>4.8</b>
4096 × 4096	Demofusion [8]	0.0236	0.0374	10.56	9.23	30.41	2.1
	DiffuseHigh [18]	0.0215	0.0298	11.40	9.81	32.70	3.9
	I-Max [9]	<u>0.0208</u>	<u>0.0275</u>	<b>11.79</b>	9.95	32.69	<u>4.0</u>
	HiFlow [4]	<b>0.0203</b>	<b>0.0245</b>	<u>11.65</u>	<b>10.12</b>	<b>32.74</b>	<b>4.3</b>
	ResDiT(Ours)	0.0217	<u>0.0252</u>	11.46	<u>9.97</u>	<u>32.71</u>	<b>4.3</b>

Table 1. **Quantitative comparisons with baselines.** The best results are highlighted in **bold**, and the second-best results are underlined. Compared to approaches that rely on base-resolution images for high-resolution generation, ResDiT achieves competitive, near state-of-the-art performance at 3072 × 3072 resolution. A slight performance drop is observed at 4096 × 4096, which is further analyzed in the experimental section.

In this way, the first patch starts at  $t_1 = 0$ , the last patch ends at  $t_N + h = H$ , and the stride between neighboring patches is smaller than  $h$ , ensuring both full coverage of the axis and a positive overlap between adjacent patches with only a small number of partitions.

**Gaussian Weighting Splicing.** When splicing the overlapped regions of partitioned patches, instead of treating each patch equally, we apply a Gaussian weighting strategy to integrate the overlapping features in a weighted manner. It further alleviates boundary artifacts by enabling smoother feature transitions between neighboring patches. Formally, for a token located at  $\mathbf{p}$  within an overlap region, let  $\mathcal{W}(\mathbf{p})$  denote the set of attention windows that cover  $\mathbf{p}$ . For each patch  $i \in \mathcal{W}(\mathbf{p})$ , we assign a Gaussian weight:

$$w_i(\mathbf{p}) = \exp\left(-\frac{\|\mathbf{p} - \mathbf{c}_i\|_2^2}{2\sigma^2}\right), \quad (5)$$

where  $\mathbf{c}_i$  is the centre of patch  $i$  and  $\mathcal{W}(\mathbf{p}) \in (0, 1]$ . The final fused feature of token  $\mathbf{p}$  is obtained:

$$\mathbf{f}(\mathbf{p}) = \frac{\sum_{i \in \mathcal{W}(\mathbf{p})} w_i(\mathbf{p}) \mathbf{f}_i(\mathbf{p})}{\sum_{i \in \mathcal{W}(\mathbf{p})} w_i(\mathbf{p})}. \quad (6)$$

### 3.4. Patch-wise Spectral Fusion

To effectively integrate the outputs of the two branches, we first note that they contribute complementary information: the global branch captures reliable low-frequency layout structures, while the local branch excels at high-frequency details. This motivates us to perform fusion in the frequency domain, where such components can be cleanly separated and recombined. Specifically, we suppress high-frequency components in the global branch and low-frequency components in the local branch, enabling the two sources to merge in a naturally complementary manner.

Our fusion operates at the patch level, as the frequency composition varies significantly across spatial re-

gions—high frequencies dominate textured or edge-rich areas, while low frequencies are more prevalent in smooth or homogeneous regions. Patch-wise spectral fusion allows the model to adaptively modulate frequency components for each region, yielding more flexible and effective integration than applying a global frequency filter.

Specifically, we partition the global output feature  $\mathbf{x}_g$  from the first branch using the same minimum-overlap partitioning strategy:

$$\mathbf{x}_g = \{\mathbf{x}_g^1, \dots, \mathbf{x}_g^i, \dots\}, \quad (7)$$

which correspond one-to-one with the local patch outputs. For each aligned patch pair  $(\mathbf{x}_g^i, \mathbf{x}_l^i)$ , we apply the Fast Fourier Transform [3] (FFT) to obtain their spectral representations:

$$\hat{\mathbf{x}}_g^i = \mathcal{F}(\mathbf{x}_g^i), \quad \hat{\mathbf{x}}_l^i = \mathcal{F}(\mathbf{x}_l^i). \quad (8)$$

We reconstruct the spatial-domain patch by applying the inverse FFT to the fused spectrum:

$$\mathbf{x}^i = \mathcal{F}^{-1}(\mathbf{M} \odot \hat{\mathbf{x}}_g^i + (1 - \mathbf{M}) \odot \hat{\mathbf{x}}_l^i), \quad (9)$$

where  $\mathbf{M}$  is a binary mask that filters the frequency spectrum, and  $\mathcal{F}^{-1}$  denotes the inverse FFT that maps features back to the spatial domain.

## 4. Experiments

### 4.1. Implementation Details

**Experimental settings.** If not specified, we deploy ResDiT on FLUX.1-dev [19], an advanced open-source model based on DiT architecture. The sampling steps are set to 35, and the guidance scale is set to 3.5. Following prior work [25, 26], we set the normalized frequency cutoff of the spectral mask to 0.2. Recent research has observed the coarse-to-fine nature of diffusion denoising [41]; inspired

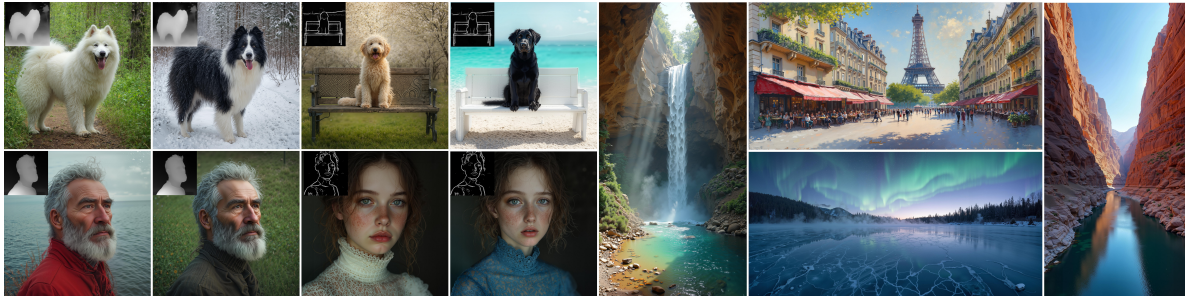


Figure 5. ResDiT seamlessly integrates with ControlNet, enabling precise structure-controlled generation of the images at resolutions of  $3072 \times 3072$ . Furthermore, ResDiT supports arbitrary aspect ratios, the images at resolutions of  $2048 \times 4096$  and  $4096 \times 2048$ .

by this, we use a global branch for the first 10 timesteps, a local branch for the last 15 timesteps, and Patch-wise Spectral Fusion for the remaining steps. This strategy is similar to those proposed in prior work [13, 44]. All experiments are conducted on RTX 4090 GPU.

**Baselines.** We compare our method with training-free high-resolution generation approaches, Demofusion [8], Diffuse-High [18], I-Max [9] and HiFlow [4]. They all adopt a two-stage paradigm: first generate a base-resolution image and then perform high-resolution extrapolation. All methods are evaluated using their official implementations.

**Evaluation.** We collect 500 high-quality captions and generate images corresponding to each caption. We selected Kernel Inception Distance [2] (KID), Inception Score [34] (IS), and CLIP Score [30] as our evaluation metrics. KID measures the similarity between the generated high-resolution images and the original resolution images. IS assesses the diversity and definition of the generated images, and CLIP Score represents the prompt-following capability. The KID is calculated between generated images and 2K real high-quality images sourced from LAION-Aesthetics-v2 6.5plus [35]. To further provide the concrete evaluation, we also adopted patch KID and patch IS as our evaluation metrics. Besides, we conduct a User Study for further evaluation. Twenty participants independently rated images on a scale from 1 to 5 for their image visual quality based on 40 randomly selected prompts per method. The average user scores are reported. A comparison of generation latency between the original method and the main baselines is provided in the appendix.

## 4.2. Comparison to State-of-the-Art Methods

**Qualitative comparison.** We qualitatively compare our method with baseline approaches in Fig. 4. They all rely on base-resolution images as guidance for high-resolution generation. Specifically, Demofusion preserves the overall image structure but introduces significant noise and severe detail loss. DiffuseHigh improves upon this by reducing artifacts, yet still struggles with fine-detail fidelity. I-max de-

livers richer details while maintaining structural accuracy, but suffers from localized blurring. The state-of-the-art HiFlow achieves superior consistency with base-resolution references, yielding visually sharper results, yet at the cost of over-smooth textures and diminished fine details. For instance, zoomed regions reveal unnatural smoothness on the child’s face, impoverished tree trunk textures in the background, and blurred distant mountain contours. In contrast, our method simultaneously preserves accurate global structure, recovers rich fine-scale details, and delivers superior visual realism without requiring any base-resolution input as a reference.

**Quantitative comparison.** We report the quantitative evaluation results in Tab. 1. ResDiT achieves a notably high CLIP score, demonstrating strong image–text alignment. At a resolution of  $3072 \times 3072$ , it also attains superior KID and IS scores, reflecting improved image quality and diversity, benefiting from its ability to introduce rich fine-grained details while maintaining coherent global structure. However, at  $4096 \times 4096$  resolution, a moderate drop in KID and IS scores is observed. We attribute this to the inherent difficulty of single-stage high-resolution generation. In contrast, most competing approaches adopt a two-stage strategy that heavily relies on base-resolution images generated by the model itself. As a result, their high-resolution outputs closely follow the distribution of the base-resolution images, leading to favorable scores. In comparison, ResDiT performs direct sampling in the high-resolution noise space without guidance from lower-resolution priors. While this design allows genuine high-resolution synthesis with richer and more diverse details, it also introduces a distributional shift from the original model’s generation space, which can lead to slightly inferior KID and IS metrics.

**Additional Qualitative Results.** Fig. 5 (left) illustrates ResDiT’s seamless compatibility with control modules such as ControlNet [42], using depth and HED edge maps as structured priors under different textual prompts. The generated outputs show strong spatial alignment with the control inputs, confirming that ResDiT effectively preserves

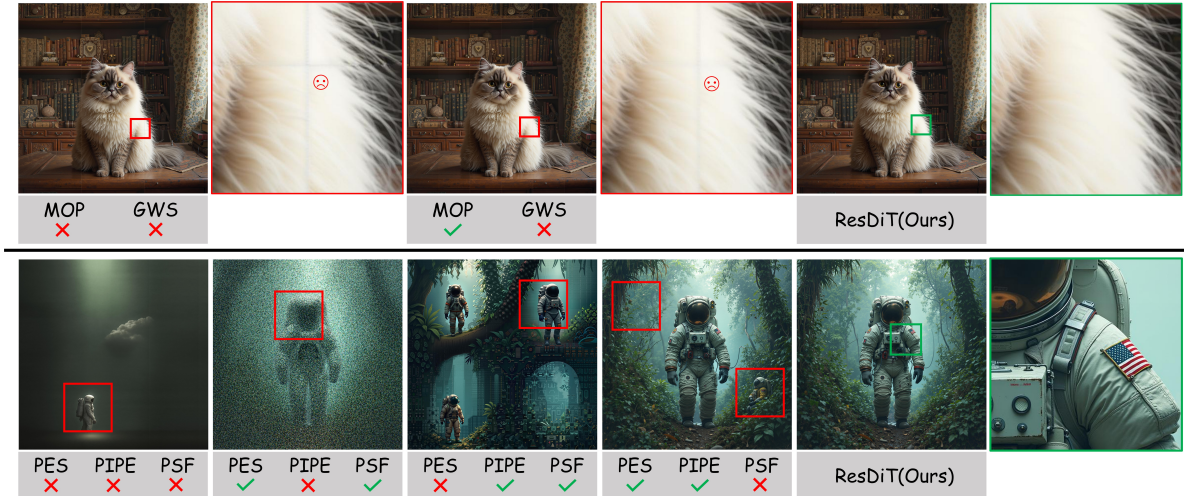


Figure 6. **Ablating each component of ResDiT.** Patch Partitioning & Splicing confirms its role in preserving boundary fidelity and local detail. Without PES, the global structure becomes inconsistent. It ensures scalable coherence, while PIPE adds rich fine details. PSF then synergistically combines both strengths, yielding superior generation quality.

and enhances control consistency without compromising generation quality. The right side further demonstrates that ResDiT naturally supports arbitrary aspect ratios, while maintaining rational global layouts and delivering rich, high-fidelity local details, demonstrating its flexibility and robustness across diverse resolution configurations. More qualitative results are included in the appendix.

### 4.3. Ablation Study

**Position Embedding Rectification & Patch-wise Spectral Fusion.** We validate the contribution of each component in our pipeline, as shown in the bottom part of Fig. 6. We validated the respective contributions of PE Scaling (PES), Patch-wise Independent PE (PIPE), and Patch-wise Spectral Fusion (PSF). Removing PSF, replacing frequency-domain fusion with spatial-domain addition and averaging, led to repeated generation artifacts and increased image blurriness, confirming that frequency-domain fusion effectively leverages the complementary strengths of PES and PIPE while mitigating their drawbacks. Removing PES and reverting to the original positional encoding caused complete structural collapse, showing that PES is essential for modeling global structure at high resolution. When PIPE was removed, the global layout remained reasonable under PES, but fine details were severely degraded, demonstrating PIPE’s importance in enhancing local fidelity. Finally, removing all three resulted in both global structural collapse and local details distortion. Only the joint use of PES, PIPE, and PSF yields globally coherent and visually sharp results, highlighting their complementary roles in global layout control and fine-detail generation. Quantitative ablation experiments are included in the appendix.

**Patch Partitioning & Splicing.** We conduct ablation study to evaluate the effectiveness of Minimum-Overlap Partitioning (MOP) and Gaussian Weighting Splicing (GWS). As shown in the upper part of Fig. 6, removing both components, clear boundary artifacts emerge. Although our method maintains global information exchange, it inevitably introduces boundary discontinuities due to patch-level fusion and attention mechanisms. Using MOP alone greatly reduces segmentation artifacts, achieving results close to ours. However, for high-resolution generation, fine details remain crucial. Zooming into local regions reveals that artifacts persist, overlapping alleviates their visibility but increases their number and damages delicate details. In contrast, combining them effectively eliminates boundary artifacts, ensuring smooth transitions and consistent spatial coherence across global and local features. Quantitative ablation experiments are included in the appendix.

## 5. Conclusion

We introduce ResDiT, a training-free framework that extends pre-trained Diffusion Transformers to high-resolution generation without relying on base-resolution images. We investigated intrinsic causes of high-resolution failures in DiTs and found that positional embeddings govern spatial layout while attention receptive-field scale controls detail fidelity. By restructuring the attention mechanism into global and local branches, ResDiT effectively preserves large-scale structural coherence while enriching fine-grained visual details. Extensive experiments demonstrate that ResDiT achieves competitive or superior performance at high resolutions, validating its effectiveness as a simple yet powerful solution for high-resolution diffusion generation.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62173045), the Beijing Natural Science Foundation under Grant F2024203115, the Young Scientists Fund of NSFC (Grant No. 62406035), the China Postdoctoral Science Foundation under Grant Number 2024M750255, and BUPT Excellent Ph.D. Students Foundation (CX20241088).

## References

- [1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2, 3
- [2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7
- [3] E Oran Brigham. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988. 6
- [4] Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025. 2, 3, 6, 7
- [5] Boyuan Cao, Jiaxin Ye, Yujie Wei, and Hongming Shan. Ap-ldm: Attentive and progressive latent diffusion model for training-free high-resolution image generation. *arXiv preprint arXiv:2410.06055*, 2024. 2, 3
- [6] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey, 2024. 2
- [7] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. 2
- [8] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6159–6168, 2024. 2, 3, 6, 7
- [9] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. 2024. 2, 3, 6, 7
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3
- [12] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European conference on computer vision*, pages 39–55. Springer, 2024. 2
- [13] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3, 7
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3
- [15] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023. 2
- [16] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European conference on computer vision*, pages 196–212. Springer, 2024. 2, 3
- [17] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36:70847–70860, 2023.
- [18] Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4338–4346, 2025. 2, 3, 6, 7
- [19] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 2, 3, 6
- [20] Zhihang Lin, Mingbao Lin, Meng Zhao, and Rongrong Ji. Accdiffusion: An accurate method for higher-resolution image generation. In *European Conference on Computer Vision*, pages 38–53. Springer, 2024. 2, 3
- [21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3
- [22] Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. Linfusion: 1 gpu, 1 minute, 16k image. *arXiv preprint arXiv:2409.02097*, 2024. 2
- [23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3
- [24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3
- [25] Yu Lu and Yi Yang. Freelong++: Training-free long video generation via multi-band spectralfusion. *arXiv preprint arXiv:2507.00162*, 2025. 6
- [26] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37:131434–131455, 2024. 6

- [27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3
- [29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7
- [31] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. *Advances in Neural Information Processing Systems*, 37:111131–111171, 2024. 2
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3
- [34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7
- [35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 7
- [36] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 4
- [37] Tobias Vontobel, Seyedmorteza Sadat, Farnood Salehi, and Romann Weber. Hiwave: Training-free high-resolution image generation via wavelet-based diffusion sampling. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025. 3
- [38] Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3944–3953. IEEE, 2025. 2, 3
- [39] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Diffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023. 2
- [40] Haosen Yang, Adrian Bulat, Isma Hadji, Hai X Pham, Xiaotian Zhu, Georgios Tzimiropoulos, and Brais Martinez. Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2459–2468, 2025. 2, 3
- [41] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. *Advances in Neural Information Processing Systems*, 37:55342–55369, 2024. 6
- [42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 7
- [43] Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024. 2, 3
- [44] Feng Zhou, Pu Cao, Yiyang Ma, Lu Yang, and Jianqin Yin. Exploring position encoding in diffusion u-net for training-free high-resolution image generation. *arXiv preprint arXiv:2503.09830*, 2025. 2, 3, 7