FreSca: Scaling in Frequency Space Enhances Diffusion Models

Chao Huang[1], Susan Liang[1], Yunlong Tang[1], Jing Bi[1], Li Ma[2], Yapeng Tian[3], Chenliang Xu[1]
[1]University of Rochester, [2]HKUST, [3]The University of Texas at Dallas

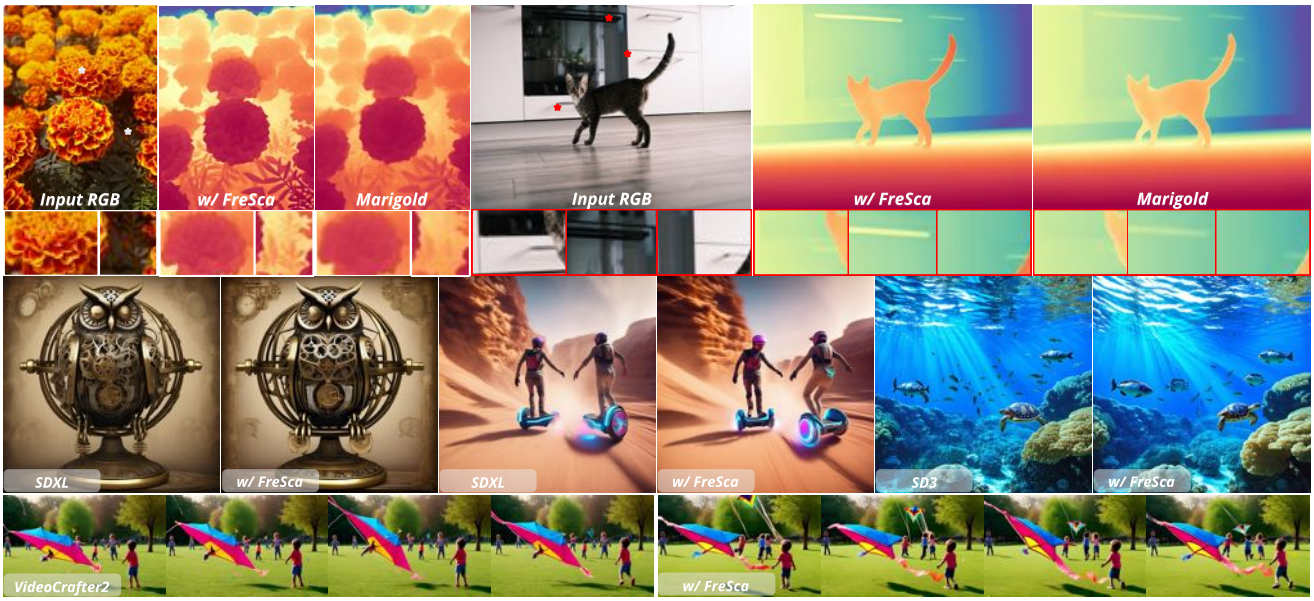

Figure 1. **`FreSca`: A plug-and-play enhancement for diffusion models.** Without retraining, `FreSca` refines Marigold [10] depth predictions to recover fine details (top); enables precise, prompt-aligned generation over SD3 [5] (middle) ; and boosts motion, detail, and temporal consistency in VideoCrafter2 [3] video generation (bottom). Project page: https://wikichao.github.io/FreSca/.

Abstract

Latent diffusion models (LDMs) have achieved remarkable success in a variety of image tasks, yet achieving fine-grained, disentangled control over global structures versus fine details remains challenging. This paper explores frequency-based control within latent diffusion models. We first systematically analyze frequency characteristics across pixel space, VAE latent space, and internal LDM representations. This reveals that the "noise difference" term, $\Delta\epsilon_t$, derived from classifier-free guidance at each step t, is a uniquely effective and semantically rich target for manipulation. Building on this insight, we introduce `FreSca`, a novel and plug-and-play framework that decomposes $\Delta\epsilon_t$ into low- and high-frequency components and applies independent scaling factors to them via spatial or energy-based cutoffs. Essentially, `FreSca` operates without any model retraining or architectural change, offering model- and task-agnostic control. We demonstrate its versatility and effectiveness in improving generation quality and structural emphasis on multiple architectures (e.g., SD3, SDXL) and across applications including image generation, editing, depth estimation, and video synthesis, thereby unlocking a new dimension of expressive control within LDMs.

1. Introduction

Latent diffusion models (LDMs) [12] generate high-quality images [5, 11] but offer limited nuanced control beyond initial conditioning. Users often desire direct, disentangled modulation of image characteristics, such as texture/shape balance or artistic style, which current methods often require complex changes or retraining.

The frequency domain, generally excelling at separating global structures (low-frequencies) from fine details (high-frequencies), offers a powerful paradigm for image manipulation. We hypothesize that integrating frequency-domain operations within LDM internals can unlock more intuitive

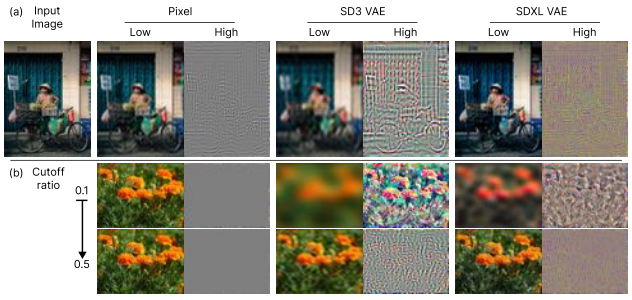

Figure 2. (a) Frequency decomposition of image and SD3 [5]/SDXL [11] VAE encodings with $r_0 = 0.05$ (pixel) and $r_0 = 0.5$ (latent). (b) Cutoff-radius sensitivity in pixel vs. latent.

control. This raises key questions: how do frequency characteristics translate to the VAE latent space, and which LDM component is best for such operation?

Our investigation compares frequency representations in pixel space, VAE latent space (as shown in Fig. 2), and internal LDM states. We identify the classifier-free guidance (CFG) derived "noise difference" term, $\Delta\epsilon_t$ [7], as a semantically rich and highly effective target for frequency manipulation. Building on this, we propose `FreSca`, a versatile, plug-and-play method. `FreSca` decomposes $\Delta\epsilon_t$ into low- and high-frequency components at each denoising step and applies distinct scaling factors, enabling independent control over global structures and fine details. It supports various frequency cutoffs and, by operating on the common noise space, is model- and task-agnostic, unlike prior methods [2, 14]. We demonstrate its efficacy on models like SDXL [11] and tasks including image generation, editing, depth estimation, and video synthesis [1, 3, 10]. Our core contributions are the identification of $\Delta\epsilon_t$ as an optimal control target and the versatile `FreSca` approach.

2. Method

This section details our approach. We first analyze frequency characteristics in pixel versus VAE latent spaces and within various diffusion model representations to identify an optimal target for manipulation. We then introduce `FreSca`, our method for fine-grained frequency control in latent diffusion models (LDMs).

Preliminaries. LDMs encode images I into latents $\mathbf{x} = \mathcal{E}(I)$ using a VAE encoder $\mathcal{E}$ and decode them with $\mathcal{D}$. The denoising network $\epsilon_\theta(\mathbf{x}_t, t)$ operates on noisy latents $\mathbf{x}_t$ at timestep t to predict the added noise ϵ_t, reversing a noise corruption process over T steps.

2.1. Frequency Analysis in VAE Latent and Diffusion Spaces

Frequency Decomposition. We define a frequency decomposition for a signal u (either an RGB image I or VAE latent $\mathbf{x}$). Given its 2D Fourier transform $U = \mathcal{F}(u)$ (Eq. (1)), and a cutoff ratio $r_0 \in [0, 1]$ determining a cutoff radius $R_c = r_0 \cdot \min(H/2, W/2)$ for spatial dimensions $H \times W$, we define binary low-pass M_l and high-pass M_h masks (Eq. (2)).

$$U = \mathcal{F}(u), \quad u = \mathcal{F}^{-1}(U). \tag{1}$$

$$M_l(k_x, k_y) = \begin{cases} 1, & \text{if } \sqrt{k_x^2 + k_y^2} \le R_c, \\ 0, & \text{otherwise,} \end{cases} \qquad M_h(k_x, k_y) = 1 - M_l(k_x, k_y). \tag{2}$$

The low- and high-frequency components are

$$u_l = \mathcal{F}^{-1}\big(M_l \odot U\big), \quad u_h = \mathcal{F}^{-1}\big(M_h \odot U\big). \tag{3}$$

Pixel vs. VAE Latent Frequencies. Applying this to images I and their VAE encodings $\mathbf{x}$ (see Fig. 2(a)) reveals that latent high-frequencies ($\mathbf{x}_h$) retain more abstract semantic patterns and exhibit different sensitivity to r_0 compared to pixel-space details (Fig. 2(b)). This suggests the VAE latent space is a richer domain for semantic frequency manipulation.

Targeting Diffusion Model Internals. For conditional generation, LDMs use Classifier-Free Guidance (CFG) [7], where the effective noise prediction is:

$$\epsilon_t = \epsilon_\theta(\mathbf{x}_t, t) + \omega \cdot \Delta\epsilon_t, \quad \text{with } \Delta\epsilon_t = \epsilon_\theta(\mathbf{x}_t, \boldsymbol{c}, t) - \epsilon_\theta(\mathbf{x}_t, t). \tag{4}$$

We investigate applying frequency decomposition (Eqs. (1) to (3)) to three candidate representations within the diffusion process: the noisy latents $\mathbf{x}_t$, the combined noise prediction ϵ_t, and the noise difference term $\Delta\epsilon_t$. Our experiments (visualized in Fig. 3(a)) reveal that manipulating $\Delta\epsilon_t$ yields the most semantically meaningful and controllable results. The time-averaged $\Delta\bar{\epsilon}$ also exhibits clearer semantic structures compared to averages of other candidates (Fig. 3(b)).

Step-wise Frequency Dynamics. Analysis of spectral profiles over denoising steps (Fig. 4) indicates that $\Delta\epsilon_t$ evolves from a more low-pass characteristic at early stages towards a broader spectrum later. Its magnitude also generally increases as t decreases, suggesting guidance becomes more

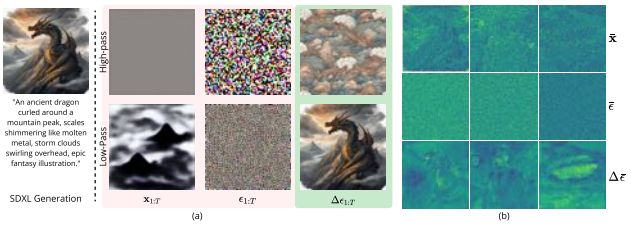

Figure 3. (a) Results of frequency decomposition on various diffusion representations; (b) Temporal average over T steps for each representation, highlighting the semantic richness of the noise-difference term. Please zoom in for better visibility.

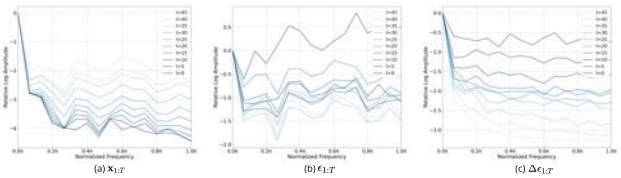

Figure 4. Relative log amplitudes of Fourier over all T denoising steps for (a) the latent variables $\mathbf{x}_t$, (b) the noise prediction ϵ_t, and (c) the noise-difference term $\Delta\epsilon_t$. Each curve corresponds to a timestep, illustrating how low and high frequencies changes in each representation.Please zoom in for better visibility.

influential in refining details during later steps. This dynamic nature motivates adaptive frequency manipulation.

2.2. FreSca: Versatile Frequency Scaling

Based on these findings, we propose FreSca, which modifies the noise difference term $\Delta\epsilon_t$ by independently scaling its low- and high-frequency components. Let $U_t = \mathcal{F}(\Delta\epsilon_t)$. The modified term $\hat{\Delta\epsilon_t}$ (illustrated in Fig. 5) is:

$$\hat{\Delta\epsilon_t} = \mathcal{F}^{-1}\big(l \cdot M_l \odot U_t + h \cdot M_h \odot U_t\big), \tag{5}$$

where l and h are scaling factors for low- and high-frequency bands, respectively. This $\hat{\Delta\epsilon_t}$ then replaces $\Delta\epsilon_t$ in Eq. (4). FreSca offers flexibility (e.g., detail enhancement with $h > 1, l = 1$; smoothing with $l > 1, h < 1$), faithfulness (original CFG if $l = h = 1$), and generality across models and tasks.

Dynamic Cutoff Determination. The cutoff radius R_c for M_l, M_h can be set dynamically at each timestep t:

1. **Spatial-Ratio Cutoff:** $R_c(t) = r_0 \cdot \min(H_t/2, W_t/2)$, where H_t, W_t are spatial dimensions of U_t.

$$R_c(t) = r_0 \cdot \min(H_t/2, W_t/2). \tag{6}$$

2. **Energy-Based Cutoff:** $R_c(t)$ is the smallest radius R such that cumulative spectral magnitude within R reaches a fraction r_0 of the total energy $E_{\text{tot}}(t) = \sum_{k_x,k_y} |U_t(k_x, k_y)|$.

$$R_c(t) = \min\Big\{ R \in \mathbb{N}_0 \mid \sum_{\sqrt{k_x^2+k_y^2} \le R} |U_t(k_x, k_y)| \ge r_0 E_{\text{tot}}(t) \Big\}. \tag{7}$$

3. Experiment

We evaluate FreSca 's effectiveness across various generative tasks, including text-to-image, monocular depth estimation, text-guided image editing, and text-to-video generation.

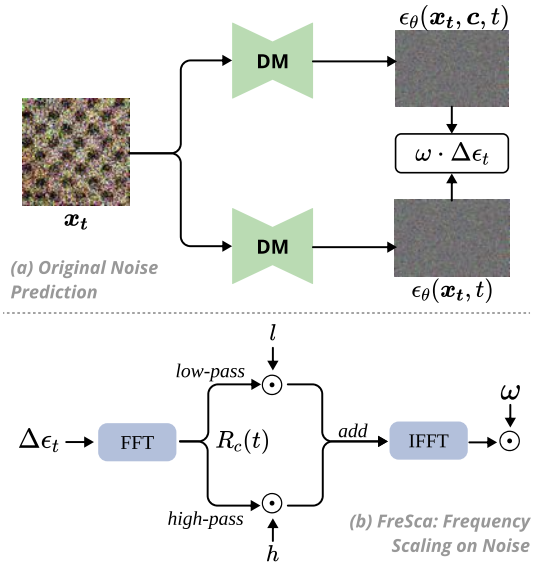

Figure 5. Overview of FreSca. We introduce scaling factors l and h to decompose the control mechanisms in the Fourier domain.

3.1. Monocular Depth Estimation

Monocular depth estimation is crucial for 3D scene understanding. We enhance Marigold [10], a diffusion-based depth estimation model, by integrating FreSca to boost its high-frequency noise components ($h > 1, l = 1$). As shown in Tab. 1, FreSca consistently **outperforms Marigold baselines** (with or without ensemble) on DIODE, KITTI, and ETH3D, recovering finer structures and sharper edges, as illustrated in Fig. 1.

3.2. Text-to-Video Generation

FreSca extends beyond static images to video generation. Integrating FreSca into VideoCrafter2 [3], an open-source video diffusion model, **improves video quality and fidelity at no additional cost**. As shown in Figs. 1 and 6, FreSca enhances motion coherence, preserves intricate details, and mitigates text-video misalignment. This highlights FreSca 's versatility across diverse diffusion models and tasks.

3.3. Text-to-Image Generation

FreSca demonstrates **model-agnostic versatility** by integrating with both SDXL (U-Net backbone) and SD3 (multi-

Table 1. Zero-Shot Depth Estimation (AbsRel↓, δ_1↑). FreSca boosts Marigold. ens stand for ensemble, **bold** indicating the best results.

Method	DIODE [15]		KITTI [6]		ETH3D [13]	
	AbsRel↓	$\delta 1$ ↑	AbsRel↓	$\delta 1$ ↑	AbsRel↓	$\delta 1$ ↑
Marigold (w/o ens)	31.0	77.2	10.5	90.4	7.1	95.1
Marigold (w/ ens)	30.8	77.3	9.9	91.6	6.5	**96.0**
Marigold w/ FreSca	**30.2**	**77.8**	**9.8**	**91.7**	**6.4**	95.9

Figure 6. FreSca enhances VideoCrafter2's [3] video generation quality at no additional cost.

modal diffusion transformer). As shown in Fig. 7, FreSca consistently enhances prompt fidelity and image quality, reducing distortions in both setups. An **energy-based cutoff strategy** (see Fig. 8) also yields generations that more closely match prompts compared to a spatial-ratio cutoff.

3.4. Text-guided Image Editing

We integrate FreSca into existing training-free image editing frameworks, LEdits++ [1] and Edited-Friendly DDPM Inversion [8], on the TEdBench [9] dataset. FreSca seamlessly plugs into these methods without architectural changes. Quantitatively (Tab. 2), FreSca consistently

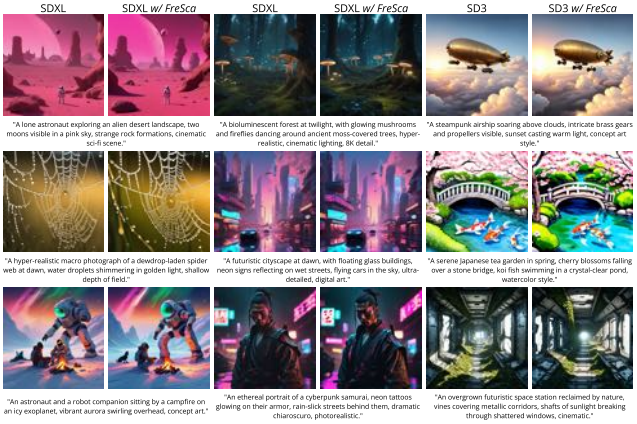

Figure 7. Samples generated by SDXL [11] and SD3 [5] with or without FreSca. Please zoom in for better readability.

Figure 8. Ablation of cutoff strategies: (a) original SDXL output; FreSca applied with (b) spatial-ratio cutoff and (c) energy-based cutoff (both $h = 1.5$).

Table 2. Image editing results evaluated by both generative metrics (FID-30k and CLIP-text) and human-centric VLM metrics (Success Rate and Quality).

	FID-30k ↓	CLIP-text (%) ↑	Success Rate (%) ↑	Quality
Edited-Friendly DDPM [8]	255.5	31.35	75.0	**4.23**
DDPM [8] w/ FreSca	**253.4**	**31.54**	**80.0**	4.18
LEdits++ [1]	255.3	31.08	72.5	4.08
LEdits++ [1] w/ FreSca	**255.0**	**31.34**	**72.5**	**4.18**

Figure 9. Editing results from LEdits++ [1] with or without FreSca.

boosts CLIP-text scores and reduces FID, demonstrating that selective high-frequency amplification strengthens the target edit while preserving image fidelity. Additionally, we perform evaluation using the large vision–language model InternVL2.5-8B [4]. Qualitatively, Fig. 9 further illustrates these enhancements.

4. Conclusion

This paper introduced FreSca, a novel, **model-agnostic framework** for fine-grained, disentangled control over latent diffusion models through frequency-domain manipulation. By applying scaled adjustments to the semantically rich classifier-free guidance noise difference $\Delta\epsilon_t$ with dynamic cutoffs, FreSca offers **practical creative control** across diverse models (SDXL, SD3, Marigold, VideoCrafter2) and tasks (image generation, editing, depth estimation, video synthesis). This plug-and-play approach not only enhances visual attributes but also deepens the understanding of frequency components in LDMs. Future work can explore advanced spectral techniques and learned control strategies.

References

[1] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 2, 4

[2] Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Broadway: Boost your text-to-video generation model in a training-free way. *arXiv preprint arXiv:2410.06241*, 2024. 2

[3] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 1, 2, 3, 4

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[5] 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 *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 4

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[8] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024. 4

[9] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 4

[10] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[11] 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. 1, 2, 4

[12] 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*, pages 10684–10695, 2022. 1

[13] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 3

[14] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024. 2

[15] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 3