HRDiT: Training-Free High-Resolution Image Generation with Off-the-Shelf Diffusion Transformer Models

Yu Xue[1], Haoxuan Qu[1]*, Zhuoling Li[1], Hongbin Xu[2], Jianxiong Yin[3], Simon See[3], Hossein Rahmani[1], and Jun Liu[1]

[1] Lancaster University, Lancaster, UK
[2] South China University of Technology, Guangzhou, China
[3] NVIDIA AI Tech Centre, Singapore
{y.xue9,h.qu5,z.li81,h.rahmani,j.liu81}@lancaster.ac.uk,
hongbinxu1013@gmail.com, {jianxiongy,ssee}@nvidia.com

Abstract. Training-free text-to-high-resolution image generation has recently attracted growing research attention. However, existing studies on this task primarily focus on adapting off-the-shelf U-Net-based diffusion models to high resolutions, with limited progress on adapting off-the-shelf Diffusion Transformer (DiT) models despite their strong text-to-image generation capabilities at limited resolutions. In this work, we find two key challenges particularly hindering the application of off-the-shelf DiT models for high-resolution image synthesis in a training-free manner, namely, spatial disorder and long generation time. To address these challenges, we propose a novel method tailored to adapt off-the-shelf DiT models for high-resolution image synthesis. Extensive experiments show the efficacy of our method.

Keywords: Text-to-high-resolution image generation · Training-free · Off-the-shelf diffusion transformer models

1 Introduction

Text-to-high-resolution image generation aims to directly synthesize high-quality, large-resolution images from text prompts. It benefits diverse applications such as digital art creation [33], advertising [12], and game development [30]. To tackle this task, existing methods can be broadly categorized into *training-based* and *training-free* approaches. *Training-based* methods [9,14,26,48] optimize new generators specifically for high-resolution outputs. However, they generally require substantial computational resources and large-scale high-resolution datasets for effective training [6, 46], limiting their practicality. Given these limitations, recently, *training-free* methods have emerged as an appealing alternative. Instead of optimizing new generators, these methods aim to unlock the potential of off-the-shelf text-to-image diffusion models typically trained at limited resolutions

* Corresponding author.

(e.g., 1024×1024), driving them to effectively generate higher-resolution images in a training-free manner. Due to their simplicity and flexibility, such approaches have attracted growing attention recently [13, 32, 40, 47].

Among existing training-free approaches, the early dominance of U-Net architectures in off-the-shelf text-to-image diffusion models has led most of them [13, 32, 40, 47] to focus on adapting U-Net-based diffusion models to high resolutions. However, the landscape of off-the-shelf diffusion models has recently shifted: Diffusion Transformer (DiT)-based models now substantially outperform their U-Net counterparts at limited resolutions [28], driving the rapid emergence of various powerful DiT-based off-the-shelf text-to-image models, such as Stable Diffusion 3 [8] and FLUX [20]. However, despite this shift, training-free approaches for adapting off-the-shelf DiT models to high-resolution generation remain largely underexplored. Motivated by this research gap, in this work, we aim to narrow down this gap, *exploring effective adaption of off-the-shelf DiT models for high-resolution image generation in a training-free manner*.

To achieve this, we first attempt to adapt off-the-shelf DiT models to higher resolutions either directly or by applying the existing DiT-tailored method I-Max [7]. Profiling across a wide range of generation tasks, we observe that, regardless of the adaptation strategy, off-the-shelf DiT models encounter two critical challenges that severely hinder their practicality at high resolutions. We illustrate these challenges in Fig. 1 and summarize them as follows. (1) Spatial disorder: both adaptation strategies lead to the generation of images with noticeable spatial disorder (highlighted by red boxes in Fig. 1), which markedly degrades overall image quality. (2) Long generation time: when adapted for high-resolution image synthesis, off-the-shelf DiT models also suffer from very long generation time, further limiting their practicality. To effectively adapt off-the-shelf DiT models to high resolutions, we therefore aim to address these two challenges in the remainder of this work.

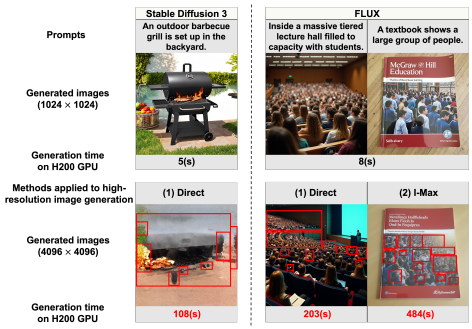

Fig. 1: Results of (1) *directly* adapting off-the-shelf text-to-image DiT models to high-resolution (4096×4096) generation, and (2) applying the DiT-tailored method I-Max [7]. At limited resolution (1024×1024), off-the-shelf DiT models generate high-quality images quickly. However, when adapted to higher resolutions through *direct* adaptation or I-Max, they exhibit spatial disorder (see red boxes and zoom-in) and incur long generation times (red text). Note that off-the-shelf DiT models can be *directly* adapted since they mechanically accept higher-resolution initial noises, while I-Max is, by design, applicable to FLUX but not Stable Diffusion 3.

To this end, theoretical and empirical analyses are first drawn on to try to identify the key underlying causes of these challenges. The theoretical analysis

(detailed in Sec. 3.1) reveals that, a key reason behind the *spatial disorder* issue can lie in the "limited expressiveness" of off-the-shelf DiT model's positional embedding mechanism when adapted to high resolutions. Meanwhile, the empirical profiling shows that, multi-head attention computations are the dominant source of the *long generation time* faced by DiT models at high resolutions. For example, when directly adapting mainstream DiT models (such as FLUX and Stable Diffusion 3) to generate 8K (8192×8192) images, the multi-head attention computations alone account for over 90% of the overall generation time for both models (as illustrated in Fig. 2), taking 1,385 and 646 seconds per image, respectively.

Building on these insights, to effectively adapt off-the-shelf DiT models to high resolutions, in this work, we propose a novel training-free framework, **H**igh-**R**esolution **DiT** (**HRDiT**). It comprises two components, each corresponding to one of the aforementioned challenges based on its identified underlying cause. To tackle the *spatial disorder* challenge, inspired by [11, 24], HRDiT involves a *Spatial Position Alignment* (SPA) component, which adjusts the positional embedding mechanism of off-the-shelf DiT models via two complementary operations (Bundle and Slide) to help it preserve its proper functionality when extended to high resolutions. To tackle the *long generation time* problem, inspired by [44], HRDiT further integrates a *Head-adaptive Attention Pruning* (HAP) component to effectively prune redundant computations in the multi-head attention, thereby substantially reducing runtime cost for high resolution generation. With these two components, HRDiT enables high-quality and efficient high-resolution image synthesis from off-the-shelf DiT models trained only at limited resolutions, entirely in a training-free manner.

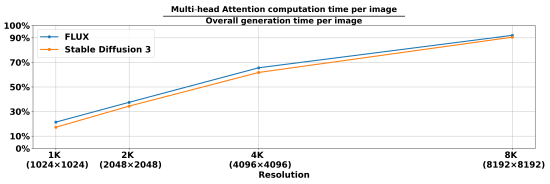

Fig. 2: Proportion of time that mainstream off-the-shelf DiT models (FLUX [20] and Stable Diffusion 3 [8]) spent on multi-head attention computations per image, relative to that on the overall generation process per image, across different image resolutions. These models are adapted to high resolutions in the same *direct* manner as described in the caption of Fig. 1. As shown, at high resolutions, multi-head attention computations alone dominate the overall generation time of these off-the-shelf DiT models.

2 Related Work

Training-free Text-to-high-resolution Image Generation, which aims to unlock the high-resolution generation ability of off-the-shelf text-to-image diffusion models originally trained at limited resolutions, has recently gained significant attention [2–4, 6, 7, 13, 15, 16, 18, 19, 21, 25, 31, 32, 37, 38, 40, 41, 47, 49, 50]. To tackle this task, some works, such as DemoFusion [6], DiffuseHigh [18], and

HiFlow [3], focus primarily on modifying the diffusion process of diffusion models. Some others instead focus more on the architectures of diffusion models. For instance, ScaleCrafter [13] employed dilation to expand U-Net's receptive field. HiDiffusion [47] dynamically adjusted feature map resolutions in the outer U-Net blocks to tackle the object repetition issue. FreeScale [32] integrated restrained dilated convolutions and multi-scale fusion into the U-Net to enhance both global and local generation fidelity. Beyond U-Net-based designs, the architecture of DiT models has also been explored, e.g., by I-Max [7]. Yet, we find that such investigations still leave two key challenges (i.e., spatial disorder and long generation time) that critically limit the effective application of off-the-shelf DiT models at high resolutions unaddressed.

In this work, we focus on the architectural perspective, and specifically target these two key challenges that significantly hinder DiT models from effective high-resolution generation. By identifying their key causes and correspondingly integrating proper strategies to address them, our framework enables off-the-shelf DiT models to achieve high-quality, efficient high-resolution synthesis without any retraining.

3 The Proposed HRDiT Framework

Here, we aim to effectively adapt off-the-shelf DiT models to high-resolution image synthesis, by addressing two critical challenges that severely hinder their practical use at high resolutions: *spatial disorder* and *long generation time*. Drawing on analyses, we identify the key underlying causes of each challenge. Correspondingly, we propose **HRDiT**, a novel training-free framework integrated with two components, SPA and HAP, respectively addressing aforementioned challenges based on identified underlying causes. Below, we detail these components.

3.1 Spatial Position Alignment (SPA)

To handle the *spatial disorder* challenge, we first examine its underlying cause, drawing inspiration from [11]. Specifically, in DiT models, the positional embedding mechanism typically plays a crucial role in expressing positional information, distinguishing spatial locations within the generated image, and thus preventing spatial disorder [1, 22, 35]. This naturally raises a question: could the spatial disorder observed in off-the-shelf DiT models during high-resolution generation stem from the possible "limited expressiveness" of their positional embedding mechanisms, when extended to higher resolutions? To explore this, we first formalize the positional embedding mechanisms in off-the-shelf DiT models, which will then serve as the foundation for the subsequent analysis.

Formalization of position embedding mechanism. To facilitate a general analysis across different mainstream off-the-shelf DiT models, we aim to first formalize their position embedding mechanisms in a unified manner. In specific,

we analyse that the positional embedding mechanisms of mainstream off-the-shelf DiT models, such as FLUX [20] and Stable Diffusion 3 [8], though varying in detailed designs, can be unified under the following formulation:

$$\mathbf{q_i} = g_q(\mathbf{x_i}, f_{pe}^{tok}(i)), \mathbf{k_i} = g_k(\mathbf{x_i}, f_{pe}^{tok}(i)), \mathbf{v_i} = g_v(\mathbf{x_i}, f_{pe}^{tok}(i)), \tag{1}$$

where $\mathbf{x_i}$ denotes the i-th token in the input hidden state of an attention block, and $\mathbf{q_i}$, $\mathbf{k_i}$, $\mathbf{v_i}$ are the corresponding query, key, and value vectors derived from $\mathbf{x_i}$ via the functions $g_q(\cdot,\cdot)$, $g_k(\cdot,\cdot)$, and $g_v(\cdot,\cdot)$, respectively. To inject spatial awareness, the position embedding mechanisms feed the token-level positional signal $f_{pe}^{tok}(i)$ once into each of the three functions $g_q(\cdot,\cdot)$, $g_k(\cdot,\cdot)$, and $g_v(\cdot,\cdot)$. Additional details on this formulation are in Supplementary.

Denote T the total number of visual tokens in the input hidden state. After obtaining the query, key, and value vectors for all T tokens (i.e., $\{\mathbf{q_t}\}_{t=0}^{T-1}$, $\{\mathbf{k_t}\}_{t=0}^{T-1}$, and $\{\mathbf{v_t}\}_{t=0}^{T-1}$) via Eq. 1, the attention in off-the-shelf DiT models can then be performed in a spatial-aware manner as:

$$\mathbf{c_{i,j}} = \frac{e^{\mathbf{q_i}\mathbf{k_j}^\intercal}}{\sum_{t=1}^{T} e^{\mathbf{q_i}\mathbf{k_t}^\intercal}}\mathbf{v_j}, \tag{2}$$

where $\mathbf{c_{i,j}}$ denotes the contribution of token j to the attention output at position i, and e is the Euler number. For simplicity, we omit the stabilization factor $\sqrt{d}$, where d is the dimensionality of each key vector, in Eq. 2. At this point, we arrive at a unified formalization of how positional embedding mechanisms in mainstream off-the-shelf DiT models are designed to inject spatial awareness into their attention operations, which underpins the analysis below.

Expressiveness analysis. Based on our above unified formalization of positional embedding mechanisms in mainstream off-the-shelf text-to-image DiT models, we next ask whether existing theoretical tools can be drawn on to analyze the expressiveness of these mechanisms when they are extended to high resolutions with substantially more tokens, i.e., much larger T. To this end, we observe that the pseudo-dimension-based analysis developed in prior work [11] for relative positional embeddings can, when combined with our formalization, be adapted to perform the desired analysis under mild assumption, across mainstream off-the-shelf DiT models including FLUX and Stable Diffusion 3, *regardless of whether the positional embeddings are relative or not*. As elaborated below, this adapted theoretical analysis reveals that these mechanisms can indeed suffer from limited expressiveness when extended to high resolutions.

To set up the adapted analysis, we introduce the following notations. (1) By combining Eq. 1 and Eq. 2, we first express $\mathbf{c_{i,j}}$ via a single function g as:

$$\mathbf{c_{i,j}} = g(\mathbf{x_i}, \mathbf{x_j}, f_{pe}(i,j)), \tag{3}$$

where g represents the complete computation performed by off-the-shelf DiT models to obtain $\mathbf{c_{i,j}}$ from the input hidden-state tokens $\mathbf{x_i}$ and $\mathbf{x_j}$. Expressing $\mathbf{c_{i,j}}$ in this way, the positional information involved in computing $\mathbf{c_{i,j}}$ can be compactly represented by a single *pairwise positional signal* denoted as $f_{pe}(i,j)$, rather than by two separate token-level positional signals $f_{pe}^{tok}(i)$ and $f_{pe}^{tok}(j)$ (see supplementary for more details). This compact representation facilitates the subsequent analysis. (2) We further denote $S_{pe} = \{ f_{pe}(i,j) \mid i \in \{0,\ldots,T-$

$1\},\ j \in \{0,\ldots,T-1\}\}$ as the set of all possible pairwise positional signals. With these notations established, we now proceed to the adapted theoretical analysis.

Proposition 1. *Let dis_g denote a distance measure quantifying the discrepancy between two pairwise positional signals in the output space of the function g (see Supplementary for exact definition of dis_g). By definition of g in Eq. 3, dis_g then measures the discrepancy between two pairwise positional signals, after they are respectively propagated through g during attention computation. Using dis_g, we partition all elements in the set S_{pe} into $h(T)$ mutually exclusive subsets such that any two elements within the same subset satisfy $dis_g \leq \epsilon$, while any two elements from different subsets satisfy $dis_g > \epsilon$, where ϵ is a predefined threshold. Intuitively, the resulting $h(T)$ characterizes the positional embedding mechanism's ability to distinguish between different pairwise positional signals in S_{pe} after attention computation. A larger $h(T)$, indicating that S_{pe} can be divided into a greater number of well-separated subsets under dis_g, corresponds to a stronger distinction capability of the positional embedding mechanism. Yet meanwhile, $h(T)$ can be proved to be upper-bounded by:*

$$h(T) \ \leq \ \lambda \cdot \Big(\sup_{S_{pe}} |g(\mathbf{x_i}, \mathbf{x_j}, f_{pe}(i,j))| \Big)^{2\xi}, \tag{4}$$

where ξ denotes the pseudo-dimension of the function class $\mathcal{G} = \{\, g(\cdot,\cdot,f_{pe}(i,j)) \mid i \in \{0,\ldots,T-1\},\ j \in \{0,\ldots,T-1\}\,\}$, and $\lambda = (\frac{e}{\epsilon})^{2\xi} \cdot 2^{4\xi+1}$. Notably, λ can be regarded as a constant when the threshold ϵ is fixed.

The proof of Proposition 1 is provided in Supplementary. Note that, ideally, for the positional embedding mechanism to function properly, $h(T)$, which measures the mechanism's ability to express different pairwise positional signals in a distinguishable manner after attention computation, should reach its maximum possible value, namely the size $|S_{pe}|$ of the set S_{pe} (i.e., satisfy $h(T) = |S_{pe}|$). Otherwise, at least two pairwise positional signals in the set S_{pe} would have cross-distances below ϵ after attention computation, making them indistinguishable and thereby inducing spatial disorder. However, Proposition 1 suggests that $h(T) = |S_{pe}|$ may not hold at high resolutions. This is because, (1) as resolution increases, the number of tokens T grows rapidly, and $|S_{pe}|$ typically scales accordingly (see Supplementary for details). (2) Yet meanwhile, Eq. 4 shows that the attainable $h(T)$ is upper-bounded by the supremum of $g(\cdot)$ (up to a constant factor λ and a constant power 2ξ). Taken together, these two factors indicates that, as resolution increases, achieving $h(T) = |S_{pe}|$ would necessitate the aforementioned upper bound of $h(T)$ to scale at least as fast as $|S_{pe}|$. The associated analysis, however, suggests that this scaling may fail to occur in practice. Taking FLUX as an example, when an analysis similar to that in [24] is applied to 4K image generation, fewer than 10% of the test samples are found to satisfy Eq. 4 after substituting $h(T)$ with $|S_{pe}|$, and this ratio can further decrease at higher resolutions. The above indicate that, when adapted for high-resolution generation, the positional embedding mechanisms in off-the-shelf text-to-image DiT models can exhibit limited expressiveness, failing to adequately distinguish pairwise positional signals, thus leading to observed spatial disorder.

Addressing limited expressiveness through SPA. Building on the above analysis, we aim to address the limited expressiveness of DiT models' positional embedding mechanisms to handle the spatial disorder problem. Achieving this addressing, however, is non-trivial, particularly because our goal is to do so in a training-free and universally compatible manner across diverse mainstream off-the-shelf DiT models. To tackle this challenge, we note that, though different mainstream DiT models may differ in their specific implementations of the positional function f_{pe}, they all share a common structure: f_{pe} takes token indices (i.e., i and j) as inputs. This observation suggests that if we can achieve the addressing by directly manipulating these token indices i and j before they are fed into f_{pe}, the addressing process can be carried out in a universally compatible manner. To this end, we take inspiration from [17, 24], adopt a similar approach, and integrate a training-free component, termed Spatial Position Alignment (SPA), into our framework to achieve the addressing in such a way.

Specifically, SPA consists of two key operations: **bundle** and **slide**. In the **bundle** operation, to facilitate $h(T)$ in attaining $|S_{pe}|$ (i.e., to satisfy $h(T) = |S_{pe}|$) and thus help the positional embedding mechanism to function properly even at high resolutions, we aim to reduce $|S_{pe}|$. To achieve this by manipulating token indices before they are fed into the positional function f_{pe}, we group tokens into several bundles and feed the corresponding bundle indices, rather than individual token indices, into f_{pe}. In this way, as the number of grouped bundles is typically much smaller than the number of tokens, we can constrain the input diversity of f_{pe} and thus reduce $|S_{pe}|$, which reflects the output diversity of f_{pe} based on the definition of S_{pe}. In turn, as elaborated in the above analysis associated to Proposition 1, this helps the positional embedding mechanism better preserve the distinguishability among positional signals in S_{pe} and restore its proper functionality. However, despite such benefit, bundling also introduces an inherent limitation: positional distinctions among tokens within the same bundle become indistinguishable, potentially weakening fine-grained spatial discrimination in generated images. To tackle this, SPA integrates a complementary **slide** operation, which further equips the framework with fine-grained positional distinguishability among tokens within each bundle. Together, these two operations effectively rectify the positional embedding mechanism, facilitating off-the-shelf DiT models to preserve spatial coherence even when adapted to high resolutions. Below, we detail these two operations.

(1) The bundle operation. In this operation, we start from the original positional function $f_{pe}(i, j)$, whose inputs are the token indices themselves (i.e., i and j). Our goal is to group tokens into bundles and feed f_{pe} with the corresponding bundle indices instead. Formally, this rewrites $f_{pe}(i, j)$ as $f_{pe}(\phi_{bundle}(i), \phi_{bundle}(j))$, where $\phi_{bundle}(i)$ and $\phi_{bundle}(j)$ denote the indices of the bundles containing the i-th and j-th tokens, respectively. For simplicity, we here bundle tokens sequentially according to their token indices, so that the resulting bundle indices approximately preserve the original positional order among the T tokens. In other words, tokens belonging to bundles with larger bundle indices naturally possess larger token indices than those tokens in bundles with smaller bundle indices.

Under this setup, $\phi_{bundle}(\cdot)$ is defined as follows:

$$\phi_{bundle}(i) = \begin{cases} 0, & 0 \le i < N_1, \\ \left\lceil \dfrac{i+1-N_1}{N} \right\rceil, & N_1 \le i < T, \end{cases} \tag{5}$$

where N is a hyperparameter, and N_1 ranges from 1 to N. N_1 and N denote the number of tokens in the first bundle and in each middle bundle (excluding the first and last ones), respectively. Here, the first bundle is intentionally not enforced to reach the same size as the middle ones, which facilitates the subsequent **slide** operation. Furthermore, given N_1 and N, the size of the last bundle can be automatically determined as $T - N_1 - \left\lfloor \frac{T-N_1}{N} \right\rfloor \times N$, so it does not need to be explicitly specified.

An illustrative example of the application of the $\phi_{bundle}(\cdot)$ function is provided in Fig. 3(a). As shown, applying $\phi_{bundle}(\cdot)$ reduces the diversity of each input to $f_{pe}(\cdot,\cdot)$, from the total number of token indices T to the total number of bundle indices, which can be computed as $\left\lceil \frac{T+N-N_1}{N} \right\rceil$. Accordingly, when computing $|S_{pe}|$ based on f_{pe} with bundle indices as its inputs, we can replace T with $\left\lceil \frac{T+N-N_1}{N} \right\rceil$, yielding a significantly smaller $|S_{pe}|$. For instance, in the FLUX model, instead of $|S_{pe}| = 2T - 1$, we now have $|S_{pe}| = 2 \times \left(\left\lceil \frac{T+N-N_1}{N} \right\rceil\right) - 1$ after applying the bundle operation. Following the analysis associated with Proposition 1, this reduction enhances the distinguishability among elements in S_{pe}, and thus helps restore the proper functionality of the positional embedding mechanism.

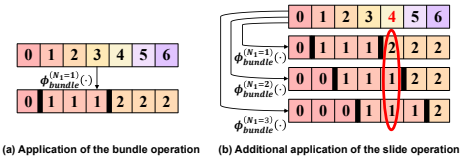

Fig. 3: Illustration of the **bundle** and **slide** operations, with $T = 7$ and $N = 3$. (a) The **bundle** operation with $N_1 = 1$: using the mapping function $\phi_{bundle}(\cdot)$ with $N_1 = 1$ (i.e., $\phi_{bundle}^{(N_1=1)}(\cdot)$), the $T = 7$ token indices from 0 to 6 are mapped to $\left\lceil \frac{T+N-N_1}{N} \right\rceil = 3$ bundle indices (0 to 2), reducing the input diversity of f_{pe} when its inputs are bundle indices rather than token indices. (b) Additionally applying the **slide** operation introduces $N = 3$ variants of the mapping function, each with a distinct N_1 value. Progressively varying N_1 (from 1 to 3) produces bundles with sliding boundaries across successive bundling procedures (the thick black line segments sliding rightward).

(2) The slide operation. However, applying the bundle operation alone cannot fully resolve the spatial disorder problem, as it eliminates positional distinctions among tokens within the same bundle, thereby weakening fine-grained spatial discrimination. To tackle this issue, a **slide** operation is further integrated to recover intra-bundle positional distinctions, with the key idea of using *sliding bundle boundaries*, as elaborated below.

Specifically, given the bundle size N for each middle bundle, the slide operation constructs N variants of the mapping function $\phi_{bundle}(\cdot)$, denoted as $\{\phi_{bundle}^{(N_1=1)}(\cdot), \ldots, \phi_{bundle}^{(N_1=N)}(\cdot)\}$, where each variant adopts a distinct first-bundle size N_1. As illustrated in Fig. 3(b), constructed in this way, these variants pro-

duce bundles whose boundaries progressively slide as N_1 varies from 1 to N. Importantly, with such sliding, by collectively considering all N mapping functions rather than any single one in isolation, the desired intra-bundle position distinctions can be achieved among tokens. For example, as shown in the red ellipse of Fig. 3(b), the token with index 4 is distinctively (uniquely) represented by the set of $N = 3$ bundle indices $\{2, 1, 1\}$ derived respectively from the $N = 3$ mapping functions, and no any other token corresponds to this same set. We also prove in Supplementary that this representation approach always uniquely represents every token index. The above means that, the collective use of all mapping functions constructed in the slide operation ensures clear positional distinctions among all tokens. Then, as the final step of the slide operation, to realize this collective use in a training-free manner, we compute:

$$\mathbf{c}_{\mathbf{i},\mathbf{j}}^{\mathbf{SPA}} = \frac{1}{N}\sum_{n=1}^{N} g\Big(\mathbf{x_i}, \mathbf{x_j}, f_{pe}\Big(\phi_{bundle}^{(N_1=n)}(i), \phi_{bundle}^{(N_1=n)}(j)\Big)\Big), \tag{6}$$

where $\mathbf{c}_{\mathbf{i},\mathbf{j}}^{\mathbf{SPA}}$ (replacing $\mathbf{c}_{\mathbf{i},\mathbf{j}}$ in Eq. 3) denotes the contribution of token j to the attention output at position i after incorporating SPA into our framework. Intuitively, Eq. 6 applies $g(\cdot)$ multiple times in parallel, once under each mapping function, and then averages their corresponding results. In the n-th application of $g(\cdot)$, the only modification introduced by SPA relative to the off-the-shelf DiT model is that the token indices i and j are replaced by their corresponding bundle indices $\phi_{\text{bundle}}^{(N_1=n)}(i)$ and $\phi_{\text{bundle}}^{(N_1=n)}(j)$ before being passed to the positional function f_{pe}, while all other parts remain unchanged. The mapping from token indices to bundle indices is illustrated in Fig. 3.

In summary, to tackle the *spatial disorder* problem, SPA simply needs to replace $\mathbf{c}_{\mathbf{i},\mathbf{j}}$ with $\mathbf{c}_{\mathbf{i},\mathbf{j}}^{\mathbf{SPA}}$ in the attention computation of off-the-shelf DiT models at high resolutions. It is worth noting that, since the N token-to-bundle index mapping procedures in Eq. 6 can be executed in parallel, SPA generally incurs only a minor increase in overall generation time (e.g., from 203 to 212 seconds on average per 4K (4096×4096) image generation, as reported in Tab. 2).

3.2 Head-adaptive Attention Pruning (HAP)

Above, we handle the *spatial disorder* challenge. Yet, as illustrated in Fig. 1, the *long generation time* problem still limits the practicality of adapting off-the-shelf DiT models for high-resolution synthesis. To address this problem, the underlying cause is first profiled. As shown in Fig. 2, the profiling reveals that, when off-the-shelf DiT models are directly applied to higher testing resolutions, *multi-head attention* computations rapidly dominate the overall generation time due to the quadratic complexity of attention. This profiling motivates a natural strategy for accelerating DiT's high-resolution generation process: focus on multi-head attention as the *dominant cause* of the *long generation time* problem and prune "redundant" attention computations that contribute little to final generation quality.

To realize this idea, we first examine whether existing vision transformer pruning techniques in image can be well-suited to perform "redundant" attention

pruning in our training-free text-to-high-resolution image generation problem. We however find them ill-suited. Many existing pruning methods [36, 51] rely on additional training, which directly contradicts the training-free constraint of our problem. To circumvent this issue, recent works [27, 43] introduce local-window-based attention, which prunes computations by focusing attention to local regions (windows), enabling reducing runtime in a training-free manner. Yet, these methods remain ill-suited when facing our problem, i.e., when handling high resolutions, in which long-range interactions are essential for maintaining generation fidelity [23], local-window-based approaches [27] typically still require additional training to prevent noticeable quality degradation, again conflicting with our training-free constraint.

Given these incompatibilities, addressing the *long generation time* of off-the-shelf DiT models at high resolutions calls for inspiration beyond conventional vision pruning. In particular, motivated by recent progress in attention pruning for long-context text generation [44], we adapt this idea to the training-free text-to-high-resolution image generation setting. Specifically, we integrate an attention pruning component into HRDiT, termed *Head-adaptive Attention Pruning* (HAP), which enables effective attention pruning in our context without additional training. Overall, HAP builds on the local-window-based paradigm to take advantage of its potential training-free applicability. Meanwhile, HAP seeks to further address the paradigm's tendency to cause noticeable generation quality degradation at high resolutions when equipped with no additional training.

For HAP to achieve the latter, we draw on a key observation from prior works [5, 39, 44]: once trained, different heads in a Transformer's multi-head attention tend to assume distinct yet relatively stable roles, each focusing on a particular *attention scope*. Specifically, for a given head, when computing the attention output at position i, only tokens within a certain scope around that position contribute meaningfully, and this scope may vary across heads rather than being uniformly local. For example, some heads specialize in long-range interactions and operate over broad scopes, whereas others focus on local structures within narrow neighborhoods. Building on this inspiration, we further validate that a similar head-dependent attention-scope pattern also emerges in our high-resolution image generation setting.

Motivated by the above, a natural way to prune redundant attention computations when scaling off-the-shelf DiT models to high-resolution generation is hypothesized to be the adaptive allocation of attention windows across heads according to each head's inherent attention scope, rather than the uniform application of identical local windows to all heads as in conventional local-window-based strategies. Ideally, such head-adaptive window allocation allows more accurate identification of truly redundant computations, thereby enabling attention pruning that preserves high generation quality while substantially improving efficiency. However, determining appropriate attention scopes for each head in practice remains non-trivial. To address this, inspired by [44], a two-step preparatory procedure is adopted in HAP to derive head-specific scopes from the off-the-shelf DiT model before inference. Once derived, these head-specific scopes can then

be applied as per-head attention windows during inference, enabling substantial runtime reduction while preserving generation fidelity. The two preparatory steps adopted in HAP are detailed below.

Preparatory Step 1: appropriateness quantification for each candidate scope. To determine an appropriate attention scope for each head, the effect of each candidate scope on computational cost and generation quality is first quantified. Formally, let N_{head} denote the total number of attention heads in an off-the-shelf DiT model, and assign each head N_{scope} candidate scopes, where N_{scope} is a hyperparameter. For each candidate scope $n_{\text{scope}} \in \{1, \ldots, N_{\text{scope}}\}$ of each head $n_{\text{head}} \in \{1, \ldots, N_{\text{head}}\}$, two indicators are defined: $I_c(n_{\text{head}}, n_{\text{scope}})$, representing the computational cost of this head under the given scope, and $I_q(n_{\text{head}}, n_{\text{scope}})$, representing the degradation in generation quality when this head operates under the given scope instead of the full attention scope covering all T tokens. For simplicity, given a head with T input tokens, its N_{scope} candidate scopes are uniformly defined to span all possible coverage ratios of these tokens. Specifically, under the n_{scope}-th candidate, the attention output at each position is computed by summing only the contributions from the $\frac{n_{\text{scope}}}{N_{\text{scope}}} \times T$ tokens neighboring that position, with all T tokens being covered only when $n_{\text{scope}} = N_{\text{scope}}$. Based on this setup, we next describe how $I_c(n_{\text{head}}, n_{\text{scope}})$ and $I_q(n_{\text{head}}, n_{\text{scope}})$ are computed.

For $I_c(n_{\text{head}}, n_{\text{scope}})$, note that the computational complexity of producing the attention output at each position scales linearly with the number of tokens participating in that position's attention computation. Accordingly, the total attention computational cost for the n_{head}-th head under the n_{scope}-th scope, i.e., $I_c(n_{\text{head}}, n_{\text{scope}})$, can be simply quantified by the total number of participating tokens involved in its attention computations, aggregated over all positions. Deriving $I_q(n_{\text{head}}, n_{\text{scope}})$, however, is more difficult. Naively, a brute-force estimation of $I_q(n_{\text{head}}, n_{\text{scope}})$ across all N_{head} heads and N_{scope} candidate scopes could require at least $N_{\text{head}} \times (N_{\text{scope}} - 1)$ DiT model passes, each followed by a loss computation on the generated image to evaluate its impact on generation quality (see Supplementary for details). This can make the preparatory stage by itself already computationally prohibitive. To accelerate the above brute-force estimation procedure, a natural attempt shall be to use losses computed on intermediate features as efficient proxies for losses computed on the final generated images [45]. However, we find that, at high resolutions, such intermediate signals can correlate poorly with final image quality, making them unreliable estimators of $I_q(n_{\text{head}}, n_{\text{scope}})$. To overcome the limitations of the above strategies, inspired by [44], HAP adopts its efficient estimation strategy, where careful mathematical derivations show that a single model pass under the full attention scope suffices to *efficiently yet reliably* estimate $I_q(n_{\text{head}}, n_{\text{scope}})$ across all heads and candidate scopes. Formally, under this estimation approach, by leveraging a Taylor expansion and accounting for the softmax operation in DiT attention, $I_q(n_{\text{head}}, n_{\text{scope}})$

can be estimated using only a single model pass, as

$$I_q(n_{\text{head}}, n_{\text{scope}}) \approx \sum_{(u,v)\in S_{\text{omit}}^{(n_{\text{scope}})}} \Big(\frac{\partial L}{\partial A(u,v)} \cdot \big(- A(u,v) \big) \quad + \sum_{w\in\{1,\dots,T\}\setminus\{v\}} \frac{\partial L}{\partial A(u,w)} \cdot \frac{A(u,w)\cdot A(u,v)}{1 - A(u,v)} \Big), \tag{7}$$

where L is the loss computed on the image generated from this single model pass, $A \in \mathbb{R}^{T\times T}$ is the attention score matrix in this model pass corresponding to the n_{head}-th head, and $A(u,v)$ denotes its (u,v)-th element. Here, $S_{\text{omit}}^{(n_{\text{scope}})}$ denotes the set of index pairs (u,v) that fall outside the n_{scope}-th attention scope, i.e., pairs for which when applying the n_{scope}-th attention scope and computing the attention output at position u, the contribution of token v (i.e., $\mathbf{c}_{\mathbf{u},\mathbf{v}}^{\mathbf{SPA}}$) is omitted. In practice, to stabilize the estimation, I_q is further averaged over different input prompts rather than estimated from a single input prompt. Despite this, computing $I_q(n_{\text{head}}, n_{\text{scope}})$ remains lightweight, taking only less than an hour in total across all heads and candidate scopes during the one-time preparatory stage (see Tab. 4 in Supplementary).

Preparatory Step 2: appropriate scope determination. At this point, quantitative estimates have been obtained for how different candidate scopes, when assigned to each head in DiT, affect both computational cost and generation quality. The work left is then to assign an appropriate scope to each head such that the collective scope assignment achieves the best trade-off between efficiency and generation quality.

Formally, let $r_c \in [0,1]$, as a hyperparameter, denote the target ratio of total computational cost for all multi-head attention operations, measured relative to that under full attention (i.e., $r_c = 1$ corresponds to using full attention). The problem of determining the optimal collective scope assignment can then be formulated as the optimization problem of minimizing the total generation quality degradation across all heads under this computational constraint:

$$\begin{aligned} &\arg\min_{(n_{\text{scope}}^{(1)},\dots,n_{\text{scope}}^{(N_{\text{head}})})} \sum_{n_{\text{head}}=1}^{N_{\text{head}}} I_q(n_{\text{head}}, n_{\text{scope}}^{(n_{\text{head}})}) \\ &\text{s.t.} \sum_{n_{\text{head}}=1}^{N_{\text{head}}} I_c(n_{\text{head}}, n_{\text{scope}}^{(n_{\text{head}})}) \le r_c \times \sum_{n_{\text{head}}=1}^{N_{\text{head}}} I_c(n_{\text{head}}, N_{\text{scope}}). \end{aligned} \tag{8}$$

where $n_{\text{scope}}^{(n_{\text{head}})}$ denotes the n_{scope}-th candidate scope of the n_{head}-th head. Though this optimization problem appears complex, following MoA [44], it can be reformulated as an integer programming problem, enabling Eq. 8 to be efficiently solved using linear solvers (e.g., Gurobi [10]) within minutes.

Through the above two preparatory steps, we obtain a collective scope assignment that can substantially enhance generation efficiency while maintaining high generation quality, requiring only a few DiT passes and a single call to an existing linear solver. Once derived, this assignment can be precompiled into GPU kernels (e.g., via FlashInfer [42]). During inference, the precompiled as-

Table 1: Results on 2K and 4K image generation. Notably, by their design, I-Max is applicable to FLUX but not to Stable Diffusion 3 (SD3), whereas FreCaS is applicable to SD3 but not to FLUX.

Resolution	Model Type	Method	FID ↓	FID_p ↓	KID ↓	KID_p ↓	CLIP Score ↑	Latency(s) ↓
2k	U-Net backbone	SDXL [29] + FreeScale [32]	73.23	65.04	0.0093	0.0132	30.57	26
	DiT-based backbone	SD3 [8] (*Direct*)	81.34	70.88	0.0215	0.0194	30.23	17
		SD3 [8] + DemoFusion [6]	72.21	59.24	0.0132	0.0125	31.40	35
		SD3 [8] + DiffuseHigh [18]	70.57	57.39	0.0119	0.0112	31.59	20
		SD3 [8] + FreCaS [49]	73.27	63.55	0.0139	0.0132	31.41	16
		SD3 [8] + HiFlow [3]	70.49	57.35	0.0126	0.0110	31.62	18
		SD3 [8] + Ours	**67.90**	**52.59**	**0.0081**	**0.0073**	**31.81**	**11**
		FLUX [20] (*Direct*)	73.96	69.56	0.0173	0.0164	29.63	47
		FLUX [20] + DemoFusion [6]	68.95	60.53	0.0127	0.0120	30.83	94
		FLUX [20] + DiffuseHigh [18]	67.40	58.63	0.0115	0.0108	31.01	52
		FLUX [20] + I-Max [7]	68.37	56.96	0.0121	0.0106	30.67	85
		FLUX [20] + HiFlow [3]	67.45	56.79	0.0117	0.0105	30.73	49
		FLUX [20] + Ours	**64.42**	**51.55**	**0.0074**	**0.0067**	**31.27**	**35**
4k	U-Net backbone	SDXL [29] + FreeScale [32]	74.85	68.89	0.0112	0.0157	30.49	248
	DiT-based backbone	SD3 [8] (*Direct*)	84.29	74.45	0.0243	0.0231	30.18	108
		SD3 [8] + DemoFusion [6]	73.39	60.75	0.0141	0.0133	31.16	229
		SD3 [8] + DiffuseHigh [18]	71.88	58.76	0.0130	0.0126	31.35	115
		SD3 [8] + FreCaS [49]	79.30	72.82	0.0226	0.0213	30.32	86
		SD3 [8] + HiFlow [3]	71.39	58.40	0.0133	0.0118	31.59	112
		SD3 [8] + Ours	**68.63**	**53.04**	**0.0089**	**0.0079**	**31.79**	**58**
		FLUX [20] (*Direct*)	75.68	72.73	0.0192	0.0188	29.57	203
		FLUX [20] + DemoFusion [6]	71.18	62.38	0.0150	0.0142	30.48	649
		FLUX [20] + DiffuseHigh [18]	69.23	59.13	0.0132	0.0125	30.97	215
		FLUX [20] + I-Max [7]	70.11	59.83	0.0142	0.0133	30.52	484
		FLUX [20] + HiFlow [3]	68.36	57.43	0.0124	0.0114	30.62	209
		FLUX [20] + Ours	**64.91**	**51.84**	**0.0078**	**0.0074**	**31.17**	**116**

signment can then be applied to eliminate redundant attention computations in a cost-free manner, thereby effectively addressing the *long generation time* issue encountered by off-the-shelf DiT models when adapted to high resolutions.

3.3 Overall Testing

For testing at the high resolution, before testing, HRDiT first performs the one-time preparatory stage described in Sec. 3.2, to determine an appropriate collective scope assignment plan across different attention heads. For per-input generation (testing), HRDiT adopts the determined scope plan to enhance efficiency (handling the *long generation time* problem). Notably, in both the one-time preparatory stage and the per-input generation process, HRDiT computes the per-token contribution $\mathbf{c}_{\mathbf{i,j}}^{\mathbf{SPA}}$ via Eq. 6 to address the *spatial disorder* issue.

4 Experiments

4.1 Experimental Settings & Implementation Details.

To evaluate the efficacy of our HRDiT framework, we apply it to several popular off-the-shelf DiT models, including Stable Diffusion 3 (SD3) [8] and FLUX [20]. Following the settings in [3, 32], we conduct experiments on both 2K (2048×2048) and 4K (4096×4096) image generation. For text prompts used in our

evaluation, following [6,18], we randomly sample 1,000 captions from the LAION-5B dataset [34]. To assess image quality, following [18], we report evaluation metrics including Fréchet Inception Distance (FID), Kernel Inception Distance (KID), their patch-based counterparts (FID_p and KID_p), and the CLIP score. In addition, we report the per-generation runtime to evaluate efficiency. We set the hyperparameters $N_{\text{scope}} = 50$, $r_c = 0.1$. We conduct our experiments on NVIDIA H200 GPU. More experimental settings and implementation details are in Supplementary.

4.2 Main Results

In our experiments, we compare against three categories of methods: (1) *Direct*, which directly scales off-the-shelf DiT models to higher-resolution generation by passing them higher-resolution initial noises; (2) recent training-free text-to-high-resolution image generation methods applicable to DiT models, including the DiT-tailored method I-Max [7], and methods that focus on modifying the diffusion process of diffusion models [3, 6, 18, 49]; (3) the state-of-the-art training-free text-to-high-resolution image generation method that is tailored to U-Net-based diffusion models (i.e., FreeScale [32]). For FreeScale, following its original setting, we apply it to the U-Net-based model Stable Diffusion XL (SDXL) [29]. We present the main quantitative results in Tab. 1. As shown, across both 2K and 4K generation tasks, our method consistently outperforms all baselines on all five metrics assessing generation quality, showing its strong ability to produce high-quality, high-resolution images. Moreover, when applied to either Stable Diffusion 3 (SD3) or FLUX, our method achieves substantial efficiency gains over all existing approaches under the same DiT backbone, validating its efficacy in addressing the long generation time challenge. Qualitative comparisons in Fig. 4 further confirm that existing methods often yield high-resolution images with spatial disorder (marked by red boxes), while our method preserves strong spatial coherence, further showing its efficacy.

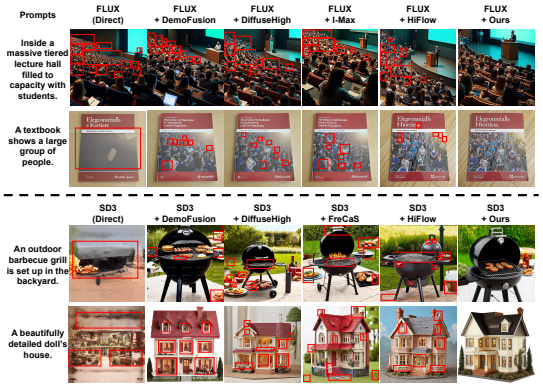

Fig. 4: Qualitative results of the 4K images generated by our method and by other methods (zoom-in for better view). More qualitative results are in Supplementary.

Table 3: Additional results on 8K image generation.

Resolution	Model Type	Method	FID ↓	FID_p ↓	KID ↓	KID_p ↓	CLIP Score ↑	Latency(s) ↓
8k	U-Net backbone	SDXL [29] + FreeScale [32]	76.51	70.50	0.0123	0.0176	30.44	3958
	DiT-based backbone	SD3 [8] (*Direct*)	86.75	77.68	0.0258	0.0251	30.01	822
		SD3 [8] + DemoFusion [6]	76.07	65.22	0.0207	0.0206	30.96	2720
		SD3 [8] + DiffuseHigh [18]	73.30	60.97	0.0166	0.0161	31.19	889
		SD3 [8] + FreCaS [49]	81.19	76.06	0.0241	0.0232	30.16	635
		SD3 [8] + HiFlow [3]	73.72	61.77	0.0179	0.0173	31.44	832
		SD3 [8] + Ours	**69.51**	**54.47**	**0.0096**	**0.0083**	**31.71**	**454**
		FLUX [20] (*Direct*)	78.47	74.59	0.0212	0.0201	29.34	1708
		FLUX [20] + DemoFusion [6]	75.72	68.71	0.0201	0.0187	30.03	5749
		FLUX [20] + DiffuseHigh [18]	71.02	63.87	0.0169	0.0158	30.69	1835
		FLUX [20] + I-Max [7]	72.54	65.34	0.0196	0.0183	30.23	4488
		FLUX [20] + HiFlow [3]	70.67	60.79	0.0142	0.0136	30.43	1721
		FLUX [20] + Ours	**65.73**	**53.55**	**0.0089**	**0.0084**	**31.04**	**827**

4.3 Additional Experiments and Ablation Studies

Additional experiments on 8K (8192×8192) generation. To assess scalability, in addition to 2K and 4K, we further evaluate our framework on 8K generation. As shown in Tab. 3, our method consistently achieves highest generation quality and fastest generation speed across both Stable Diffusion 3 and FLUX, further showing its efficacy at even higher resolutions. Qualitative 8K results are in Supplementary.

Impact of key components in the HRDiT framework. In our framework (**Ours (full)**), we integrate two key components: SPA and HAP. To assess the contribution of each component, we test two variants, on 4K image generation task using FLUX as the generation model. In the first variant (**Ours (w/o SPA)**), we replace $\mathbf{c}_{\mathbf{i,j}}^{\mathbf{SPA}}$ with the original $\mathbf{c_{i,j}}$ in Eq. 3 as the per-token contribution. In the second variant (**Ours (w/o HAP)**), we do not apply the determined attention scope assignment plan to prune "redundant" attention computations. Instead, we direct employ the full attentions scope for all heads. As shown in Tab. 2, our full framework significantly outperforms the first variant in generation quality. Meanwhile, it also achieves substantially faster generation speed than the second variant, with negligible degradation in generation quality. Together, these results validate the efficacy of both key components in our framework. **More ablation studies are in Supplementary.**

Table 2: Evaluation on key components in HRDiT.

Method	FID ↓	FID_p ↓	KID ↓	KID_p ↓	CLIP Score ↑	Latency(s) ↓
FLUX [20] (*Direct*)	75.68	72.73	0.0192	0.0188	29.57	203
Ours (w/o SPA)	75.85	72.83	0.0195	0.0192	29.55	110
Ours (w/o HAP)	64.82	51.81	0.0076	0.0074	31.19	212
Ours (full)	64.91	51.84	0.0078	0.0074	31.17	116

5 Conclusion

In this paper, we have proposed HRDiT, a novel training-free framework for text-to-high-resolution image generation. HRDiT integrates two components, SPA and HAP, to address two key challenges, spatial disorder and long generation time, thereby enabling off-the-shelf DiT models to be effectively adapted for high-resolution synthesis. Extensive experiments show HRDiT's efficacy.

References

1. Bai, Y., Li, H., Huang, Q.: Positional encoding field. arXiv preprint arXiv:2510.20385 (2025)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: International Conference on Machine Learning. pp. 1737–1752. PMLR (2023)
3. Bu, J., Ling, P., Zhou, Y., Zhang, P., Wu, T., Dong, X., Zang, Y., Cao, Y., Lin, D., Wang, J.: Hiflow: Training-free high-resolution image generation with flow-aligned guidance. arXiv preprint arXiv:2504.06232 (2025)
4. Cao, B., Ye, J., Wei, Y., Shan, H.: Ap-ldm: Attentive and progressive latent diffusion model for training-free high-resolution image generation. arXiv preprint arXiv:2410.06055 (2024)
5. Chen, Y., Zhang, P., Su, R., Ding, H., Stoica, I., Liu, Z., Zhang, H.: Fast video generation with sliding tile attention. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=U74MOXPEJd
6. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: Demofusion: Democratising high-resolution image generation with no $$$. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6159–6168 (2024)
7. Du, R., Liu, D., Zhuo, L., Qi, Q., Li, H., Ma, Z., Gao, P.: I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. arXiv preprint arXiv:2410.07536 (2024)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
9. Guo, L., He, Y., Chen, H., Xia, M., Cun, X., Wang, Y., Huang, S., Zhang, Y., Wang, X., Chen, Q., et al.: Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In: European conference on computer vision. pp. 39–55. Springer (2024)
10. Gurobi Optimization, L.: Gurobi Optimizer Reference Manual (2025), https://www.gurobi.com, last accessed 2026/06/29
11. Han, C., Wang, Q., Peng, H., Xiong, W., Chen, Y., Ji, H., Wang, S.: Lm-infinite: Zero-shot extreme length generalization for large language models. arXiv preprint arXiv:2308.16137 (2023)
12. Hartmann, J., Exner, Y., Domdey, S.: The power of generative marketing: Can generative ai create superhuman visual marketing content? International Journal of Research in Marketing **42**(1), 13–31 (2025)
13. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In: The Twelfth International Conference on Learning Representations (2023)
14. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. In: International Conference on Machine Learning. pp. 13213–13232. PMLR (2023)
15. Huang, L., Fang, R., Zhang, A., Song, G., Liu, S., Liu, Y., Li, H.: Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In: European conference on computer vision. pp. 196–212. Springer (2024)
16. Issachar, N., Yariv, G., Benaim, S., Adi, Y., Lischinski, D., Fattal, R.: Dype: Dynamic position extrapolation for ultra high resolution diffusion. arXiv preprint arXiv:2510.20766 (2025)

17. Jin, H., Han, X., Yang, J., Jiang, Z., Liu, Z., Chang, C.Y., Chen, H., Hu, X.: LLM maybe longLM: Selfextend LLM context window without tuning. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=nkOMLBIiI7
18. Kim, Y., Hwang, G., Zhang, J., Park, E.: Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 4338–4346 (2025)
19. Koh, S., Cha, S., Oh, H., Lee, K., Kim, D.J.: Scalediff: Higher-resolution image synthesis via efficient and model-agnostic diffusion. arXiv preprint arXiv:2510.25818 (2025)
20. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024), last accessed 2026/06/29
21. Lai, H.P., Nguyen, P., Tran, A.: Pixelrush: Ultra-fast, training-free high-resolution image generation via one-step diffusion. arXiv preprint arXiv:2602.12769 (2026)
22. Lee, Y., Yoon, T., Sung, M.: Groundit: Grounding diffusion transformers via noisy patch transplantation. Advances in Neural Information Processing Systems **37**, 58610–58636 (2024)
23. Leroy, V., Revaud, J., Lucas, T., Weinzaepfel, P.: Win-win: Training high-resolution vision transformers from two windows. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=N23A4ybMJr
24. Li, Z., Rahmani, H., Ke, Q., Liu, J.: Longdiff: Training-free long video generation in one go. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17789–17798 (2025)
25. Lin, Z., Lin, M., Zhao, M., Ji, R.: Accdiffusion: An accurate method for higher-resolution image generation. In: European Conference on Computer Vision. pp. 38–53. Springer (2024)
26. Liu, C., Hou, L., Zheng, M., Tao, X., Wan, P., Zhang, D., Gai, K.: Boosting resolution generalization of diffusion transformers with randomized positional encodings. arXiv preprint arXiv:2503.18719 (2025)
27. Liu, S., Tan, Z., Wang, X.: Clear: Conv-like linearization revs pre-trained diffusion transformers up. arXiv preprint arXiv:2412.16112 (2024)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
30. Qin, J.: How does text-to-image ai affect indie game designers and artists? Journal of Innovation and Development **5**, 107–111 (12 2023). https://doi.org/10.54097/f7of9f8k
31. Qiu, H., Yu, N., Huang, Z., Debevec, P., Liu, Z.: Cinescale: Free lunch in high-resolution cinematic visual generation. arXiv preprint arXiv:2508.15774 (2025)
32. Qiu, H., Zhang, S., Wei, Y., Chu, R., Yuan, H., Wang, X., Zhang, Y., Liu, Z.: Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. arXiv preprint arXiv:2412.09626 (2024)
33. Ren, J., Li, W., Chen, H., Pei, R., Shao, B., Guo, Y., Peng, L., Song, F., Zhu, L.: Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. Advances in Neural Information Processing Systems **37**, 111131–111171 (2024)
34. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-

scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
35. Shen, T., Yu, J., Zhou, D., Li, D., Barsoum, E.: E-mmdit: Revisiting multimodal diffusion transformer design for fast image synthesis under limited resources. arXiv preprint arXiv:2510.27135 (2025)
36. Tang, Y., Han, K., Wang, Y., Xu, C., Guo, J., Xu, C., Tao, D.: Patch slimming for efficient vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12165–12174 (2022)
37. Vontobel, T., Sadat, S., Salehi, F., Weber, R.: Hiwave: Training-free high-resolution image generation via wavelet-based diffusion sampling. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
38. Wu, H., Shen, S., Hu, Q., Zhang, X., Zhang, Y., Wang, Y.: Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3944–3953. IEEE (2025)
39. Wu, W., Wang, Y., Xiao, G., Peng, H., Fu, Y.: Retrieval head mechanistically explains long-context factuality. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=EytBpUGB1Z`
40. Yang, H., Bulat, A., Hadji, I., Pham, H.X., Zhu, X., Tzimiropoulos, G., Martinez, B.: Fam diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2459–2468 (2025)
41. Yang, Z., Shen, G., Hou, L., Liu, M., Wang, L., Tao, X., Wan, P., Zhang, D., Chen, Y.C.: Rectifiedhr: Enable efficient high-resolution image generation via energy rectification. arXiv e-prints pp. arXiv–2503 (2025)
42. Ye, Z., Chen, L., Lai, R., Lin, W., Zhang, Y., Wang, S., Chen, T., Kasikci, B., Grover, V., Krishnamurthy, A., Ceze, L.: Flashinfer: Efficient and customizable attention engine for LLM inference serving. In: Eighth Conference on Machine Learning and Systems (2025), `https://openreview.net/forum?id=RXPofAsL8F`
43. Yuan, Z., Zhang, H., Pu, L., Ning, X., Zhang, L., Zhao, T., Yan, S., Dai, G., Wang, Y.: Ditfastattn: Attention compression for diffusion transformer models. Advances in Neural Information Processing Systems **37**, 1196–1219 (2024)
44. Zhang, G., Fu, T., Huang, H., Ning, X., , Chen, B., Wu, T., Wang, H., Huang, Z., Li, S., Yan, S., et al.: Moa: Mixture of sparse attention for automatic large language model compression. arXiv preprint arXiv:2406.14909 (2024)
45. Zhang, H., Su, R., Yuan, Z., Chen, P., Shen, M., Fan, Y., Yan, S., Dai, G., Wang, Y.: Ditfastattnv2: Head-wise attention compression for multi-modality diffusion transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16399–16409 (2025)
46. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23464–23473 (2025)
47. Zhang, S., Chen, Z., Zhao, Z., Chen, Y., Tang, Y., Liang, J.: Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In: European Conference on Computer Vision. pp. 145–161. Springer (2024)
48. Zhang, S., Liang, S., Tan, Y., Chen, Z., Li, L., Wu, G., Chen, Y., Li, S., Zhao, Z., Chen, C., et al.: Ledit: Your length-extrapolatable diffusion transformer without positional encoding. arXiv preprint arXiv:2503.04344 (2025)

49. Zhang, Z., Li, R., Zhang, L.: Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=TsBDfe8Ra5
50. Zhao, M., Yan, B., Yang, X., Zhu, H., Zhang, J., Liu, S., Li, C., Zhu, J.: Ultraimage: Rethinking resolution extrapolation in image diffusion transformers. arXiv preprint arXiv:2512.04504 (2025)
51. Zheng, C., Zhang, K., Yang, Z., Tan, W., Xiao, J., Ren, Y., Pu, S., et al.: Savit: Structure-aware vision transformer pruning via collaborative optimization. Advances in Neural Information Processing Systems **35**, 9010–9023 (2022)