LSSGen: Leveraging Latent Space Scaling in Flow and Diffusion for Efficient Text to Image Generation

Supplementary Material

1. Complexity Analysis

We analyze the computational complexity implications of resolution scaling in diffusion models. When image resolution increases by a factor of k, the computational requirements grow quadratically as $O(k^2)$ due to the direct relationship between resolution and total pixel count. This scaling effect becomes particularly significant in transformer-based architectures like Diffusion Transformer (DiT) [4], where self-attention [6] operations exhibit quadratic complexity with respect to sequence length. For generation tasks beyond 1024^2 resolution, this computational burden becomes prohibitive. Our progressive dynamic resolution approach addresses this limitation by performing initial probabilistic paths at lower resolutions, thereby substantially reducing overall computational demands. The computational complexity manifests in both transformer-based and CNN architectures, though with different scaling characteristics. For transformer models utilizing self-attention mechanisms with complexity $O(n^2)$ [6], where n represents the number of image patches, the total computational complexity scales as:

$$C(H, W) = O((H \times W) \cdot (H \times W)) = O((H \times W)^2)$$

where $C(H, W)$ denotes the computational cost for an image of height H and width W. This formulation yields two crucial insights:

1. Reducing patch count through our dynamic resolution strategy provides quadratic computational savings in self-attention operations.
2. The quartic relationship between resolution and computational requirements (in both FLOPs and memory) demonstrates the theoretical efficiency of our progressive resolution approach compared to fixed-resolution methods.

2. Case Study

2.1. Effect of Initial Noise Intensity Parameter

The parameter σ_{init} controls the initial noise intensity in the Latent Space Scaling Generation (LSSGen) generation stages. Fig. S1 illustrates the visual impact of different σ_{init} values. The results demonstrate that higher σ_{init} values produce images with enhanced detail sharpness. Among all configurations examined, $\sigma_{\text{init}} = 0.75$ yields optimal quality across the tested parameter range. Our method performs upsampling operations in latent space rather than pixel space, which preserves the structural integrity of the generated images across different parameter settings. This observation suggests a configurable quality-speed tradeoff: a slight reduction in image detail fidelity can yield significant inference acceleration. Thus, σ_{init} serves as an effective control parameter that enables practitioners to balance generation quality against computational efficiency according to application requirements.

2.2. Comparative Analysis of Progressive Scaling Approaches in Flow Model

Fig. S2 presents a qualitative comparison of various progressive scaling approaches. We evaluate the baseline FLUX.1-dev [1], the pixel-space approach MegaFusion [7], and our proposed method LSSGen. The visual results demonstrate that LSSGen preserves fine-grained details with superior fidelity compared to MegaFusion, which exhibits characteristic blur artifacts inherent to pixel-space scaling transformations. This comparison substantiates the efficacy of latent-space manipulation in preserving high-frequency components during multi-resolution synthesis. The enhanced perceptual quality is particularly evident in complex textures and sharp boundaries, where our approach maintains structural coherence across different resolution scales. These observations align with our quantitative metrics that indicate significant improvements in both computational efficiency and generation quality.

2.3. Comparative Analysis of LSSGen in Distillated Flow Model

Fig. S3 presents a quantitative evaluation contrasting the non-scaling baseline approach with our proposed LSSGen framework. We systematically compare FLUX.1-schnell [1] against LSSGen across multiple generation tasks. The empirical results demonstrate that LSSGen consistently produces outputs with enhanced perceptual quality characterized by superior detail preservation and edge definition. This quality enhancement stems from our latent space scaling operations that maintain high-frequency information through the generation process. The latent-space manipulations introduce precisely controlled detail amplification that manifests as enhanced definition in textural elements. This characteristic proves advantageous for most generative tasks, though practitioners should note potential over-sharpening effects when synthesizing human facial features, where the enhanced detail reproduction may accentuate fine wrinkles beyond natural appearance.

Figure S1. Comparison between different σ_{init} setting on SD3.5-medium [2].

Most significantly, our progressive upsampling methodology achieves a substantial inference time reduction from 4.77s to 4.36s (an 1.1x acceleration) while maintaining generation quality integrity. This performance optimization exemplifies the efficiency even further with timestep-distilled models, indicating the broadness of our methods. The demonstrated compatibility with model distillation techniques underscores the generalizability of our latent space transformations across varied architectural configurations and computational constraints.

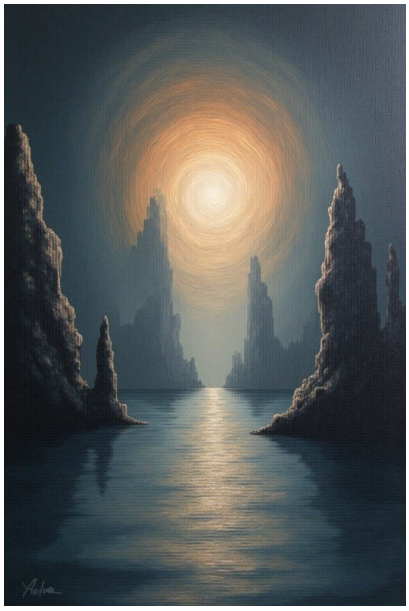
FLUX.1-dev	FLUX.1-dev-MegaFusion	LSS-FLUX.1-dev (ours)
		
a photo of a bench and a snowboard.		
		
An epic painting of Gandalf the Black summoning thunder and lightning in the mountains.		
		
An oil painting of a latent space.		

Figure S2. More comparison between different progressive approaches on FLUX.1-dev [1] with 1024^2 resolution.

2.4. Comparative Analysis of LSSGen in Diffusion Models

Fig. S4 presents a quantitative evaluation contrasting various approaches with our proposed LSSGen framework. We systematically compare SDXL [5], Self-Cascade [3], and MegaFusion [7] against LSSGen across multiple prompts. The results demonstrate that LSSGen produces outputs with quality comparable to Self-Cascade while offering significant advantages in inference speed and universal applicability. In contrast, MegaFusion exhibits consistent blurriness across all generated samples, highlighting the limitations of pixel-space transformation methods. The superior performance of our approach stems from effective latent space manipulations that preserve semantic structure and fine details during the scaling process. This enables LSSGen to maintain perceptual quality while achieving computational efficiency that makes it practical for real-world applications requiring both high-quality outputs and responsive generation times.

2.5. Details of Timesteps in LSSGen

This section provides a detailed description of the input parameters for our proposed LSSGen, as presented in Algorithm 1. These parameters allow for precise control over the progressive generation process, enabling users to balance computational efficiency and final image quality. Each parameter is defined as follows:

`min_resolution`: This integer value specifies the initial, lowest resolution at which the generative process begins. The first stage of the pipeline synthesizes a latent tensor at this resolution from pure noise. We use 512 in FLUX.1-dev.

`target_resolution`: Defines the final, desired resolution of the output image. The LSSGenframework progressively upscales the latent representation through multiple stages until this target resolution is reached.

`base_resolution`: Serves as a reference resolution for the dynamic step calculation. When the shorten_steps flag is enabled, any generation stage operating at a resolution lower than base_resolution will use a proportionally reduced number of denoising steps.

`base_steps`: The baseline number of denoising steps performed by the diffusion pipeline (e.g., FLUX.1-dev and SDXL are 50 steps) for any stage operating at or above the base_resolution.

`init_noise_level`: This floating-point value corresponds to the initial noise coefficient (σ_{init}) discussed in our methodology (Section 4.1). It governs the ratio between the signal from the upsampled latent and the stochastic noise injected at the beginning of a new stage. Based on our analysis, this is typically set to 0.75 for optimal quality. The parameter can be lower if efficiency is desired.

`shorten_steps`: A boolean flag that enables (True) or disables (False) the intermediate step reduction strategy. When enabled, this optimization accelerates the initial, lower-resolution stages by reducing their denoising step count, significantly improving overall inference speed with minimal impact on quality.

Together, these parameters provide fine-grained control over the speed-quality trade-off within the LSSGen framework, making it adaptable to various hardware constraints and use cases.

References

[1] Black Forest Labs. FLUX.1: Text-to-image Generation Model. https://github.com/black-forest-labs/flux, 2024. Released: August 2, 2024. 1, 3, 6, 8, 9

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024. 2

[3] 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 *Proceedings of the European Conference on Computer Vision*, pages 39–55, 2024. 4

[4] 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. 1

[5] 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. In *Proceedings of the International Conference on Learning Representations*, 2024. 4, 7

[6] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[7] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 1, 4

Algorithm 1 LSSGen: Latent Space Scaling Generation

Input:
  `min_resolution` // *Initial generation resolution*
  `target_resolution` // *Final desired resolution*
  `base_resolution` // *Reference resolution for step calculation*
  `base_steps` // *Number of diffusion steps for the base resolution*
  `init_noise_level` // *Initial noise factor* σ_{init}
  `shorten_steps` // *Boolean to enable step reduction*

Initialize:
`stages` ← Define progressive scaling stages from `min_resolution` to `target_resolution`
`latents` ← `RandomNoiseGenerator`(`min_resolution`)

for `stage_res` in `stages` **do**
    if `stage_res` > `min_resolution` **then**
        `upsampled_latents` ← `Upsampler`(`latents`)
        `noise` ← `RandomNoise`(`stage_res`)
        `latents` ← `upsampled_latents` * (1 - `init_noise_level`) + `noise` * `init_noise_level`
    end if
    // *Calculate steps for the current stage*
    if `shorten_steps` and `stage_res` < `base_resolution` **then**
        `steps` ← `base_steps` / int(`base_resolution` / `stage_res`)
    else
        `steps` ← `base_steps`
    end if
    `latents` ← `DiffusionPipeline`(`latents`, `steps`, `stage_res`)
end for
`images` ← `VAE_Decode`(`latents`)
return `images`

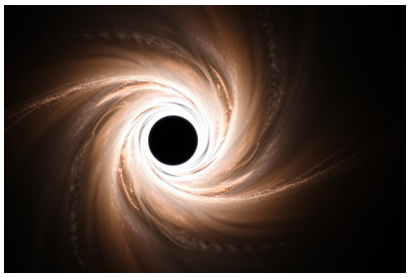
Prompt	FLUX.1-schnell	LSS-FLUX.1-schnell (ours)
A photo of a black hole.		
On a beautiful snowy mountain top, stands a sign with 'LSSGen' on it.		
A beautiful lady holding a sign saying 'LSSGen'.		
A giant rocket launching from a huge chocolate cake.		

Figure S3. More comparison between different progressive approaches on FLUX.1-schnell [1] with 1024^2 resolution.

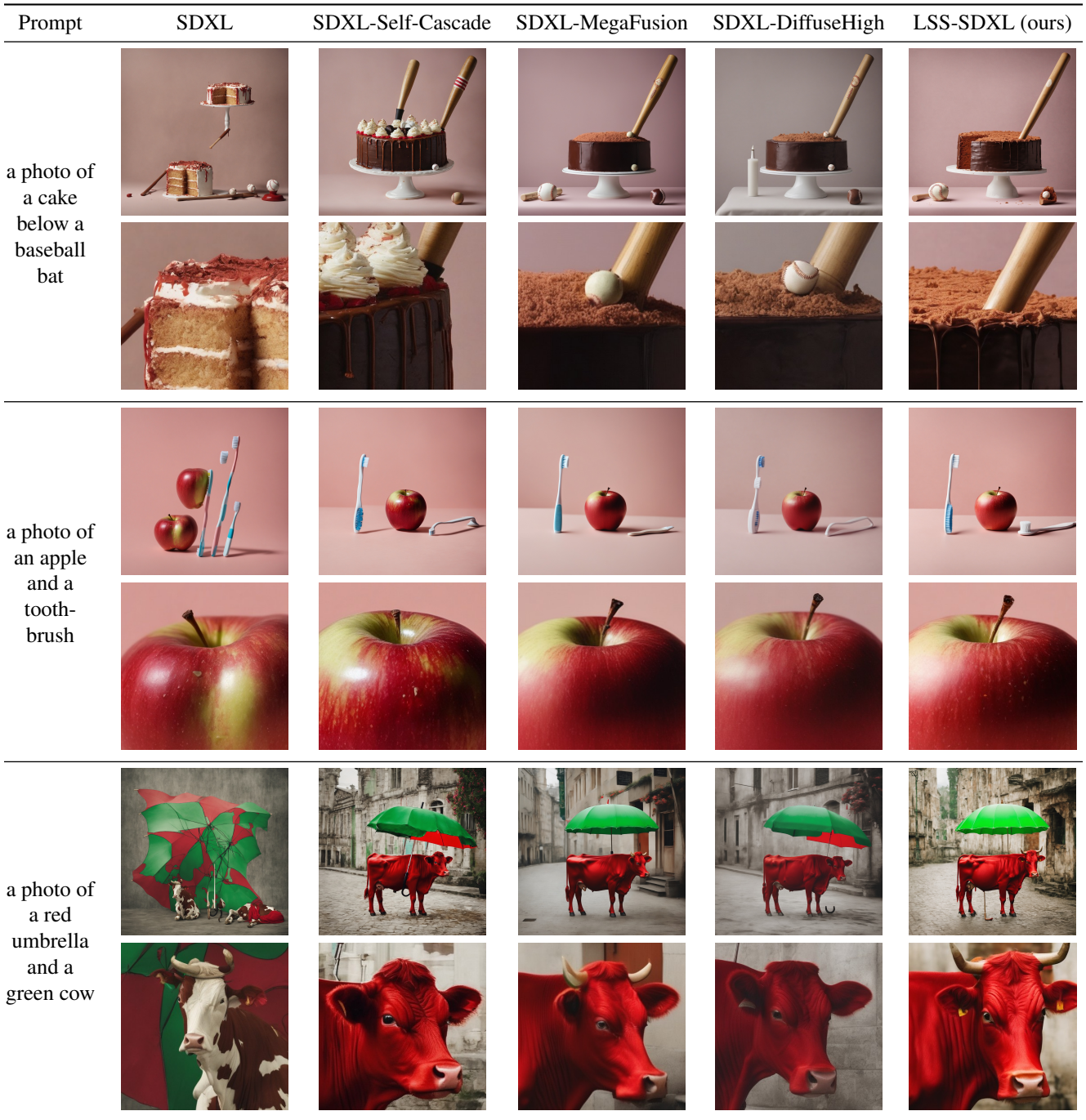


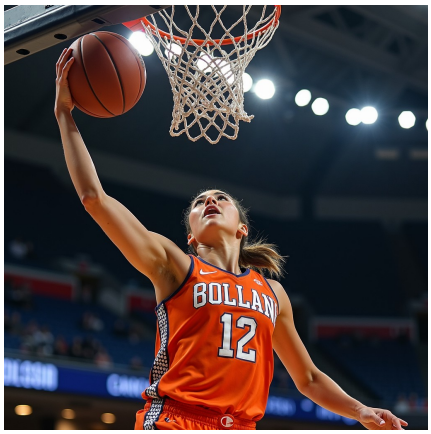
Figure S4. More comparison between different progressive approaches on SDXL [5], where the top row represents the original 2048^2 output images and bottom row represents cropped results.

detailed pen and ink drawing of a massive complex alien space ship above a farm in the middle of nowhere

an anthopomorphic pink donut with a mustache and cowboy hat standing by a log cabin in a forest with an old 1970s orange truck in the driveway

a cartoon of aboy playing with a tiger

Medium shot of a friendly male barista with curly hair and an apron, smiling in a cozy, warm-lit caf



Action shot of a female basketball player mid-dunk, with an intense expression in a brightly lit arena

Full-body shot of a woman with pink hair in a neon-lit, rainy Tokyo street at night, wearing a futuristic jacket

photo of a bear wearing a suit and tophat in a river in the middle of a forest holding a sign that says I cant bear it

A futuristic city at night, where the skyscrapers are built from living, glowing organic material,

An intricate city of marble towers and ornate bridges built upon a dense layer of clouds,

Figure S5. More generated samples of LSSGen on FLUX.1-dev [1] at 1024^2 resolution.

tilt shift aerial photo of a cute city made of sushi on a wooden table in the evening

dark high contrast render of a psychedelic tree of life illuminating dust in a mystical cave

fox sitting in front of a computer in a messy room at night. On the screen is a 3d modeling program with a line render of a zebra

cat patting a crystal ball with the number 7 written it in black marker

Cute adorable little goat, unreal engine, cozy interior lighting, art station, detailed digital painting, cinematic, octane rendering

Close-up portrait of an old fisherman with a kind, wrinkled face and white beard, soft window lighting

Fantasy portrait of a non-binary elf with silver hair and violet eyes on a throne, in a grand hall with moonlight

An intricate city of marble towers and ornate bridges built upon a dense layer of clouds,

A cozy village of round-doored homes built into lush green hills

Figure S6. More generated samples of LSSGen on FLUX.1-dev [1] at 2048^2 resolution.