Toward Diffusible High-Dimensional Latent Spaces: A Frequency Perspective

Supplementary Material

This is the supplementary material for the paper titled "Toward Diffusible High-Dimensional Latent Spaces: A Frequency Perspective". The outline is listed as below.

A – Explanation of Authenticity and Amplitude
B – Quantitative Evaluation of the Importance of Different Frequencies in Latent/RGB Spaces
C – Comparison with Low-resolution Pretraining
D – Limitation and Future Work

A. Explanation of Authenticity and Amplitude

The final goal of our method is to improve the *authenticity* of *generated* high-frequency components which the decoder relies heavily on according to the Finding 1. For this purpose, we further find the encoder underrepresents high-frequency latent embeddings (Finding 3), which leads to underfitting of LDMs on high-frequency bands *during training*. Thus LDMs generate *unauthentic* high-frequency embeddings in inference, resulting in bad reconstruction fidelity. Note that our goal is *not* synthesizing high-amplitude high-frequency signals. Instead, we want to make the generated high-frequency signals more authentic. Authenticity and amplitude are two distinct axes. We try to find a method to improve the amplitude of high-frequency components *simply* in training to mitigate the imbalance issue. In inference, we still make LDMs synthesize high-frequency embeddings with proper amplitude, but with high authenticity. That is also the reason why we still need to finetune the model on full-frequency bands after FreqWarm.

This also answers another question: *why don't we see a better reconstruction performance by applying our method only to the reconstruction process (without diffusion models)?* High amplitude does not lead to better reconstruction. Instead, high authenticity is the deciding factor. The latent embeddings encoded directly from input images are undoubtedly authentic (despite low-amplitude), so direct filtering in RGB space does not lead to improvement.

B. Quantitative Evaluation of the Importance of Different Frequencies in Latent/RGB Spaces

In Sec. 3 of the main paper, we conduct analysis on encoder and decoder responses to input with different frequencies. To further support our conclusion, we quantitatively evaluate the quality of images reconstructed from different frequencies. We use CLIP [4] score to measure the semantic alignment between the decoded images and original images. As shown in Tab. A, when the threshold changes from 0.03 to 0.20, the CLIP similarity of images reconstructed from the high-frequency embeddings are consistently higher than those reconstructed from low-frequency embeddings. This result quantitatively validates our first finding that the decoder relies on the high-frequency bands in the latent space more than the low-frequency bands.

Threshold r_0	0.03	0.05	0.20
Low Frequency ($r < r_0$)	29.6	30.5	31.2
High Frequency ($r > r_0$)	67.4	65.8	58.4

Table A. Quantitative evaluation of images reconstructed from high-frequency latent embeddings and low-frequency latent embeddings. r_0 is the distance to the center on frequency profile after FFT, which is a proxy of frequency. We use CLIP similarity between images as the metric.

Threshold r_0	0.03	0.05	0.20
Low Frequency ($r < r_0$)	87.5	91.1	98.9
High Frequency ($r > r_0$)	24.2	21.9	20.3

Table B. Quantitative evaluation of high-frequency and low-frequency components in RGB images. r_0 is the distance to the center on frequency profile after FFT, which is a proxy of frequency. We use CLIP similarity between images as the metric.

We also implement a similar quantitative evaluation on RGB images directly. As shown in Tab. B, the images recovered from low-frequency bands are way closer to the original images, which suggests that most information exists in the low-frequency signals in the RGB space (our second finding).

C. Comparison with Low-resolution Pretraining

In our method, we remove a portion of high-frequency components in RGB images for warm-up in frequency domain. A similar strategy is pretraining the diffusion model using low-resolution images, and then finetuning the model using high-resolution images. We implement this strategy by using 256×256 resolution for pretraining and using 512×512 resolution for finetuning. In contrast, our method always trains models with the 512×512 resolution and controls the frequency by explicitly setting up a frequency threshold.

As shown in Tab. C, the low-resolution pretraining strategy can improve the performance decently. However, it still lags behind our method by a large margin. Though low-resolution images have more low-resolution signals

Model	Strategy	gFID ↓	IS ↑
Wan2.2-AE-f16c48 [5]	None	43.67	33.48
Wan2.2-AE-f16c48 [5]	Low-res Pretrain	39.10	35.09
Wan2.2-AE-f16c48 [5]	FreqWarm	**29.56**	**46.16**
LTX-AE-f32c128 [2]	None	24.18	61.60
LTX-AE-f32c128 [2]	Low-res Pretrain	22.04	67.33
LTX-AE-f32c128 [2]	FreqWarm	**18.05**	**76.06**
DC-AE-f32c128 [1]	None	13.84	85.40
DC-AE-f32c128 [1]	Low-res Pretrain	12.14	88.82
DC-AE-f32c128 [1]	FreqWarm	**9.42**	**108.80**

Table C. Comparison with low-resolution pretraining. We use USiT-H [3] as the diffusion model for all experiments. The orange rows are the results of our method.

compared with their high-resolution counterparts, the high-frequency components that may prevent the encoding process are not fully removed. In contrast, our method explicitly sets up a hard threshold based on the frequency distribution, leading to a latent space with stronger high-frequency embeddings which is easier for diffusion models to learn.

D. Limitation and Future Work

Despite the notable gains from our method, we still have to point out some limitations. Our method narrows the reconstruction–generation trade-off in high-dimensional latents. However, it does not eliminate the conflict. More studies are needed to further figure out other reasons for this phenomenon. In addition, we do not evaluate our findings on other modalities (*e.g.*, videos) and model architectures (*e.g.*, autoregressive models). We will continue to expand our analysis to a wider range of task settings.

We also propose some promising future research directions based on our work.

- **Learned frequency curricula.** Replace fixed thresholds with adaptive schedules and integrate anti-alias downsampling in encoders to reduce band interference.
- **Autoencoder–denoiser co-design.** Jointly optimize VAEs and diffusion transformers under explicit frequency budgets to avoid the impact of extremely high-frequency signals in autoencoder training stage.
- **Broader modalities and scales.** Validate on modern text-to-video systems and longer sequences where spatiotemporal high-frequency components are rarer but crucial.

References

[1] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[2] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2

[3] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 2

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[5] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2