TOKENSCALE: TRAINING-FREE HIGH-RESOLUTION IMAGE GENERATION IN TOKEN SPACE

Anonymous authors
Paper under double-blind review



Figure 1: Super-resolution images of various sizes generated by TokenScale, showcasing high clarity and rich details. More cases are provided in Fig. 18, and all prompts are listed in the appendix.

ABSTRACT

Diffusion Transformer (DiT)-based flow models achieve strong synthesis performance at training resolution, yet their extension to high-resolution (HR) generation frequently suffers from blurring and spatial layout collapse. We attribute this degradation to the token level. As resolution increases, image tokens grow quadratically while text tokens remain fixed, thereby diluting image–text attention and disrupting prompt alignment. In addition, persistent low-resolution ref-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

erence guidance can restrict the synthesis of new HR details. To address these issues, we propose TokenScale, a training-free token-level framework. TokenScale interpolates complete image-token vectors obtained from low-resolution generation, rather than individual positions on the unpacked latent grid, and renoises the expanded representation to initialize high-resolution sampling. During an initial denoising window, selective reference fusion corrects clean predictions in spatially coherent high-deviation regions, leaving unselected tokens unchanged and imposing no further reference correction afterward. This design supports structural consistency while allowing later denoising to synthesize details beyond the low-resolution content. Extensive experiments demonstrate that TokenScale substantially improves HR image generation quality, outperforms state-of-the-art training-free methods, and can be integrated with positional-encoding-based approaches for further gains.

1 INTRODUCTION

Text-to-image (T2I) diffusion and flow models (Rombach et al., 2022; Podell et al., 2023; Chen et al., 2024; Black Forest Labs, 2024; Peebles & Xie, 2023; Esser et al., 2024; Liu et al., 2023) have substantially advanced visual content creation. However, directly extending generation from their native resolution (e.g., 1024^2) to 2048^2 or 4096^2 can introduce repeated structures, blurred details, and weakened text alignment. These failures highlight resolution generalization as a central challenge for pretrained generative models. Fine-tuning on high-resolution data incurs substantial costs (Ren et al., 2024), motivating training-free approaches that extend the capabilities of existing models.

Existing approaches explore positional-encoding adaptation and coarse-to-fine generation. DyPE (Issachar et al., 2025), UltraImage (Zhao et al., 2025), and SigMa (Zhao et al., 2026) adapt spatial encodings for resolution extrapolation. Cascade methods, including DemoFusion (Du et al., 2024a), DiffuseHigh (Kim et al., 2025), I-Max (Du et al., 2024b), and HiFlow (Bu et al., 2025), reuse low-resolution outputs or trajectories to initialize and guide high-resolution sampling. The low-resolution outputs used by these methods provide useful global structure, but maintaining strong reference constraints throughout high-resolution sampling can restrict the synthesis of new details. This raises the question of how to transfer that structure effectively while limiting subsequent reference intervention.

We examine this challenge at the token level in Diffusion Transformer (DiT)-based flow models. With fixed tokenization, image-token counts grow quadratically with resolution, while text-token counts remain unchanged for a fixed prompt. The enlarged image-key set shares the softmax normalization with text keys in joint attention. On FLUX.1[dev] (Black Forest Labs, 2024), image-to-text attention mass decreases as resolution increases (Fig. 2, left), alongside dispersed spatial responses to prompt words in direct high-resolution generation (Fig. 2, right). This suggests establishing localized text–image correspondence becomes more difficult on the enlarged grid. By contrast, low-resolution generation already organizes scene content under text conditioning. This motivates transferring the resulting token representations to initialize high-resolution sampling, while using the reference for limited correction rather than continuous guidance.

In this work, we introduce TokenScale, a training-free token-space framework for high-resolution generation. Starting from the final latent of low-resolution generation, TokenScale interpolates complete image-token vectors on their spatial grid, rather than individual positions on the unpacked latent grid. The expanded representation is re-noised to initialize high-resolution sampling at a selected restart level. During an initial denoising window, selective reference fusion identifies spatially coherent deviations between clean-prediction tokens and a fixed reference, and corrects only the selected regions. No further reference correction is applied afterward, allowing later denoising to synthesize details beyond the low-resolution content. This framework applies to FLUX.1[dev] and Qwen-Image (Wu et al., 2025). Its token-space operations complement positional-encoding adaptations, allowing integration with methods such as UltraImage.

In summary, the contributions of this work are as follows: (i) we propose TokenScale, a training-free token-space framework for extending pretrained DiT-based flow models to high-resolution generation; (ii) we combine whole-token expansion with spatially selective, early reference fusion to

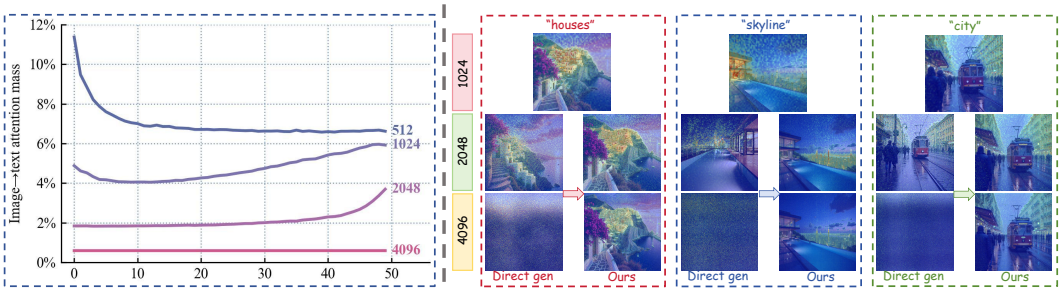

Figure 2: Attention behavior across resolutions. Left: image-to-text attention mass, i.e., the total softmax weight that each image token assigns to text tokens in joint attention, in the first three multimodal blocks, measured from 512^2 to 4096^2. Right: spatial attention maps for selected prompt words, comparing direct high-resolution generation with TokenScale.

transfer low-resolution structure without persistent reference constraints; and (iii) comprehensive experiments demonstrate the framework's superiority over state-of-the-art training-free methods and its flexibility in coupling with positional-encoding-based approaches for further quality gains.

2 Related Work

Text-to-image Models. Text-to-image generation has progressed from pixel-space diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021) to latent diffusion models (Rombach et al., 2022; Podell et al., 2023), which operate in the compressed VAE latent space. The backbones then shifted to Transformers (Peebles & Xie, 2023; Chen et al., 2024), where image latents are patchified into token sequences. Recent large-scale models train these backbones with flow-matching objectives (Lipman et al., 2023; Liu et al., 2023): SD3 (Esser et al., 2024) couples rectified-flow training with a multimodal DiT, and FLUX.1[dev] (Black Forest Labs, 2024) and Qwen-Image (Wu et al., 2025) represent the state-of-the-art open-source family. In these models, image tokens and text tokens interact through joint attention, so the token sequence is the natural unit for studying resolution generalization.

Training-Free High-Resolution Generation. Because collecting high-resolution data and fine-tuning models are costly, training-free methods have attracted wide attention. Existing approaches mainly follow two routes. Cascade methods reuse low-resolution outputs as structural references (Du et al., 2024a; Kim et al., 2025; Du et al., 2024b; Bu et al., 2025), but persistent image- or trajectory-level guidance can over-constrain the high-resolution branch and suppress newly generated details. A complementary line adapts positional encodings or attention to improve resolution extrapolation, including DyPE (Issachar et al., 2025), UltraImage (Zhao et al., 2025), ResDiT (Ma et al., 2026), and SigMa (Zhao et al., 2026). These methods address positional extrapolation but not the starting state of the enlarged token sequence. TokenScale targets this initialization and applies only lightweight, early reference correction, so it is complementary to these approaches.

3 Method

3.1 Preliminaries

Rectified flow Liu et al. (2023) interpolates linearly between a clean latent $\mathbf{x}_0$ and independent Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$: at noise level $\sigma \in [0, 1]$, $\mathbf{x}_\sigma = (1-\sigma)\mathbf{x}_0 + \sigma\boldsymbol{\epsilon}$. Generation proceeds from noise to data: the noise schedule $\{\sigma_t\}$ decreases with t, from $\sigma_T = 1$ (pure noise) to $\sigma_0 = 0$ (clean data). At step t, let $\mathbf{x}_t$ denote the latent state at noise level σ_t and $\mathbf{c}$ the text condition. The pretrained network f_θ predicts the velocity $\mathbf{v}_t = f_\theta(\mathbf{x}_t, \sigma_t, \mathbf{c})$, and sampling follows the Euler update, for $t = T, \ldots, 1$,

$$\mathbf{x}_{t-1} = \mathbf{x}_t + (\sigma_{t-1} - \sigma_t)\mathbf{v}_t. \tag{1}$$

Under this formulation, the same velocity estimate yields an approximate clean latent $\hat{\mathbf{x}}_{0,t}$ through one-step extrapolation from the current noisy state:

$$\hat{\mathbf{x}}_{0,t} = \mathbf{x}_t - \sigma_t\mathbf{v}_t. \tag{2}$$

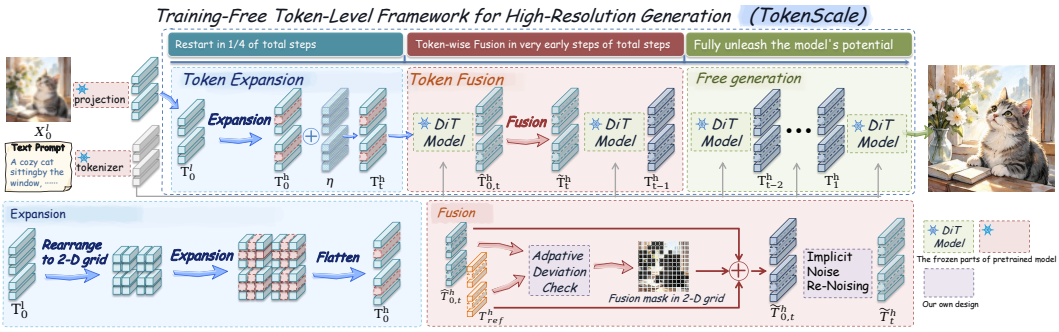

Figure 3: Overview of TokenScale: Image-token embeddings obtained from the final low-resolution latent are expanded and re-noised to initialize high-resolution sampling. Selective token fusion is applied only during an initial post-restart window.

We distinguish the VAE latent state from its image-token representation. Let H denote the latent side length, c_{ch} the channel count, and s the patch side length. Patchification $\mathcal{P}$ rearranges the latent into $N = (H/s)^2$ non-overlapping patch vectors of dimension $p = s^2 c_{\mathrm{ch}}$, with $\mathcal{P}^{-1}$ the inverse rearrangement. Let d be the image-token embedding dimension and $\mathbf{W}_{\mathrm{in}} \in \mathbb{R}^{d\times p}$, $\mathbf{b}_{\mathrm{in}} \in \mathbb{R}^d$ the pretrained input projection parameters. With $\mathbf{1}_N$ the length-N all-ones vector, the tokenization map $\mathcal{T}$ maps a latent $\mathbf{x}$ to its image-token representation in $\mathbb{R}^{N\times d}$:

$$\mathcal{T}(\mathbf{x}) := \mathcal{P}(\mathbf{x})\mathbf{W}_{\mathrm{in}}^{\top} + \mathbf{1}_N \mathbf{b}_{\mathrm{in}}^{\top}. \tag{3}$$

This affine input projection is shared across resolutions. Image tokens are the embeddings immediately after this projection, not intermediate attention outputs. After its own projection, the text condition becomes L text tokens $\mathbf{C} \in \mathbb{R}^{L\times d}$ interacting with the image tokens through joint attention. Sampling states and velocity predictions remain in the VAE latent space.

3.2 Token Expansion

The dispersed attention observed in direct high-resolution generation (Fig. 2, right) motivates transferring the semantic–spatial information established during low-resolution generation. We transfer this structure through image-token embeddings obtained from the final low-resolution latent, rather than intermediate attention features. Concretely, let $\mathbf{x}_0^{\ell}$ be the final low-resolution latent and $\mathbf{T}^{\ell} = \mathcal{T}(\mathbf{x}_0^{\ell}) \in \mathbb{R}^{N\times d}$ its tokens (Eq. (3)). Let $\mathbf{E} \in \mathbb{R}^{M\times N}$ denote the *expansion matrix* and $\mathbf{T}^{h} \in \mathbb{R}^{M\times d}$ the expanded sequence, acting directly on the flattened low-resolution tokens:

$$\mathbf{T}^{h} = \mathbf{E}\,\mathbf{T}^{\ell}, \qquad \mathbf{t}_i^{h} = \sum_{j\in\mathcal{N}(i)} E_{ij}\,\mathbf{t}_j^{\ell}, \tag{4}$$

where $\mathcal{N}(i)$ collects the neighbors of location i and E_{ij} the entry of $\mathbf{E}$. We choose $\mathbf{E}$ sparse and row-stochastic, with the i-th row supported only on $\mathcal{N}(i)$ and unit row sums. With the four bilinear weights of $\mathcal{N}(i)$ as each row's nonzeros, the expansion coincides with bilinear interpolation on the two-dimensional token grid. Unlike latent-grid interpolation, which acts across the feature dimensions of each token and destroys the carried information, token expansion applies each spatial weight to an entire token vector, thereby endowing the model with an initialization that carries semantics and features, rather than merely latent-pixel information. A direct comparison of the two initializations is provided in Appendix B.

The expanded tokens are clean, so we re-noise them to the restart noise level to continue the cascade at high resolution. Let n_s denote the restart step with noise level σ_{n_s}, and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independent Gaussian noise on the target latent grid. Its tokens are $\boldsymbol{\eta} = \mathcal{T}(\boldsymbol{\xi})$ via Eq. (3). The noised token sequence $\mathbf{T}_{n_s}^{h}$ is

$$\mathbf{T}_{n_s}^{h} = (1 - \sigma_{n_s})\mathbf{T}^{h} + \sigma_{n_s}\boldsymbol{\eta}. \tag{5}$$

Since $\mathcal{T}$ is affine and $\mathbf{E}$ has unit row sums, this re-noising is in fact equivalent to latent-space noising (proof in Appendix C.3). Denoising then resumes from the corresponding latent state $\mathcal{T}^{-1}\big(\mathbf{T}_{n_s}^{h}\big)$ via Eq. (1), repeated at each resolution transition.

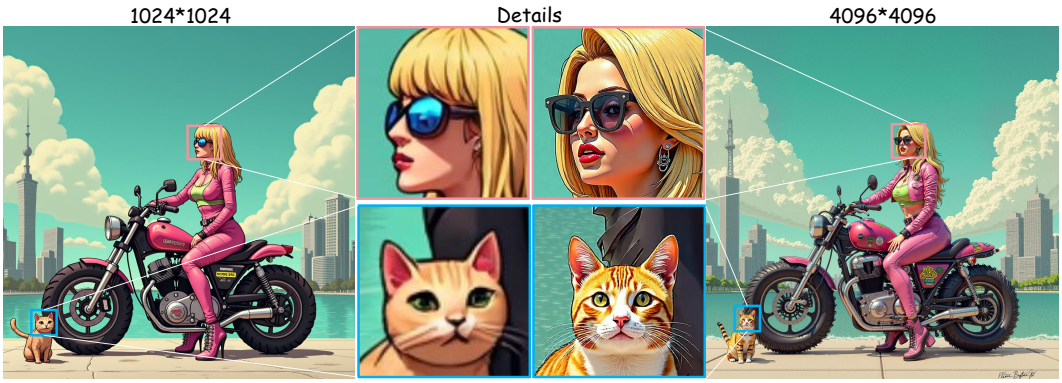

Figure 4: The 1K base image and the corresponding 4K output generated by TokenScale, with enlarged views of matching local regions. More 1K-to-4K showcases are provided in Fig. 17.

3.3 Token Fusion

Token expansion provides a structured initialization but does not prevent subsequent layout drift. To limit such drift while allowing further detail refinement, we apply selective reference correction in token space during approximately the first 10% of the high-resolution denoising steps after restart. We construct a reference latent $\mathbf{R}_{\text{ref}}$ by decoding the low-resolution final latent into pixel space, bilinearly upsampling it to the target resolution, and re-encoding it with the VAE. Its token representation $\mathbf{T}_{\text{ref}} = \mathcal{T}(\mathbf{R}_{\text{ref}}) \in \mathbb{R}^{M\times d}$ is held fixed during denoising as the structural reference for the high-resolution predictions.

At step t, let $\hat{\mathbf{T}}_{0,t} = \mathcal{T}(\hat{\mathbf{x}}_{0,t})$ be the token representation of the clean prediction (Eq. (2)). For token index i, the deviation between the prediction and the corresponding reference is $\delta_{t,i} = \left\|\hat{\mathbf{T}}_{0,t,i} - \mathbf{T}_{\text{ref},i}\right\|_2$. Let $m_{t,i} = \mathbf{1}_{\delta_{t,i}>\tau}$ be the fusion mask for threshold $\tau \geq 0$, and $\beta \in [0,1]$ the fusion strength. The corrected tokens are then

$$\tilde{\mathbf{T}}_{0,t,i} = \hat{\mathbf{T}}_{0,t,i} + \beta\, m_{t,i} \left(\mathbf{T}_{\text{ref},i} - \hat{\mathbf{T}}_{0,t,i}\right), \tag{6}$$

Each selected token is blended with its reference using a single scalar weight applied to the entire token vector, while unselected tokens receive no direct reference correction. Spatially coherent deviations are used to identify candidate regions for structural correction, rather than as a direct measure of prediction correctness. Because the input projection is affine, (Eq. (6)) can be implemented by applying the same mask and scalar blending weights to the corresponding patchified latent vectors, without inverting the projection; the construction is given in Appendix C.3.

Adaptive threshold. The scale of token deviations varies across prompts and denoising steps, making a fixed absolute threshold difficult to use consistently. We therefore upgrade the raw mask into a structural-region mask in three steps: spatial smoothing, quantile thresholding, and connected-component filtering, so that the mask selects spatially coherent deviation regions and adapts to the current noise level. As shown in Fig. 5, the resulting mask stably locates the appropriate token regions for fusion across representative cases. The detailed algorithm and the corresponding experiments are provided in Appendix D.



Figure 5: Adaptive fusion masks on four representative cases: the method stably locates the appropriate token regions for fusion.

Implicit-noise re-noising. The fused clean prediction must be incorporated at the current noise level σ_t. Let $\mathbf{T}_t = \mathcal{T}(\mathbf{x}_t)$ denote the current token sequence and $\tilde{\mathbf{T}}_t$ the corrected sequence. Retaining the implicit residual determined by $\mathbf{x}_t$ and its original clean prediction $\hat{\mathbf{x}}_{0,t}$ gives

$$\tilde{\mathbf{T}}_t = \mathbf{T}_t + (1-\sigma_t)\left(\tilde{\mathbf{T}}_{0,t} - \hat{\mathbf{T}}_{0,t}\right), \tag{7}$$

This update preserves the implicit residual of the current state–prediction pair without injecting fresh random noise. The model is then re-evaluated on the corrected latent state $\mathcal{T}^{-1}(\tilde{\mathbf{T}}_t)$ at σ_t, followed by one Euler step (Eq. (1)). Early

Table 1: Quantitative comparison with state-of-the-art baselines. The best results are highlighted in **bold**, and the second best are underlined.

Resolution	Method	FID ↓	KID ↓	CLIP Score ↑	CLIP-IQA ↑	FGResQ ↑
2048^2 (2K)	DemoFusion (Du et al., 2024a)	59.78	0.0081	**35.26**	0.668	0.5834
	DiffuseHigh (Kim et al., 2025)	59.67	**0.0062**	34.62	0.695	0.5839
	I-Max (Du et al., 2024b)	59.38	0.0089	34.64	0.564	0.6030
	HiFlow (Bu et al., 2025)	60.40	0.0080	34.55	0.671	0.6055
	TokenScale(Ours)	**57.14**	0.0076	34.53	**0.742**	**0.6061**
4096^2 (4K)	DemoFusion (Du et al., 2024a)	63.80	0.0097	**34.82**	0.683	0.5779
	DiffuseHigh (Kim et al., 2025)	59.46	0.0088	34.44	0.583	0.5773
	I-Max (Du et al., 2024b)	60.55	0.0092	34.42	0.406	0.5999
	HiFlow (Bu et al., 2025)	60.48	0.0089	34.54	0.529	0.5986
	TokenScale(Ours)	**59.15**	**0.0072**	34.31	**0.735**	**0.6024**

correction stabilizes the inherited structure. The remaining steps are devoted to detail refinement, where continued reference constraint would suppress new content, so fusion is limited to this window and the model then generates freely.

However, as $\mathcal{T}^{-1}$ is not parameterized in the DiT network, we use an equivalent algorithm (Appendix C.3); Algorithm 1 summarizes the method and Fig. 4 shows the 1K-to-4K results.

4 Experiments

4.1 Implementation Details

Baselines. We compare with representative open-source training-free high-resolution generation methods, including DemoFusion (Du et al., 2024a), DiffuseHigh (Kim et al., 2025), I-Max (Du et al., 2024b), and HiFlow (Bu et al., 2025). DemoFusion and DiffuseHigh build on SDXL, while I-Max, HiFlow, and our method build on the same FLUX.1[dev]. For a fair comparison, all methods use their official inference settings and the same Diffusion-4K prompts (Zhang et al., 2025).

Evaluation. FID (Heusel et al., 2017)/KID (Bińkowski et al., 2018) measure distribution discrepancy, CLIP Score (Hessel et al., 2021) assesses prompt consistency, and CLIP-IQA (Wang et al., 2023) quantifies perceptual quality. As conventional metrics saturate on high-quality outputs, we additionally report **FGResQ** (Sheng et al., 2026), a fine-grained quality-assessment metric for super-resolution outputs. Test images are generated from the 1K prompts of Diffusion-4K (Zhang et al., 2025). For FID and KID, Aesthetic-4K serves as the reference set of high-quality natural images.

Experimental settings. In the comparison and ablation experiments, each stage uses a 32-step schedule. High-resolution sampling skips the first eight scheduled updates and starts at $n_s = 24$ under our descending index convention. Token expansion uses bilinear interpolation, and token fusion is applied during the first three executed updates with strength $\beta = 0.2$. Following the cascade of Sec. 3.2, images are generated level by level in the order 1k $\rightarrow$ 2k $\rightarrow$ 4k. The restart-step hyperparameter is analyzed in Sec. 4.4, and the remaining hyperparameters in Appendix D. The parameter settings of all other standalone experiments are summarized in Appendix F. We mainly use FLUX.1[dev] as the base model. In Sec. 4.5, we also apply our method to Qwen-Image (Wu et al., 2025) and couple it with methods such as UltraImage (Zhao et al., 2025).

4.2 Comparison to state-of-the-art methods

Quantitative comparison. Tab. 1 reports FID, KID, CLIP Score, CLIP-IQA, and FGResQ. Apart from CLIP Score, our method ranks in the top two on every metric at both resolutions, and ranks first on all four at 4096. The CLIP-IQA margin over I-Max and HiFlow, the two FLUX.1[dev]-based baselines, is substantial, confirming a clear advantage in sharpness and perceptual quality. CLIP Scores remain lower than those of the baselines, while FGResQ provides complementary evidence of the perceptual-quality advantage. TokenScale also has the lowest per-step inference time and peak memory among the FLUX.1[dev]-based methods at 4K resolution (Tab. 3).

Qualitative comparison. As shown in Fig. 6, DemoFusion and DiffuseHigh preserve the overall structure, but their local details are of low quality. I-Max sometimes produces noisy and blurred results. HiFlow's results are generally acceptable, but zoomed-in regions still lack clarity and rich-

Table 2: Pairwise preference comparison under FGResQ Quality Ranking. We report the win rates and average pairwise preference probabilities (p_{ours}) of **Ours** against four competing baselines, over 1,000 paired comparisons per method and resolution.

Competitor	**2K: Win %** ↑	p_{ours} ↑	p_{opp} ↓	**4K: Win %** ↑	p_{ours} ↑	p_{opp} ↓
HiFlow (Bu et al., 2025)	**70.0**	0.4987	0.4373	**73.6**	0.5266	0.4135
I-Max (Du et al., 2024b)	**79.4**	0.5412	0.3881	**82.2**	0.5764	0.3620
DemoFusion (Du et al., 2024a)	**83.6**	0.6241	0.3104	**88.2**	0.7044	0.2505
DiffuseHigh (Kim et al., 2025)	**87.1**	0.6680	0.2765	**91.0**	0.7316	0.2234

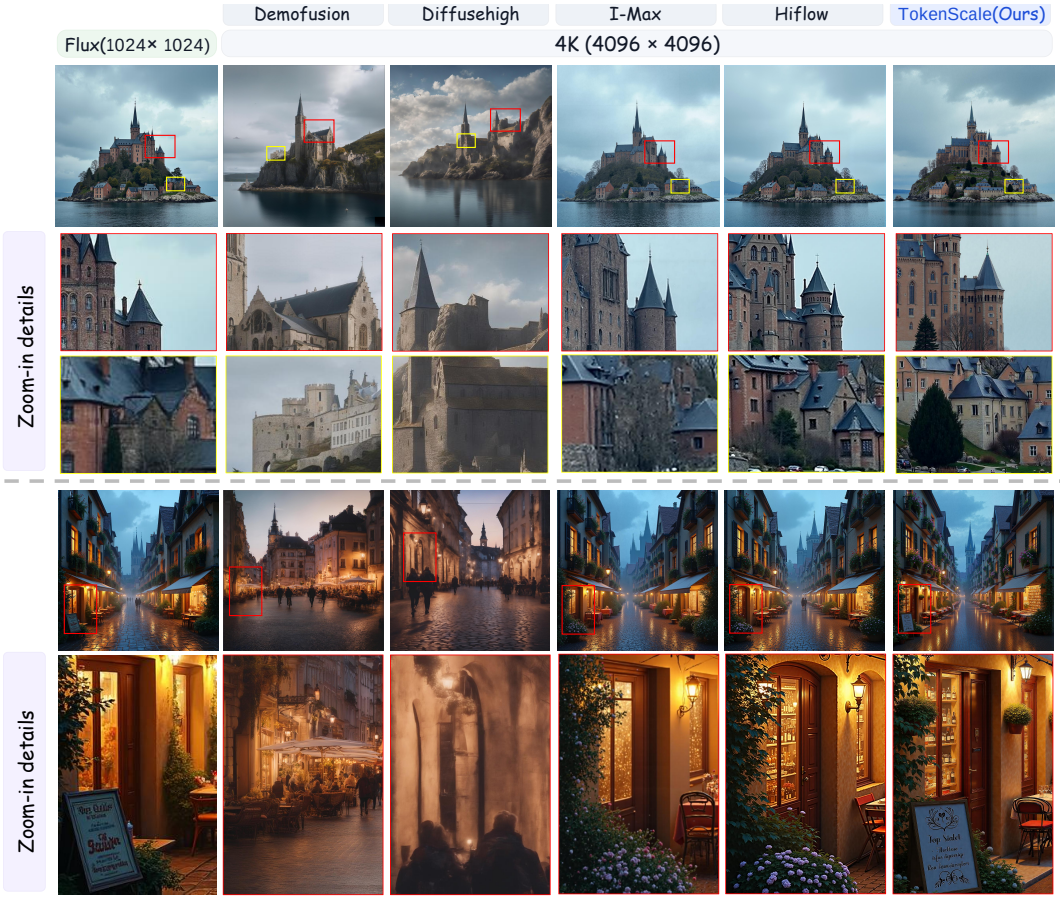

Figure 6: Qualitative comparison with training-free baselines: baselines show low-quality local details, noise, or blur, while TokenScale produces content-rich high-resolution details.

ness due to excessive low-resolution guidance. In contrast, TokenScale produces content-rich high-resolution scenes with high aesthetic quality and semantic consistency, especially in local details, surpassing previous models. More qualitative comparisons are provided in Appendix E.2.

Fine-grained pairwise preference. We further let FGResQ judge 1,000 pairwise comparisons per baseline, following the ranking-mode protocol of Appendix F. As shown in Tab. 2, our method wins the majority at both resolutions (up to 91.0% at 4K), with p_{ours} consistently above the opponent's, confirming that our advantage persists where conventional metrics saturate.

4.3 Ablation Study

We ablate the two core modules at 4096. As shown in Fig. 7, with TE alone, the images already exhibit plausible global structure, but several problems remain, such as unnatural repeated petals, dense spots in the background, and unnatural tablecloth patterns. TF eliminates these unnatural high-

Table 3: Inference time and memory comparison.

Method	Per-Step (s) ↓	Mem (GB) ↓
I-Max (Du et al., 2024b)	35.76	47.3
HiFlow (Bu et al., 2025)	30.23	47.1
TokenScale(Ours)	**26.53**	**46.3**

Table 4: Quantitative results of ablation study.

Config.	FID	KID	CLIP-IQA	FGResQ
-TE, -TF	268.34	0.0287	0.234	0.2376
-TF	60.45	0.0073	0.722	0.6004
TokenScale	**59.15**	**0.0072**	**0.735**	**0.6024**

frequency repetitions, stabilizing super-resolution generation on top of TE (Tab. 4). The ablation study of 2K image is provided in Appendix E.1.



Figure 7: Ablation study at 4K resolution. Best viewed zoomed in.

4.4 The analysis of the restart step.

Attention-map evolution over a full 100-step generation (Fig. 8, top) shows that within the first 1/3, tokens already carry the low-frequency information, while after 1/3 the model starts fine-tuning details; we therefore restart before this point, inheriting the low-frequency structure while leaving detail generation free. However, restarting before the first 1/6 adds too much noise that excessively masks the token information. Restarting within 1/6 to 1/3 of the total steps yields the best results (Fig. 8, bottom), consistent with this analysis.

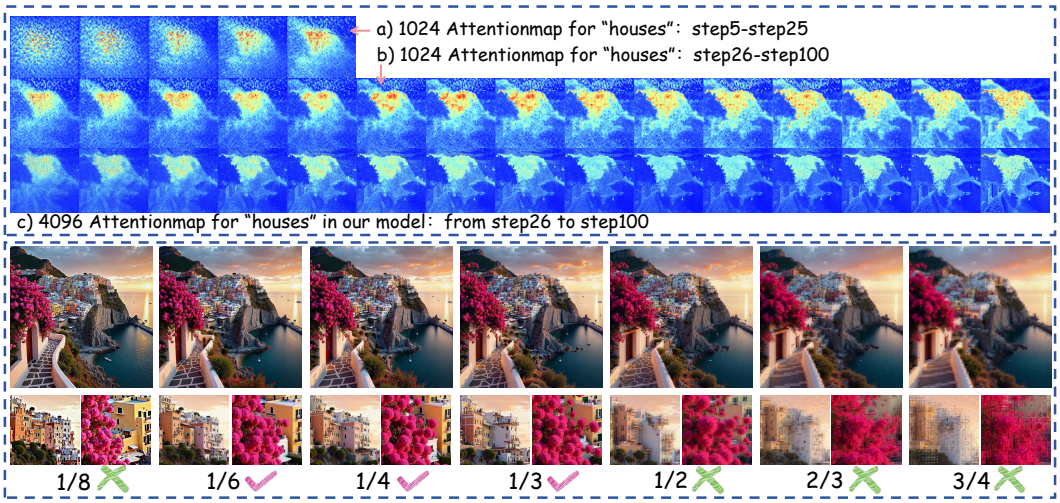

Figure 8: Top: evolution of attention maps over a full 100-step generation. Bottom: results under different restart steps, where restarting within 1/6 to 1/3 of the total steps yields the best quality.

4.5 APPLICATIONS

Application in Qwen-Image. Our method also applies to DiT-based T2I models with rotary position encoding, such as Qwen-Image (Wu et al., 2025). Incorporating our technique yields richer and more detailed content (Fig. 9, top).

Coupling with UltraImage. TokenScale can be integrated with UltraImage (Zhao et al., 2025) by using token expansion to initialize sampling from a half-noise level. During 4096^2 generation, TF is disabled and UltraImage's positional-encoding and attention mechanisms are retained. In our FLUX.1[dev] (Black Forest Labs, 2024) experiments, UltraImage alone produces degenerate outputs at 4096^2, whereas token-expanded initialization yields coherent results (Fig. 9, bottom right).

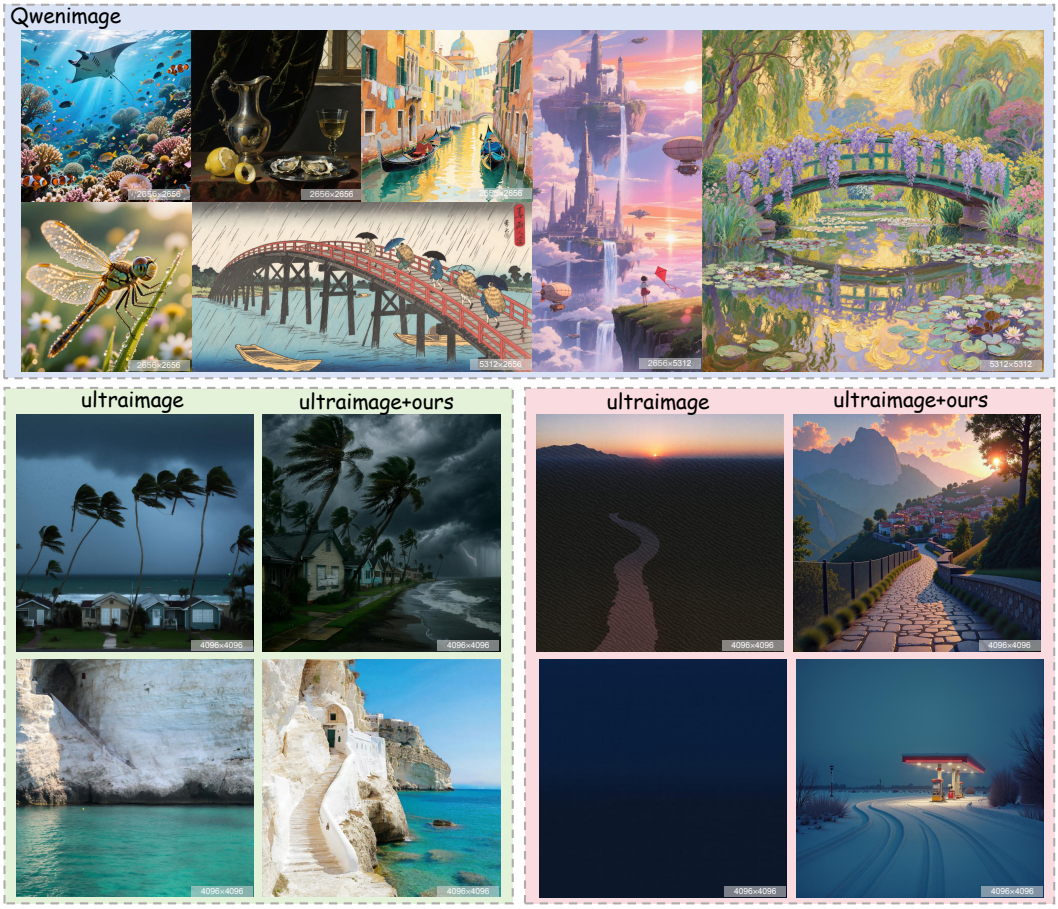

Figure 9: Qualitative results in applications. Top: TokenScale on Qwen-Image. Bottom: coupling with UltraImage, on Qwen-Image (left) and on FLUX.1[dev] (right).

5 CONCLUSION

We presented TokenScale, a training-free token-level framework that enables high-resolution generation with high fidelity and rich details from pre-trained DiT-based flow models. It expands fused low-resolution tokens into the high-resolution token space and applies lightweight, early reference fusion, stabilizing inherited structure without over-constraining the model. Extensive experiments show it surpasses state-of-the-art training-free methods and transfers across Qwen-Image and positional-encoding-based approaches. Limitations on irregular textures are discussed in Appendix I.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language of the manuscript and to assist with the implementation and debugging of the evaluation code. We have not used generative AI tools for designing or proposing the method, for deriving the theoretical results, or for conducting the experiments. We have reviewed all AI-assisted work: all AI-assisted code was verified and tested by the authors, and all experimental results and claims were checked manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies training-free high-resolution image generation using publicly available pretrained models and benchmark data. The study does not recruit human participants or collect new personal data. Like other image-generation techniques, TokenScale could be misused to produce misleading visual content, including generated images presented as authentic photographs. Improving image fidelity does not remove these risks. Responsible downstream use should distinguish generated content from authentic imagery and respect privacy and the rights of affected individuals.

REPRODUCIBILITY STATEMENT

All experiments use publicly available checkpoints—FLUX.1[dev] as the base model, Qwen-Image for the application study (Sec. 4.5), and SDXL for the DemoFusion and DiffuseHigh baselines—with all baselines run from their official implementations under official inference settings. We use the Diffusion-4K prompt set (the prompts of all figures are listed in Appendix J), with Aesthetic-4K as the reference set for FID and KID. All results fit on a single GPU with at least 48GB memory (peak 46.3GB at 4096^2, Tab. 3). The complete settings are summarized in Sec. 4.1 and Appendix F, including the NTK-aware techniques and the usage of FGResQ. The per-stage procedure is Algorithm 1, the latent-space equivalences of expansion and fusion are in Appendix C.3, the adaptive mask and fusion hyperparameters in Appendix D, and the token-space noise derivation in Appendix C.1. Code will be released upon acceptance.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, pp. 1737–1752, 2023.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

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.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

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*, pp. 6159–6168, 2024a.

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. *arXiv preprint arXiv:2410.07536*, 2024b.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

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 *International Conference on Learning Representations*, 2024.

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*, pp. 7514–7528, 2021.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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*, 2024.

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.

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

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, pp. 4338–4346, 2025.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. In *Advances in Neural Information Processing Systems*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yiyang Ma, Feng Zhou, Xuedan Yin, Pu Cao, Yonghao Dang, and Jianqin Yin. ResDiT: Evoking the intrinsic resolution scalability in diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40612–40621, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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. `https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/`, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

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.

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*, 2025.

Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. UltraPixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Xiangfei Sheng, Xiaofeng Pan, Zhichao Yang, Pengfei Chen, and Leida Li. Fine-grained image quality assessment for perceptual image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8914–8922, 2026.

Shuwei Shi, Wenbo Li, Yuechen Zhang, Jingwen He, Biao Gong, and Yinqiang Zheng. ResMaster: Mastering high-resolution image generation via structural and fine-grained guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6887–6895, 2025.

Jianyi Wang, Kelvin C. K. 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, pp. 2555–2563, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23464–23473, 2025.

Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhao Chen, Yao Tang, and Jiajun Liang. HiDiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, 2024.

Bingxuan Zhao, Qing Zhou, Yu Wang, Chuang Yang, and Qi Wang. σ: Sigmoid modulation for ultra high resolution diffusion. In *International Conference on Machine Learning*, 2026.

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.

APPENDIX

In the Appendix, we provide additional attention visualizations of the 4096 generation (Section A), a comparison of token expansion and latent-grid interpolation (Section B), the derivations and proofs of the method (Section C), the hyperparameter analysis of token fusion (Section D), supplementary experiments including ablation, qualitative comparisons, showcase cases, and seed stability (Section E), more experimental details (Section F), the attention-evolution analysis of guidance (Section G), the extended related work (Section H), a discussion of limitations (Section I), and the prompts used for the figures (Section J).

Contents

A Attention Visualization 14

B Token Expansion versus Latent-Grid Interpolation 14

C Derivations in the Method 15

C.1 Token-Space Noise Distribution 15

C.2 Implicit Noise Re-Noising 15

C.3 Equivalent Forms of Token-Space Expansion and Fusion 16

C.4 High-Resolution Stage Pseudocode 17

D Hyperparameter Analysis 17

D.1 Fusion Strength and Fusion Steps 17

D.2 Adaptive Threshold 19

E Supplementary Experiments for Comparison and Ablation 19

E.1 2048 Ablation 19

E.2 Additional Qualitative Comparisons 19

E.3 Additional Showcase Cases 20

E.4 Stability across Different Seeds 20

F More Experimental Details 20

G How Guidance Shapes Attention Evolution 21

H Extended Related Work 28

I Limitation 28

J Prompts for Figures 30

A Attention Visualization

Fig. 10 visualizes additional attention maps for more prompt words during the 4096 generation with TokenScale, supplementing the representative words in Fig. 2 (right). Across all cases, the attention remains anchored to the semantically corresponding regions and does not disperse over the enlarged token grid, confirming that localized text–image correspondence is preserved throughout high-resolution denoising.

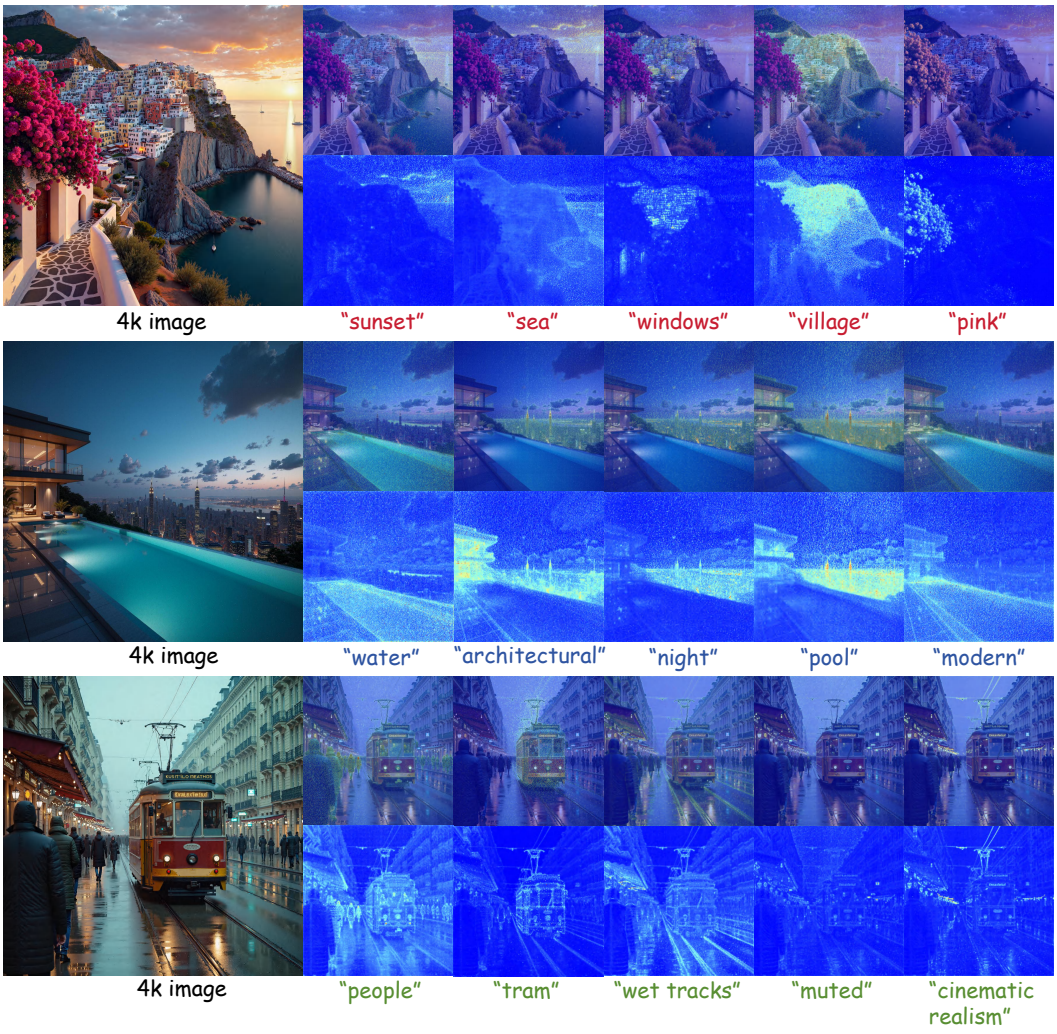

Figure 10: Supplement to Fig. 2 (right): original images and attention heatmaps for more words.

B Token Expansion versus Latent-Grid Interpolation

Fig. 11 compares the two initialization strategies on a representative case: the low-resolution reference (left), the result of latent-grid interpolation (middle), and that of token expansion (right), with enlarged details shown below. Latent-grid interpolation only upsamples the low-resolution content: the result stays blurry and adds no new detail. Token expansion, in contrast, inherits the low-resolution token information and structure intact, and the model then builds on this initialization to produce a markedly clearer image with substantially more detail than the low-resolution reference—a refinement that is possible only because attention operates on well-formed tokens carrying the semantic features, rather than on raw latent pixels.

Moreover, latent-grid interpolation can produce a wrong initialization. In Fig. 11, the moon in the latent-space result is absent from the low-resolution reference: it likely originates from the clouds on

the right side of the reference image. Interpolating at the latent-pixel level blurs this cloud region into an ambiguous pattern, and after re-noising the model misinterprets it as a moon. Token expansion avoids such structural hallucinations, because the interpolation weights act on entire token vectors and preserve the semantic features and their neighborhood relations. Both paths share the same restart noise level and denoising schedule, with token fusion disabled, so the differences stem only from the initialization.



Figure 11: Comparison of the two initialization strategies: the low-resolution reference (left), latent-grid interpolation (middle), and token expansion (right), with enlarged details shown below.

C Derivations in the Method

C.1 Token-Space Noise Distribution

Token-space noise distribution. We first show that the effective noise distribution in token space is $\mathcal{N}(\mathbf{b}_{\text{in}}, \mathbf{W}_{\text{in}}\mathbf{W}_{\text{in}}^{\top})$. Consider a clean latent $\mathbf{x}_0$ and its patchified representation $\mathbf{P} = \mathcal{P}(\mathbf{x}_0)$. The corresponding token sequence is $\mathbf{T}^{\ell} = \mathbf{P}\mathbf{W}_{\text{in}}^{\top} + \mathbf{b}_{\text{in}}$. If we add noise in the latent space,

$$\mathbf{x}_\sigma = (1-\sigma)\mathbf{x}_0 + \sigma\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

then the token sequence becomes

$$\mathbf{T}_\sigma^{\ell} = \mathcal{P}(\mathbf{x}_\sigma)\mathbf{W}_{\text{in}}^{\top} + \mathbf{b}_{\text{in}} = (1-\sigma)\mathbf{T}^{\ell} + \sigma\boldsymbol{\eta},$$

where

$$\boldsymbol{\eta} = \mathcal{P}(\boldsymbol{\xi})\mathbf{W}_{\text{in}}^{\top} + \mathbf{b}_{\text{in}}.$$

Since $\mathcal{P}(\boldsymbol{\xi})$ is merely a rearrangement of i.i.d. standard Gaussian entries, it remains $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Hence

$$\boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{b}_{\text{in}}, \mathbf{W}_{\text{in}}\mathbf{W}_{\text{in}}^{\top}\right),$$

which is exactly the token-space noise distribution in Eq. (5). Directly sampling from this high-dimensional, low-rank Gaussian is impractical.

C.2 Implicit Noise Re-Noising

We derive the adjustment form of Eq. (7) used in the main text. All token sequences below are the tokenizations of the corresponding latent states (Sec. 3.1). The *implicit noise* is defined in the latent space as

$$\boldsymbol{\varepsilon}_{\text{impl},t} = \frac{\mathbf{x}_t - (1-\sigma_t)\,\hat{\mathbf{x}}_{0,t}}{\sigma_t},$$

i.e., the noise realization implied by the current state $\mathbf{x}_t$ and the *original* clean prediction $\hat{\mathbf{x}}_{0,t}$. Tokenizing it gives the corresponding *noise token*,

$$\boldsymbol{\eta}_{\text{impl},t} \;:=\; \mathcal{T}\big(\boldsymbol{\varepsilon}_{\text{impl},t}\big) = \frac{\mathbf{T}_t - (1-\sigma_t)\,\hat{\mathbf{T}}_{0,t}}{\sigma_t},$$

where the second equality follows from the affinity of the tokenization: $\mathbf{T}_t - (1-\sigma_t)\hat{\mathbf{T}}_{0,t} = \sigma_t \mathcal{P}(\boldsymbol{\varepsilon}_{\text{impl},t})\mathbf{W}_{\text{in}}^\top + \sigma_t \mathbf{b}_{\text{in}}$, since the bias terms cancel as $\mathbf{b}_{\text{in}} - (1-\sigma_t)\mathbf{b}_{\text{in}} = \sigma_t \mathbf{b}_{\text{in}}$. Re-noising the fused clean latent (constructed in Appendix C.3) at the *current* noise level σ_t and tokenizing yields

$$\tilde{\mathbf{T}}_t = (1-\sigma_t)\,\tilde{\mathbf{T}}_{0,t} + \sigma_t\,\boldsymbol{\eta}_{\text{impl},t},$$

where the noise token is computed from the *original* prediction $\hat{\mathbf{x}}_{0,t}$ rather than the fused one. Substituting $\sigma_t \boldsymbol{\eta}_{\text{impl},t} = \mathbf{T}_t - (1-\sigma_t)\hat{\mathbf{T}}_{0,t}$ into the re-noising step yields the adjustment form,

$$\tilde{\mathbf{T}}_t = \mathbf{T}_t + (1-\sigma_t)\big(\tilde{\mathbf{T}}_{0,t} - \hat{\mathbf{T}}_{0,t}\big).$$

The corrected state thus differs from $\mathbf{T}_t$ only by the fused-content correction, scaled by $(1-\sigma_t)$, and the noise realization of the original trajectory is preserved exactly. Since the tokenization defined in Sec. 3.1 is affine, each token-space update above is exactly the tokenization of the corresponding latent-space update, so the two views are equivalent.

C.3 Equivalent Forms of Token-Space Expansion and Fusion

Token expansion. Token expansion (Eq. (4)) is defined in token space, where the expanded and re-noised sequence $\mathbf{T}_{n_s}^{\hbar}$ (Eq. (5)) serves directly as the initial input to the DiT network. Denoising, however, also requires the latent state at the same noise level, and the pretrained DiT provides no inverse of the tokenization map. We therefore implement an equivalent expansion in the latent space: we expand the *patchified* low-resolution latent with the same expansion matrix, add noise, and unpack the result, and we show that the tokenization of this latent state coincides exactly with the token-space expansion.

Let $\mathbf{P}^\ell = \mathcal{P}(\mathbf{x}_0^\ell) \in \mathbb{R}^{N\times p}$ be the patchified low-resolution latent, so that $\mathbf{T}^\ell = \mathbf{P}^\ell \mathbf{W}_{\text{in}}^\top + \mathbf{1}_N \mathbf{b}_{\text{in}}^\top$ (Eq. (3)). With noise $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ on the high-resolution latent grid, the token-space path produces

$$\begin{aligned}
\mathbf{T}_\sigma^{\hbar} &= (1-\sigma)\,\mathbf{E}\,\mathbf{T}^\ell + \sigma\,\mathcal{T}(\boldsymbol{\xi}) \\
&= \Big[(1-\sigma)\,\mathbf{E}\,\mathbf{P}^\ell + \sigma\,\mathcal{P}(\boldsymbol{\xi})\Big]\mathbf{W}_{\text{in}}^\top + \Big[(1-\sigma)\,\mathbf{E}\mathbf{1}_N + \sigma\,\mathbf{1}_M\Big]\mathbf{b}_{\text{in}}^\top \\
&= \Big[(1-\sigma)\,\mathbf{E}\,\mathbf{P}^\ell + \sigma\,\mathcal{P}(\boldsymbol{\xi})\Big]\mathbf{W}_{\text{in}}^\top + \mathbf{1}_M \mathbf{b}_{\text{in}}^\top,
\end{aligned}$$

where the last equality uses $\mathbf{E}\mathbf{1}_N = \mathbf{1}_M$, i.e., the unit row sums of $\mathbf{E}$.

The latent-space path expands the patchified latent with the same matrix, adds noise, and unpacks:

$$\mathbf{P}_\sigma^{\hbar} = (1-\sigma)\,\mathbf{E}\,\mathbf{P}^\ell + \sigma\,\mathcal{P}(\boldsymbol{\xi}), \qquad \mathbf{x}_\sigma^{\hbar} = \mathcal{P}^{-1}\big(\mathbf{P}_\sigma^{\hbar}\big).$$

Tokenizing the resulting latent yields

$$\mathcal{T}\big(\mathbf{x}_\sigma^{\hbar}\big) = \mathbf{P}_\sigma^{\hbar}\mathbf{W}_{\text{in}}^\top + \mathbf{1}_M \mathbf{b}_{\text{in}}^\top = \Big[(1-\sigma)\,\mathbf{E}\,\mathbf{P}^\ell + \sigma\,\mathcal{P}(\boldsymbol{\xi})\Big]\mathbf{W}_{\text{in}}^\top + \mathbf{1}_M \mathbf{b}_{\text{in}}^\top,$$

which is exactly $\mathbf{T}_\sigma^{\hbar}$. Hence the two paths coincide, and the latent-space construction is the valid implementation adopted in Algorithm 1.

Token fusion. The fusion of Eq. (6) is likewise defined in token space, but the affine tokenization is not invertible, so the fused tokens cannot be pushed back to the latent grid by inverting $\mathcal{T}$. We instead fuse the *patchified* latent vectors directly, which realizes the token-space fusion exactly. Let $\hat{\mathbf{P}}_{0,t} = \mathcal{P}(\hat{\mathbf{x}}_{0,t})$ and $\mathbf{P}_{\text{ref}} = \mathcal{P}(\mathbf{R}_{\text{ref}})$ be the patchified clean prediction and reference, and let $\mathbf{m}_t$ be the structural-region mask (Appendix D.2). The fused clean latent is constructed in the latent space as

$$\tilde{\mathbf{P}}_{0,t} = \hat{\mathbf{P}}_{0,t} + \beta\,\mathbf{m}_t \odot \big(\mathbf{P}_{\text{ref}} - \hat{\mathbf{P}}_{0,t}\big), \qquad \tilde{\mathbf{x}}_{0,t} = \mathcal{P}^{-1}\big(\tilde{\mathbf{P}}_{0,t}\big),$$

where $\odot$ multiplies row i by the scalar $m_{t,i}$ across all its coordinates. Because the tokenization is affine (Eq. (3)) and the per-row scalar blending commutes with the right-multiplication by $\mathbf{W}_{\text{in}}^\top$,

$$\mathcal{T}\big(\tilde{\mathbf{x}}_{0,t}\big) = \Big[\hat{\mathbf{P}}_{0,t} + \beta\,\mathbf{m}_t \odot \big(\mathbf{P}_{\text{ref}} - \hat{\mathbf{P}}_{0,t}\big)\Big]\mathbf{W}_{\text{in}}^\top + \mathbf{1}_M \mathbf{b}_{\text{in}}^\top = \hat{\mathbf{T}}_{0,t} + \beta\,\mathbf{m}_t \odot \big(\mathbf{T}_{\text{ref}} - \hat{\mathbf{T}}_{0,t}\big) = \tilde{\mathbf{T}}_{0,t},$$

so the patch-space construction realizes the token-space fusion of Eq. (6) without inverting the tokenization. This is the construction used in Algorithm 1.

C.4 High-Resolution Stage Pseudocode

Algorithm 1 One high-resolution stage of TokenScale. The velocity field and the Euler update live in the VAE latent space; TE and TF are both implemented in the latent space via the equivalences of Appendix C.3. t is the global denoising index, with σ_t decreasing from $\sigma_T = 1$ to $\sigma_0 = 0$; under the base setting of Sec. 4.1, $T = 32$, $n_s = 24$, and $F = 3$. The multi-stage cascade (Sec. 3.2) repeats this routine level by level, taking the final latent of each stage as the input of the next.

$\mathbf{x}_{n_s} \leftarrow \mathcal{P}^{-1}\big((1-\sigma_{n_s})\,\mathbf{E}\,\mathcal{P}(\mathbf{x}_0^{\ell}) + \sigma_{n_s}\,\boldsymbol{\xi}\big)$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ {restart latent via TE, equivalent to the token-space re-noising (Appendix C.3)}
$\mathbf{R}_{\mathrm{ref}} \leftarrow$ Decode-Bilinear-Encode$(\mathbf{x}_0^{\ell})$; $\mathbf{T}_{\mathrm{ref}} \leftarrow \mathcal{T}(\mathbf{R}_{\mathrm{ref}})$; $\mathbf{P}_{\mathrm{ref}} \leftarrow \mathcal{P}(\mathbf{R}_{\mathrm{ref}})$ {reference tokens and patchified reference, Sec. 3.3}
for $t = n_s, \ldots, 1$ **do**
  $\mathbf{v}_t \leftarrow f_\theta(\mathbf{x}_t, \sigma_t, \mathbf{c})$ {velocity field, Eq. (1)}
  $\hat{\mathbf{x}}_{0,t} \leftarrow \mathbf{x}_t - \sigma_t\,\mathbf{v}_t$ {clean prediction, Eq. (2)}
  if $t > n_s - F$ **then**
    $\hat{\mathbf{T}}_{0,t} \leftarrow \mathcal{T}(\hat{\mathbf{x}}_{0,t})$ {to token space for the deviation mask}
    $\delta_{t,i} \leftarrow \big\|\hat{\mathbf{T}}_{0,t,i} - \mathbf{T}_{\mathrm{ref},i}\big\|_2$ {deviation $\delta_{t,i}$}
    $\mathbf{m}_t \leftarrow \mathrm{AdaptiveMask}(\delta_t)$ {structural-region mask, Appendix D.2}
    $\tilde{\mathbf{x}}_{0,t} \leftarrow \mathcal{P}^{-1}\big(\hat{\mathbf{P}}_{0,t} + \beta\,\mathbf{m}_t \odot (\mathbf{P}_{\mathrm{ref}} - \hat{\mathbf{P}}_{0,t})\big)$, where $\hat{\mathbf{P}}_{0,t} = \mathcal{P}(\hat{\mathbf{x}}_{0,t})$ {patch-space fusion, equivalent to Eq. (6) (Appendix C.3)}
    $\mathbf{x}_t \leftarrow \mathbf{x}_t + (1-\sigma_t)\,\big(\tilde{\mathbf{x}}_{0,t} - \hat{\mathbf{x}}_{0,t}\big)$ {Eq. (7)}
    $\mathbf{v}_t \leftarrow f_\theta(\mathbf{x}_t, \sigma_t, \mathbf{c})$ {re-evaluate after correction}
  end if
  $\mathbf{x}_{t-1} \leftarrow \mathbf{x}_t + (\sigma_{t-1} - \sigma_t)\,\mathbf{v}_t$ {Euler step, Eq. (1)}
end for
return $\mathbf{x}_0$ {clean latent of this stage}

D Hyperparameter Analysis

D.1 Fusion Strength and Fusion Steps

Fig. 12 ablates the two fusion hyperparameters on a representative case: the fusion strength β with the fusion step count fixed at 3 steps (top), and the fusion step count with β fixed at 0.5 (bottom).

Fusion strength. When the strength is too low ($\beta < 0.2$), unreasonable artifacts can appear and the detail structure shifts away from the reference, e.g., the excessive petals at $\beta = 0.1$. With $\beta \in [0.2, 0.5]$, the model maintains fine details together with a reasonable low-frequency structure. When the strength is too high, however, the strong correction over-constrains the image and breaks its structure.

Fusion step count. Generation is best when the fusion step count is around $1/10$ of the denoising steps after restart. Increasing the step count causes blurring and distortion: in later steps the model has already generated much information that is absent from the low-resolution reference, so continued reference correction strongly constrains the model and damages the result.

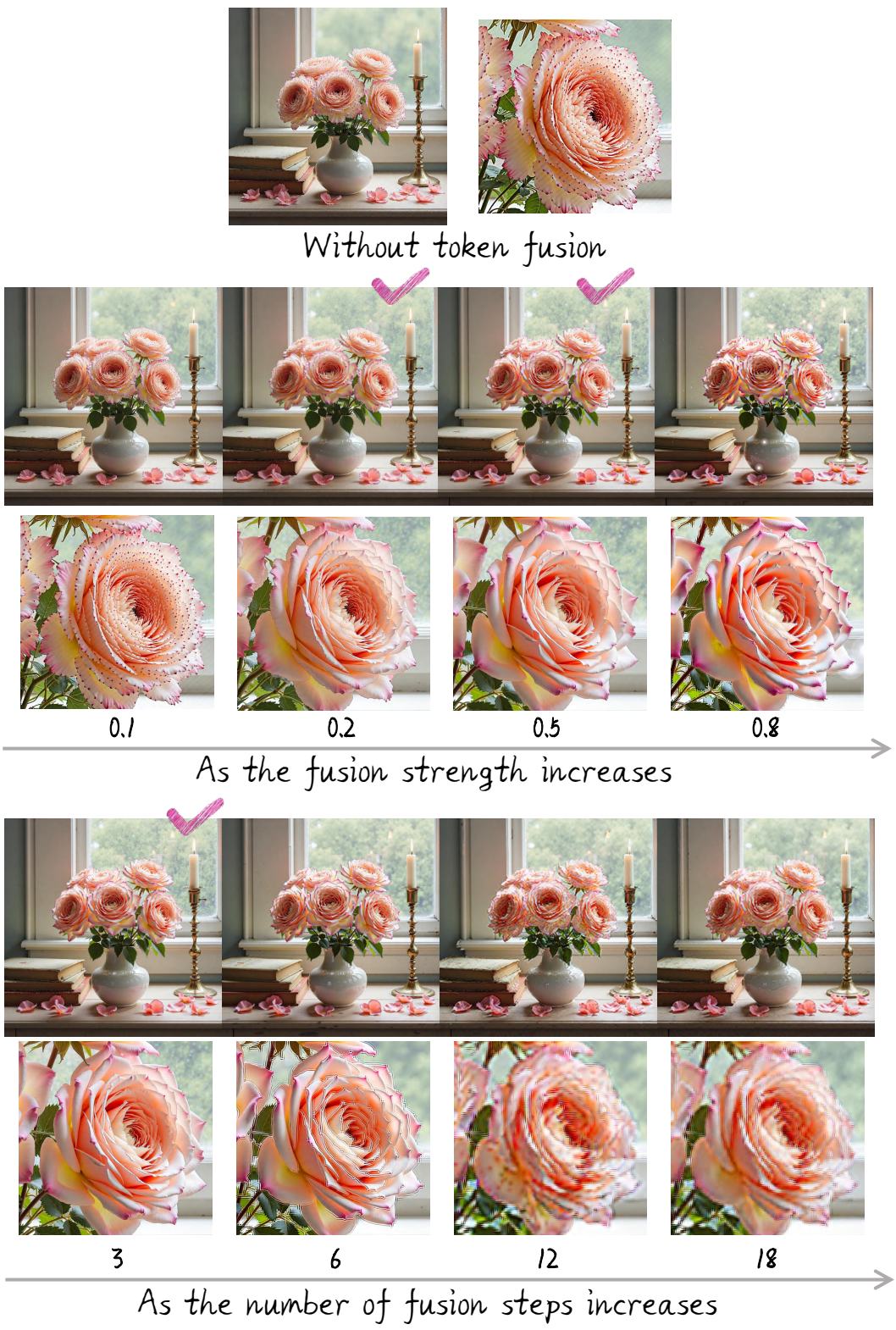

Figure 12: Ablations of the fusion hyperparameters. Top: fusion strength β with the step count fixed at 3 steps. Bottom: fusion step count with β fixed at 0.5.

D.2 Adaptive Threshold

A fixed threshold turns out to be unsuitable: the distribution of per-token deviations $\delta_{t,i}$ shifts with the prompt and the scene, so a single absolute value cannot transfer across images. (For instance, a zero-false-alarm choice such as $\tau = 2 \times \mathrm{med}_i\, \delta_{t,i}$ fuses only about 4% of tokens—too few to lock the structure.) The key observation is that real structural errors are *spatially coherent*: the model renders a whole object or texture region incorrectly, whereas noise fluctuations are independent across tokens. We therefore upgrade the single-point criterion into a *structural-region* criterion in three steps:

1. **Spatial smoothing.** We convolve the deviation map with a 3×3 box kernel, $\tilde{\delta} = \mathrm{Box}_{3\times3} \circledast \delta$. Isolated noise spikes are suppressed by roughly $3\times$ through neighborhood averaging, while coherent deviation regions are preserved—the criterion changes from "error points" to "error blobs".
2. **Quantile thresholding.** We set $\tau_t = q_{0.6}(\tilde{\delta})$, the 60th percentile of the smoothed deviations. The fused fraction is then controlled directly by the quantile ($\approx 40\%$ of tokens), independent of the absolute scale of δ—hence stable across prompts—and the quantile shrinks together with σ_t as denoising proceeds, giving automatic annealing (measured τ_t decreases from 2.32 to 1.65).
3. **Connected-component filtering.** After thresholding, we remove connected components with area smaller than 8 tokens to clean residual speckle. The final mask $\mathcal{M}$ is exactly the set of structural regions; the fusion formula is unchanged: tokens in $\mathcal{M}$ are updated as in Eq. (6), i.e., $\tilde{\mathbf{T}}_{0,t,i} = (1-\beta)\,\hat{\mathbf{T}}_{0,t,i} + \beta\,\mathbf{T}_{\mathrm{ref},i}$.

Algorithm 2 Adaptive structural-region mask

Require: deviation map δ over tokens, quantile $p = 0.6$, minimum component area $a_{\min} = 8$
Ensure: structural-region mask $\mathcal{M}$
$\tilde{\delta} \leftarrow \mathrm{Box}_{3\times3} \circledast \delta$ {spatial smoothing}
$\tau \leftarrow q_p(\tilde{\delta})$ {quantile threshold}
$\mathcal{M} \leftarrow \{n : \tilde{\delta}_n > \tau\}$
$\mathcal{M} \leftarrow \{C \in \mathcal{M} : |C| \geq a_{\min}\}$ {connected-component filtering}
return $\mathcal{M}$

This adaptive scheme has two benefits: (i) the fused fraction stays stable across prompts and steps, and (ii) the mask avoids scattered points and captures only large-area low-frequency structure. Fig. 13 shows the complete comparison on the five cases: the columns correspond to fixed thresholds $\tau = 1, 3, 5, 7, 9$ and our adaptive method. Fixed values are unstable across cases, whereas our method consistently selects the low-frequency structural regions, with the fused ratio staying around 40% for every case—verifying the effectiveness of the adaptive design.

E Supplementary Experiments for Comparison and Ablation

This section supplements the ablation study and the comparison experiments of the main text with additional cases at both 2048 and 4096, further demonstrating the effectiveness of TokenScale.

E.1 2048 Ablation

Due to the page limit, we postpone the 2048 ablation to this appendix. Since the 2048 super-resolution task is relatively simple, token expansion alone already produces very good results without the token fusion module (Fig. 14).

E.2 Additional Qualitative Comparisons

Fig. 15 presents additional qualitative comparisons at 2048: even though the 2048 super-resolution task is simpler, our model still achieves higher clarity and better details than the baselines. Fig. 16 further compares TokenScale with the FLUX.1[dev]-based methods I-Max and HiFlow at 4096, where our results show richer details, with enlarged views provided for inspection.



Figure 13: Complete comparison of fusion masks across five cases (rows): fixed thresholds $\tau = 1, 3, 5, 7, 9$ (columns 1–5) versus our adaptive method (last column). Fixed values are unstable across cases, while our method stably selects low-frequency structure with a fused ratio of about 40% in every case.

E.3 Additional Showcase Cases

Fig. 17 shows 1K-to-4K showcase comparisons, and Fig. 18 presents more showcase cases. Due to the file size and page limits, additional ultra-high-resolution 4K images generated by our model are packed into the supplementary material.

E.4 Stability across Different Seeds

We further evaluate the stability of our method across random seeds: for the same prompt, our method consistently produces structurally stable, high-quality results under different seeds, with only natural stochastic variation (Fig. 19).

F More Experimental Details

This section collects additional experimental details in three parts: the summary of all experiment settings, the inference techniques enabled for extrapolated resolutions, and the usage of the FGResQ metric.

Restart attention-map observation. We run the full 1024 generation with 100 steps, and start the super-resolution stage from the $1/4$ restart ratio; the fusion is applied in 6 steps with strength 0.2 and the adaptive threshold.

Restart-ratio illustration. To conveniently compare different restart ratios, the schedule is shortened to 72 steps, with the super-resolution stage restarted at $1/8$, $1/6$, $1/4$, $1/3$, $1/2$, $2/3$, and

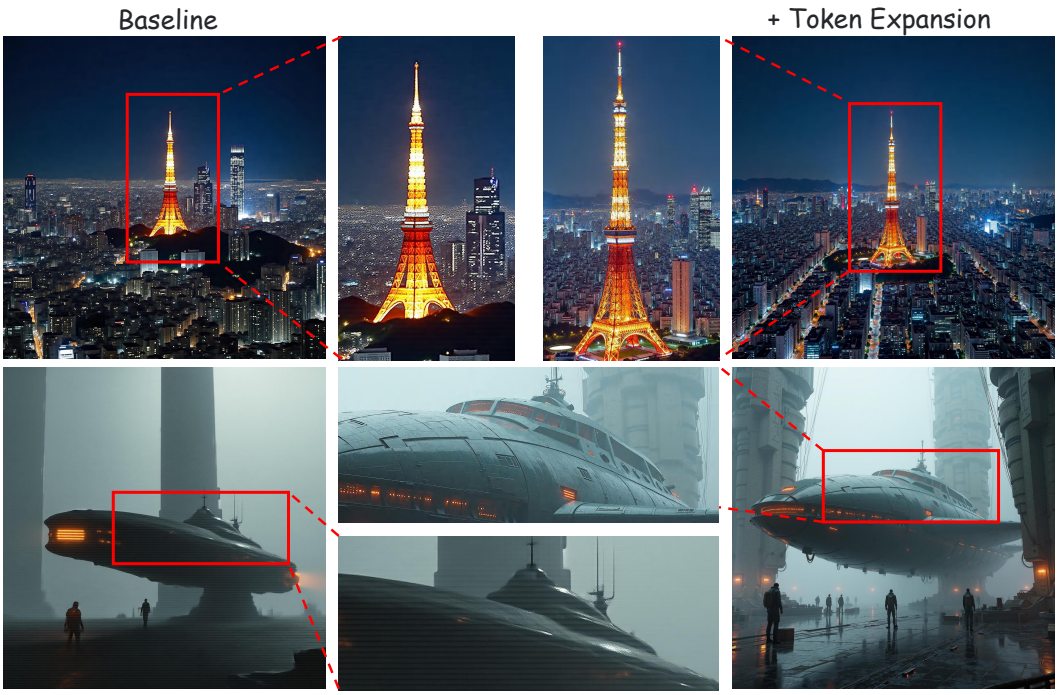

Figure 14: Ablation study at 2048 resolution. Best viewed zoomed in.

$3/4$; the corresponding fusion step counts are $5, 5, 4, 4, 3, 3, 2$, with strength 0.2 and the adaptive threshold.

Resolution extrapolation techniques. To keep the Rectified Flow model stable when sampling at resolutions beyond its training size, we also enable the inference-time techniques suggested by previous works: NTK-aware scaled RoPE (Peng & Quesnelle, 2023), balancing the entropy shift of self-attention (Jin et al., 2023), and balancing the image/text sequence length ratio for MMDiT (Du et al., 2024b). These techniques are applied in all high-resolution stages and are independent of TokenScale's token expansion and fusion.

Evaluation with FGResQ. As conventional metrics saturate on high-quality outputs, we additionally adopt FGResQ (Sheng et al., 2026), a fine-grained image quality assessment model for image restoration, trained on the FGRestore dataset with over 30K fine-grained pairwise preference annotations collected under a just-noticeable-difference filtering. FGResQ supports two inference modes. The *single-image scoring* mode outputs a quality score in $[0, 1]$ (higher is better), which we use for the aggregate comparison in Tab. 1. The *pairwise quality ranking* mode takes a pair of images and outputs the probabilities of three outcomes—Image 1 better, Image 2 better, or roughly equal—so that near-equivalent pairs are not forced into a strict ranking. In Tab. 2 we use this mode as follows: for each method and resolution, we run 1,000 paired comparisons between our method and the baseline on images generated from the same prompts, and report the win rate, i.e., the fraction of pairs in which our method is judged better, together with the average preference probabilities p_{ours} and p_{opp} of the two "better" classes. The residual probability mass corresponds to near-equal judgments, which should be read together with the win rate: a large win rate with a small margin between p_{ours} and p_{opp} indicates that many winning pairs are themselves subtle, whereas a persistent margin over a large tie mass shows a robust quality advantage.

G How Guidance Shapes Attention Evolution

TokenScale rests on the claim that light, early guidance stabilizes the inherited structure while releasing the model's native generative capacity, and that persistent strong guidance over-constrains generation. Image-quality comparisons support this claim only indirectly. Here we test it at the mechanism level through attention evolution: we track how, over the 4096 denoising stage, the

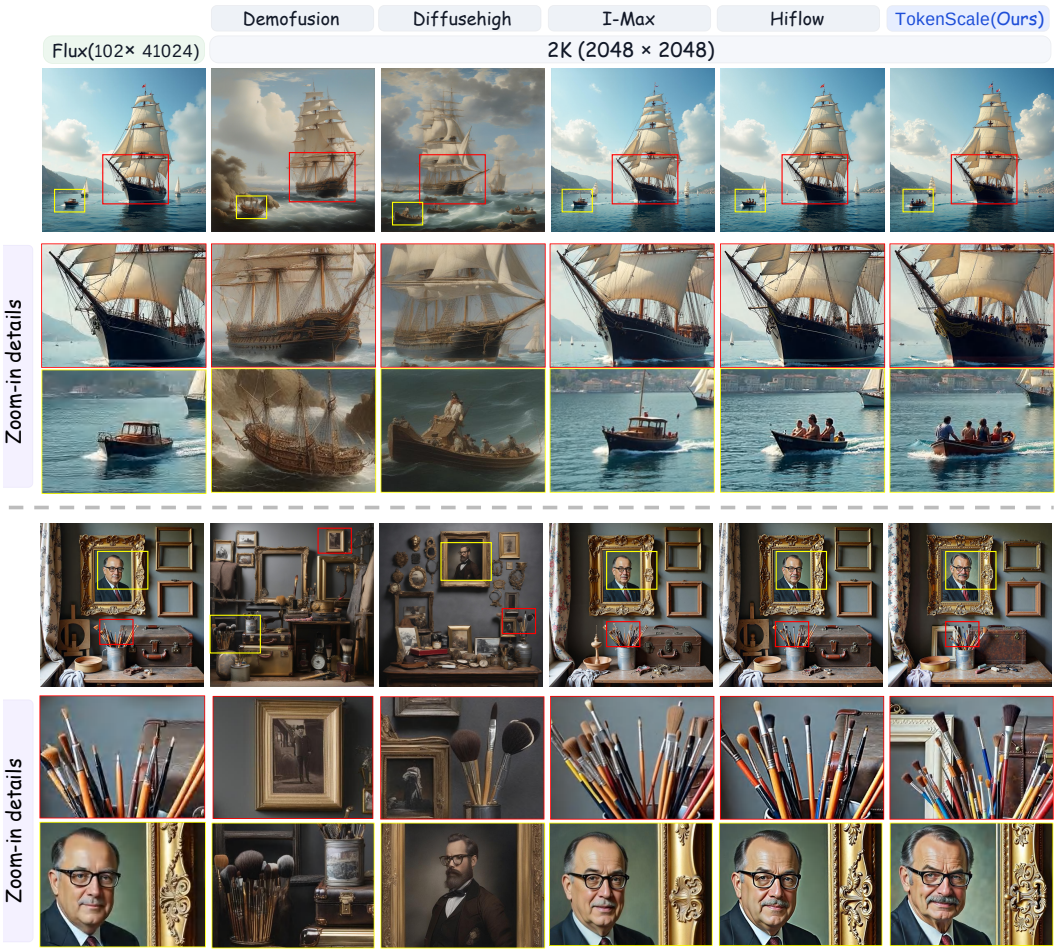

Figure 15: Additional qualitative comparison with training-free baselines at 2048.

model's attention spreads across spatial frequencies, and ask whether light guidance lets the model attend to structure beyond what the low-resolution reference can express.

New-band energy. We quantify each attention map by its *new-band energy*, the fraction of its spectral power outside the band that 1024 content can express. In FLUX.1[dev]'s VAE, a 1024 image corresponds to 64×64 latent tokens and a 4096 image to 256×256, so the axial Nyquist frequency of the 64×64 grid is 32 cycles per map and any attention map upsampled from 1024 content has its spectrum entirely within the band

$$B = \{f_x \le 32 \ \wedge \ f_y \le 32\}.$$

For a word w, let $p \in \mathbb{R}^{256\times256}$ be its normalized cross-attention map at a step of the 4096 stage (averaged over 24 heads and over double layers L0–18), with $\sum p = 1$. Removing the DC component $p' = p - 1/N$ and computing the power spectrum $P(f_x, f_y) = |\,\mathrm{rfft2}(p')|^2$ for $f_x, f_y \in [0, 128]$ (in cycles per map), we define

$$E_{\text{new}} = \frac{\sum_{(f_x,f_y)\notin B} P(f_x, f_y)}{\sum P} \ \in \ [0, 1].$$

The metric is dimensionless and scale-free, hence comparable across models. High E_{new} means the attention autonomously generates new structure that 1024 content cannot express, i.e., genuine 4096-scale detail. Low E_{new} means the attention stays in re-narration and anchoring of the 1024 content. In particular, pure upsampling of 1024 content yields $E_{\text{new}} \approx 0$.

Experimental setup. We use $T = 60$ with start step 15: the 4096 stage has 45 steps with $\sigma \in [0.754, 1/60]$, and TokenScale and HiFlow are compared step by step at aligned σ values.

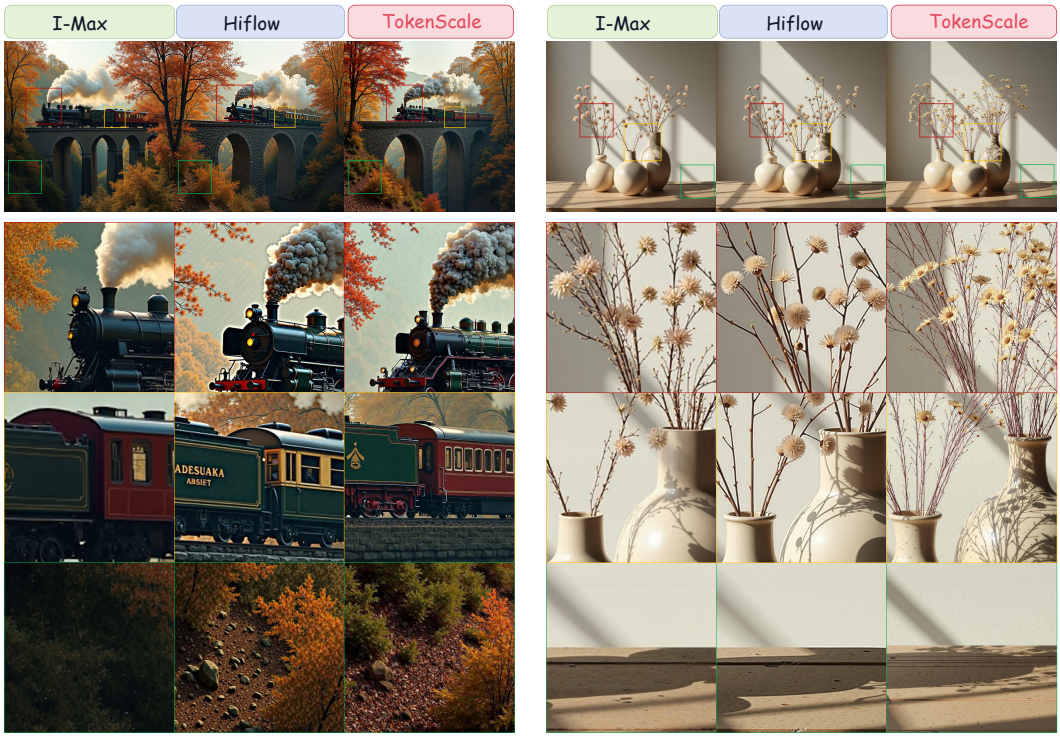

Figure 16: Additional qualitative comparison with training-free baselines at 4096.

HiFlow is tested under two configurations: one uses the same restart step as ours, and the other follows the parameters of its original paper. For each of the three prompts, the prompt words are tracked, producing one map per word per step. The shared 1024 attention (summed over all steps) upsampled bilinearly $\times 4$ serves as the reference line: its new-band energy is measured at 0.005–0.058, validating the band definition—pure upsampling carries almost no new-band energy. The two methods share the same 1024 base (the 1024-stage attention maps have IoU $= 1.00$), so the differences come entirely from the 4096 stage.

Guidance strength. We compare four guidance curves: our method with the 4-step guidance, our method with global guidance, and HiFlow under its two settings. Our light guidance attains higher new-band energy than the other curves on most words, and this advantage persists into the final steps, whereas HiFlow's curves drop rapidly toward the upsampled-reference level (0.005–0.058). This suggests that light guidance lets the model attend to higher-frequency details, whereas over-guidance constrains its capability. The light token-space guidance also brings two practical benefits: we can improve quality by restarting at the optimal restart step (Sec. 4.4) and by increasing the number of sampling steps, without the over-constraint that additional pixel-domain guidance imposes on the image.

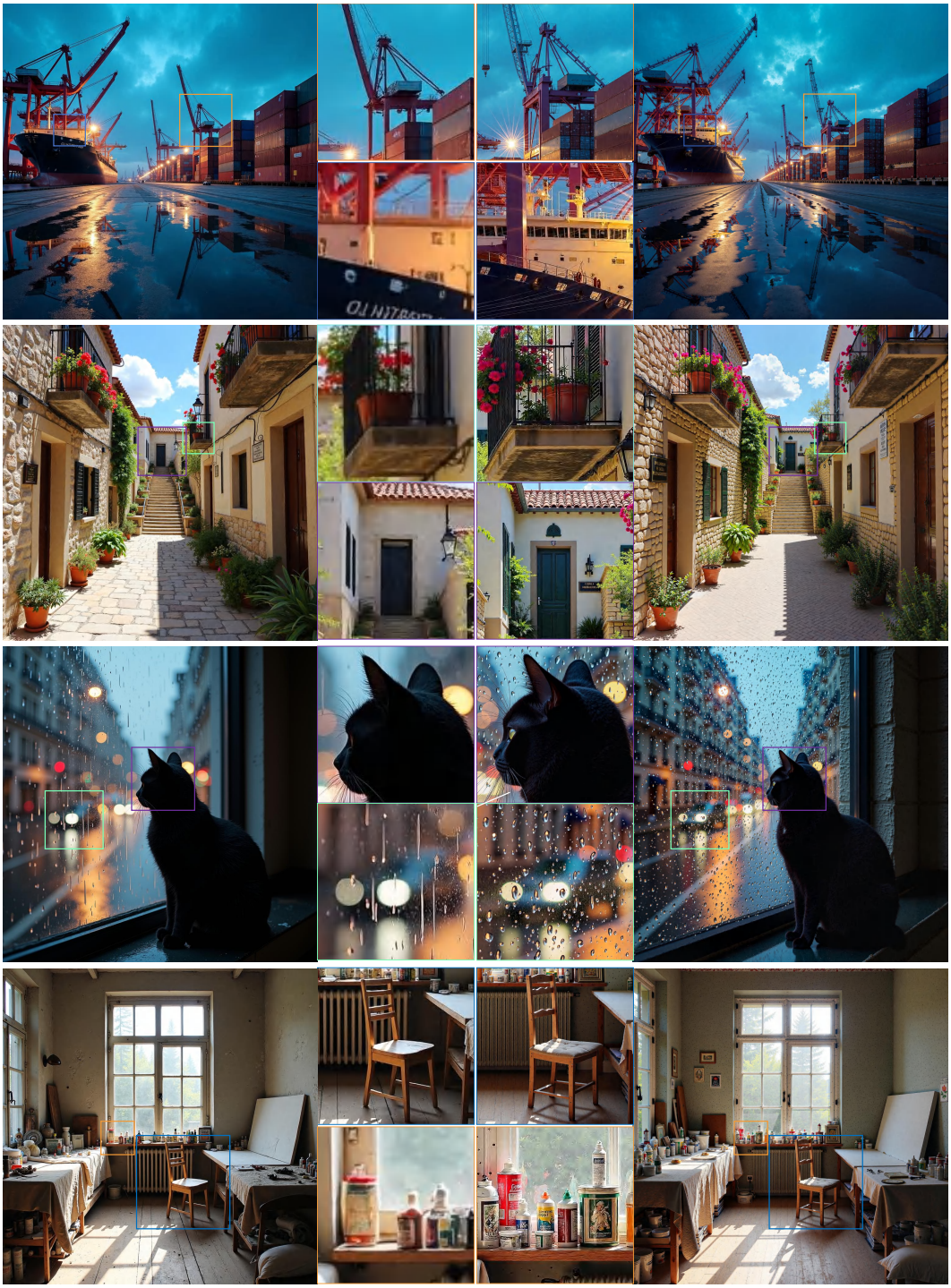

Figure 17: Additional showcase: 1K base images and the corresponding 4K outputs generated by TokenScale, with enlarged views of matching local regions.

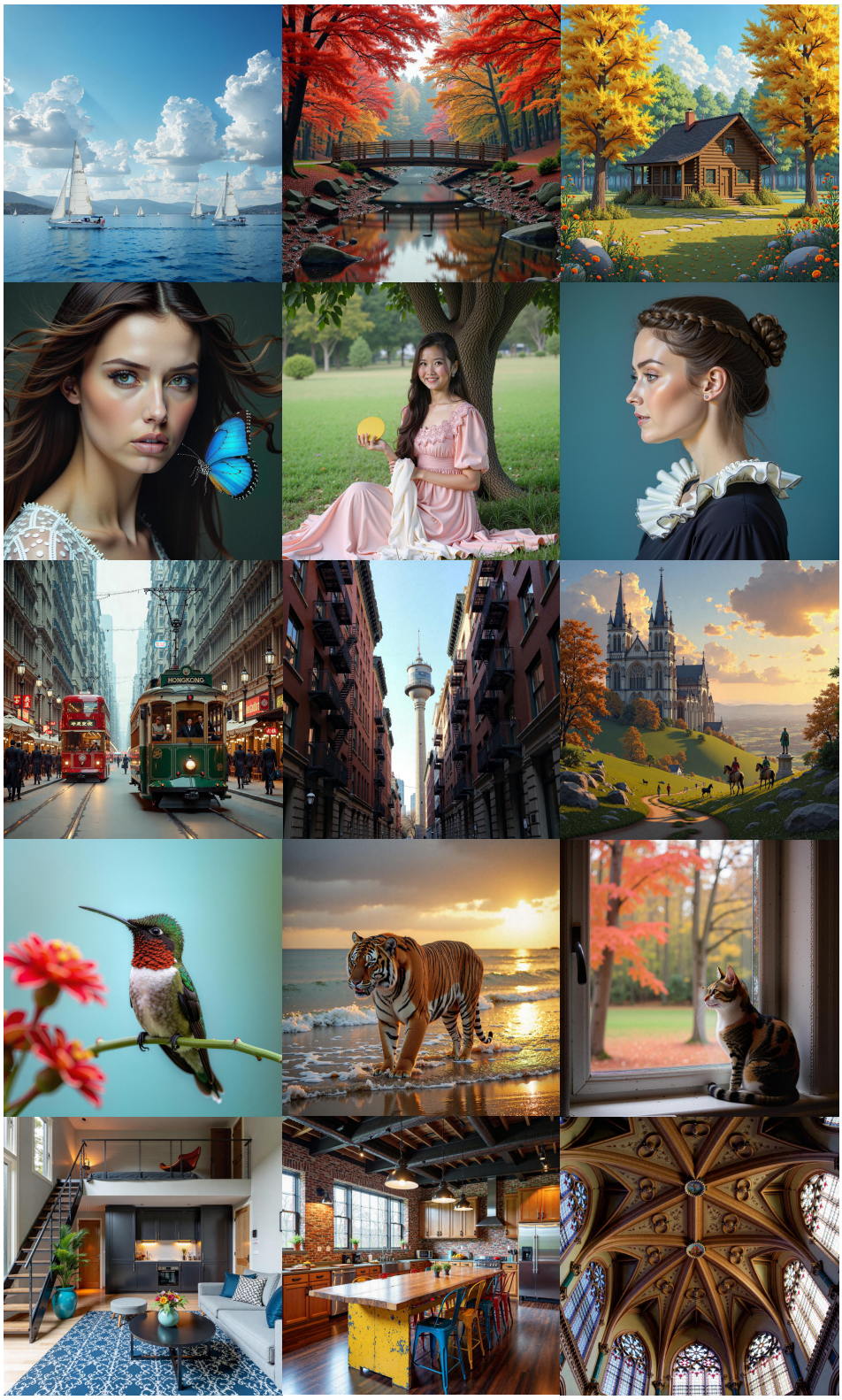

Figure 18: Additional showcase: more 4K images generated by TokenScale.



Figure 19: Stability of TokenScale across different random seeds for the same prompt.

Full results. Averaging over σ per word, TokenScale attains higher new-band energy on most words. Below, for each of the three prompts, we provide the full prompt text with the words tracked in the new-band energy analysis highlighted with a light-yellow marker, followed by the generated candidates.

> Prompt 1. A luxury rooftop infinity pool overlooking a glowing modern skyline at night, reflections on water, photorealistic architectural scene.

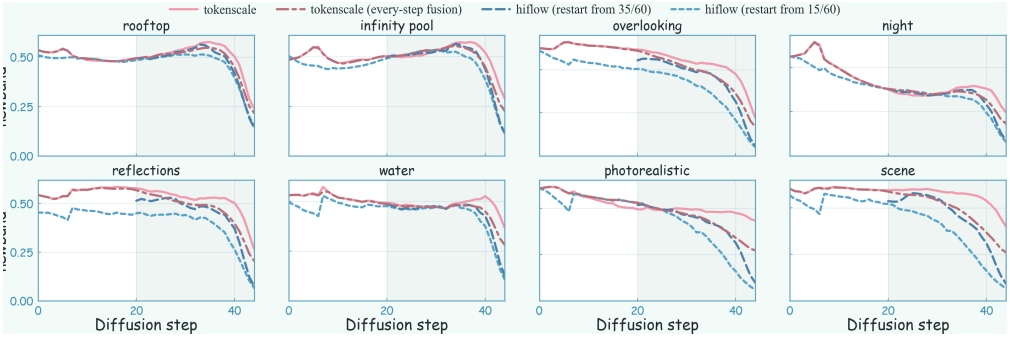

Figure 20: Generated candidates for Prompt 1.

> Prompt 2. A romantic Mediterranean cliffside village at sunset, white houses covered in pink bougainvillea, stone stairways winding down toward the sea, glowing windows, small boats on sparkling water, orange and rose-colored sky, rich depth, cinematic realism.

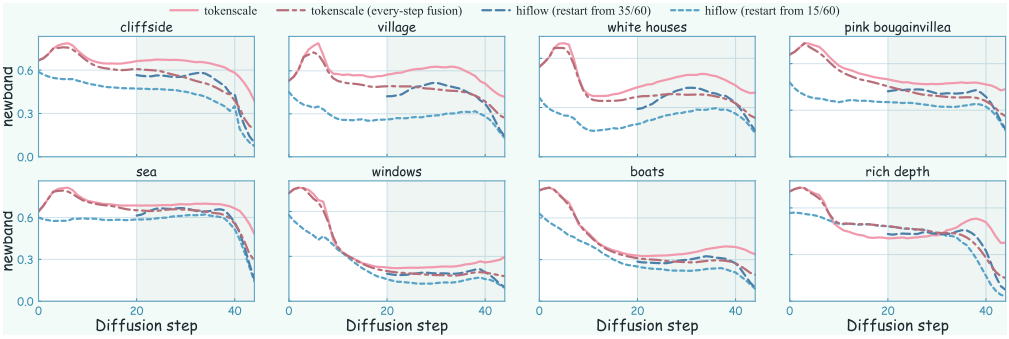

Figure 21: Generated candidates for Prompt 2.

Prompt 3. A rainy tram stop in a Central European city , old tram approaching , wet tracks , people in coats , muted winter palette , cinematic realism .

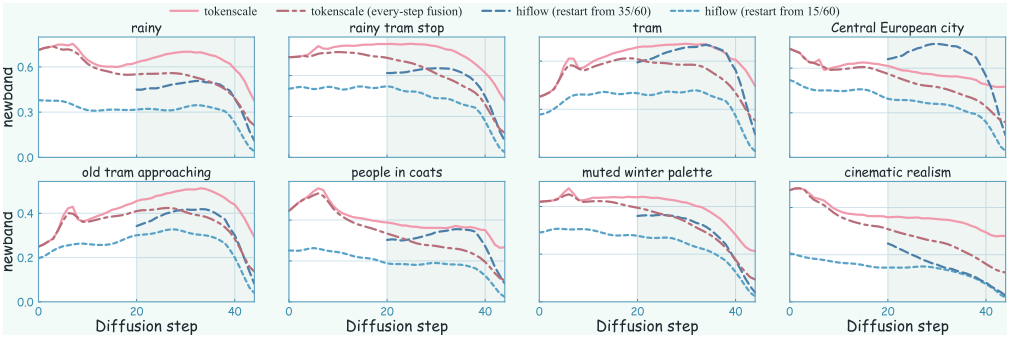

Figure 22: Generated candidates for Prompt 3.

Boundary. This metric measures the structural novelty of attention maps (mechanism evidence) and does not directly measure image quality.

We additionally provide the figures for the two practical benefits discussed above: the restart-ratio comparison (Fig. 23) and the effect of the number of sampling steps (Fig. 24).

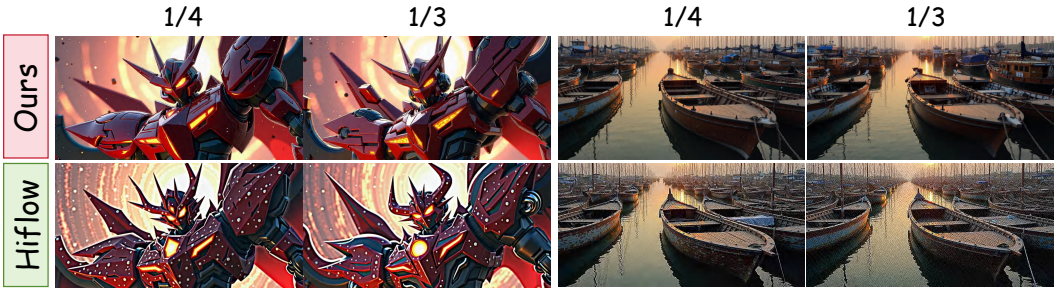

Figure 23: Restarting the super-resolution stage earlier. Top: our method, where restarting at an early stage yields the best quality. Bottom: with persistent guidance, restarting this early degrades the image.

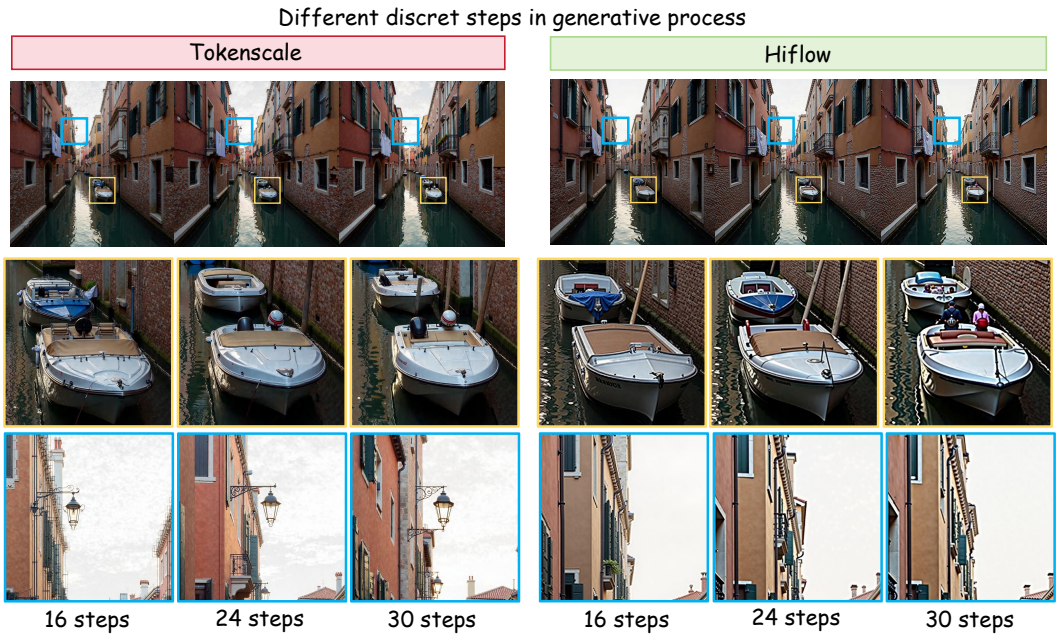

Figure 24: Effect of discrete sampling steps. Ours: quality improves monotonically as the number of steps increases. Prior work: more steps accumulate more per-step image-domain guidance and degrade the image.

H Extended Related Work

U-Net-era training-free methods. Beyond the two routes of Sec. 2, an earlier line of training-free methods was designed for U-Net-based diffusion models. MultiDiffusion and SyncDiffusion fuse or synchronize multiple diffusion paths over the canvas (Bar-Tal et al., 2023; Lee et al., 2023); ScaleCrafter, HiDiffusion, and FouriScale modify dilated receptive fields, feature maps, or frequency components (He et al., 2024; Zhang et al., 2024; Huang et al., 2024); and FreeScale and ResMaster preserve cross-scale information through feature fusion (Qiu et al., 2025; Shi et al., 2025). These methods rely on the convolutional U-Net structure and do not transfer to DiT-based flow models, which motivates the token-space treatment adopted in this work.

Origins of the positional-encoding route. The positional-encoding route of Sec. 2 descends from context-length extrapolation in language models: position interpolation (Chen et al., 2023), NTK-aware RoPE scaling (Peng & Quesnelle, 2023), and YaRN (Peng et al., 2024), which DyPE, UltraImage, SigMa, and ResDiT adapt to the RoPE of image DiTs. We enable the NTK-aware variant in all high-resolution stages (Appendix F).

I Limitation

The main limitation of TokenScale lies in complex irregular textures and noise-like random details, such as glaciers, starry skies, cave walls, and coral reefs, which lack explicit semantics. Fig. 25 (left) shows such cases, where the model does not know how the stalactites or coral reefs should be super-resolved, so it keeps stacking details without a faithful structure, producing unnatural results. In contrast, for semantically clear objects such as the mermaid (Fig. 25, right), the model has a strong prior for generating sharp details and can fully exploit this capability at high resolution. This issue is further amplified by the reference-based design: token expansion inherits the over-simplified structure of the low-resolution reference, and token fusion pulls strongly deviating tokens back toward it, suppressing the natural evolution of such textures. This difficulty is not specific to TokenScale: such regions are intrinsically noise-like for the image itself, and predicting them is inherently hard in image generation. Token fusion alleviates this issue to some extent, but does not fully resolve it.



Figure 25: Limitation cases. Left: noise-like random details without explicit semantics, such as stalactites and coral reefs, where the model keeps stacking details without a faithful structure. Right: semantically clear objects such as the mermaid, where the model exploits its strong prior and generates sharp details.

J PROMPTS FOR FIGURES

Table 5: The image generation prompts for each figure are listed sequentially, following the order in which the figures appear in the paper (Part 1, 19 prompts).

Figure	Text Prompt
Fig. 1	A vintage-style bus floats amid colorful clouds, outfitted with futuristic propulsion devices and glowing lights, displaying a mix of retro charm and sci-fi elements.
	A vibrant street scene features people in 19th-century attire walking alongside a tree with autumn leaves, horse-drawn carriages, and a green trolley car, set against historic buildings under a cloudy sky.
	A woman in a flowing red dress floats gracefully underwater, her long hair billowing around her as she covers her mouth with one hand, surrounded by clear blue water and a sandy seabed.
	A grand white mosque features multiple large domes and towering minarets, surrounded by a vast, polished marble courtyard with intricate floral designs. The sky is clear with scattered clouds, enhancing the mosque's majestic appearance.
	A woman with voluminous red hair and bold red lips poses confidently, wearing an off-the-shoulder light blue top.
	Colorful wooden houses with sloped roofs line the waterfront, reflecting in calm water under a dramatic sky.
	Two fishing boats, one blue and one white, dock at a harbor with reflections shimmering on the calm water, surrounded by a muted grey sky.
	A frying pan filled with two pieces of golden-brown fish atop a bed of rice, garnished with cherry tomatoes, avocado slices, fresh cilantro, and a dollop of sauce, accompanied by a small bowl of creamy dressing and a blue cloth nearby on a cracked peach-colored surface.
	A vibrant, detailed landscape featuring a small town with red-brick buildings, green trees, and a rural backdrop. Prominently displayed in the background is a grand temple-like structure and a circular building, with a railroad track featuring a vintage streetcar running through the scene. Workers are seen in a field, and livestock grazes nearby, under a blue sky with fluffy clouds.
	A charming coastal cottage surrounded by vibrant flowers and palm trees, with a pathway leading to the beach and a clear blue sky overhead.
Fig. 2	A romantic Mediterranean cliffside village at sunset, white houses covered in pink bougainvillea, stone stairways winding down toward the sea, glowing windows, small boats on sparkling water, orange and rose-colored sky, rich depth, cinematic realism
	A rooftop garden above a modern city at blue hour, lush greenery and flowers surrounding reflective pools, elegant lights along the pathways, glowing skyscrapers in the background, twilight sky fading from blue to purple, luxurious cinematic atmosphere
	A rainy tram stop in a Central European city, old tram approaching, wet tracks, people in coats, muted winter palette, cinematic realism
Fig. 4	A blonde woman in a vibrant pink and green futuristic outfit sits on a sleek, heavily customized motorcycle adorned with stickers, while a curious cat watches nearby against a backdrop of a modern city skyline under a blue sky with clouds.
Fig. 6	A rustic wooden table with fresh fruit, bread, and flowers, natural window lighting, classical still life photography, highly detailed
	A bustling European old town square at dusk, warm street lamps, outdoor cafés, cobblestone streets, people walking, cinematic urban photography, ultra-detailed
Fig. 7	A rustic wooden table with fresh fruit, bread, and flowers, natural window lighting, classical still life photography, highly detailed
	A luxury department store atrium during holiday season, glass elevators, decorative lights, polished surfaces, rich interior detail
Fig. 8	A rustic wooden table with fresh fruit, bread, and flowers, natural window lighting, classical still life photography, highly detailed

Table 6: The image generation prompts for each figure are listed sequentially, following the order in which the figures appear in the paper (Part 2, 11 prompts).

Figure	Text Prompt
Fig. 9	A thriving coral reef city teeming with life, table corals and sea fans in every color, schools of anthias fish swirling like confetti, a manta ray gliding overhead, sun rays piercing the blue water, tiny cleaner shrimp and clownfish among anemones, vibrant underwater photography, dense ecological detail
	Extreme macro photograph of a resting dragonfly, compound eyes reflecting a field of wildflowers, delicate translucent wing veins catching golden backlight, water droplets beading on its abdomen, dew-covered grass stem, creamy bokeh background, intricate biological detail
	Dutch Golden Age oil painting still life, a pewter ewer, a peeled lemon with curling rind, oysters on a silver plate, a half-filled wine glass catching window light, dark velvet backdrop, intricate reflections in polished metal, dramatic chiaroscuro, luminous paint layers, museum masterpiece quality
	Delicate watercolor painting of Venice at golden hour, gondolas moored along a narrow canal, reflections dissolving into loose washes of gold and turquoise, laundry lines between ochre facades, a distant dome, pigment blooms and soft edges, wet-on-wet technique, romantic travel sketchbook style
	Ukiyo-e woodblock print in the style of Hokusai, an enormous indigo wave with clawing white foam looming over three fishing boats whose crews bow low, snow-capped Mount Fuji and flying seabirds in the distance, intricate woodgrain lines and washi paper fibers visible, subtle ink registration offsets, traditional palette of deep indigo, ochre and ivory, dynamic and dramatic composition
	Epic anime background art, a floating sky city with waterfalls pouring off its edges into the clouds, airships and flying machines circling its spires, a lone girl with a red kite standing on a grassy cliff in the foreground, sunset colors of pink orange and lavender, lens flare, breathtaking scale, in the style of Makoto Shinkai
	Impressionist oil painting of Monet's garden at Giverny, the arched Japanese footbridge draped in cascading wisteria spanning a pond crowded with water lilies, willow branches trailing into the water, layered dappled reflections of a golden sky between the blossoms, visible thick brushstrokes and canvas texture, palette of lavender rose sage and gold, lush and serene
	A tropical storm approaching a coastal town, palm trees bending in the wind, shuttered houses, dark sea, high weather drama
	A seaside monastery carved into white cliffs, long stairs, turquoise water below, sunlit stone textures, Mediterranean realism
	A winding stone pathway leads down from a hilltop village, flanked by a chain-link fence, with vibrant houses nestled below and dramatic mountains rising in the background against a colorful sunset sky.
	A quiet snowbound gas station on a remote road, single light glowing, tire tracks, deep blue evening tones, cinematic isolation

Table 7: The image generation prompts for each figure are listed sequentially, following the order in which the figures appear in the paper (Part 3, 15 prompts).

Figure	Text Prompt
Fig. 10	A cute and adorable fluffy puppy wearing a witch hat in a halloween autumn evening forest, falling autumn leaves, brown acorns on the ground, halloween pumpkins spiderwebs, bats, a witch's broom.
	A majestic, metallic giraffe walking through a futuristic city park, its long neck towering above the plants and creatures.
	A city at night, illuminated by neon lights, with cars zooming between towering skyscrapers, and a giant holographic owl in the sky.
	A mystical elf with emerald green eyes and a crown of leaves, standing in a forest glade bathed in soft moonlight.
Fig. 12	A still life of a few roses with layered petals in soft blush pink and cream in a small ceramic vase on a wooden table, occupying only a modest corner of the scene, beside a stack of vintage books, a brass candlestick and scattered petals, soft diffused window light, petals in blush pink, cream, coral and dusty lavender, delicate watercolor washes blended with fine oil-painting detail, elegant artistic composition, high detail, beautiful color palette
Fig. 14	A vibrant cityscape illuminated at night, featuring the Tokyo Tower as a central highlight, surrounded by a sea of skyscrapers and a starry sky.
	A futuristic spaceship with sleek, dark metallic features and glowing orange lights stands stationed on a misty landing platform, surrounded by towering structures and shadowy figures, while technicians work with equipment below.
Fig. 15	A large sailing ship with billowing sails dominates the foreground, while smaller boats and sailboats navigate the calm waters; a distant coastal town and rolling hills are partially obscured by soft blue hues and light clouds.
	A cluttered artist's corner features an ornate gold frame, a painting of a man with glasses, several empty picture frames, a vintage suitcase, and various brushes in a tin can, all set against a textured gray wall and draped with patterned fabric
Fig. 16	A vintage steam locomotive crossing an old stone viaduct through an autumn forest, billowing white steam, golden and crimson foliage, warm sunlight breaking through the trees, nostalgic cinematic atmosphere, rich warm-cool color contrast
	A minimalist still life of handmade ceramic vases and dried flowers arranged on a weathered wooden table, soft window light casting long shadows, muted earthy Morandi-inspired palette, quiet elegance, fine-art composition
Fig. 17	A seaport container yard at blue hour, cranes, stacked containers, wet ground reflections, industrial precision, cinematic realism
	A narrow cobblestone alleyway flanked by rustic stone buildings adorned with vibrant flower pots, leading to a staircase illuminated by sunlight and a blue sky dotted with clouds above.
	A black cat sitting by a rainy window at night, city lights blurred outside, moody cinematic lighting, cozy and emotional atmosphere
	A cluttered artist's studio with north-facing windows, oil paint tubes, unfinished canvases, dusty wooden floor, textured fabrics, soft natural light

Table 8: The image generation prompts for each figure are listed sequentially, following the order in which the figures appear in the paper (Part 4, 22 prompts).

Figure	Text Prompt
Fig. 18	Sailing boats dot the calm water under a clear blue sky filled with fluffy white clouds, with a distant shoreline and gentle hills in the background.
	Vibrant red leaves contrast against tranquil water reflecting trees in autumn colors, with a wooden bridge spanning a small creek amidst scattered rocks and fallen leaves.
	A cozy wooden cabin nestled among tall, vibrant yellow and green trees, surrounded by wildflowers and a clear blue sky.
	A portrait of a woman with flowing hair, gazing intensely, alongside a vibrant blue butterfly.
	A young woman in a soft pink dress with a lace collar sits gracefully under a tree, holding a white cloth and a yellow circle, surrounded by greenery and a serene landscape.
	A woman's profile with a braided hairstyle, wearing a dark dress with a white, ruffled collar, set against a blue background.
	A green tram labeled "Hongkong" and a red double-decker bus navigate a bustling street lined with tall buildings, illuminated by streetlights, while pedestrians in vintage clothing walk along the pavement.
	Narrow city street lined with tall buildings, featuring a mix of architectural styles, fire escapes, and a prominent water tower, leading towards a distant skyline.
	A picturesque landscape with a gothic cathedral featuring tall spires, surrounded by vibrant trees and rolling hills; travelers on horseback and pedestrians traverse the lush greenery, while a statue and animals dot the serene setting under a colorful sky at sunset.
	A vibrant hummingbird with a green head and fiery red throat perched on a green stem, surrounded by red flowers against a soft blue background.
	A tiger stands in shallow water at the beach, reflecting golden sunlight as waves lap at its legs, with a calm ocean and a cloudy sunset in the background.
	A calico cat sits on a windowsill, gazing outside at a forest of autumn trees with vibrant orange and red leaves covering the ground.
	A modern loft features a spacious, open-plan living area with a grey sofa, circular coffee tables, and a blue patterned rug, complemented by tropical flowers in a turquoise vase. A staircase leads to a lofted sleeping area with a wooden chair, adjacent to a sleek kitchen with black cabinets and stainless steel appliances, illuminated by warm lighting.
	A spacious, industrial-style kitchen features wooden cabinets, stainless steel appliances, and a large dining table with a distressed yellow base surrounded by colorful metal chairs. Bright pendant lights hang from the ceiling, illuminating the room's rustic charm. Large windows let in natural light, casting a warm glow on the dark wood floor.
	Intricate vaulted ceiling adorned with elaborate woodwork, surrounded by vibrant stained glass windows, and featuring a central star-shaped ornament and a colorful mural at the apex.
Fig. 19	A futuristic cityscape with sleek skyscrapers, elaborate bridges, and a vibrant blue sky, surrounded by lush greenery and turquoise waters, while a red and white spacecraft zips through the scene.
Fig. 20	A luxury rooftop infinity pool overlooking a glowing modern skyline at night, reflections on water, photorealistic architectural scene.
	A romantic Mediterranean cliffside village at sunset, white houses covered in pink bougainvillea, stone stairways winding down toward the sea, glowing windows, small boats on sparkling water, orange and rose-colored sky, rich depth, cinematic realism.
	A rainy tram stop in a Central European city, old tram approaching, wet tracks, people in coats, muted winter palette, cinematic realism.
Fig. 23	A dramatic bust portrait of a red Gundam-style battle mecha, glossy crimson and scarlet armor plating with weathered scratches and panel lines, glowing amber eyes, exposed mechanical joints on the shoulders, soft rim lighting against a dark smoky backdrop, volumetric light rays, fine dust floating in the air, hyper-detailed photorealistic sci-fi rendering
	A harbor filled with traditional wooden boats at sunrise, ropes, nets, peeling paint, glassy water, soft pastel light, documentary-style realism
Fig. 24	A narrow Venetian canal at first light, old stone facades, laundry hanging above the water, small ripples reflecting pastel sky, quiet atmosphere, cinematic realism