Snap Inc.
Carnegie Mellon University

Improving the Diffusability of Autoencoders

Ivan Skorokhodov Sharath Girish Benran Hu Willi Menapace Yanyu Li Rameen Abdal Sergey Tulyakov Aliaksandr Siarohin

ICML International Conference On Machine Learning 2025

Summary

- We analyze the spectral properties of modern autoencoders (AEs) and observe that their latents have inflated high frequency components
- We propose a simple regularization to align RGB and latent spectra via only 10–20K of fine-tuning iterations
- This greatly improves "**diffusability**": the generation quality of the downstream latent diffusion model (LDM) increases by 20%+

Motivation

CogVideoX-AE and Wan2.1-AE are *very* similar AEs (in terms of architecture and reconstruction quality), but they lead to very different LDM quality:

Autoencoder	PSNR	LPIPS	FID	KL/dim	DiT-XL/2 FDD	DiT-XL/2 FVD
CogVideoX AE	34.95	0.073	2.96	3.53	381.30	160.88
Wan2.1 AE	35.24	0.057	2.30	9.03	242.56	95.44

We name such property of the latent space "**diffusability**"?

Why latent spectral properties are important?

Diffusion Models (DMs) are spectral autoregressive models [1] and they generate low frequencies first and then high frequencies on top of them:

As AR models, DMs are prone to error accumulation [2], messing up high frequency components. It's not a problem for pixel-space DMs since the human eye is oblivious of high frequencies anyway:

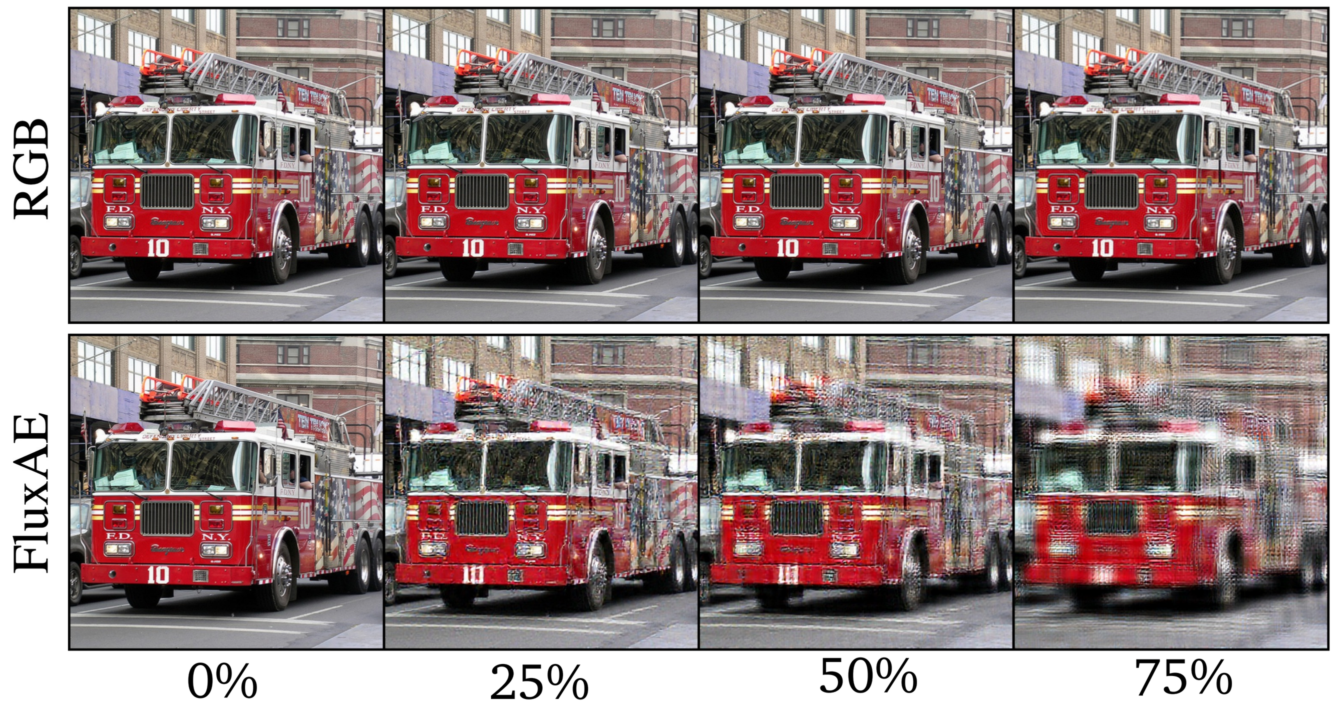
This image looks very realistic…

…even though its HF details are messed up

But for latent diffusion, autoencoders (AEs) can store important visual stuff in high frequency components of their latents!

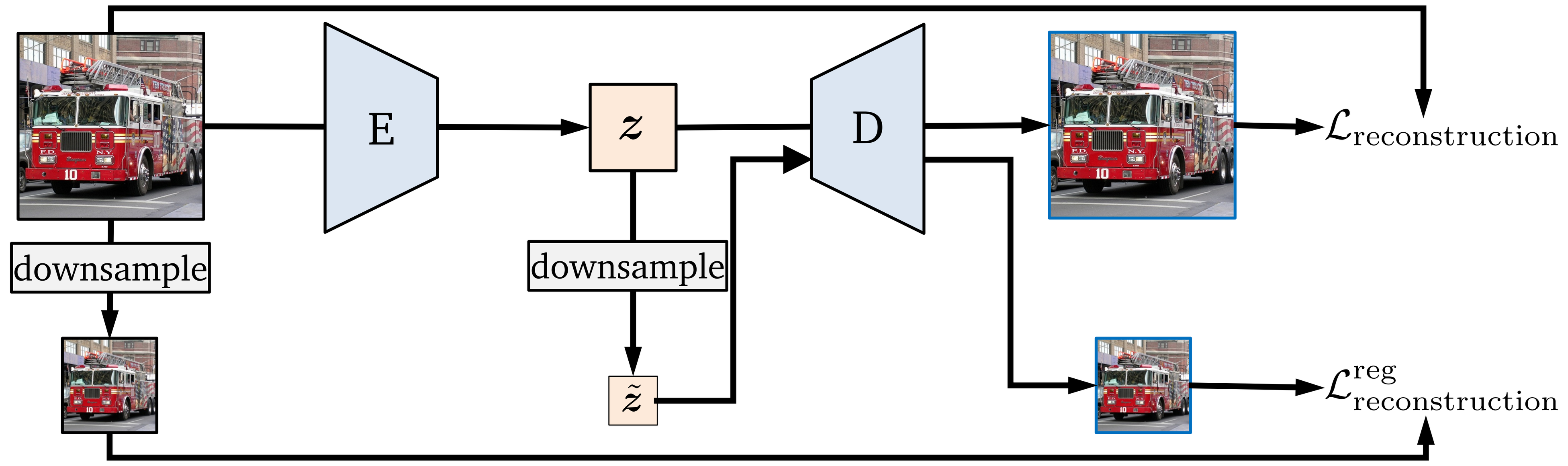
Scale Equivariance Regularization

We can see if an AE stores anything important in high frequencies by chopping them off and reconstructing the result:

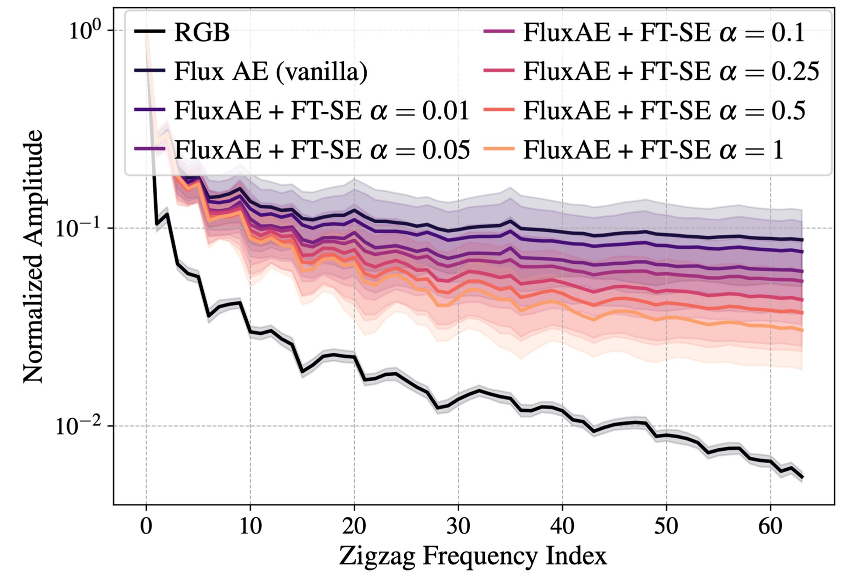


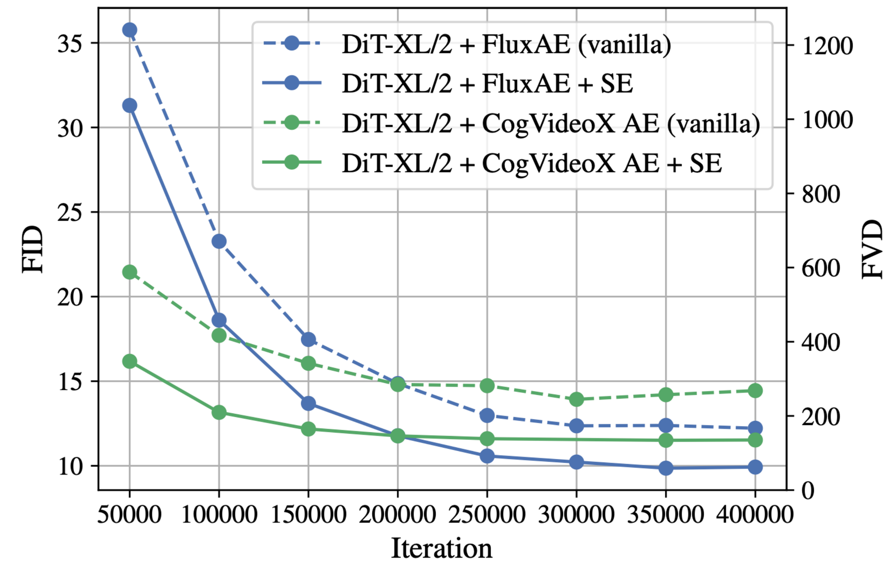
Chopping off high-frequency components for RGB and Flux AE latent representations

Cutting out high frequencies is mathematically equivalent (more or less) to downsampling! So, we can rectify the spectrum via a simple regularization:



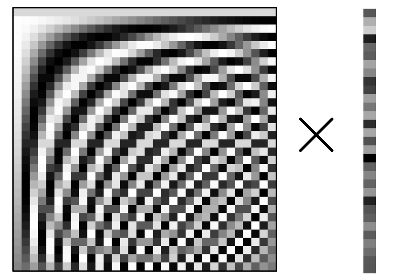
Such regularization rectifies the spectrum and improves LDM convergence:





Discrete Cosine Transform (DCT)

frequency amplitudes =



$$X_{\text{2D-DCT}} = D \cdot X \cdot D^\top$$

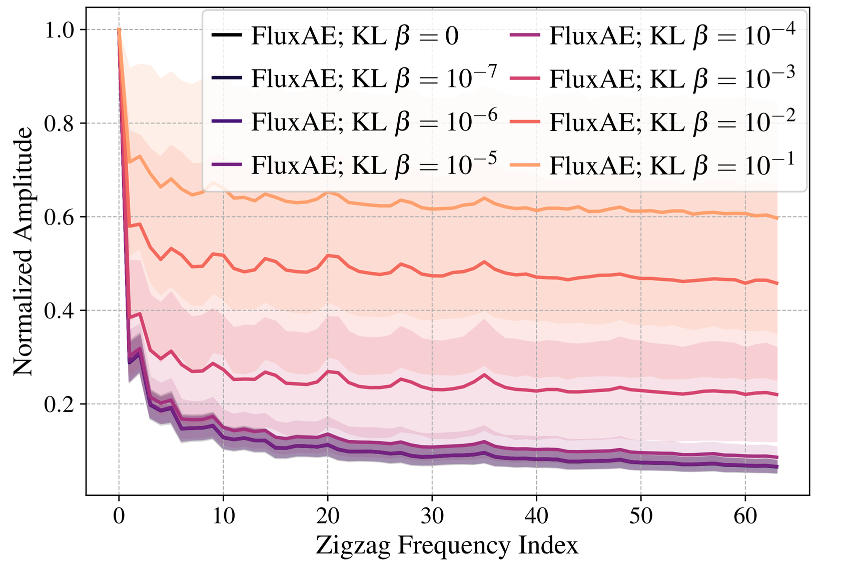
1D DCT is just a multiplication of a 1D signal vector with the DCT matrix

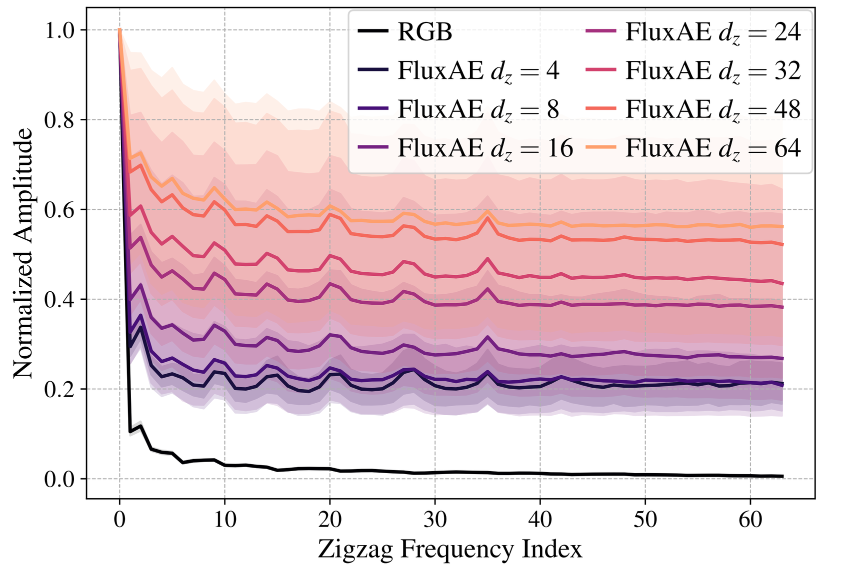
2D DCT is just 1D DCT applied twice over rows and then columns!

What affects the AE spectral properties?

Increasing KL regularization strength or the latent channel size inflates the amplitudes of the high frequency components of the latents:

Method	DiT-S/2 FDD	DiT-L/2 FDD	PSNR
FluxAE (vanilla)	992.05	415.87	30.20
+ KL β = 0	968.26	472.08	29.97
+ KL β = 10^{-7}	1018.6	425.35	30.29
+ KL β = 10^{-6}	1095.2	612.12	19.66
+ KL β = 10^{-5}	940.13	403.99	29.21
+ KL β = 10^{-4}	974.67	404.61	30.22
+ KL β = 10^{-3}	982.91	425.24	29.51
+ KL β = 10^{-2}	1946.5	1737.47	10.82
+ KL β = 10^{-1}	929.58	472.74	23.72
+ SE (ours)	924.28	369.15	30.37





Increasing the KL strength helps smaller models, but sacrifices the reconstruction quality, training stability and performs worse for larger DMs. Our SE regularization is universally helpful without such downsides.

References

[1] Ning et al., "DCTdiff: Intriguing Properties of Image Generative Modeling in the DCT Space", ICML 2025
[2] Li et al., "On Error Propagation of Diffusion Models", ICLR 2024
[3] Mitchell et al., "Neural Isometries: Taming Transformations for Equivariant ML", NeurIPS 2024

https://github.com/snap-research/diffusability