UltraImageGen: Efficient Ultra-High-Resolution Image Generation with Hierarchical Local Attention

Yuyao Zhang and Yu-Wing Tai

Dartmouth College, USA

https://github.com/PeterYYZhang/UltraImageGen/

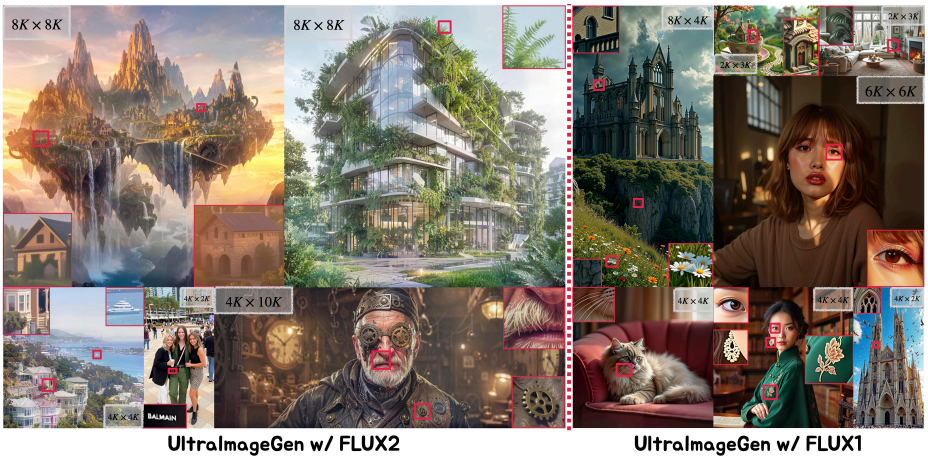

Fig. 1: Ultra-high-resolution images generated by UltraImageGen with different base models at $8K \times 8K$, $4K \times 10K$, $8K \times 4K$, $6K \times 6K$, $4K \times 4K$, $4K \times 2K$, $2K \times 3K$. Zoomed-in regions highlight the fine-grained details (See supplementary materials for original PDF). Notice that the pretrained FLUX 1 and FLUX 2 are only capable of generating sub 2MP ($< 1K \times 2K$) and sub 4MP ($< 2K \times 2K$) images.

Abstract. Ultra-high-resolution image generation is increasingly vital for applications requiring fine-grained textures and global structural fidelity, yet state-of-the-art text-to-image diffusion models such as FLUX and SD3 remain confined to sub 2MP ($< 1K \times 2K$) resolutions due to the quadratic complexity of attention mechanisms and the scarcity of high-quality high-resolution training data. We present **UltraImageGen**, a novel framework that introduces hierarchical local attention with low-resolution global guidance, enabling efficient, scalable, and semantically coherent image synthesis at ultra-high resolutions. Specifically, high-resolution latents are divided into hardware aligned fixed-size local windows to reduce attention complexity from quadratic to near-linear, while a low-resolution latent with scaled positional embeddings injects global semantics as an anchor. A lightweight LoRA adaptation bridges global and local pathways during denoising, ensuring consistency across

structure and detail. To maximize efficiency and achieve scalable ultra-high-resolution generation, we repermute token sequence in window-first order, so that the GPU-friendly dense local blocks in attention calculation equals to the fixed-size local window in 2D regardless of resolution. Together UltraImageGen reliably scales pretrained models to resolutions higher than $8K$ with more than $10\times$ speed up and significantly lower memory usage. Extensive experiments demonstrate that UltraImageGen achieves superior quality while maintaining computational efficiency, establishing a practical paradigm for advancing ultra-high-resolution image generation. Our project page and code are available here.

Keywords: Efficient Attention, Ultra-High-Resolution Generation, Diffusion Models, Scalable Diffusion Paradigm.

1 Introduction

Ultra-high-resolution image generation is becoming increasingly crucial for both creative and practical applications. From digital art and advertising to scientific visualization and virtual environments, users demand outputs with fine-grained textures, faithful structures, and seamless details at $4K$ resolution or higher. However, despite the rapid progress of text-to-image (T2I) diffusion models [2,5,7,8,12,29,30,38,39,44,54,55,58], most state-of-the-art systems remain confined to resolutions below $1K \times 1K$. This limitation fundamentally restricts their usability in scenarios where ultra-high fidelity is paramount.

The root of this challenge lies in two interdependent bottlenecks. First, scaling model capacity and datasets to higher resolutions requires tremendous resources, as high-quality $2K$–$8K$ training data are scarce and expensive to curate. Second, attention-based diffusion architectures suffer from quadratic growth in token complexity with respect to image resolution, making naive scaling computationally prohibitive. For example, generating a $4K$ image would require tens of thousands of tokens, leading to impractical memory and runtime costs. As a result, existing approaches either resort to costly retraining on synthetic high-resolution datasets [22, 33, 42, 50, 64, 75], or adopt training-free upscaling methods [4, 15, 18, 25, 26, 34, 40, 48, 52] that often compromise efficiency and stability. Consequently, the field still lacks a solution that delivers *both* ultra-high-resolution fidelity and computational scalability.

In this work, we introduce **UltraImageGen**, a novel framework that rethinks ultra-high-resolution synthesis through the lens of efficiency and scalability. Inspired by the way artists construct large-scale murals, we decompose the generation task hierarchically: local windows capture fine textures through lightweight attention, while a low-resolution guidance image preserves global semantic structure. Technically, UltraImageGen introduces three innovations. First, a *hierarchical local attention* mechanism restricts computation to fixed-size windows, reducing quadratic cost to near-linear scaling while yielding over $10GB$ memory savings and more than $10\times$ faster inference compared to dense attention. Second, a *global guidance pathway* employs a low-resolution latent with

scaled RoPE positional anchors to maintain long-range dependencies and ensure semantic coherence across windows. Third, a *parameter-efficient joint denoising framework* integrates global and local pathways via LoRA-adapted projections trained only on 256×256–$1K \times 1K$ data, enabling direct scaling to $4K$ or higher without requiring any native high-resolution training. These design choices translate into two key outcomes: (1) *scalable fidelity*, achieved by hierarchically fusing global composition with local detail to extend beyond pretrained resolutions, and (2) *computational efficiency*, enabled by reducing computation, memory, and runtime while reusing pretrained weights with lightweight adaptation.

Extensive experiments validate these contributions: our work reliably scales to $4K \times 4K$ and higher resolution synthesis with commodity training resolutions, achieving sharper textures, richer details, and stronger global coherence than existing approaches. Quantitatively, it surpasses dense attention baselines with over $10\times$ faster inference and lower memory usage, and it matches or outperforms state-of-the-art methods trained with native 4K data across FID, IS, and CLIP Score. Together, these results show UltraImageGen as a practical and effective solution for advancing ultra-high-resolution text-to-image generation.

We summarize our contributions as follows:

- We introduce a new paradigm for ultra-high-resolution image generation, which combines window-first local attention and low-resolution global anchors to achieve resolution agnostic synthesis far beyond pretraining limits, without requiring native high-resolution training data.
- Our framework entirely eliminates the need for high-resolution training data, instead leveraging positionally aligned low-resolution guidance to preserve global coherence.
- We demonstrate substantial efficiency improvements ($> 10\times$ speedup in $8K$, less memory usage) while achieving state-of-the-art quality at $4K, 8K$ and higher, offering a practical and generalizable solution for real-world ultra-high-resolution generation.

2 Related Work

Text-to-Image Generation. Recent years have witnessed rapid progress in text-to-image (T2I) generation [2, 3, 5, 7, 8, 12, 16, 29, 30, 38, 39, 44, 54, 55, 58, 73], where diffusion models now achieve near-photorealistic synthesis at resolutions up to $1K \times 1K$. The dominant paradigm relies on cross-attention mechanisms and DiT architecture [38], as seen in the PixArt series [6–8], or multi-modal diffusion transformers (MMDiT) such as FLUX.1.0-dev [2]. These architectures have established strong foundations for controllable and high-quality T2I synthesis. However, despite their success, existing methods struggle to scale beyond pretrained resolutions due to quadratic attention costs and the lack of large-scale high-resolution data. Our work builds directly on this line of research, but departs from the prevailing direction by focusing on resolution scalability and computational efficiency rather than retraining larger models.

High-resolution image synthesis. Real-world applications increasingly demand resolutions of $4K$ or higher, sparking methods in pushing beyond the $1K \times 1K$ barrier. Several works, including PixArt-Σ and SANA 1.5 [55], have achieved $4K$ synthesis through extensive high-resolution pretraining, while others [22, 33, 42, 50, 56, 60, 61, 64, 75] pursue fine-tuning or training-from-scratch approaches using curated datasets. For example, [64] used wavelet supervision to enhance detail clarity, and [61] proposed lightweight fine-tuning for adapting to higher resolutions. Although effective, these methods remain constrained by the scarcity of high-quality high-resolution data and substantial GPU requirements, and since the lack of high quality data as lower resolutions, the performance sometimes degrades and sensitive to aspect ratio [60, 64]. More recent training-free strategies [4, 15, 18, 25, 26, 34, 40, 48, 52, 71, 74] avoid data collection by leveraging pretrained models directly. While promising, these approaches often inherit significant runtime and memory overhead, limiting accessibility for broader use. Other recent methods [4, 13, 15, 20, 24, 27, 71, 74] further explore training-free or architecture-specific high-resolution synthesis. [15, 71] regularizes the denoising trajectory, while [4, 74] adopts a coarse-to-fine strategy to synthesize high-frequency content, but they either struggle with long-range consistency or require careful stage alignment. RAGSR [20] focuses on single-image super-resolution with region-text alignment, but its image self-attention still scales with full-resolution tokens. ΔConvFusion [13] replaces self-attention with convolutional modules through distillation, and FCDM [27] trains an attention-free diffusion backbone from scratch at low resolutions. UltraGen [24] targets video generation with a dual-branch hierarchical architecture, and [72] focuses on ultra-high-resolution image editing. In contrast, UltraImageGen directly reuses pretrained backbones and combines hierarchical local-window attention with scaled-RoPE LR anchors, enabling stable and efficient ultra-high-resolution generation without native HR training data, multi-stage repair, or new architectures.

Attention Acceleration. As image and video resolutions increase, the quadratic complexity of attention becomes the dominant bottleneck. Works including the SANA series [54, 55, 76] leverage linear attention to reduce complexity. While such methods achieve satisfactory performance, the non-injective property and loss of attention spikiness [19, 37, 69] inherent in linear attention lead to confusion and inconsistent local details in real-world scenarios. For softmax attention, system-level optimizations such as Flash Attention [9, 10, 47] exploit GPU features for faster execution, while quantization [65, 66] and sparsity-based designs [11, 31] reduce computational load. Architectural innovations, such as LongFormer [1] and SwinFormer [35], employ local attention patterns, and more recent works [28, 32, 43, 53, 57, 59, 62, 63, 67, 70] propose block sparsification or compression strategies for diffusion transformers. Although these techniques yield noticeable acceleration, the gains remain insufficient for ultra-high-resolution synthesis, and compression often risks degrading fine-grained fidelity. Methods like [32, 43, 70] aim to achieve faster generation while preserving local image quality using local neighbor attention. However, none of them fully maintain long-range dependencies: CLEAR [32] and GRAT [43] require increasingly large

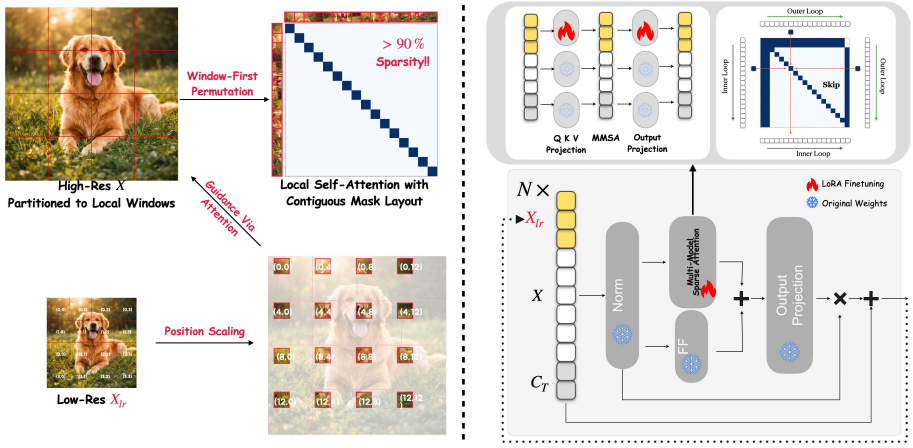

Fig. 2: Schematic of UltraImageGen's attention block modifications. The left column illustrates that high-resolution image latents X are partitioned into local windows (red) and flattened in window-first order that each window attends to itself. Simultaneously, a low-resolution guidance latent X_{lr} (yellow) provides global context to each window via position scaling. The right column shows the joint-denoising process and the attention kernel, that X, X_{lr} are processed together with LoRA applied on the X_{lr} part. The attention mask (upper right) enforces local and guidance-specific interactions and tile skipping in attention calculation to enable efficient and coherent ultra-high-resolution generation. No additional high-resolution training data are needed.

local windows/grouping as resolution increases, while STA [70] sacrifices some long-range consistency, also [32, 70] do not support ultra-high-resolution generation without other techniques like SDEdit [36]. In contrast, our method explicitly maintains long-range consistency at low resolution through hierarchical local-window attention, achieving stronger acceleration ($> 4\times$) and better detail preservation when scaling to 4K and higher, without relying on excessively large local windows or compromising global structure.

3 Method

Figure 2 illustrates the overall framework of UltraImageGen, highlighting the key modifications introduced in the attention blocks to support ultra-high-resolution generation. High-resolution image latents X are partitioned into local windows (red grids), where tokens attend only within their window to capture fine-grained detail efficiently. A low-resolution guidance latent X_{lr} (yellow), enhanced with scaled positional anchors, injects global semantics and preserves long-range consistency across windows. The attention mask is designed to enforce both local and guidance-specific interactions while also enabling tile-skipping to avoid unnecessary computation. Together, these mechanisms allow UltraImageGen to generate coherent and detail-rich $4K$ images with near-linear complexity and without requiring any native high-resolution training data. This schematic provides the foundation for the following technical components.

The Multi-Modal Diffusion Transformer (MMDiT) Preliminaries. The MMDiT employed in state-of-the-art models [2, 3, 5, 16], represents the current benchmark architecture for text-to-image generation. MMDiT processes two distinct token modalities: a noisy image token sequence $X \in \mathbb{R}^{N \times d}$ and a text token sequence $C_T \in \mathbb{R}^{M \times d}$. MMDiT processes image and text tokens through a unified Multi-Modal Attention (MMA) mechanism. Specifically, image and text tokens are concatenated as $[C_T; X]$ and processed via self-attention.

Within the MMA framework, spatial information is encoded using Rotary Position Embedding (RoPE) [49], which applies rotational transformations to capture relative positional relationships. This is mathematically represented as: $X_{i,j} = X_{i,j} \cdot \mathrm{Rot}(i, j)$, where $\mathrm{Rot}(i, j)$ denotes the rotation matrix corresponding to position (i, j) in the 2D spatial grid. The MMA mechanism $\mathrm{MMA}([C_T; X])$ is formally defined as:

$$\mathrm{Softmax}\left(\frac{([Q_T(C_T), Q(X)])\,([K_T(C_T), K(X)]^T)}{\sqrt{d}}\right)[V_T(C_T), V(X)]$$

where Q, Q_T, K, K_T, and V, V_T represent the query, key, and value projection-matrices applied to the position-encoded image and text tokens X, C_T. This formulation enables bidirectional attention across all token modalities.

Efficient Local Window Attention with Window-First Permuted Token Sequence. Transformer-based generative models rely on self-attention, whose computational and memory costs scale quadratically with the number of tokens. For an image of size $H \times W$, the number of spatial tokens is $N = H \cdot W$, leading to $O(N^2)$ complexity. Scaling from 1024×1024 (4096 tokens[1]) to 4096×4096 (65536 tokens) increases the cost by a factor of 256, which is prohibitive even with optimized kernels such as FlashAttention [9, 10, 47].

To overcome this limitation, we partition the high-resolution latent $X \in \mathbb{R}^{H \times W}$ into non-overlapping windows $x_i \in \mathbb{R}^{l \times l}$, with window size l bounded by the pretrained resolution (e.g., 1024). Self-attention is computed independently within each window, reducing complexity from $O(N^2)$ to $O\left(\lceil \frac{H}{l} \rceil \cdot \lceil \frac{W}{l} \rceil \cdot (l^2)^2\right) \approx O(N \cdot l^2)$. In practice, we set $l = 16$ (corresponding to 256×256 windows) to strike a balance between accuracy and efficiency: larger windows bring little additional benefit in quality, while smaller windows underutilize GPU kernels, which are optimized for tile sizes around 128×128[2]. An ablation study in Section 5 further analyzes this trade-off. With this design, scaling to $4K$ resolution ($N = 65536$ tokens) reduces the attention complexity from $O(65536^2)$ to $O(65536 \cdot 16^2)$, yielding a 256× reduction. To maintain boundary consistency, we also allow limited cross-window attention along adjacent regions. Moreover, because the per-window computation cost remains constant, total runtime scales

[1] A VAE downsampling factor of 16 yields $\frac{1024}{16}^2 = 4096$ tokens, similar scaling for other resolutions.

[2] With 16× VAE downsampling, 256×256 windows yield 16×16 tokens, creating attention matrices that efficiently utilize GPU kernels. Smaller window size produces tokens less than the granularity of 128.

nearly linearly with image resolution, making the generation process both efficient and memory-friendly via integration with the tiling strategy in FlashAttention [9, 10].

A key aspect of our approach is the **window-first token permutation**: we reorder the flattened image token sequence in a window-first fashion (instead of the default raster-scan order) so that each block in the attention map corresponds directly to a 256×256 spatial window in the original image. This ensures that every attention window operates within a fixed RoPE positional range of 256×256 which is within the relative range which the model was pretrained regardless of the image resolution, thereby enabling ultra-high-resolution generation without extrapolating beyond learned positional encodings.

Maintaining Global Semantics via Low-Resolution Guidance. Partitioning a high-resolution image into local windows risks creating discontinuities and losing global semantic coherence. The key insight is that global long-range dependencies primarily influence the overall structure and layout of the generated image, but have minimal effect on fine local details. We generate a low-resolution guidance image $X_{lr} \in \mathbb{R}^{h \times w}$ to provide global context. The corresponding position indices are denoted as (m, n), where $m \in \{0, 1, \ldots, h-1\}$ and $n \in \{0, 1, \ldots, w-1\}$. We define the scaling ratio as $\rho = \frac{H}{h}$ (empirically set to 4 for an optimal balance of performance and efficiency). We then scale the low-resolution image's position indices by this ratio, mapping them to $(\tilde{m}, \tilde{n}) = (\rho \cdot m, \rho \cdot n)$. This effectively projects X_{lr} to the same spatial scale as the high-resolution image X. Each token $X_{lr}[\tilde{m}, \tilde{n}]$ acts as an anchor point, providing contextual information. The resulting attention structure is as follows: each high-resolution window attends to its own local and adjacent tokens together with the corresponding scaled region in X_{lr}, while X_{lr} tokens attend globally among themselves. In concrete terms, we formulate the attention mask M by allowing the following interactions: text tokens attend to themselves and to X; tokens in X attend locally and both text and X_{lr}; and tokens in X_{lr} attend to each other. This design ensures semantic consistency across spatially distant regions. The framework naturally supports recursive generation: the high-resolution output can serve as low-resolution guidance for an even higher resolution, enabling arbitrarily large-scale synthesis with stable memory and computation.

Parameter-Efficient Joint Denoising. To integrate the low-resolution guidance X_{lr}, we concatenate it with the standard text–image sequence to form $[C_T, X, X_{lr}]$. Since X_{lr} is still an image modality, we can reuse the pretrained VAE and DiT blocks for its processing. However, scaling the positional indices of X_{lr} alters the frequency characteristics of the RoPE embeddings. To adapt to this change, we fine-tune only the query, key, and value projections (Q, K, V) that process X_{lr} using LoRA [23], yielding adapted matrices $\tilde{Q}, \tilde{K}, \tilde{V}$. Importantly, the original Q, K, V are still used for the high-resolution windows X, ensuring that pretrained generative capabilities for local content are preserved. The unified attention mechanism $\mathrm{MMA}([X; X_{lr}])$, simplified by omitting text

tokens, is then:

$$\text{Softmax}\left(\frac{\left([Q(X), \tilde{Q}(X_{lr})]\right)\left([K(X), \tilde{K}(X_{lr})]^T\right)\cdot M}{\sqrt{d}}\right)[V(X), \tilde{V}(X_{lr})],$$

where X_{lr} are the scaled guidance tokens and M is an attention mask enforcing local window and guidance-specific interactions.

We train this framework with the standard flow matching objective, extended to include the low-resolution guidance:

$$L_{\text{FM}}(\theta; C_T, X_{lr}) = \mathbb{E}_{\substack{t\sim\mathcal{U}(0,1),\\ X_t\sim p(X|t,C_T,X_{lr})}}\left[\left\|f_\theta(X_t, t, C_T, X_{lr}) - u_t(X_t; C_T, X_{lr})\right\|^2\right].$$

In our implementation, X_{lr} is generated at 256×256, while X is trained at $1K \times 1K$. Since the model is merely adapting to a novel attention pattern guided by low-resolution inputs, rather than learning new high-resolution feature, **no native 4K data** are needed. This design yields three key advantages: (1) training is efficient, since only a small set of LoRA parameters is updated; (2) adaptation can be performed entirely on commodity-resolution data; and (3) ultra-high-resolution synthesis is achieved without the need for prohibitively expensive $4K$ training datasets.

Inference Acceleration with Low-Res Guidance. Since the low-resolution guidance image already encodes the global structure of the target output—differing primarily in high-frequency detail—we leverage it to warm-start the high-resolution sampling process. Specifically, we first upsample the low-resolution reference to match the target resolution, apply sharpening, and add noise scaled to an intermediate timestep t, yielding $X_t^{\text{ref}} = \text{Interpolate}(X_{\text{lr}})$. The high-resolution latent is then initialized as $X_{\text{hr}}^t = \alpha\, X_T + (1-\alpha)\, X_t^{\text{ref}}$, where $\alpha = \frac{t}{T}$ controls the noise ratio and X_T denotes pure Gaussian noise. By injecting the low-frequency components directly, this initialization allows us to bypass the early denoising stages, substantially accelerating inference.

4 Experiments

4.1 Implementation Details

Experimental Setup. We adopt FLUX.1-dev [2] as our text-to-image backbone, utilizing the Hugging Face Diffusers library for implementation. Parameter-efficient fine-tuning is conducted using LoRA from the PEFT library, with both rank and alpha equal to 16. LoRA layers are only applied to the Query, Key, and Value projection layers across the DiT blocks. We adopt pyriology scheduler with default learning rate of 1 and weight decay of 0.01, and set bias correction and safe-guard warmup to true. During inference, we set the NTK factor of RoPE embedding to be 10 and disabled the dynamic shifting of the scheduler following [4, 15]. Our custom attention mechanisms are implemented following

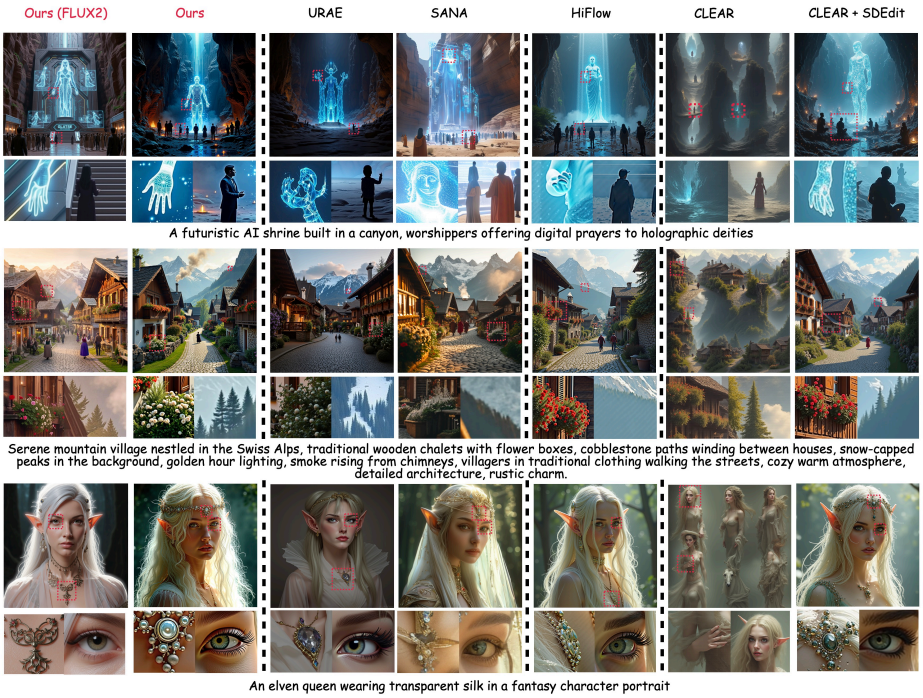

Fig. 3: 4K comparison with leading baselines based on FLUX.1 (except for SANA). Zoom in to observe the fine details. UltraImageGen produces results with superior fidelity. Additional comparisons with other baselines are provided in the Appendix. The second column section shows methods finetuned on 4K data, the third column section shows training-free method, the right column section shows another efficient attention method. Our method achieves sharper results than other methods. Notably, although HiFlow produces good results but it shows diagonal artifacts. We also provide additional results for our method based on FLUX.2, which yields more natural textures.

the principles of FlashAttention-2 [9] and SageAttention [65]. The model was trained for 20,000 steps on two NVIDIA RTX A6000 Ada GPUs (48GB VRAM each), using a per-GPU batch size of 1 with gradient accumulation over 4 steps. All subsequent experiments were conducted on this same hardware configuration. For the fine-tuning dataset, we generated 10,000 images at a resolution of 1024×1024 using the base FLUX.1-dev model, prompted by a randomly sampled collection of high-quality text descriptions. The kernel size is designed to be Q-block=128 and K-block=64 (128 for blocksparse Flash Attention) which suits the granularity supported for [65] on NVIDIA A6000 Ada. All experiments were conducted on one NVIDIA RTX Pro 6000 Blackwell GPU (96GB VRAM).

Metrics and Evaluation Protocol. To ensure a comprehensive and diverse evaluation, we generated a benchmark suite of 1,000 high-quality prompts across various categories using GPT-4o. We assess performance using a standard set of metrics: CLIP Score [41] for prompt-image alignment, CLIP-IQA [51], Fréchet

Inception Distance (FID) [21], and Inception Score (IS) [45] for image quality. The FID score is computed against a reference set of 10,000 real images from the LAION-High-Resolution [46] dataset. To specifically evaluate the fidelity of local details, we also report patch-based versions of these metrics, $\text{FID}_{\text{patch}}$ and IS_{patch}, calculated on local image patches. For all comparisons, competing training-free methods were evaluated on the same FLUX.1-dev base model, following their official implementations to ensure fairness.

4.2 Comparison with State-of-the-Art Methods

We compare our results with methods that can also generate 4K resolution images. These include data-intensive, training-based methods: SANA [54], Diffusion-4K [64], FLUX+URAE [61], UltraFLUX [60], and UltraPixel [42]; a super-resolution baseline: FLUX+BSRGAN [68]; and several training-free approaches: DemoFusion [14], DiffuseHigh [26], I-MAX [15], and HiFlow [4] and an efficient attention method CLEAR [32] (with the 4K ability comes from SDEdit [36] and the optimal setting of window size 32).

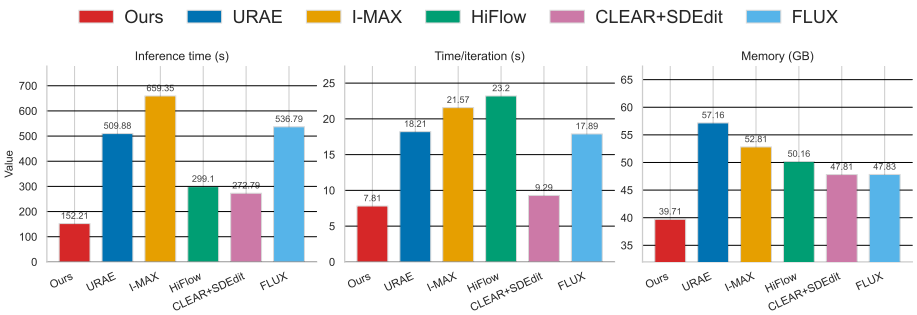

Fig. 4: Comparison of latency (sec; time to generate one 4K latents (65536 tokens)), time per iteration (sec), and memory (GB) and across FLUX.1-dev-based 4K generation methods. Our method achieves significant speedup and memory reduction.

Qualitative Comparison. Figure 1 presents qualitative results of UltraImageGen at various resolutions from $2K$ to $> 8K$, demonstrating superior text-to-image generation with high-fidelity details and coherent global structure. A more detailed comparison at 4K is shown in Figure 3, where we include a representative subset of leading baselines for clarity; comprehensive comparisons against the fourteen methods, as well as additional 2K and 4K results, are provided in the Appendix. At 4K resolution, our method renders anatomically correct hands with exceptional detail as shown in the first row, while SANA produces a blurry image, HiFlow introduces diagonal artifacts, and I-MAX fails to generate the correct number of fingers. In the second row, UltraImageGen consistently preserves textural clarity, such as tree structures, surpassing competitors while maintaining global coherence. The third row highlights highly detailed eye and jewelries generation at 4K, where our method clearly captures sharper details

of eyelashes, and jewelries with fine structure. Training free methods like HiFlow [4] in the fourth row frequently shows diagonal artifacts, while another efficient attention method [32] failed to preserve global consistency as shown in the fifth columns. These examples demonstrate UltraImageGen's ability to synthesize images that combine fine-grained local details with coherent large-scale composition, validating its performance across diverse high-resolution scenarios.

Table 1: Quantitative comparison at 4K × 4K resolution. The best result is highlighted in bold, while the second-best result is underlined. UltraImageGen consistently perform competitive results even compared to training-based methods. We also include our work with stronger base model FLUX2 here (in Italic) for reference.

Method	$\text{CLIP}_{\text{IQA}}\uparrow$	FID ↓	$\text{FID}_{\text{patch}}\downarrow$	IS ↑	$\text{IS}_{\text{patch}}\uparrow$	$\text{CLIP}_{\text{score}}\uparrow$
SANA	<u>0.4457</u>	76.31	74.27	16.68	**14.02**	0.3197
Diffusion-4K	0.3012	121.85	120.59	14.39	10.77	0.2844
UltraPixel	0.4421	77.42	70.94	16.98	13.26	**0.3251**
URAE	0.4369	<u>67.39</u>	<u>62.56</u>	17.11	12.39	0.3204
FLUX+BSRGAN	0.3897	71.39	63.45	17.08	12.87	0.3210
UltraFlux	0.4371	75.89	65.12	17.01	13.24	0.3165
DemoFusion	0.4392	74.89	66.37	16.23	13.02	0.3187
DiffuseHigh	0.4221	81.54	73.35	16.15	13.08	0.3175
HiDiffusion	0.4021	172.46	217.49	9.43	8.12	0.3129
FreCas	0.2704	213.44	322.08	7.98	6.84	0.2805
I-MAX	0.4381	70.33	65.67	16.50	12.69	0.3211
HiFlow	0.4407	69.18	63.72	<u>17.13</u>	13.43	0.3113
CLEAR	0.3463	131.15	120.79	14.18	11.92	0.2913
CLEAR+SDEdit	0.4352	88.23	87.19	16.84	<u>13.78</u>	0.3221
Ours	**0.4505**	**67.03**	**61.78**	**17.21**	13.31	<u>0.3231</u>
Ours(FLUX2)	***0.4594***	***61.38***	***59.82***	***18.02***	***14.93***	***0.3267***

Quantitative Comparison. Table 1 presents a quantitative comparison of our method against the nine state-of-the-art methods at 4K resolution where most of them are based on the same FLUX.1-dev backbone. The results show that our method achieves highly competitive performance across all metrics, even outperforming training-based methods that rely on extensive 4K-resolution training data. Table 2 demonstrates consistent performance on the GenEval [17] benchmark, confirming that UltraImageGen's output quality is resolution-agnostic.

Efficiency and Scalability Analysis. The efficiency gains of our approach are summarized in Figure 4 and Figure 5. For a fair comparison, we focus on methods built upon the FLUX.1-dev backbone without architectural modifications. Under this setting, our method achieves a 1.7× to 4.3× inference speedup over its counterparts (with all methods exhibiting greater than 2.3× speedup per denoising iteration). Although CLEAR [32] (augmented with SDEdit [36]) attains a comparable per-iteration speed, our method surpasses it in overall inference time by substantially reducing the number of required denoising steps. In terms of memory consumption, our method is highly efficient: excluding the base model's footprint $\sim 32\,GB$), it requires only $7.3\,GB$ of additional VRAM,

Table 2: Performance of our method from 1K to 4K on the GenEval benchmark. Results show that output quality is maintained across resolutions despite training on 1K data, demonstrating scalability and resolution-agnostic design.

Model	Overall	1 Obj	2 Obj	Counting	Colors	Position	Attr. Binding
FLUX.1-Dev	0.66	0.98	0.81	0.74	0.79	0.22	0.45
Ours(2K)	0.67	0.99	0.83	0.74	0.81	0.22	0.45
Ours(3K)	0.67	0.98	0.82	0.73	0.81	0.23	0.46
Ours(4K)	0.67	0.98	0.83	0.75	0.79	0.24	0.45

whereas a naive implementation with FlashAttention consumes $15.6\,GB$. This significant reduction in memory overhead enables inference on consumer-level GPUs without module offloading, and facilitates more GPU-friendly fine-tuning.

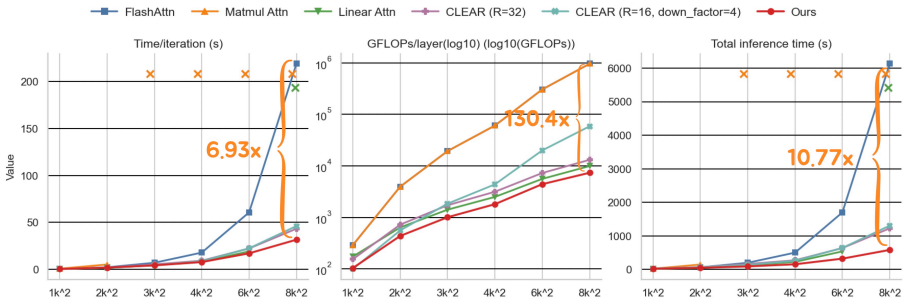

Fig. 5: Efficiency comparisons across resolutions (1K to 8K). We report per-iteration latency (**Left**), floating-point operations per attention layer (**Middle**), and total inference time (**Right**) for our method, naive matrix multiplication attention, Linear Attention, FlashAttention-2, and two linearlized attention configurations from CLEAR. At 8K resolution, our method achieves a 6.93× per-iteration speedup, 130.4× reduction in attention computation, and 10.77× total inference speedup over FLUX with FlashAttention, while consistently outperforming CLEAR and Linear Attention across all metrics. The "X" mark means out-of-memory (OOM). Linear Attention reaches OOM at 8K resolution, while the naive one reaches OOM at 3K resolution.

Figure 5 further illustrates the scalability of our approach. As the target resolution increases from 1K to 8K, our method achieves a 10.77× reduction in latency and a 130.4× reduction in floating-point computation within the attention layer compared to FlashAttention-2 [9]. Compared with another linearized attention mechanism [32], our approach remains consistently faster and more memory-efficient. We select two optimal configurations from their paper—CLEAR with a window radius of 32 and a radius of 16 with a downsample factor of 4—and as shown in the line chart, our method yields faster inference and lower computational cost in the attention layer as resolution scales, while also achieving superior generation quality as demonstrated in Table 1 and Figure 3. We attribute these efficiency gains to two key design choices: (1) our dense-block-style local window attention is more hardware-friendly, with window sizes that align well with GPU architectures; and (2) we fully leverage low-resolution guidance as an initializa-

tion signal, in contrast to CLEAR, which employs low-resolution guidance solely as a starting point for SDEdit due to its limited scalability across resolutions. For linear attention, while also having relatively low GFLOPs, it requires scaling the hidden dimension with sequence length to maintain performance [19,37,69], and it is not able to produce meaningful results via directly substituting the softmax based attentions in pretrained models [32]. In contrast, our method uses local window attention, which does not need to scale with resolution, enabling consistent memory and latency efficiency as resolution increases. Since linear attention requires the hidden dimension to scale with resolution, previous works would need to fine-tune on high-resolution data for each target resolution, whereas our framework achieves ultra-high-resolution synthesis using only 1K-resolution training data without any performance drop. Collectively, these results validate the scalability, efficiency, and robustness of our proposed framework.

5 Ablation Study

To validate our core design choices, we conduct a series of ablation studies analyzing the impact of window size, guidance resolution ratio, attention patterns, and token permutation strategies.

Table 3: Ablation experiments investigating the effects of window size and the high-to-low resolution ratio demonstrated a negligible impact on model performance. Therefore, to prioritize computational efficiency, the configuration utilizing the smallest window size and the largest ratio was chosen for all subsequent experiments.

Method	**ws256 ρ4**	**ws256 ρ2**	**ws512 ρ4**	**ws1024 ρ4**
FiD ↓	67.23	67.19	66.31	67.83
$\text{FiD}_{\text{patch}}$ ↓	61.46	62.28	63.48	62.61

Window size and resolution ratio. We conduct ablation study on window size (ws) and ratio between high-res and low-res image (ρ). We experiment with window sizes from 256, 512, and 1024, and $\rho = 2, 4$. As shown in Table 3, the variations in FiD scores across these configurations were not statistically significant. To maximize computational efficiency without compromising performance, we selected the smallest window size (ws=256) and the largest ratio (ρ=4). It is worth noting that smaller window sizes or larger ratios are constrained by the fixed granularity of the kernel depending on the hardware design of the GPU.

Attention Pattern We conducted an ablation study to evaluate the impact of different text-to-image attention mechanisms. Case 1 examines the effect of attending to low-resolution tokens X_{lr}, comparing (a) high-resolution tokens X attending to both textual embeddings and X_{lr} versus (b) attending only to textual embeddings. Case 2 investigates the effect of attending to neighboring windows, comparing (a) each window of X attending to its neighbors versus (b) attending only to itself.

As shown in Figure 6, including low-resolution tokens in Case 1 produces consistently stable results, whereas excluding them causes discontinuities and

multi-face artifacts, highlighting the importance of low-resolution guidance. In Case 2, attending to neighboring windows effectively eliminates boundary artifacts, while incurring only additional computational cost of less than 2%.

Fig. 6: Ablation Study on attention scale. Images from left to right corresponds to cases that 1) High-res X attends to both text and X_{lr}. 2) X attends to only Text. 3) Each window of X attends to its neighbors. 4) X only attends to itself. 1) and 2) demonstrates the effectiveness of our low-res guidance, and 3) and 4) illustrates that allowing each window to attend to part of the neighbors will solve the boundary issues.

Robustness on Different Models To demonstrate the robustness of UltraImageGen, we conducted additional experiments on FLUX.2-series. We provide both qualitative (Figure 3 and Figure 1) and quantitative (Figure 1). (Table 1 results for FLUX.2-Klein-9B, which is a time-distilled 4 step model. We therefore achieve **sub 30-second generation** for 4K images and **sub 2-minute** for 8K images, enabling maximum efficiency in ultra-high-resolution content creation.

Content in the Supplementary Materials. We will include (1) More 8K, 4K and 2K resolution galleries, and the PDF file for the Teasor figure. (2) a detailed quantitative 4K resolution comparison against all 14 baselines, (3) demo for our method's compatability with different styled LoRA models, robustness to model sizes and aspect ratios, (4) more analysis and ablations.

6 Conclusion

We present UltraImageGen, a novel framework enabling pre-trained diffusion transformers to generate ultra-high-resolution images *without requiring additional high-resolution training data*. UltraImageGen introduces a hierarchical attention mechanism that partitions high-resolution latents into fixed-size local windows while maintaining global coherence through low-resolution guidance with scaled positional anchors. The framework comprises three key components: (i) efficient local window attention that reduces quadratic complexity to near-linear scaling; (ii) global semantic preservation via low-resolution guidance latents with position scaling; and (iii) parameter-efficient joint denoising through LoRA-based adaptations trained solely on commodity resolutions. Extensive experiments demonstrate that UltraImageGen achieves superior visual quality at 4K resolution while delivering over 10× *speedup* in $8K$ resolution and *less memory usage* compared to dense attention baselines, establishing a practical paradigm for scalable ultra-high-resolution text-to-image generation.

References

1. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020) 4
2. Black Forest Labs: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025) 2, 3, 6, 8
3. Black Forest Labs: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025), accessed: 2026-06-28 3, 6
4. 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. In: Advances in Neural Information Processing Systems (2025) 2, 4, 8, 10, 11
5. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025) 2, 3, 6
6. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: European Conference on Computer Vision. pp. 74–91. Springer (2024) 3
7. Chen, J., Luo, S., Xie, E.: Pixart-δ: Fast and controllable image generation with latent consistency models. In: ICML 2024 Workshop on Theoretical Foundations of Foundation Models (2024) 2, 3
8. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J.T., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024) 2, 3
9. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (2024) 4, 6, 7, 9, 12
10. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In: Advances in Neural Information Processing Systems (2022) 4, 6, 7
11. Deng, Y., Song, Z., Yang, C.: Attention is naturally sparse with gaussian distributed input. CoRR (2024) 4
12. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: mastering text-to-image generation via transformers. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. pp. 19822–19835 (2021) 2, 3
13. Dong, Z., Zhou, C., Deng, W., Wei, P., Ji, X., Lin, L.: Can we achieve efficient diffusion without self-attention? distilling self-attention into convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17401–17410 (2025) 4
14. 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 (2024) 10
15. Du, R., Liu, D., Zhuo, L., Qin, 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) 2, 4, 8, 10
16. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025) 3, 6

17. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023) 11
18. 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) 2, 4
19. Han, D., Pu, Y., Xia, Z., Han, Y., Pan, X., Li, X., Lu, J., Song, S., Huang, G.: Bridging the divide: Reconsidering softmax and linear attention. Advances in Neural Information Processing Systems **37**, 79221–79245 (2024) 4, 13
20. He, H., Bai, Y., Lan, R., Duan, X., Sun, L., Chu, X., Xia, G.S.: Ragsr: Regional attention guided diffusion for image super-resolution. arXiv preprint arXiv:2508.16158 (2025) 4
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 10
22. 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) 2, 4
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022) 7
24. Hu, T., Zhang, J., Su, Z., Yi, R.: Ultragen: High-resolution video generation with hierarchical attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 4923–4931 (2026) 4
25. 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) 2, 4
26. 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) 2, 4, 10
27. Kwon, T., Bianchi, L., Wittke, L., Watine, F., Carrara, F., Ye, J.C., Weber, R., Azevedo, V.: Reviving convnext for efficient convolutional diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43675–43685 (2026) 4
28. Lai, X., Lu, J., Luo, Y., Ma, Y., Zhou, X.: Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference. In: The Thirteenth International Conference on Learning Representations (2025) 4
29. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024) 2, 3
30. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024) 2, 3
31. Liu, L., Qu, Z., Chen, Z., Tu, F., Ding, Y., Xie, Y.: Dynamic sparse attention for scalable transformer acceleration. IEEE Transactions on Computers **71**(12), 3165–3178 (2022) 4
32. Liu, S., Tan, Z., Wang, X.: CLEAR: Conv-like linearization revs pre-trained diffusion transformers up. In: Advances in Neural Information Processing Systems (2025) 4, 5, 10, 11, 12, 13

33. Liu, S., Yu, W., Tan, Z., Wang, X.: Linfusion: 1 gpu, 1 minute, 16k image. arXiv preprint arXiv:2409.02097 (2024) 2, 4
34. Liu, X., He, Y., Guo, L., Li, X., Jin, B., Li, P., Li, Y., Chan, C.M., Chen, Q., Xue, W., Liu, J., Liu, S.: HiPrompt: Tuning-free higher-resolution generation with hierarchical MLLM prompts. International Journal of Computer Vision **134**, 147 (2026) 2, 4
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 4
36. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022) 5, 10, 11
37. Meng, W., Luo, Y., Li, X., Jiang, D., Zhang, Z.: Polaformer: Polarity-aware linear attention for vision transformers. In: The Thirteenth International Conference on Learning Representations (2025) 4, 13
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 2, 3
39. 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. In: International Conference on Learning Representations (2024) 2, 3
40. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16893–16903 (2025) 2, 4
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 9
42. 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) 2, 4, 10
43. Ren, S., Yu, Q., He, J., Yuille, A., Chen, L.C.: Grouping first, attending smartly: Training-free acceleration for diffusion transformers. arXiv preprint arXiv:2505.14687 (2025) 4
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3
45. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016) 10
46. 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) 10

47. Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., Dao, T.: Flashattention-3: fast and accurate attention with asynchrony and low-precision. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. pp. 68658–68685 (2024) 4, 6
48. Shi, S., Li, W., Zhang, Y., He, J., Gong, B., Zheng, Y.: Resmaster: Mastering high-resolution image generation via structural and fine-grained guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6887–6895 (2025) 2, 4
49. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 6
50. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. In: The Twelfth International Conference on Learning Representations (2024) 2, 4
51. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023) 9
52. 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. pp. 3944–3953. IEEE (2025) 2, 4
53. Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., et al.: Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. In: Forty-second International Conference on Machine Learning (2025) 4
54. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S.: SANA: efficient high-resolution text-to-image synthesis with linear diffusion transformers. In: The Thirteenth International Conference on Learning Representations (2025) 2, 3, 4, 10
55. Xie, E., Chen, J., Zhao, Y., YU, J., Zhu, L., Lin, Y., Zhang, Z., Li, M., Chen, J., Cai, H., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In: Forty-second International Conference on Machine Learning (2025) 2, 3, 4
56. Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., Li, Z.: Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4230–4239 (2023) 4
57. Xu, R., Xiao, G., Huang, H., Guo, J., Han, S.: Xattention: Block sparse attention with antidiagonal scoring. In: Proceedings of the 42nd International Conference on Machine Learning (2025) 4
58. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. Advances in Neural Information Processing Systems **36**, 41693–41706 (2023) 2, 3
59. Yang, S., Xi, H., Zhao, Y., Li, M., Zhang, J., Cai, H., Lin, Y., Li, X., Xu, C., Peng, K., et al.: Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. arXiv preprint arXiv:2505.18875 (2025) 4
60. Ye, T., Fei, S., Zhu, L.: Ultraflux: Data-model co-design for high-quality native 4k text-to-image generation across diverse aspect ratios. arXiv preprint arXiv:2511.18050 (2025) 4, 10
61. Yu, R., Liu, S., Tan, Z., Wang, X.: Ultra-resolution adaptation with ease. In: International Conference on Machine Learning (2025) 4, 10

62. 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) 4
63. 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) 4
64. 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) 2, 4, 10
65. Zhang, J., Huang, H., Zhang, P., Wei, J., Zhu, J., Chen, J.: Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In: International Conference on Machine Learning (2025) 4, 9
66. Zhang, J., Wei, J., Zhang, P., Zhu, J., Chen, J.: Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In: International Conference on Learning Representations (2025) 4
67. Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., Chen, J.: Spargeattn: Accurate sparse attention accelerating any model inference. In: International Conference on Machine Learning (2025) 4
68. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: IEEE International Conference on Computer Vision. pp. 4791–4800 (2021) 10
69. Zhang, M., Bhatia, K., Kumbong, H., Re, C.: The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In: The Twelfth International Conference on Learning Representations (2024) 4, 13
70. Zhang, P., Chen, Y., 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) 4, 5
71. 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 (2025) 4
72. Zhang, Y., Huang-Menders, A., Tai, Y.W.: Hieredit: Region-aware hierarchical diffusion for efficient high-resolution editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43546–43557 (2026) 4
73. Zhang, Y., Li, J., Tai, Y.W.: LayerCraft: Enhancing text-to-image generation with cot reasoning and layered object integration. In: Advances in Neural Information Processing Systems (2025) 3
74. 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) 4
75. Zheng, Q., Guo, Y., Deng, J., Han, J., Li, Y., Xu, S., Xu, H.: Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7571–7578 (2024) 2, 4
76. Zhu, L., Huang, Z., Liao, B., Liew, J.H., Yan, H., Feng, J., Wang, X.: Dig: Scalable and efficient diffusion models with gated linear attention. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7664–7674 (2025) 4