SEGA: Spectral-Energy Guided Attention for Resolution Extrapolation in Diffusion Transformers

Javad Rajabi **Kimia Shaban** **Koorosh Roohi** **David B. Lindell** **Babak Taati**

University of Toronto Vector Institute

{rajabi, lindell, taati}@cs.toronto.edu

Project page: https://rajabi2001.github.io/sega/

Figure 1: **Gallery of SEGA.** SEGA unlocks the high-resolution generation capabilities of pre-trained T2I models (Flux [1] and Qwen [2]), producing high-quality images. Best viewed zoomed in.

Abstract

Diffusion transformers (DiTs) have emerged as a dominant architecture for text-to-image generation, yet their performance drops when generating at resolutions beyond their training range. Existing training-free approaches mitigate this by modifying inference-time attention behavior, often through Rotary Position Embeddings (RoPE) extrapolation combined with attention scaling. However, these strategies apply a uniform and content-agnostic scaling across RoPE components with distinct frequency characteristics, inducing a trade-off between preserving global structure and recovering fine detail. We introduce **SEGA**, a training-free method that dynamically scales attention across RoPE components according to the latent's spatial-frequency structure at each denoising step. This adaptive scaling improves both structural coherence and fine-detail fidelity. Experiments show that SEGA consistently improves high-resolution synthesis across multiple target resolutions, outperforming state-of-the-art training-free baselines.

Preprint.

1 Introduction

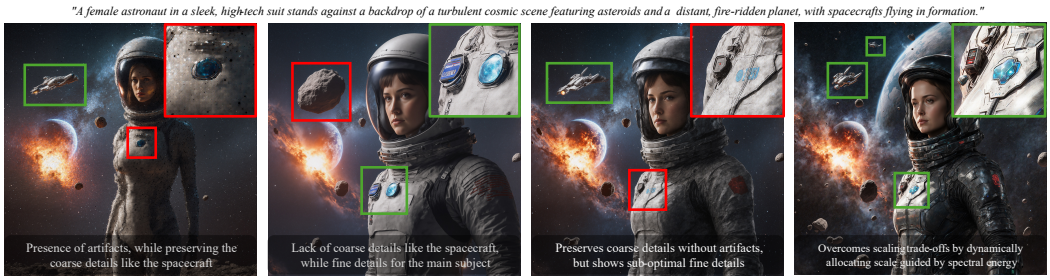
Diffusion transformers (DiTs) [3, 4] have become the dominant approach to text-to-image (T2I) generation, producing images with a level of quality that would have been hard to imagine just a few years ago. Despite considerable improvements, existing T2I models remain largely constrained by the resolution ranges used during training, typically between 1024^2 and 2048^2 resolutions, limiting their practical applicability [5, 6, 7]. Consequently, extrapolating beyond this training resolution at inference time often leads to notable quality degradation and even structural breakdown. A straightforward solution is to train or fine-tune models at the target resolution [8, 9]. However, such approaches are practically limited by the scarcity of high-resolution data, the quadratic cost of longer token sequences, and the need for model-specific fine-tuning. These bottlenecks have motivated growing interest in training-free high-resolution synthesis from pre-trained models [10, 11, 12, 13].

Existing training-free methods for high-resolution image generation generally fall into two categories: (i) direct inference [14, 15, 16, 17] and (ii) multi-stage guidance-based approaches [18, 19, 20, 5, 6]. Direct inference methods attempt to extend pretrained models to higher resolutions by modifying the denoising process or adjusting components such as positional encoding and attention without additional training. In contrast, multi-stage approaches first generate a base-resolution image and then use it to guide high-resolution synthesis. Although often effective, these methods introduce additional complexity and depend heavily on the quality of the low-resolution prediction. More importantly, they fundamentally cast high-resolution generation as a super-resolution problem, relying on external guidance rather than improving the model's intrinsic ability to extrapolate to higher resolutions.

In this work, we focus on direct-inference methods for resolution extrapolation in DiTs and address a fundamental failure mode related to positional encoding. When extrapolating pre-trained DiTs to high-resolution synthesis, the relative positional offsets in Rotary Position Embeddings (RoPE) [21] deviate significantly from those observed at training time, causing the attention weights to become overly diluted across the expanded token grid. This weakens spatial discrimination in attention and leads to degraded outputs such as blurred textures, repetitive patterns, and structural breakdowns. To counter this, previous approaches, adapted from long-context language modeling, combine RoPE extrapolation with a uniform attention scaling to restore spatial focus [22]. Specifically, they scale the resulting attention values uniformly across the positional encoding components. While this uniform attention scaling improves image quality, it applies the same adjustment across RoPE components with different frequency characteristics, treating short-wavelength components that govern fine-grained texture identically to long-wavelength components that shape global structure. As illustrated in Figure 2, static scaling induces an inherent trade-off, yielding different failure modes across global structure and fine-grained detail. The problem is further compounded by two distinct variations in the latent's spectral characteristics. First, the spectral distribution evolves throughout denoising, with the relative contributions of low- and high-frequency bands shifting noticeably as the image resolves from noise to a structured form. Second, the spectral distribution differs across images, depending on their content and structural complexity (e.g., a foggy lake versus a bustling outdoor market). Consequently, a static, uniform scaling at inference time cannot accommodate these variations.

Building on this view, we introduce **SEGA** (**S**pectral-**E**nergy **G**uided **A**ttention), a training-free, content-aware method that dynamically adapts attention scaling to the latent's spectral structure by deriving per-component scaling magnitudes at each denoising step. Our method is motivated by a simple but consequential observation: RoPE components are coupled to spatial frequencies, as shown in Figure 2. SEGA uses the energy in each corresponding spatial frequency band to determine the scaling applied to each RoPE component: those associated with low-energy bands receive stronger scaling to preserve positional discrimination at those frequencies, whereas components associated with high-energy bands receive weaker scaling to avoid over-amplifying already prominent features. A scalar then controls how strongly this scaling is applied, based on the spectrum's entropy. The result is an attention scaling that adapts to both the content of the current latent and its evolution across denoising steps, resolving the trade-off induced by fixed global scaling.

Extensive experiments show that SEGA consistently improves structural coherence and fine-detail fidelity and achieves superior performance across baselines and resolution settings, including ultra-high resolutions exceeding 36 million pixels. SEGA introduces no learnable parameters, requires no fine-tuning or architectural changes, and integrates directly into standard RoPE-based pipelines, making it a minimal yet effective solution for stable high-resolution synthesis across a wide range of extrapolated resolutions, as shown in Figure 1.



(a) Low frequency scaling only (b) High frequency scaling only (c) Uniform scaling (d) SEGA

Figure 2: **Trade-offs in attention scaling at 4096^2.** RoPE components are coupled to spatial frequencies: low-frequency components support coarse detail and structure, whereas high-frequency components support fine detail and texture. Static scaling fails to balance this trade-off, leading to different failure modes in (a)–(c). SEGA (d) resolves them by dynamically allocating scaling according to spectral energy. Green and red boxes indicate successful and failed regions, respectively.

2 Related Work

2.1 High-Resolution Image Synthesis

Preserving both global structure and fine-grained detail remains an open challenge in high-resolution generation. Training-based approaches address this through progressive upsampling [23, 24, 25, 26], latent-space super-resolution [27], or explicit retraining on high-resolution data or model-specific fine-tuning like Diffusion-4K [20]. By contrast, training-free methods [19, 28, 29, 30] adapt pretrained models at inference time. In U-Net architectures, methods such as DemoFusion [10], FreeScale [18], and FreCaS [19] improve high-resolution generation through patch stitching, multi-scale fusion, or cascaded sampling, but often introduce additional inference complexity. In DiTs, training-free extrapolation has largely relied on more complex strategies, often involving two-stage pipelines in which a base-resolution trajectory guides high-resolution sampling, as in I-Max [6], HiFlow [5], and ScaleDiff [31]. While effective, these methods depend on multi-stage guidance and often introduce additional complexity into the denoising process.

2.2 RoPE-based Length Extrapolation

The challenge of high-resolution generation in DiTs closely mirrors long-context extrapolation in large language models (LLMs) [32, 33], largely driven by advances in RoPE [21]. Standard training-free methods [34, 35, 22] formulate extrapolation as recalibration of RoPE's rotary frequencies. Position Interpolation [34] compresses position indices to fit longer sequences within the training range, limiting phase drift. NTK [35] adjusts the RoPE base frequency to redistribute positional variation more evenly across dimensions, thereby improving extrapolation to longer sequences. YaRN [22] builds on both by applying frequency-band-specific interpolation strategies and introducing an additional uniform attention scaling. Recent works adapt these principles to visual domains [36, 37]. DyPE [15] introduces step-wise, time-aware positional adjustments across the diffusion timesteps. UltraImage [14] alleviates repetitive artifacts by shifting the dominant frequency to align with the training resolution and employing entropy-guided attention concentration. However, these approaches largely rely on predefined heuristics or target-resolution alignments. In contrast, our method directly analyzes the spectral energy of the intermediate latent to dynamically adjust attention scaling. By amplifying high-energy bands and suppressing low-energy ones, it preserves fine-grained detail without compromising structural fidelity. See Appendix A for more detailed related work.

3 Preliminaries

Rotary Position Embedding (RoPE) Positional embeddings provide spatial priors for transformer architectures, which form the core of DiT models. They encode coordinate information into feature representations, addressing the models' inherent permutation equivariance. Among various designs, RoPE [21] is a widely used scheme that encodes relative positions through rotation in the embedding space, and it has been adopted in recent T2I models such as Flux [1] and Qwen [2].

RoPE encodes a position n by applying a series of 2D rotations to paired dimensions, each at a distinct angular frequency determined by the embedding dimension index. Given a vector $\mathbf{x} \in \mathbb{R}^D$ at position n, RoPE partitions $\mathbf{x}$ into $D/2$ two-dimensional subspaces and rotates the d-th subspace as

$$\boldsymbol{f}^{\text{RoPE}}(\mathbf{x}, n, \boldsymbol{d}) = \begin{bmatrix} \cos(n\theta_d) & -\sin(n\theta_d) \\ \sin(n\theta_d) & \cos(n\theta_d) \end{bmatrix} \begin{bmatrix} x_{2d} \\ x_{2d+1} \end{bmatrix}, \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{D/2}$ with $\theta_d = b^{-2d/D}$ for $d = 0, \ldots, D/2 - 1$ and $b = 10{,}000$. In practice, RoPE is applied to the query and key vectors before the dot product operation in the attention mechanism. Additionally, it can be shown that the dot product of two RoPE-embedded vectors depends only on their relative distance, so attention naturally encodes relative positional information. For 2D images, RoPE is typically applied axially: half of the hidden dimensions encode horizontal positions and the other half encode vertical positions, enabling independent offsets along each axis [38].

3.1 Length Extrapolation Techniques and Attention Scaling

Although RoPE provides an effective positional bias within the training, models that rely on it often degrade at unseen resolutions, where attention must operate on out-of-distribution positional offsets. Several methods have been proposed to adapt RoPE to longer sequences at inference time, given an extrapolation ratio $s = (L_{\text{target}}/L_{\text{train}})$, where $s > 1$. *Position Interpolation* (PI) [34] linearly compresses position indices via $n \mapsto n/s$ for position n, which uniformly transforms all RoPE components to θ_d/s so extrapolated positions remain within the training range. *NTK-aware* [35] instead adjusts b to $b' = b \cdot s^{D/(D-2)}$, which stretches the angular frequency of each rotary dimension θ_d. *YaRN* [22] unifies these ideas by partitioning rotary dimensions and applying a gradual interpolation-extrapolation strategy, a.k.a. *NTK-by-parts* [22]. Specifically, it smoothly interpolates the modified frequencies as $\theta'_d = (1 - \lambda_d)\frac{\theta_d}{s} + \lambda_d \theta_d$ using a ramp function $\lambda_d \in [0, 1]$.

Another key component of YaRN is *attention scaling*, applied to the logits before the softmax. Notably, this effect can be implemented through RoPE by scaling the query and key vectors after rotation, thereby changing the effective attention behavior without altering the attention mechanism itself [22]. YaRN proposes a constant logit scaling factor $\tau(s)$ to compensate for the change in attention behavior under extrapolation, modifying attention as

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\tau(s) \cdot \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \qquad \tau(s) = 0.1 \ln(s) + 1 \tag{2}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices, respectively; d_k denotes the dimensionality of the queries and keys. The scaling factor $\tau(s)$ was determined empirically for length extrapolation in language models by minimizing perplexity [22]. The same heuristic has since been adopted in image generation [16]. However, this scaling remains uniform across all RoPE frequencies. Since different RoPE dimensions exhibit distinct characteristics and contribute unevenly to spatial structure, a constant scaling factor is suboptimal; it may over-sharpen some spatial-frequency bands while over-smoothing others, motivating a dynamic scaling strategy.

4 Method

Spectral-Energy Guided Attention (SEGA) introduces content-aware dynamic scaling into DiTs by coupling lightweight spectral analysis with RoPE components. Our key insight is that RoPE scaling for high-resolution extrapolation should be content-aware rather than fixed and uniform. SEGA achieves this by deriving per-dimension scaling from the latent's spectral structure at each denoising step.

Formulation Overview. SEGA applies attention scaling through RoPE using a dimension-wise scaling term m_d. Specifically, for a token at position n along axis a, we define

$$\boldsymbol{f}^{\text{SEGA}}(\mathbf{x}, n, d) = m_d^{(a)} \cdot \boldsymbol{f}^{\text{RoPE}}(\mathbf{x}, n, d), \qquad m_d^{(a)} = m_{\text{ref}} \cdot \mathcal{M}_d^{(a)}(\mathbf{Z}), \tag{3}$$

where m_{ref} is a scalar determined by the target resolution. Here, $\mathcal{M}_d^{(a)}(\mathbf{Z})$ is our novel dynamic modulator derived from the spectral structure of the current intermediate latent $\mathbf{Z}$. It consists of two complementary components: $s_d^{(a)}(\mathbf{Z})$, a *per-dimension correction* that determines the distribution of scaling across RoPE dimensions, and $\sigma(\mathbf{Z})$, a *global amplitude factor* that sets the strength of that adjustment. The remainder of this section describes how spectral structure is extracted from $\mathbf{Z}$ (Section 4.1) and converted into $s_d^{(a)}$ and σ to assemble the final formula (Section 4.2).

4.1 Spectral Analysis of the Latent

The first stage of SEGA transforms the current latent from the spatial domain to the frequency domain to characterize the spatial frequency content. Given the latent hidden states $\mathbf{Z} \in \mathbb{R}^{N \times C}$ with $N = H \cdot W$ tokens,[1] we reshape them back to their native 2D layout, average across channels, and subtract the average value across the spatial dimensions to obtain a zero-centered 2D map $\tilde{\mathbf{M}} \in \mathbb{R}^{H \times W}$ that summarizes the spatial structure of the latent. From $\tilde{\mathbf{M}}$ we extract two complementary spectral views from a single 2D Fast Fourier Transform $\mathcal{F}_{\text{2D}}$:

- **Axis-wise profiles.** For each axis $a \in \{H, W\}$ with length L_a, we marginalize the 2D power spectrum $\left|\mathcal{F}_{\text{2D}}[\tilde{\mathbf{M}}]\right|^2$ over the orthogonal frequency axis to obtain a 1D profile $\mathcal{E}_a \in \mathbb{R}^{\lfloor L_a/2 \rfloor}$. Each profile maps spectral energy to spatial frequencies along its axis.
- **Radial profile.** We obtain $\mathcal{E}_{\text{iso}}$ by averaging the same 2D power spectrum within concentric rings. This profile discards directional information and instead provides a rotation-invariant summary of how energy is distributed across spatial scales.

These profiles then determine the scaling of each RoPE dimension. Because RoPE is applied separately along the height and width axes, the axis-wise profiles capture directional differences in spectral energy and allow the corresponding RoPE dimensions to be scaled independently, while the radial profile determines the strength of this scaling, as described in the next section.

4.2 From Spectrum to Per-Dimension RoPE Scaling

The second stage converts the spectral profiles into the modulator $\mathcal{M}(\mathbf{Z})$, which defines the per-dimension scaling applied to the rotary embeddings. This formulation consists of three components: a reference scale that anchors the scaling, a per-dimension term that scales individual dimensions, and a global gate that controls the strength of that scaling.

Reference scale. The reference scale m_{ref} is a scalar determined solely by the ratio between the target and training resolutions. Assuming $R_{\text{target}}/R_{\text{train}} \geq 1$, we adopt a power-law form,

$$m_{\text{ref}} = \left(\frac{R_{\text{target}}}{R_{\text{train}}}\right)^{\kappa}, \tag{4}$$

where $\kappa > 0$ is a small exponent chosen empirically. See Appendix H for alternative formulations.

Per-dimension correction. Each RoPE dimension governs the attention mechanism's sensitivity at a specific spatial wavelength, modifying the scaling at dimension d directly alters how sharply the model can discriminate positional offsets at that wavelength, and therefore affects the corresponding spatial frequency. This coupling motivates a per-dimension correction tied to the latent's actual spectral content. For each RoPE dimension d on axis a, we use its wavelength $T_d = 2\pi/\theta_d$ to identify the corresponding band in $\mathcal{E}_a$, retrieve the log-energy $\hat{E}_d^{(a)}$, and standardize it across dimensions as $z_d^{(a)} = (\hat{E}_d^{(a)} - \mu^{(a)})/\nu^{(a)}$, where $\mu^{(a)}$ and $\nu^{(a)}$ denote the mean and standard deviation of $\hat{E}^{(a)}$. To enforce a strict zero-sum redistribution, the final correction is defined as $s_d^{(a)} = \phi(z_d^{(a)}) - \mathbb{E}[\phi(z^{(a)})]$, where $\phi(\cdot)$ is a non-linearity, for which we use $\tanh$. By construction, $s_d^{(a)} < 0$ when dimension d falls in a band with below-average energy and $s_d^{(a)} > 0$ when it falls in a band with above-average energy, while the zero-mean property $\sum_d s_d^{(a)} = 0$ ensures that the correction adjusts the scaling across dimensions without shifting its overall average.

Global amplitude factor. To regulate the *magnitude* of the scaling introduced by the axis profiles, SEGA reduces the radial profile $\mathcal{E}_{\text{iso}}$ to a single scalar statistic that captures whether the latent's spectral energy is concentrated in a few dominant bands or spread evenly across all bands. For this purpose we adopt the *spectral flatness*, also known as the *Wiener entropy*, defined as the ratio of the

[1] For notational simplicity, we omit the batch dimension B in our formulation, as all operations are applied independently across the batch.

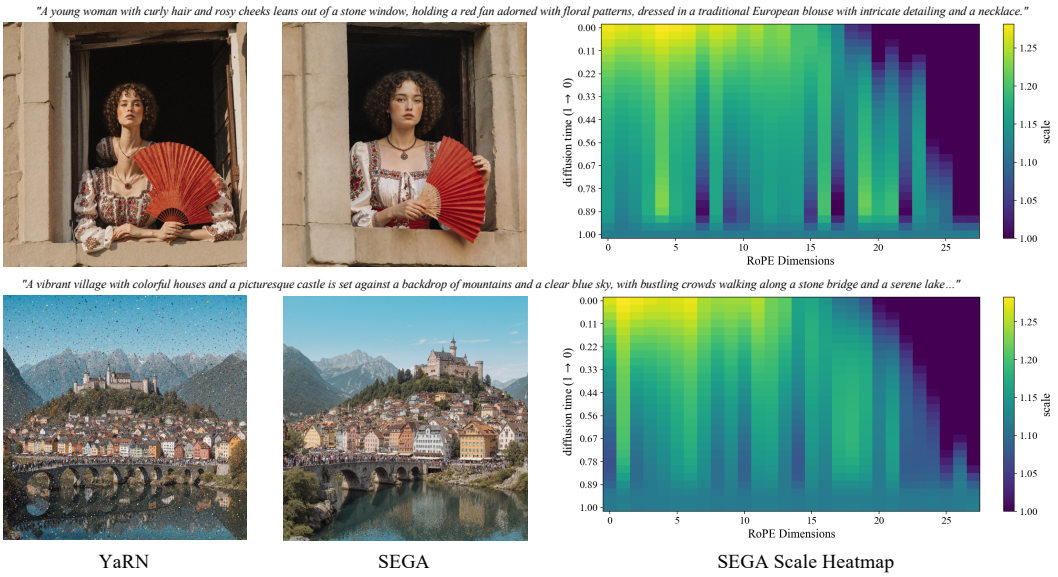


Figure 3: **SEGA scaling maps at 4096^2.** For two representative prompts, the scaling maps show how the horizontal-axis scaling magnitudes m_d change across RoPE dimensions over denoising time.

geometric mean to the arithmetic mean of a power spectrum. Applied to $\mathcal{E}_{\text{iso}}$, this yields

$$\text{SF}(\mathcal{E}_{\text{iso}}) = \frac{\exp\left(\frac{1}{n_{\text{bins}}^{(\text{iso})}} \sum_{b=0}^{n_{\text{bins}}^{(\text{iso})}-1} \ln \mathcal{E}_{\text{iso}}[b]\right)}{\frac{1}{n_{\text{bins}}^{(\text{iso})}} \sum_{b=0}^{n_{\text{bins}}^{(\text{iso})}-1} \mathcal{E}_{\text{iso}}[b]} \in (0, 1], \tag{5}$$

where $n_{\text{bins}}^{(\text{iso})}$ is the number of radial bins used to compute $\mathcal{E}_{\text{iso}}$. We then remap the spectral flatness through a simple nonlinearity to produce a scalar *amplitude factor*:

$$\sigma = 1 - \text{SF}(\mathcal{E}_{\text{iso}})^{\gamma} \in [0, 1], \tag{6}$$

where $\gamma \geq 1$ controls how quickly σ rises as the spectrum departs from flatness. Without clear spectral structure, $\sigma \to 0$ and SEGA suppresses its scaling; as structural content resolves, $\sigma \to 1$ and the correction applies at full strength.

Final scaling formula. Combining the three components, we define the modulator and the resulting per-dimension scaling $m_d^{(a)}$ along each spatial axis $a \in \{H, W\}$ as

$$\mathcal{M}_d^{(a)}(\mathbf{Z}) = 1 - \sigma \cdot s_d^{(a)}, \qquad m_d^{(a)} = m_{\text{ref}} \cdot \mathcal{M}_d^{(a)}(\mathbf{Z}). \tag{7}$$

Intuitively, m_{ref} sets the shared magnitude across RoPE dimensions, $s_d^{(a)}$ determines which dimensions are scaled above or below that reference, and σ controls the strength of this redistribution. In this way, SEGA adapts continuously to the latent's spectral content at each denoising step, sharpening attention at under-resolved frequencies and softening it at over-emphasized ones.

5 Analysis of Spectral-Energy Guided Attention

To better understand how SEGA and spectral guidance influence denoising, we analyzed scaling behavior and the attention focus during the denoising process. As shown in Figure 3, we visualized the resulting *scaling map*, a temporal representation of how the attention scaling factors m_d are distributed throughout the denoising process. When comparing the scaling maps produced for two distinct prompts, as shown, the difference is apparent. The method yields a customized scaling map for each image, effectively acting as a unique *spectral fingerprint*. This occurs because SEGA is content-aware, dynamically adapting scaling to the latent's spatial frequencies. In early steps where the latent is dominated by noise and the spectrum is relatively flat, the scaling remains near the reference scale m_{ref}. However, as distinct structural energy emerges in later steps, SEGA selectively

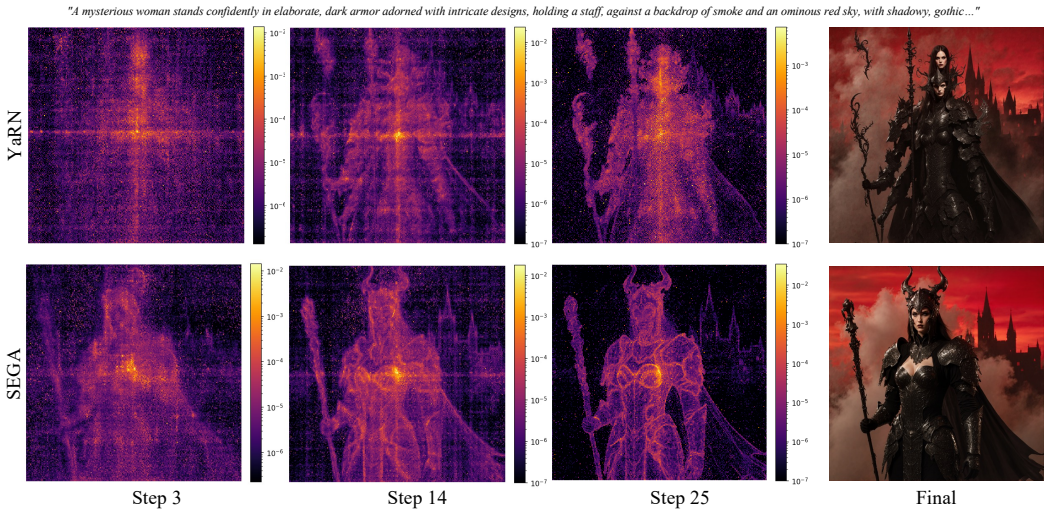


Figure 4: **Impact on Attention Evolution.** Visual comparison of attention maps for the center latent token in YaRN and SEGA across multiple denoising steps, evaluated on Flux at 4096^2.

Figure 5: **Qualitative comparison.** Results on two representative prompts for Qwen and Flux at 4096^2 resolution show that SEGA improves structural coherence and fine detail over other methods.

redistributes scaling across RoPE dimension d to sharpen focus at under-resolved spatial frequency bands while softening it at over-emphasized ones.

This content-aware spectral redistribution directly impacts the attention mechanism's stability. As visualized in Figure 4, YaRN [22], which uses fixed, uniform scaling, suffers from attention dilution, where the model loses the ability to discriminate between positional offsets. SEGA mitigates this failure mode by shaping the attention grid much earlier in the denoising process. By dynamically modulating the magnitude of rotary embeddings, our method preserves semantic locality and entity consistency that uniform scaling methods fail to maintain.

6 Experiments

Experimental Settings. We evaluated our proposed method, **SEGA** on both Flux [1] and Qwen [2]. Throughout the paper, we use NTK [35] as the default length extrapolation method for SEGA, unless explicitly stated otherwise. Across all experiments, we set γ to 1.5 and κ to 0.08.

Table 1: Comparison of SEGA against state-of-the-art baselines on Flux across four high-resolution settings on Aesthetic-4K [20]. Best and second-best results are shown in **bold** and underlined.

Method	2048 × 4096						4096 × 2048					
	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓
Base	0.39	21.64	27.32	52.96	158.86	67.31	-0.74	20.61	26.99	50.92	173.18	70.65
HiFlow	1.14	22.61	27.72	44.30	152.24	69.23	0.89	22.31	28.00	43.45	169.60	69.35
I-Max	1.11	22.71	27.60	32.58	157.82	68.21	0.87	22.54	28.30	38.71	162.30	66.92
ScaleDiff	1.17	22.84	29.03	**55.84**	157.62	72.21	**0.97**	22.51	28.58	**56.85**	162.38	77.65
PI	-0.83	19.77	22.87	38.73	235.83	182.50	-1.10	19.50	23.05	37.48	225.65	179.96
NTK	0.80	21.96	27.75	48.32	157.21	66.86	0.54	22.01	27.93	48.90	156.22	61.84
YaRN	0.97	22.63	28.48	52.30	156.97	83.35	0.30	21.82	27.99	52.10	154.44	76.42
DyPE	1.10	22.90	28.87	53.35	159.81	85.12	0.53	22.15	28.26	52.85	158.49	85.39
UltraImage	0.82	22.41	28.57	55.40	157.33	**59.19**	0.60	21.99	28.84	55.13	157.97	63.11
SEGA	**1.21**	**22.91**	**29.18**	53.65	**151.93**	64.54	0.86	**22.58**	**28.99**	53.30	**153.10**	**55.85**

Method	3072 × 3072						4096 × 4096					
	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓
Base	0.49	23.00	28.18	49.93	162.25	68.63	-0.72	20.31	25.34	26.52	183.33	121.93
HiFlow	1.26	23.22	28.48	43.45	154.32	64.72	1.26	23.17	28.40	33.37	155.56	60.70
I-Max	1.28	23.15	28.42	35.23	151.98	64.56	1.26	23.10	28.45	26.36	151.74	64.06
ScaleDiff	1.30	23.22	28.73	**53.58**	153.73	72.92	1.23	23.16	28.64	44.51	153.05	76.14
PI	0.19	21.03	25.64	48.61	200.04	155.60	-0.28	20.54	24.16	29.09	208.48	202.01
NTK	0.90	22.47	28.27	35.99	156.10	63.25	-0.29	20.82	25.52	23.85	182.29	138.61
YaRN	1.11	22.89	29.10	50.49	157.04	84.31	0.88	22.21	28.30	42.87	160.48	98.52
DyPE	1.21	23.15	29.17	51.71	156.91	82.45	1.01	22.56	28.79	43.23	156.21	97.88
UltraImage	1.17	22.65	29.30	53.56	155.92	61.61	0.61	21.74	28.16	43.61	167.04	63.91
SEGA	**1.30**	**23.26**	**29.30**	52.97	**151.08**	**43.86**	**1.26**	**23.18**	**29.22**	**45.73**	**150.05**	**51.28**

Baselines. We evaluated SEGA across both the Flux [1] and Qwen [2] architectures. We compared our method against two primary categories: direct inference techniques (NTK [35], YaRN [22], DyPE [15], and UltraImage [14]), multi-stage guidance approaches (HiFlow [5], I-Max [6], and ScaleDiff [31]). Note that the multi-stage guidance methods are exclusively evaluated on Flux to align with their official implementations. See Appendix F for additional methods.

Evaluation. We used prompts and reference images from the Aesthetic-4K [20] dataset. We also curated a "Zero-Shot" benchmark comprising detailed prompts generated by an LLM, with results provided in the Table 5. Quantitative experiments are conducted across four high-resolution configurations: 2048×4096, 4096×2048, 3072^2, and 4096^2. We evaluate image quality using FID [39] and the reference-free metrics MUSIQ (MSQ) [40], and CLIP-IQA (CQA) [41]. Semantic alignment is measured by CLIP Score (CS) [42, 43], while joint alignment and human-preferred visual quality are assessed using ImageReward (IR) [44], PickScore (PS) [45], and HPSv2 [46].

6.1 Comparison to State-of-the-Art Methods

Qualitative comparison. When extrapolated to high resolutions, current direct-inference methods (e.g., YaRN [22], DyPE [15], and UltraImage [14]) often suffer from severe structural degradation, visual artifacts, and semantic omissions. As shown in Figure 5, SEGA better preserves global structural coherence, fine-grained semantic fidelity, and overall visual quality across both the Flux and Qwen architectures, even for complex prompts.

Quantitative comparison. As shown in Table 1 and Table 2, SEGA establishes a new state-of-the-art for high-resolution image generation across both the Flux and Qwen architectures. On the Flux model, SEGA consistently achieves the highest semantic alignment and image quality across different settings. The evaluation on the Qwen model further validates these findings. Notably, at the

Table 2: Quantitative comparison on Qwen across all four resolutions on Aesthetic-4K [20].

Resolution	Method	IR↑	PS↑	HPS↑	CS↑	MSQ↑	CQA↑	FID↓	FID_p↓
2048 × 4096	Base	0.60	21.77	0.25	28.17	45.05	0.64	156.53	60.33
	DyPE	1.07	22.20	0.27	28.60	47.44	0.65	157.86	75.97
	UltraImage	0.90	21.61	0.24	28.66	47.21	0.63	158.97	**41.77**
	SEGA	**1.50**	**23.63**	**0.30**	**29.75**	**50.52**	**0.72**	**149.49**	53.07
4096 × 2048	Base	-0.40	20.84	0.23	27.70	42.93	0.63	191.89	70.43
	DyPE	0.44	21.39	0.26	28.73	48.06	0.65	159.38	81.79
	UltraImage	0.46	21.04	0.23	28.21	46.11	0.63	164.06	**40.16**
	SEGA	**1.27**	**23.09**	**0.29**	**29.94**	**51.23**	**0.74**	**147.51**	51.10
3072 × 3072	Base	0.37	23.17	0.25	28.03	43.07	0.65	162.39	58.61
	DyPE	1.13	22.38	0.28	29.03	47.12	0.70	151.71	66.44
	UltraImage	1.04	22.08	0.26	29.00	47.92	0.66	153.50	**35.80**
	SEGA	**1.49**	**23.67**	**0.32**	**29.97**	**47.95**	**0.72**	**150.31**	46.22
4096 × 4096	Base	-0.10	20.95	0.21	27.82	28.48	0.54	174.00	71.25
	DyPE	0.97	22.04	0.27	29.08	37.39	0.63	159.78	66.00
	UltraImage	0.81	21.53	0.27	28.55	34.74	0.59	167.04	65.90
	SEGA	**1.51**	**23.84**	**0.33**	**30.12**	**45.03**	**0.74**	**148.26**	**65.72**

Table 3: Ablation study on Flux at 4096×4096 resolution on Aesthetic-4K. [20]

Method	IR↑	PS↑	HPS↑	CS↑	MSQ↑	CQA↑	FID↓
NTK + Fixed Scaling	0.66	22.10	0.270	28.38	42.21	0.690	162.92
SEGA W/ Axis-only	1.15	**23.05**	0.290	28.86	44.08	0.701	152.98
SEGA W/ Global-only	1.13	22.89	0.286	28.81	43.55	0.671	153.58
YaRN + SEGA	0.94	22.21	0.280	28.23	44.36	0.707	159.80
DyPE + SEGA	1.18	22.78	0.287	28.84	45.18	0.720	154.51
SEGA	**1.26**	22.35	**0.291**	**29.22**	**45.72**	**0.725**	**150.05**

4096^2 resolution, SEGA outperforms all baseline models across every evaluated metric, setting a new benchmark for high-resolution generation.

Beyond overall image quality, SEGA exhibits robustness and consistency across a diverse range of high resolutions, including non-square aspect ratios. While other models experience significant performance drops as the resolution increases, SEGA maintains highly stable results. This shows that SEGA extends generation capabilities far beyond the training resolutions of the base models.

6.2 Ablation Study

To validate our design choices, we conduct a comprehensive ablation study on the Flux architecture at the 4096^2 resolution, as detailed in Table 3. First, we evaluate the core necessity of dynamic spectral guidance by comparing SEGA against the same baseline using NTK [35] but with fixed scaling. SEGA yields substantial improvements across all metrics, confirming that fixed scaling fails to maintain structural integrity at extreme resolutions. Next, we ablate the design of our guidance mechanism by restricting SEGA to either *Axis-only* or *Global-only* scaling. While applying either axis-specific scaling or global scaling independently provides substantial improvements over the baseline, both fall short of the complete method. Finally, we ablate our default choice of NTK [35] as the base length extrapolation method by substituting it with YaRN [22] and DyPE [15].

7 Conclusion

We presented SEGA, a training-free method for high-resolution extrapolation in DiTs that adapts RoPE components scaling to the spectral structure of the current latent. By making attention scaling frequency-aware across RoPE components, SEGA addresses a key limitation of existing uniform scaling strategies, which often trade off global coherence against fine-detail fidelity. This simple modification requires no retraining or architectural changes, yet consistently improves structure, semantics, and visual quality across resolutions and model architectures. More broadly, frequency-aware attention scaling may also benefit video and other modalities where resolution extrapolation remains challenging. We hope the spectral perspective guidance introduced here motivates further research on modifying attention behavior, particularly for resolution extrapolation, to better unlock the capacity of pretrained generative models.

References

[1] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[2] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[4] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023.

[5] Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025.

[6] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow, 2024.

[7] Luigi Sigillo, Shengfeng He, and Danilo Comminiello. Latent wavelet diffusion for ultra-high-resolution image synthesis. *arXiv preprint arXiv:2506.00433*, 2025.

[8] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

[9] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European conference on computer vision*, pages 39–55. Springer, 2024.

[10] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6159–6168, 2024.

[11] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[12] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36:70847–70860, 2023.

[13] Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 4338–4346, 2025.

[14] Min Zhao, Bokai Yan, Xue Yang, Hongzhou Zhu, Jintao Zhang, Shilong Liu, Chongxuan Li, and Jun Zhu. Ultraimage: Rethinking resolution extrapolation in image diffusion transformers. *arXiv preprint arXiv:2512.04504*, 2025.

[15] Noam Issachar, Guy Yariv, Sagie Benaim, Yossi Adi, Dani Lischinski, and Raanan Fattal. Dype: Dynamic position extrapolation for ultra high resolution diffusion. *arXiv preprint arXiv:2510.20766*, 2025.

[16] Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*, 2024.

[17] Liang Hou, Cong Liu, Mingwu Zheng, Xin Tao, Pengfei Wan, Di Zhang, and Kun Gai. Boosting resolution generalization of diffusion transformers with randomized positional encodings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4762–4770, 2026.

[18] Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16893–16903, 2025.

[19] Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024.

[20] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23464–23473, 2025.

[21] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[22] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[24] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua M Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[25] Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7569–7579, 2024.

[26] Moayed Haji-Ali, Willi Menapace, Ivan Skorokhodov, Arpit Sahni, Sergey Tulyakov, Vicente Ordonez, and Aliaksandr Siarohin. Improving progressive generation with decomposable flow matching. *arXiv preprint arXiv:2506.19839*, 2025.

[27] Jinho Jeong, Sangmin Han, Jinwoo Kim, and Seon Joo Kim. Latent space super-resolution for higher-resolution image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2355–2365, 2025.

[28] Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3944–3953. IEEE, 2025.

[29] Zhihang Lin, Mingbao Lin, Meng Zhao, and Rongrong Ji. Accdiffusion: An accurate method for higher-resolution image generation. In *European Conference on Computer Vision*, pages 38–53. Springer, 2024.

[30] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European conference on computer vision*, pages 196–212. Springer, 2024.

[31] Sungho Koh, SeungJu Cha, Hyunwoo Oh, Kwanyoung Lee, and Dong-Jin Kim. Scalediff: Higher-resolution image synthesis via efficient and model-agnostic diffusion. *arXiv preprint arXiv:2510.25818*, 2025.

[32] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

[33] Jikun Hu, Dongsheng Guo, Yuli Liu, Qingyao Ai, Lixuan Wang, Xuebing Sun, Qilei Zhang, Quan Zhou, and Cheng Luo. Pepe: Long-context extension for large language models via periodic extrapolation positional encodings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21075–21085, 2025.

[34] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[35] Bowen Peng and Jeffrey Quesnelle. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation, 2023.

[36] Min Zhao, Hongzhou Zhu, Yingze Wang, Bokai Yan, Jintao Zhang, Guande He, Ling Yang, Chongxuan Li, and Jun Zhu. Ultravico: Breaking extrapolation limits in video diffusion transformers. *arXiv preprint arXiv:2511.20123*, 2025.

[37] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025.

[38] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024.

[39] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[40] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[41] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2555–2563, 2023.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[43] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021.

[44] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[45] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

[46] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[47] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Appendix

A Detailed Related Work and Preliminaries

A.1 High-Resolution Image Synthesis

Training-Based Approaches An orthogonal line of work addresses high-resolution synthesis through fine-tuning on curated high-resolution data. Diffusion-4K [20] fine-tunes latent diffusion models on a dedicated 4K dataset using wavelet-based supervision to reinforce high-frequency fidelity, achieving strong perceptual quality at the cost of retraining and reduced architectural generalizability. Latent Wavelet Diffusion (LWD) [7] takes a lighter approach, introducing frequency-aware training objectives, including a scale-consistent VAE loss and spatially adaptive denoising supervision guided by wavelet energy maps. While these methods highlight the value of frequency-domain supervision during training, they remain tied to the fine-tuning regime and do not generalize to arbitrary unseen models or resolutions at inference time.

Training-Free Methods: U-Net Architectures Training-free high-resolution generation has been studied extensively in U-Net-based latent diffusion models. DemoFusion [10] extends pretrained models beyond their native resolution using progressive upscaling, skip residuals, and dilated sampling. FreeScale [18] introduces scale fusion with selective frequency extraction, FreCaS [19] uses frequency-aware cascaded sampling, ScaleCrafter [11] exploits dilated convolutions at inference, DiffuseHigh [13] incorporates wavelet-domain guidance, and FouriScale [30] applies Fourier-domain frequency rescaling to suppress repetitive patterns. These methods show that high-resolution generation can be improved at inference time, but their mechanisms are closely tied to U-Net-style pipelines with convolutional feature maps, decoder stages, and skip connections. SEGA instead targets RoPE-based diffusion transformers, where resolution extrapolation is governed by attention over expanded latent token grids rather than explicit multi-scale feature hierarchies.

Training-Free Methods: Diffusion Transformers Training-free methods for DiT-based high-resolution generation generally fall into two categories: *direct inference* and *multi-stage guidance* approaches. Multi-stage methods condition high-resolution sampling on guidance extracted from a base-resolution generation. I-Max [6] uses projected flows derived from native-resolution generation to stabilize coarse structure formation. HiFlow [5] extends this idea by constructing a virtual reference flow from the full low-resolution trajectory, providing initialization, direction, and acceleration guidance. ScaleDiff [31] follows a similar cascade paradigm, combining upsample–diffuse–denoise refinement with patch-level attention and latent frequency mixing. While these methods provide strong structural priors, they also tie output quality to the fidelity of the base-resolution generation.

A.2 RoPE-Based Length Extrapolation Methods

RoPE-based extrapolation is the line of work most closely related to SEGA. As reviewed in Section 3, existing methods modify the RoPE schedule θ_d, the attention scaling, or both. Below, we summarize how these strategies extend to the 2D spatial setting of image generation.

Two-Dimensional Extrapolation Structure. For image generation, RoPE is applied axially [38], with separate rotary schedules for the height and width components of each token. Let $s_H = L^{(H)}_{\text{target}}/L^{(H)}_{\text{train}}$ and $s_W = L^{(W)}_{\text{target}}/L^{(W)}_{\text{train}}$ denote the per-axis extrapolation ratios.

A.2.1 Position Interpolation (PI)

Position Interpolation [34] rescales positions linearly along each axis, $n^{(a)} \mapsto n^{(a)}/s_a$ for $a \in \{H, W\}$, which is equivalent to uniformly contracting all RoPE frequencies to θ_d/s_a. This maps extrapolated positions back into the training range and reduces phase drift at long positions. However, because the same compression is applied to all dimensions, PI treats coarse long-wavelength structure and fine short-wavelength detail identically, which can weaken high-frequency positional sensitivity at large resolutions.

A.2.2 NTK

NTK [35] instead modifies the RoPE base along each axis. The original 1D rule uses

$$b' = b \cdot s_a^{D/(D-2)}, \qquad \theta'_d = (b')^{-2(d-1)/D}. \tag{8}$$

In our experiments, this correction is too weak for 2D image extrapolation, where at high resolution, the rescaled frequencies fail to provide adequate positional discrimination in attention, leading to blurred or repetitive outputs. We therefore use a stronger variant,

$$b' = b \cdot s_a^{2D/(D-2)}, \tag{9}$$

which better preserves positional contrast across the expanded 2D token grid. Unlike PI, NTK is dimension-dependent, but it remains a fixed function of s_a and d: it does not adapt to the latent content, the sample, or the denoising state.

A.2.3 YaRN

YaRN [22] refines NTK by partitioning RoPE dimensions into frequency bands and applying tailored strategies to each. Its frequency interpolation uses a smooth ramp function $\lambda_d \in [0, 1]$ to blend between PI-style interpolation and unmodified extrapolation:

$$\theta'_d = (1 - \lambda_d)\,\frac{\theta_d}{s_a} + \lambda_d\,\theta_d, \tag{10}$$

where $\lambda_d = \lambda(r_d)$ is determined by the normalized wavelength ratio $r_d = T_d / L_{\text{train}}$, with $T_d = 2\pi/\theta_d$ the wavelength of the d-th RoPE dimension:

$$\lambda(r) = \begin{cases} 0, & \text{if } r < \alpha \\ 1, & \text{if } r > \beta \\ \dfrac{r - \alpha}{\beta - \alpha}, & \text{otherwise.} \end{cases} \tag{11}$$

Although YaRN's mixed interpolation-extrapolation strategy is highly effective in the 1D setting of LLMs, we find that it does not transfer well to 2D image generation. In our experiments, YaRN frequently produces spatial structure collapse and layout confusion like objects appear in inconsistent locations, global composition breaks down, and semantically distinct regions blend together. We attribute this to YaRN's dimension-selective frequency blending: in a 1D sequence, partially interpolating high-frequency dimensions while extrapolating low-frequency ones is well-motivated by the monotonic positional structure of text. In 2D images, however, spatial structure is encoded jointly across both axes and across multiple frequency bands simultaneously, and selectively suppressing certain frequency dimensions disrupts the 2D positional geometry in ways that do not arise in the 1D case. In contrast, NTK, which rescales all dimensions consistently via the base frequency, better preserves both coarse layout and high-level spatial structure in our experiments, making it a more reliable foundation for 2D extrapolation.

YaRN further introduces a global attention temperature correction. As discussed in Section 3.1, this is written as a logit-level factor:

$$\tau(s) = \big(0.1 \ln(s) + 1\big), \tag{12}$$

which sharpens attention distributions at extended lengths to counteract the entropy collapse that arises when positional offsets grow beyond the training range.

A.2.4 DyPE

DyPE [15] makes RoPE extrapolation timestep-adaptive. Motivated by the coarse-to-fine progression of diffusion sampling, it replaces the fixed extrapolation ratio s with a timestep-dependent schedule $s(t)$ and applies it to standard RoPE corrections. For example, a Dy-NTK variant uses

$$b'(t) = b \cdot s(t)^{D/(D-2)}, \qquad \theta'_d(t) = b'(t)^{-2(d-1)/D}. \tag{13}$$

DyPE is more adaptive than PI, NTK, and YaRN; however, its adaptation is still driven by a predefined timestep schedule rather than by the observed latent of the current sample. SEGA is complementary, it also evolves during denoising, but derives its modulation directly from the current latent's spectral structure.

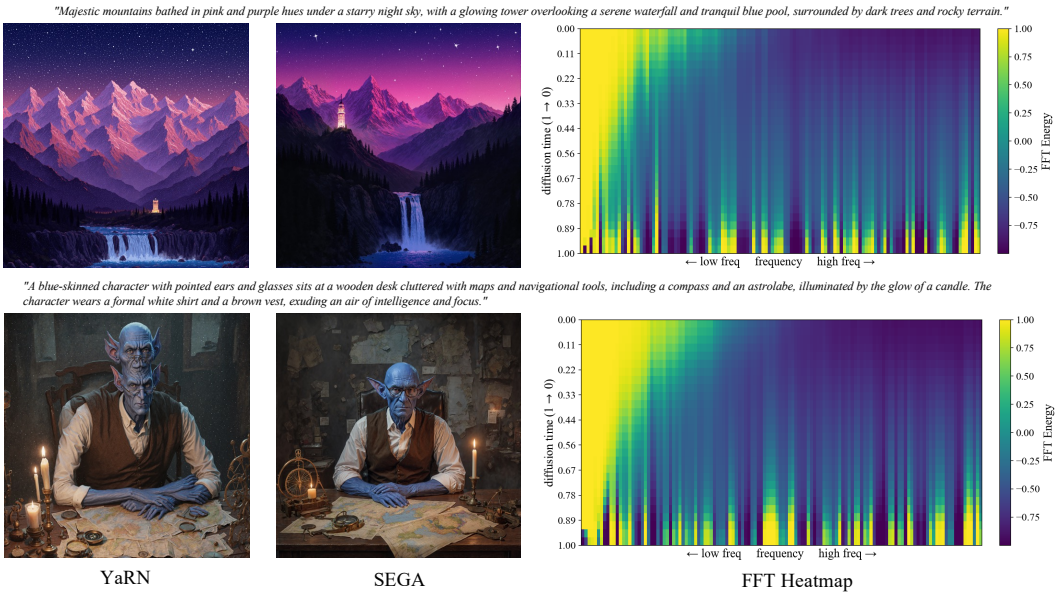


Figure 6: **Content-Aware Spectral Evolution.** The 2D power spectrum of the intermediate latents across the denoising process for two distinct prompts. The spectral energy distribution varies depending on the image content, demonstrating the necessity of a content-aware approach. Furthermore, the shifting concentration of energy, particularly in low-frequency bands where static over-scaling introduces structural artifacts (as observed in Figure 2), justifies the latent's spectral power as a dynamic guidance signal to adaptively allocate scaling, evaluated on Flux at 4096^2.

A.2.5 UltraImage

UltraImage [14] addresses two failure modes in DiT resolution extrapolation: content repetition and quality degradation. For repetition, it identifies a *dominant frequency*, a mid-band RoPE dimension whose spatial period $T_d = 2\pi/\theta_d$ aligns with the training resolution, and applies a recursive correction that reduces this frequency until its period exceeds the extrapolated extent, eliminating periodic tiling artifacts. For quality degradation, it proposes *entropy-guided adaptive attention concentration*: attention entropy $H_i = -\sum_j A_{ij} \log A_{ij}$ is computed per head and used to assign a focus factor that sharpens diffuse local attention while preserving globally concentrated patterns.

UltraImage is closely related to SEGA, as both are motivated by the view that RoPE behavior and attention degradation are central to high-resolution extrapolation in diffusion transformers. However, the two methods differ fundamentally in both diagnosis and mechanism. UltraImage identifies a discrete set of dominant RoPE frequencies whose spatial periods align with the training resolution and corrects them individually via a recursive procedure. Its attention correction is similarly discrete; an entropy score is computed per attention head and used to assign a scalar focus factor, sharpening heads that have become overly diffuse. Both corrections are therefore *sparse* and *binary* in nature. In contrast, SEGA analyzes the full spectral energy distribution of the current latent and uses it to derive a *continuous*, per-dimension scaling pattern that varies across all RoPE dimensions and both image axes. This means that every RoPE dimension receives a scaling that reflects how much spatial variation the latent currently exhibits at the corresponding frequency band, not just whether that dimension happens to coincide with a dominant period.

B Additional Analysis of Spectral-Energy Guided Attention

B.1 Content-Dependent Spectral Structure of Latent Representations

A central premise of SEGA is that the spectral energy distribution of the latent $\mathbf{Z}$ is not fixed. it varies across prompts, semantic content, and denoising timesteps, and this variation carries meaningful signal about how attention scaling should be allocated across RoPE dimensions. Figure 6 provides

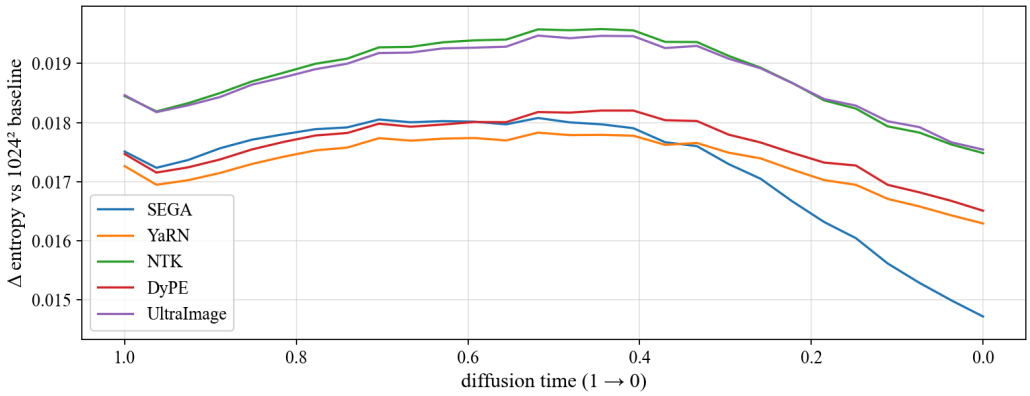


Figure 7: **Attention Entropy.** The delta of attention entropy value between different methods and the baseline image generated at 1024^2 resolution on Flux. A smaller difference indicates a closer attention structure to the baseline image generated without any RoPE extrapolation and scaling methods.

direct empirical support for this premise. Each heatmap shows the normalized 2D power spectrum of the intermediate latent tokens across the denoising trajectory, from pure noise (bottom, $t \approx 1$) to the final generated image (top, $t \approx 0$), for two prompts with markedly different visual characteristics: a landscape scene with large-scale spatial structure (top heatmap) and a portrait scene with dense local texture and fine detail (bottom heatmap).

Two observations are immediately apparent. First, the spectral energy distributions differ between the two prompts. The landscape latent develops a broader spread of energy into mid- and high-frequency bands, reflecting its detailed textures (water, foliage, rocks), whereas the portrait latent concentrates more sharply in the low-frequency region, consistent with its smoother large-scale structure. This inter-prompt variability directly motivates the content-aware design of SEGA: a fixed, globally-defined RoPE scaling, as used by YaRN and DyPE, cannot simultaneously be optimal for both spectral profiles. Applying the same frequency schedule to both prompts inevitably over-scales some bands and under-scales others, depending on where the image's structural energy actually resides.

Second, within each prompt, the spectral energy distribution evolves across the denoising trajectory. Early in denoising (bottom of each heatmap), when the latent is dominated by noise, the spectrum is highly variable across frequency bins, with no clear concentration in some specific bands. As denoising proceeds, low-frequency components emerge first and become increasingly dominant, establishing the coarse global structure of the image, while the high-frequency region remains comparatively low-energy, with its residual content varying subtly depending on the image's texture complexity. By the end of the trajectory (top of each heatmap), energy is sharply concentrated in the low-frequency region, with a smaller but content-dependent contribution in the higher bands. This temporal evolution, from an irregular noise-dominated spectrum to a structured one shaped by image content, further motivates SEGA's design of recomputing the spectral profile at each denoising step rather than fixing it at the start of sampling.

B.2 Attention Entropy Analysis

A useful signal of extrapolation quality is how closely the attention structure at high resolution resembles that of the model within its training distribution. When attention entropy deviates substantially from the baseline, the model's capacity to allocate focus appropriately is compromised, either through excessive diffusion of attention mass (high-entropy, diluted attention) or through concentration on a small number of tokens (low-entropy, collapsed attention). DyPE [15] has shown that resolution extrapolation typically induces a shift in attention entropy relative to the training distribution, and that methods which minimize this shift tend to produce higher-quality outputs.

Figure 7 reports the delta attention entropy, the difference in mean attention entropy between each extrapolation method and the baseline Flux model operating at its native 1024^2 resolution as a

function of the denoising timestep, averaged across different seeds, prompts from Aesthetic-4K [20], and all attention layers and heads. All methods are evaluated at 4096^2 resolution.

B.3 Additional Attention Evolution Results

To further illustrate how SEGA's content-aware scaling affects attention behavior at the token level, Figure 8 extends the attention map analysis from Section 5 to additional spatial locations, specifically the top-center, middle-left, and bottom-center latent tokens, comparing YaRN and SEGA across multiple denoising steps at 4096^2 resolution, consistent with the findings reported for the center token in Section 5. The consistency of this behavior across spatially diverse token positions, covering the corners, edges, and interior of the latent grid, confirms that SEGA's improvements are not localized to a particular region of the image but reflect a global improvement in attention structure throughout the high-resolution token grid.

C Additional Implementation Details

All image generation experiments were conducted using the Flux and Qwen diffusion transformer architectures. For the Flux model, we specifically utilized the `dev.Krea` checkpoint. To maintain high numerical precision without incurring unnecessary memory overhead, all model weights and latent activations were cast to `bfloat16`. The experiments, including both standard generation and high-resolution extrapolation, were executed on NVIDIA H100 GPUs. Because SEGA operates entirely at inference time and requires no parameter updates, we did not employ any training or fine-tuning infrastructure. We followed the standard inference settings provided by the official model implementations of Flux and Qwen, using their default samplers, number of denoising steps, and guidance scales. SEGA was applied at every denoising step throughout the entire trajectory, with no warmup, scheduling, or step-dependent gating beyond what is induced naturally by the spectral flatness factor.

D Limitation and Discussion

While SEGA enables stable high-resolution synthesis well beyond the native training regime, it has several limitations. First, SEGA modulates the magnitude of rotary embeddings but does not extend RoPE's positional range; it is therefore composed with an underlying length-extrapolation method (NTK in our experiments) and partially inherits its structural priors. Second, although SEGA can scale up to 8192^2, perceptual quality continues to degrade at the most extreme extrapolation factors, where the limitation is the model's intrinsic capacity rather than the positional encoding alone. Third, while SEGA itself is computationally negligible, generating multi-megapixel images remains expensive: the underlying transformer's attention cost grows quadratically with the number of tokens, making ultra-high-resolution synthesis demanding regardless of which extrapolation method is used. More broadly, SEGA shows that the latent's spectral structure can serve as a useful signal for guiding RoPE scaling at inference time, and we hope the coupling it reveals between RoPE dimensions and spatial frequencies inspires future work on inference-time adaptation of pretrained generative models.

E Societal Impact and Safeguards

Generative modeling, particularly for images and videos, has substantial potential for both beneficial and harmful use. Improvements in high-resolution generation can support creative workflows, design, visualization, and research by enabling more realistic and detailed synthesis without additional training. At the same time, increased realism may heighten risks of misuse, including disinformation, impersonation, non-consensual synthetic imagery, and amplification of existing social biases. Although SEGA does not introduce a new generative model, dataset, or training procedure, it improves the inference-time capabilities of existing text-to-image systems and may therefore amplify risks already associated with those systems. SEGA does not introduce new model-level safeguards or safety filters. Its responsible use therefore depends on the licenses, acceptable-use policies, access controls, and safety mechanisms of the underlying models and deployment platforms. In this work, we evaluate SEGA on existing models such as Flux and Qwen for research purposes. Black Forest Labs states that its Flux models and services are governed by usage policies and responsible-AI safeguards, while

Table 4: Comparison of SEGA against state-of-the-art baselines on SDXL [47] and Diffusion-4K across four high-resolution settings on Aesthetic-4K [20]. Best and second-best results are shown in **bold** and underlined.

Method	2048 × 4096						4096 × 2048					
	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓
Diffusion-4K	0.71	21.89	27.94	35.76	156.68	**55.95**	0.48	21.89	27.92	36.97	154.68	**55.03**
DemoFusion	0.41	22.03	28.73	48.39	165.57	77.62	-0.30	21.56	27.27	**53.46**	169.48	81.01
FreCas	0.93	22.34	28.89	**55.21**	156.14	100.80	0.47	21.62	28.73	53.05	158.95	99.78
FreeScale	0.89	22.23	28.78	50.94	162.86	98.66	0.69	21.92	28.97	44.61	170.32	94.87
DiffuseHigh	0.73	22.44	28.29	45.44	158.73	82.73	0.49	22.18	28.98	48.03	160.31	77.27
SEGA	**1.21**	**22.91**	**29.18**	53.65	**151.93**	64.54	**0.86**	**22.58**	**28.99**	53.30	**153.10**	55.85
Method	3072 × 3072						4096 × 4096					
	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓	IR↑	PS↑	CS↑	MSQ↑	FID↓	FID_p↓
Diffusion-4K	0.87	22.29	28.28	34.49	154.83	55.79	0.52	21.74	27.48	24.13	161.62	70.05
DemoFusion	0.89	23.00	29.35	51.45	153.65	76.56	0.88	22.99	29.45	41.27	156.94	72.86
FreCas	1.13	23.12	29.20	51.70	155.98	91.26	1.06	22.90	**29.82**	39.32	156.57	82.27
FreeScale	1.05	22.78	**29.74**	43.63	170.63	86.80	1.06	22.82	29.75	33.94	167.83	74.72
DiffuseHigh	0.91	23.17	29.59	49.48	159.33	73.97	0.93	23.16	29.21	36.44	160.43	63.19
SEGA	**1.30**	**23.26**	29.29	**51.89**	**151.08**	**43.86**	**1.26**	**23.18**	29.22	**45.72**	**150.05**	**51.28**

Table 5: Quantitative comparison at 4096^2 resolution on the zero-shot benchmark. Methods are grouped by backbone model; best and second-best results are **bolded** and underlined within each group. † denotes a closed-source proprietary model.

Backbone	Method	IR↑	PS↑	CS↑	HPS↑	MSQ↑
Flux	Base	-1.50	19.41	22.17	0.23	25.83
	NTK	-0.30	20.66	25.36	0.25	29.23
	YaRN	0.70	21.83	28.95	0.27	41.18
	DyPE	0.78	22.01	**29.65**	0.27	41.62
	UltraImage	0.42	21.36	28.35	0.28	39.29
	SEGA	**1.05**	**22.50**	29.26	**0.28**	**42.36**
Qwen	Base	0.13	21.33	28.89	0.27	30.90
	DyPE	1.13	22.58	29.55	0.29	39.10
	UltraImage	0.71	21.60	29.62	0.27	35.71
	SEGA	**1.58**	**23.86**	**30.06**	**0.30**	**45.27**
Prop.†	Nano Banana 2	1.37	23.43	30.02	0.30	42.03

Qwen provides a usage policy for its AI products and services [1, 2]. We therefore recommend using SEGA only in ways consistent with the underlying models' licenses and usage policies, together with appropriate content moderation, provenance, and misuse-monitoring mechanisms when deployed.

F Additional Quantitative Results

F.1 Generalization to Alternative Backbones

To assess the generalizability of SEGA beyond Flux-based models, Table 4 reports quantitative results on an alternative backbone across four high-resolution settings on Aesthetic-4K [20]. We compare against a broad set of state-of-the-art baselines, including methods built on SDXL [47], as well as Diffusion-4K, which relies on model fine-tuning. The baselines include fine-tuning (Diffusion-4K) and multi-stage guidance (DemoFusion, FreCas, FreeScale, DiffuseHigh). SEGA consistently achieves the best or second-best performance across the majority of metrics and resolution settings,

Table 6: Quantitative comparison at 5120^2 resolution on Aesthetic-4K [20]. Methods are grouped by backbone model; best and second-best results are **bolded** and underlined within each group.

Backbone	Method	IR↑	PS↑	CS↑	HPS↑	MSQ↑	CQA↑	FID↓
Flux	Base	-2.03	18.22	16.43	0.21	25.72	0.37	351.73
	NTK	-0.05	21.31	24.53	0.25	27.96	0.48	244.52
	YaRN	0.40	21.53	26.74	0.26	40.11	0.61	235.33
	DyPE	0.58	21.85	27.98	0.26	40.40	0.66	225.30
	UltraImage	0.24	21.34	26.39	0.27	36.25	**0.74**	251.02
	SEGA	**1.13**	**23.22**	**28.92**	**0.29**	**40.64**	0.72	**221.99**
Qwen	Base	-0.54	20.08	25.52	0.24	21.90	0.48	270.94
	DyPE	-0.37	20.34	24.72	0.24	20.58	0.41	250.45
	UltraImage	0.18	20.88	26.13	0.25	23.13	0.49	255.24
	SEGA	**1.39**	**23.94**	**29.55**	**0.30**	**41.62**	**0.74**	**218.47**

Table 7: Quantitative comparison at 6144^2 resolution on Aesthetic-4K [20]. Methods are grouped by backbone model; best and second-best results are **bolded** and underlined within each group.

Backbone	Method	IR↑	PS↑	CS↑	HPS↑	MSQ↑	CQA↑	FID↓
Flux	Base	-2.25	17.06	12.48	0.19	25.77	0.31	453.97
	NTK	-2.03	17.11	10.83	0.19	23.79	0.34	528.45
	YaRN	-0.33	20.18	24.16	0.24	35.28	0.56	288.66
	DyPE	-2.23	16.78	11.68	0.18	27.65	0.34	274.82
	UltraImage	-0.74	20.27	25.29	0.26	36.15	**0.69**	290.75
	SEGA	**0.75**	**22.47**	**27.92**	**0.27**	**43.43**	0.65	**232.18**
Qwen	Base	-0.95	19.78	23.96	0.23	22.50	0.44	279.78
	DyPE	-0.88	19.64	23.75	0.23	36.15	0.35	279.77
	UltraImage	-0.47	20.14	23.94	0.24	21.89	0.43	254.09
	SEGA	**1.36**	**23.97**	**28.75**	**0.30**	**38.77**	**0.74**	**210.21**

demonstrating that its spectral-energy-guided scaling transfers effectively across different model architectures without any architecture-specific tuning.

F.2 Zero-Shot Benchmark

A potential concern with evaluating on Aesthetic-4K [20] is that some models, particularly those with large-scale pretraining data, may have encountered images from this dataset during training, which could favor their performance on distribution-specific metrics. To mitigate this risk and assess generalization, we construct a dedicated ***zero-shot*** benchmark.

Specifically, we use an LLM to generate 200 curated, high-detail prompts covering a diverse range of scenes, lighting conditions, subjects, artistic styles, and compositional structures, with care taken to minimize overlap with the Aesthetic-4K dataset. This benchmark is designed to evaluate whether performance differences observed on Aesthetic-4K reflect genuine generalization capability or are partly attributable to dataset familiarity. We additionally include Nano Banana 2 [48], a closed-source proprietary model, in this evaluation as a reference point for the performance ceiling achievable by large-scale commercial systems.

Table 5 reports results on this zero-shot benchmark at 4096^2 resolution. SEGA achieves the best performance across all metrics on both the Flux and Qwen backbones. Notably, SEGA on the Qwen backbone achieves an ImageReward score of 1.58 and a PickScore of 23.86, approaching and in some

metrics matching or even better than the performance of Nano Banana 2 (IR: 1.37, PS: 23.43), which represents a strong closed-source commercial baseline.

F.3 Extreme Resolution: 5120^2 and 6144^2

Tables 6 and 7 extend the main evaluation to extreme resolutions of 5120^2 and 6144^2, corresponding to approximately 26 and 38 million pixels respectively, resolutions that represent a $25\times$ and $36\times$ area extrapolation factor beyond the 1024^2 training resolution of Flux. Due to the time and cost of generation at these scales, we evaluate on a randomly selected subset of 20 prompt–image pairs from Aesthetic-4K. The results show that SEGA remains substantially more consistent as resolution increases, while competing methods degrade significantly under stronger extrapolation. Its superiority is most pronounced at ultra-high resolutions, where it achieves the strongest overall performance while better preserving structural coherence and semantic fidelity.

G Additional Qualitative Results

As shown in Figure 9, SEGA performs consistently across both vertical and horizontal aspect ratios. This figure compares YaRN, DyPE, UltraImage, and SEGA on both Flux and Qwen, demonstrating that SEGA preserves the intended image geometry without stretching or distorting objects along either spatial axis. The generated images remain sharp and visually coherent, while also maintaining strong alignment with the input prompts.

On the zero-shot prompt set, we compare YaRN, DyPE, UltraImage, and SEGA on both Flux and Qwen, shown in Figure 10. The results show that SEGA avoids common high-resolution failure modes such as repeated structures, distorted layouts, and loss of semantic clarity. In particular, SEGA maintains high prompt fidelity and fine-grained visual detail without sacrificing global composition or overall image realism.

We further compare SEGA against guidance-based high-resolution approaches, as described in Appendix A. As shown in Figure 11, we compare against ScaleDiff, I-Max, and HiFlow. These guidance-based methods often improve resolution by relying on an upsampled or guided low-resolution generation, which can preserve coarse structure but may leave artifacts, uneven detail, or inconsistencies between foreground and background regions. In contrast, SEGA directly improves the high-resolution denoising process, allowing both the main subject and the surrounding scene to benefit from the same content-aware attention scaling. This leads to more realistic image components, clearer local textures, and more coherent global structure.

At higher resolutions such as 5K and 6K, SEGA continues to provide clear benefits in visual sharpness, structural consistency, and prompt alignment, as shown in Figures 12 and 13. These figures compare SEGA against direct-inference baselines, including DyPE and UltraImage, and demonstrate that SEGA remains effective even in challenging extrapolation regimes where other methods may produce severe artifacts or fail to generate a coherent image. For example, in Figure 13, DyPE fails to produce a reliable output, whereas SEGA generates a clean, consistent, and prompt-aligned image, highlighting its robustness under extreme resolution extrapolation.

Finally, Figure 14 shows fine details from an ultra-high-resolution generation produced by SEGA. The zoomed-in regions illustrate that SEGA preserves local texture and object-level detail while maintaining the broader structure of the image. This suggests that SEGA's spectral-energy-guided scaling benefits both fine-scale fidelity and global coherence, rather than improving one at the expense of the other.

H Additional Ablation: Choice of Baseline Scaling m_{ref}

The reference scale m_{ref} in Eq. 4 sets the anchor magnitude of the rotary scaling shared across all RoPE dimensions. As discussed in Sec. 4.2, m_{ref} is a function of the resolution ratio $s = R_{\text{target}}/R_{\text{train}}$ between target and training images. We consider two common formulations for this design choice:

$$m_{\text{ref}}^{\text{power}} = s^{\kappa}, \qquad m_{\text{ref}}^{\log} = 1 + \kappa \log s, \tag{14}$$

where $\kappa > 0$ is a small exponent (we use $\kappa = 0.08$ in all reported experiments). Both formulations reduce to $m_{\text{ref}} = 1$ at $s = 1$ (no extrapolation) and grow monotonically with s. The two forms behave similarly in the moderate-extrapolation regime ($s \approx 1$–2), but diverge as s grows.

Why the choice matters at high s. As the target resolution increases, the token grid grows substantially, making positional offsets harder to discriminate even with RoPE extrapolation. Attention therefore becomes increasingly prone to dilution at large extrapolation factors. A larger m_{ref} acts as a stronger anchor for positional discrimination, sharpening attention more aggressively to compensate for the expanded grid. Empirically, we find that ultra-high-resolution generation (e.g., 5120^2 or 6144^2) requires a stronger anchor than moderate extrapolation, and the power-law form provides this naturally because s^κ grows faster than $1 + \kappa \log s$. Table 8 illustrates this divergence: the two forms are nearly identical at small s, but the power-law value becomes meaningfully larger as s increases.

Table 8: Values of m_{ref} produced by the two formulations as a function of the resolution ratio s. Computed with $\kappa = 0.08$. The power-law form grows faster at large s, providing a stronger positional-discrimination anchor at extreme extrapolation factors.

s	1	2	4	8	16	32
$m_{\text{ref}}^{\text{power}}$	1.000	1.057	1.118	1.182	1.249	1.320
$m_{\text{ref}}^{\log}$	1.000	1.055	1.111	1.166	1.222	1.277

Empirical comparison. We compare the two formulations under identical SEGA settings on FLUX at 4096^2, 5120^2, and 6144^2. Table 9 shows that the two forms perform similarly at 4096^2, with a small but consistent advantage for the power-law variant. The gap widens at 5120^2 and remains clear at 6144^2, where the power-law form yields lower FID and stronger alignment across most metrics. Overall, the power-law baseline extrapolates more stably as resolution increases, matching the trend in Table 8. We therefore adopt the power-law form, which grows faster than a logarithm while remaining more moderate than a linear scaling.

Table 9: Comparison of power-law and logarithmic forms for m_{ref} on Flux. SEGA hyperparameters are held constant at $\gamma = 1.5$, $\kappa = 0.08$. Best results per resolution are in **bold**.

Resolution	m_{ref} form	CS ↑	ImageReward ↑	HPS ↑	PickScore ↑	FID ↓	MUSIQ ↑
4096×4096	logarithmic	28.46	1.23	0.29	23.10	150.16	44.65
	power-law	**29.22**	**1.26**	**0.29**	**23.18**	**150.05**	**45.73**
5120×5120	logarithmic	28.41	0.80	0.28	22.93	265.01	46.36
	power-law	**28.92**	**1.13**	**0.29**	**23.22**	**221.99**	40.64
6144×6144	logarithmic	27.54	0.67	0.27	22.37	269.58	42.92
	power-law	**27.92**	**0.75**	**0.27**	**22.47**	**232.18**	**43.43**

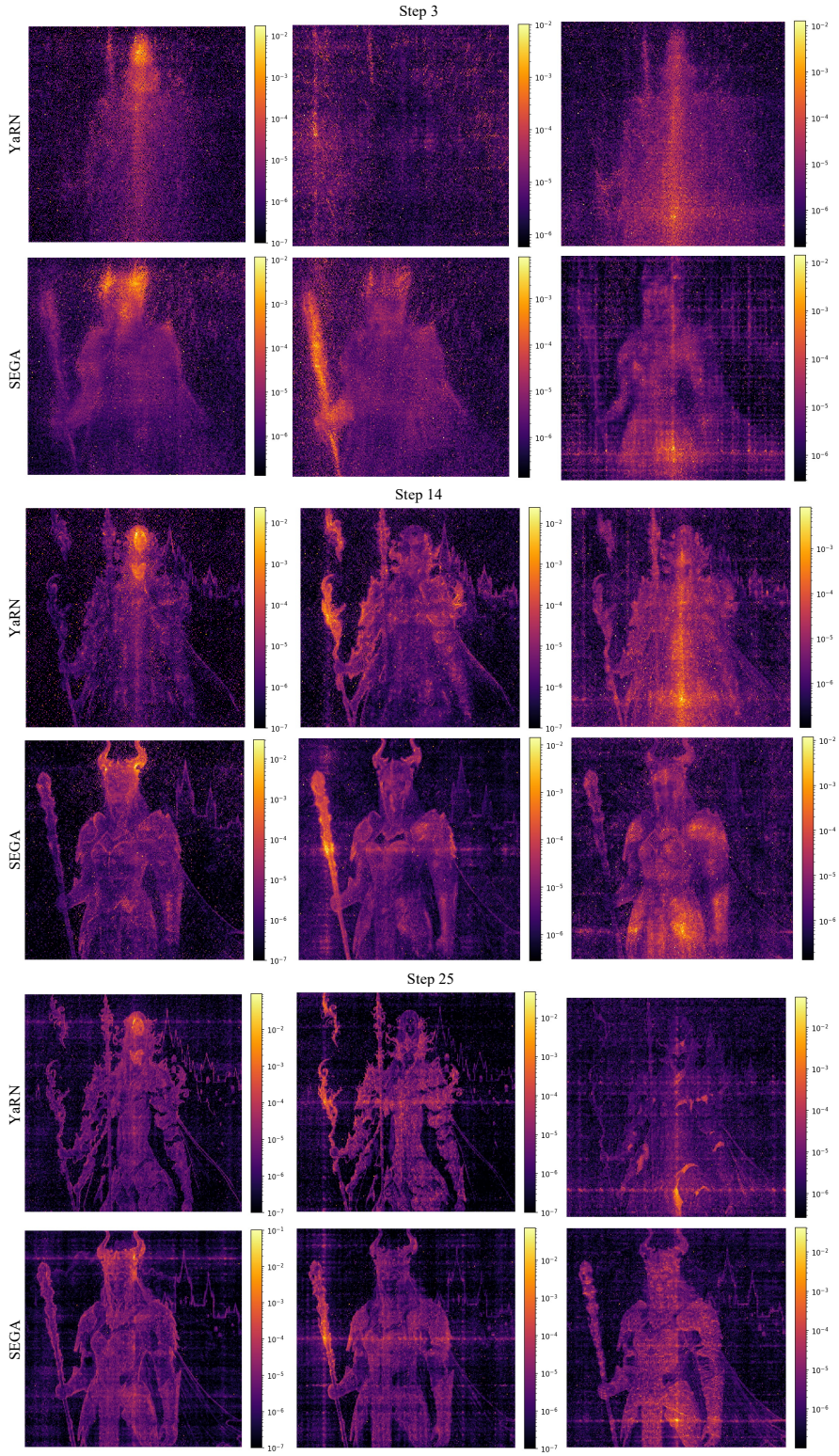


Figure 8: **Impact on Attention Evolution (Other Tokens).** Further visual comparison of attention maps for the top-center, middle-left, and bottom-center latent tokens in YaRN and SEGA across multiple denoising steps, evaluated on Flux at 4096^2.

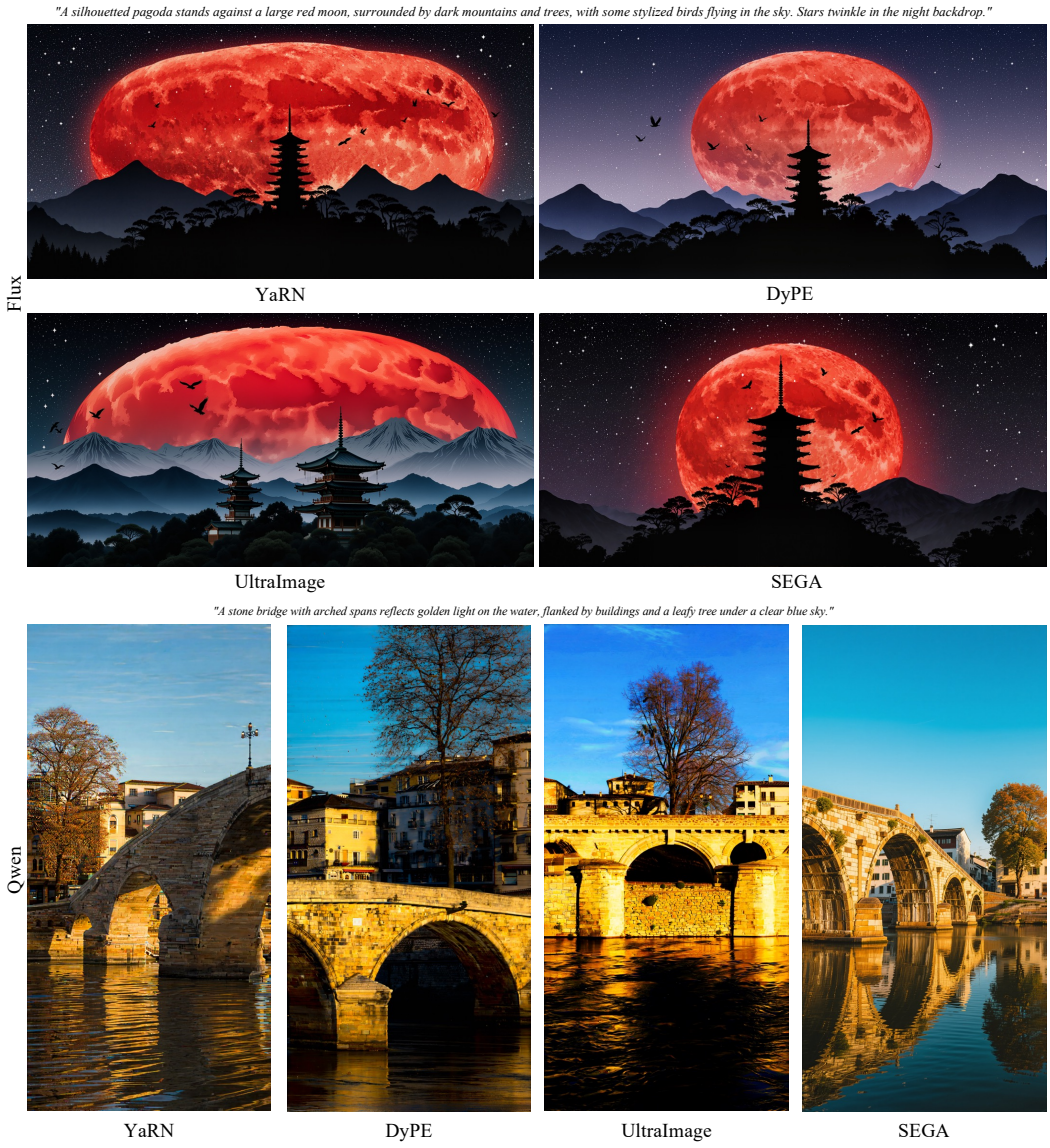


Figure 9: **Qualitative comparison (non-square resolutions).** Results on two non-square resolutions (2048×4096 and 4096×2048) on Qwen and Flux show that SEGA's ability to preserve the shape of contents in different aspect ratio.

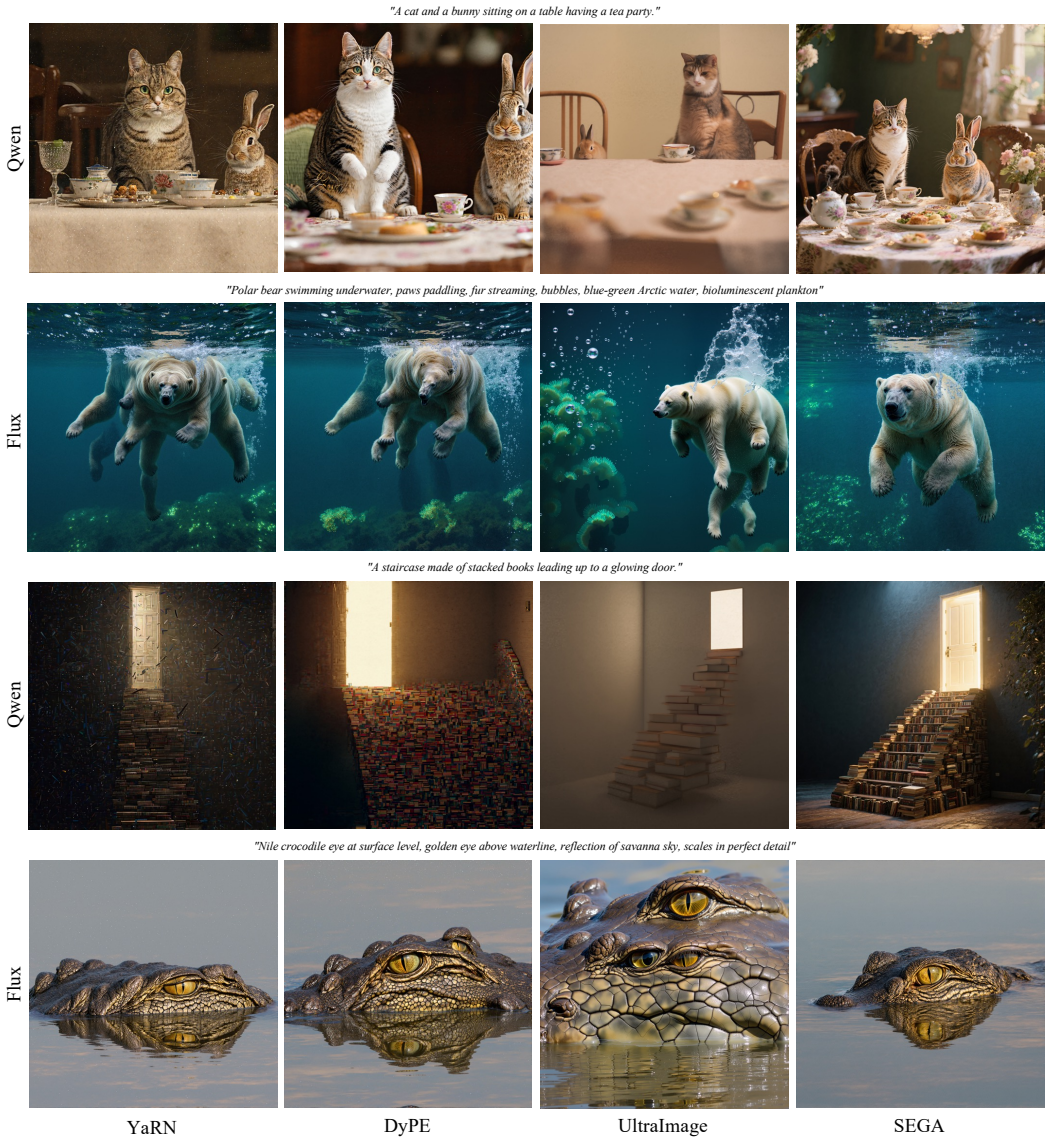


Figure 10: **Qualitative comparison (Zero-Shot Dataset).** Results on prompts from the zero-shot dataset for Qwen and Flux at 4096^2 resolution show that SEGA handles complex environments, objects and areas with reflection, contents with challenging lighting, and preserves the shapes of the objects.

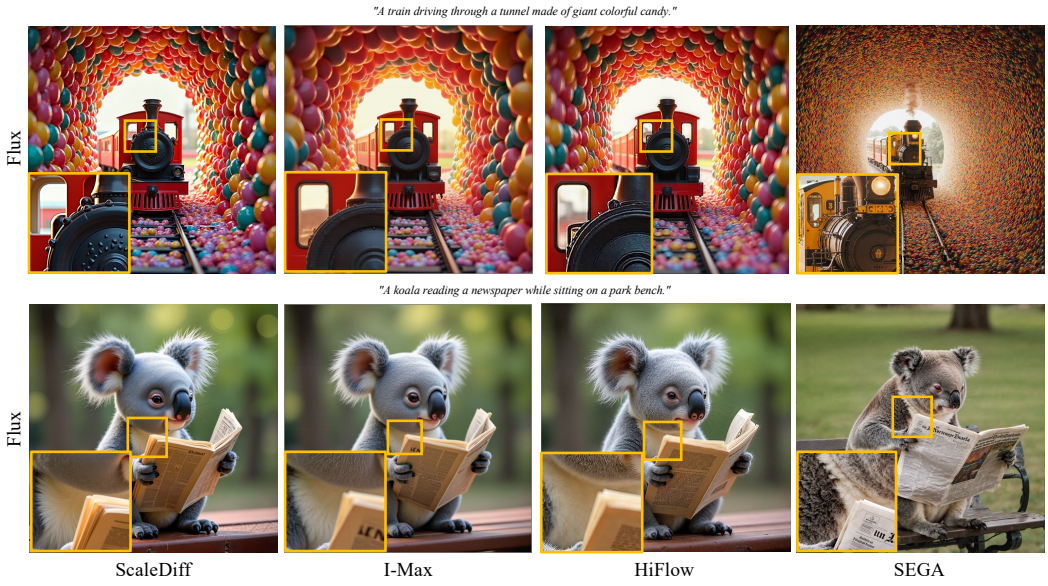


Figure 11: **Qualitative comparison (with guidance-based approaches).** Results on two representative prompts for Flux at 4096^2 resolution in comparison with top guidance-based approaches show that SEGA is not limited to the synthesized image at base resolution and provides fine details and high-quality textures.

Figure 12: **Qualitative comparison (at 5120^2 resolution).** Results on two representative prompts for Qwen and Flux at 5120^2 resolution show that SEGA elaborates on coarse and fine details as the resolution of the images increases.

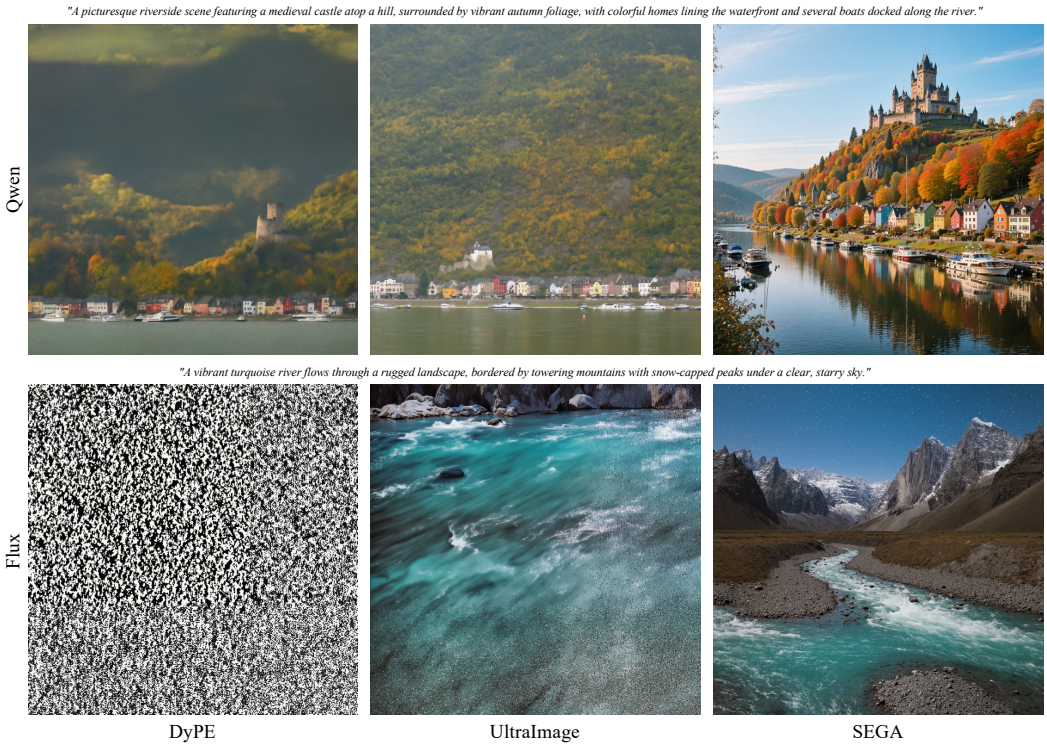


Figure 13: **Qualitative comparison (at $\mathbf{6144^2}$ resolution).** Results on two representative prompts for Qwen and Flux at 6144^2 resolution show that SEGA makes image synthesis at this resolution possible while baselines struggle with noise and collapse of global structures.

Figure 14: **Visualizing Fine-Grained Details at Extreme Resolutions.** Sample generated at 6144^2 resolution by SEGA on Qwen. The model successfully preserves high-frequency local textures and sharp structural boundaries without experiencing structural collapse or repetition artifacts typical of long-context length extrapolation.