FREQUENCY-STEERED VAEs: RESTORING FINE DETAILS IN LOW-RESOLUTION LATENTS

Anonymous authors
Paper under double-blind review

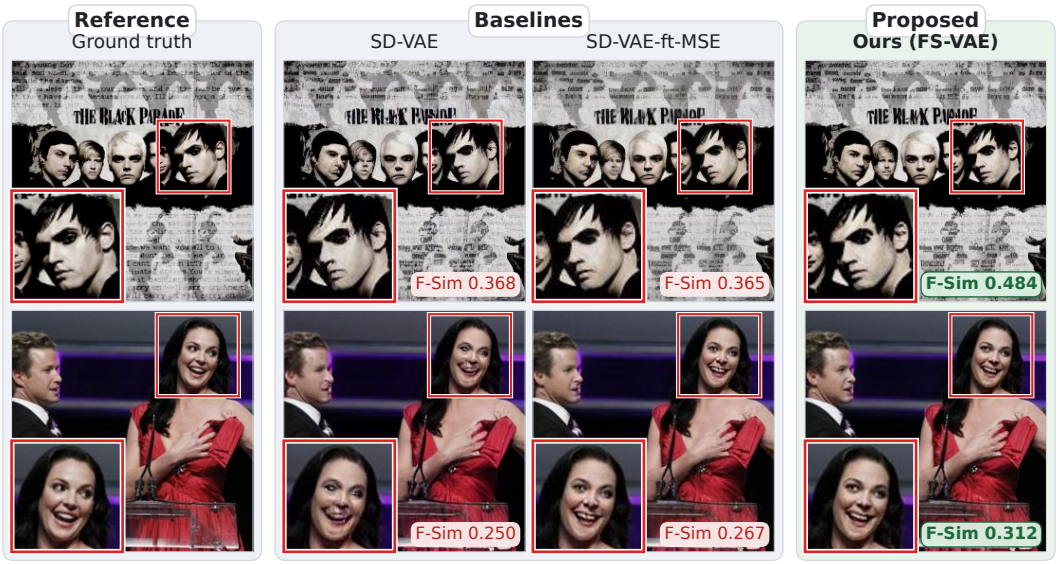

Figure 1: **FS−VAE** recovers the fine-grained detail lost at low resolution by steering frequency information during the decoding stage of the variational autoencoder (VAE), through a cross-resolution distillation setup. Results are shown for SD-VAE (Rombach et al., 2022), SD-VAE-ft-MSE (Stability AI, 2022), and FS−VAE at 256^2on TokBench (Wu et al., 2025). We report the face similarity (F-Sim) between the generated and ground truth (cropped) faces. Notice how as the face sizes decrease, the baselines struggle to reconstruct crucial details such as the shape of the eyes and teeth. FS−VAE is a light-weight plug and play upgrade where just the VAE's decoder is finetuned to recover high-frequency details from the latent at its native resolution, without expanding the channels, relying on external foundation models, or retraining the downstream diffusion model.

ABSTRACT

Latent diffusion models (LDM) rely on VAEs to compress visual inputs for computational efficiency, which is particularly consequential for very high resolution visual inputs. Through an in-depth interpretability analysis, we show that this compression disproportionately degrades high-frequency information at lower spatial resolutions, discarding fine details such as facial features or text patterns. Building on this insight, we propose a cross-resolution finetuning framework we call **FS−VAE** that uses a higher-resolution VAE as a teacher to supervise its lower-resolution counterpart, recovering lost frequency information via a band-wise *steerable-pyramid objective*. FS−VAE enables richer reconstruction from low-resolution latents without expanding channel dimensions, reference encoders, external models, or retraining downstream diffusion architectures. Furthermore, we analyze the impact of distinct loss objectives on reconstruction quality and demonstrate our approach's generalizability across diverse image and video VAEs. By preserving the original architecture and latent dimensionality, FS−VAE acts as an efficient, plug-and-play replacement for existing LDM pipelines.

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.

1 INTRODUCTION

A latent diffusion model can only generate what its variational autoencoder (VAE) can decode. Yet a VAE can produce an image that looks plausible while discarding structural and frequency information present in its input. Figure 1 makes the distinction visible: when an image is reconstructed at 256^2, facial details such as eyelashes and the edges of the mouth disappear, while the face remains recognizable at a glance. This is crucial for latent diffusion models (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2024b): their computational efficiency comes from compressing visual inputs into compact, low-dimensional latents, which the VAE decoder must then map back to pixels. If a detail is lost in that process, the perceptual quality of the reconstruction can conceal the limitation of the representation.

Although training generative models on billion-scale, high-resolution visual inputs has become the norm (Labs et al., 2025; Stability AI, 2024), application domains such as satellite imagery, security, and edge-deployed models still largely operate on low-resolution images. In these scenarios, current VAEs, pretrained on higher-resolution images, are therefore applied to low-resolution inputs for which preserving fine detail is particularly difficult. Standard reconstruction metrics, whether pixel-level (PSNR, SSIM (Wang et al., 2004)) or feature-based (LPIPS (Zhang et al., 2018)), are poor at capturing this loss, as they measure overall similarity rather than the fine details that identity depends on. TokBench (Wu et al., 2025) quantifies this degradation on $12,398$ face- and text-rich images: even for the strongest autoencoders (SD3.5 (Stability AI, 2024), FLUX.1 (Black Forest Labs, 2024b)), identity similarity between an image and its reconstruction falls from 0.96 at 1024^2 to 0.76 at 256^2, with the sharpest drop for the smallest faces.

This raises a natural question: *what exactly does the VAE discard in low-resolution visual inputs?* We investigate this by comparing VAE reconstructions across resolutions, and find that the information loss is not distributed uniformly across the frequency spectrum. The **VAE disproportionately discards high-frequency information**, such as edges, textures, and fine structural detail, while low-frequency content (coarse shape, color, layout) remain relatively well preserved. Consequently, this results in a reconstruction that retains the image's overall appearance but loses the details that make it sharp and distinctive, often leading to blurriness, texture collapse, and loss of fine detail (Figure 1).

Our investigation leads way to a further question: Can we recover the missing details without changing the latent or retraining the diffusion model? Therefore, we look for a solution that requires no additional latent channels, reference encoders, or external super-resolution model, such that it stays computationally efficient at test-time and fully compatible with currently adopted diffusion models. A model capable of enriching low-resolution latents to match the richness of their higher-resolution counterparts also offers a further benefit: achieving the same generative visual quality using more compact latents (e.g., with a downsampling factor f, a 512^2 image yields a $\frac{512}{f} \times \frac{512}{f}$ latent, while a 256^2 image yields a $\frac{256}{f} \times \frac{256}{f}$ latent, i.e., $4\times$ fewer latent tokens), which could be particularly beneficial for long-form video generative models.

To this end, we introduce Frequency-Steered VAE (**`FS-VAE`**), a cross-resolution finetuning framework. `FS-VAE` treats a VAE's high-resolution (512^2) reconstruction as a *teacher signal* for its lower-resolution (256^2) counterpart and uses a band-wise steerable-pyramid objective with a relative loss that prevents the stronger low-frequency coefficients from dominating training and specifically recovering the high-frequency content, as indicated by our investigation. We present a thorough analysis of the challenges involved in focusing solely on recovering high-frequency signal across both image and video VAEs, indicating that the identified high-frequency degradation is a general property of resolution-reduced VAE compression rather than an artifact of a particular architecture. `FS-VAE` is a general, lightweight, plug-and-play upgrade, fully compatible with existing image and video diffusion models (Fig. 1). In summary, our contributions are:

- To quantify information loss in low-resolution VAE reconstructions, we conduct a steerable pyramid band-swap test across frequency subbands. We show that reconstruction degradation is concentrated in a few mid-to-fine frequency bands and the loss is most severe in architectures with lower latent channel dimensions.
- We introduce `FS-VAE`, which recovers this detail **without** requiring channel expansion, reference encoders, external models, or retraining of any downstream diffusion model. changing the latent. Instead, we use a self-distillation approach that uses high-resolution

reconstructions as soft targets, and band-wise steerable pyramid loss designed to explicitly recover high-frequency information.
- As a flexible plug-and-play method, FS−VAE generalizes across both image and video VAEs. Fine-tuning only the decoder avoids latent drift and maintains compatibility with pretrained diffusion pipelines. Evaluations on TokBench (Wu et al., 2025), FS−VAE improves facial identity preservation and boosts face similarity by up to 10.2% on small faces while producing perceptually better image and video outputs.

2 Related Work

Latent Diffusion Models (LDMs) (Rombach et al., 2022) established the paradigm of compressing visual input into a lower-dimensional latent space using a variational autoencoder (VAE) and performing the diffusion process in this representation, substantially reducing computational cost compared to pixel-space diffusion. A broad line of work has sought to improve generative autoencoders through architectural advances, improved training objectives and recipes, large-scale and higher-quality training data, and increased representational capacity (Podell et al., 2024; Stability AI, 2024; Black Forest Labs, 2024a; 2026; Kouzelis et al., 2025).

Semantic and Representation Alignment: Several works improve latent structure by leveraging pretrained visual representations. VA-VAE (Yao et al., 2025) regularizes VAE latents toward DINO features, while RAE (Zheng et al., 2026) and AlignTok (Chen et al., 2026) build generative representations from pretrained vision foundation models. REPA-E (Leng et al., 2025) extends representation alignment to joint autoencoder and diffusion training, directly shaping the latent space for generation. Separately, EQ-VAE (Kouzelis et al., 2025) introduces equivariance regularization to encourage consistency under spatial transformations. While these approaches improve semantic structure and diffusability, they do not explicitly target the preservation of fine-grained visual details.

Increasing Latent Capacity: Increasing latent channel capacity improves fine-grained reconstruction but produces higher-dimensional representations that can be more challenging for diffusion models to learn (Yao et al., 2025; Cai et al., 2026; Lai et al., 2026). SR-VAE (Skorokhodov et al., 2025) attributes this degradation to excessive high-frequency components in expanded latent spaces and promotes scale-invariant decoding to suppress them. FreqWarm (Lai et al., 2026) provides a complementary perspective, showing that although excessive high-frequency components can interfere with other frequencies, high-frequency latent information remains important for fine-grained reconstruction. DA-VAE (Cai et al., 2026) addresses this reconstruction–diffusability trade-off by expanding latent capacity while aligning the new representation with the pretrained VAE space. Despite their effectiveness, these approaches involve changes to the architecture, the latent structure, thereby limiting their direct applicability to existing pretrained diffusion pipelines. Our goal with FS−VAE is to minimize the architectural changes required for downstream diffusion adaptation.

Frequency-Aware VAE Representations: Fourier- and wavelet-based objectives improve perceptual fidelity and fine-grained generation for VAEs (Jiang et al., 2021; Björk et al., 2022; Lin et al., 2023; Sadat et al., 2024; Kiruluta, 2025; Li et al., 2025). DeBaT (Medi et al., 2026) attributes the loss of fine-grained detail to a low-frequency bias in VAEs and decouples low and high frequencies into separate pathways. However, its frequency-specific architecture and diffusion training prevent use with pretrained models. Spectrum Matching (Ning et al., 2026) studies frequency structure for latent diffusability, introducing spectrum-matching objectives to regulate the latent spectrum and align the latent and image spectra. They improve diffusability; we target the fine-grained, identity-bearing detail lost when a VAE operates at lower resolution.

3 Investigation: What Low-Resolution VAEs Lose

Steerable pyramid decomposition: Encoding images at 256^2 rather than 512^2 leads to a notable loss of high-frequency detail during VAE reconstruction (Figure 1). Because raw RGB representations cannot isolate *how* and *where* these fine-grained structural details are lost, we analyze frequency degradation using the steerable pyramid transform (Freeman & Adelson, 1991; Simoncelli et al., 1992; Simoncelli & Freeman, 1995; Portilla & Simoncelli, 2000). The steerable pyramid is an invertible, shift-invariant multi-scale operator $\mathbf{P}$ that decomposes an image into spatial frequency subbands and orientations (Figure 2). For S spatial frequencies and O orientations, $\mathbf{P}$ yields co-

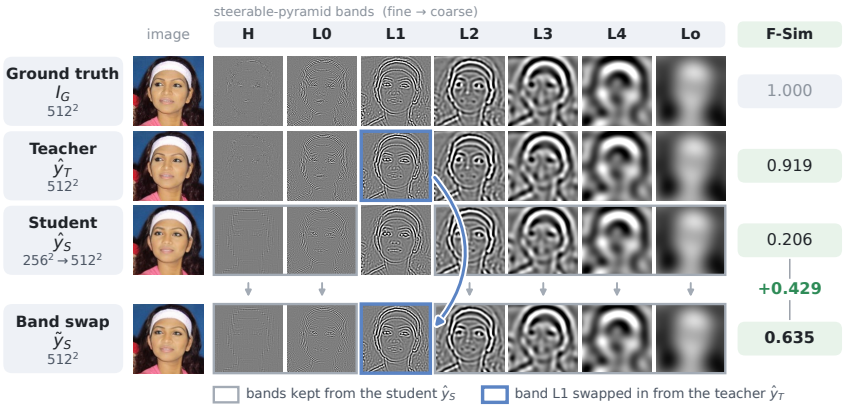

Figure 2: **Frequency Band Swap Test:** We decompose each image with a complex steerable pyramid into a highpass residual H, five oriented levels $L0$–$L4$, and a lowpass residual Lo. Then replace an individual frequency band in the student reconstruction ($\hat{y}_S$), from either the teacher reconstruction ($\hat{y}_T$), or ground truth I_G, leaving the other bands untouched (blue frames). In this example, just swapping the $L1$ band increases the absolute F-Sim by **0.43**. Repeating this for every band over 1000 images from the TokBench(Wu et al., 2025), we localize where in the spectrum low-resolution VAEs lag.

efficient maps $\{P_k\}_{k\in\mathcal{K}}$, where $|\mathcal{K}| = SO + 2$ accounts for the SO oriented subbands alongside residual high-pass (H) and low-pass (Lo) bands.

Setup: To conduct our diagnostic study, we apply steerable pyramid decompositions to 512^2 ground-truth images (I_G), their corresponding VAE reconstructions at 512^2 ($\hat{y}_T$, the teacher signal), and 256^2 VAE reconstructions upsampled to 512^2 ($\hat{y}_S$, the student signal). Each image is decomposed into a residual high-pass band H, $S = 5$ octave-spaced frequency levels ($L0$–$L4$, centered at 128, 64, 32, 16, and 8 cycles per image), and a residual low-pass band Lo (< 6 cpi). Each intermediate level is split across $O = 4$ orientations. Because the steerable pyramid transform is fully invertible, an image can be reconstructed even when a subset of its decomposed frequency subbands is replaced with subbands from another image. We leverage this and systematically swap subbands from the teacher $\hat{y}_T$ or ground-truth image I_G into the student reconstruction $\hat{y}_S$ (Figure 2). Following each swap, we quantify identity preservation using the TokBench metric (F-Sim) (Wu et al., 2025), defined as the cosine similarity between face-recognition embeddings of $\tilde{y}_S$ and I_G.

This intervention is conceptually analogous to activation patching in interpretability literature (Conmy et al., 2023; Heimersheim & Nanda, 2024), which has been widely applied to LLMs and recently extended to text-to-image and video models (Tang et al., 2023; Gandikota et al., 2024). To our knowledge, this is the first interpretability-driven causal analysis to localize frequency-band loss in low-resolution VAE reconstructions. We conduct this analysis on $1,000$ images cropped to 512^2 from the TokBench face subset, computing F-Sim across all faces in each image.

Findings: Where is face identity lost? Figure 3 plots the **absolute F-Sim gain** achieved by swapping individual student subbands in $\hat{y}_S$ with corresponding bands from $\hat{y}_T$ or I_G. We observe several key insights. First, swapping directly from ground truth I_G consistently yields higher F-Sim gains than swapping from $\hat{y}_T$, revealing that high-frequency identity information is degraded even in 512^2 teacher reconstructions. Second, for 4-channel VAEs like SD-VAE (`ft-MSE`) (Stability AI, 2022) and SDXL (Podell et al., 2024), performance gains concentrate heavily in the mid-to-fine subbands $L1$ and $L2$. For example, swapping $L1$ from $\hat{y}_T$ raises F-Sim by $+0.147$ ($+0.208$ from I_G), while $L2$ yields gains of $+0.118$ ($+0.132$ from I_G). In contrast, $L0$ and $L3$ contribute only $\sim +0.031$, while H, $L4$, and Lo provide negligible improvements—confirming that information loss at 256^2 resides primarily in mid-to-fine frequencies. Third, 16-channel VAEs such as SD3.5 (Stability AI, 2024) and FLUX.1 (Black Forest Labs, 2024b) exhibit a similar profile, however, with more finer bands ($L0$ and $L1$). Their gains are smaller due to their higher latent capacity.

Takeaways from Interpretability Analysis of VAEs

1. The loss of information between the 256^2 and 512^2 reconstructions can be precisely localized in the frequency domain.
2. Most information is lost in the mid- to high-frequency bands of low-resolution VAEs.
3. The loss of high frequency information is more pronounced in lower capacity VAEs.

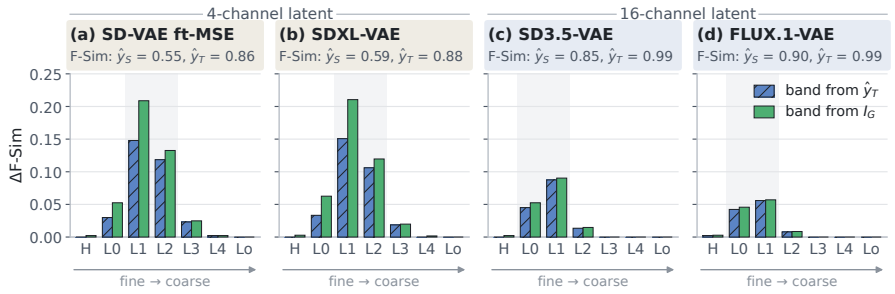

Figure 3: **Average F-Sim gains** on 1000 face images from TokBench (Wu et al., 2025) obtained from swapping bands between the original, higher-resolution image (I_G) or it's corresponding VAE reconstruction ($\hat{y}_T$) and the lower-resolution counterpart ($\hat{y}_S$). Notice how mid- to higher frequency bands such as $L1$ and $L2$ are the bands where signal is lost in a lower-resolution, 4-channel latent VAEs , while $L0$ and $L1$ is where most information is lost in low-resolution 16-channel VAEs.

4 METHOD

The band-swap test pinpoints exactly where information is lost in low-resolution latents and demonstrates how recovering this from high-resolution latents improves reconstruction quality. This analysis lays the foundation for our proposed frequency-domain distillation loss, detailed below.

4.1 STAGE 1: SELF-DISTILLATION IN THE FREQUENCY BANDS

Our goal is to transfer high-frequency details from a high-resolution teacher to a lower-resolution student via self-distillation. However, standard VAE fine-tuning relies heavily on spatial pixel losses such as MSE or ℓ_1, which treat all frequencies uniformly and fail to enforce targeted recovery of specific frequency bands (Stability AI, 2022; Esser et al., 2021). Thus, we introduce a frequency- and orientation-aware distillation loss computed directly on the steerable pyramid decompositions of the teacher and student reconstructions.

As noted in Sec. 3, $\hat{y}_T$ and $\hat{y}_S$ denote the VAE reconstructions of the teacher and student models respectively. We extract each subband k's magnitude map of $\hat{y}_T$, denoted as $a_k^T = \left|P_k(\hat{y}_T)\right|$ and of student's reconstruction $\hat{y}_S$, denoted as $a_k^S = \left|P_k(\hat{y}_S)\right|$, where $|\cdot|$ is a complex modulus operator. These magnitude maps isolate spatial energy across different frequencies and serve as core components for our distillation loss ($\mathcal{L}_{\text{dist}}$). For each sub-band k, we take the pixel wise error on the magnitude maps between the teacher and student, normalized by the teacher's energy in that band:

$$r_k = \frac{\left\langle\, |a_k^T - a_k^S| \,\right\rangle}{\left\langle\, a_k^T \,\right\rangle + \varepsilon}, \tag{1}$$

where $\langle\cdot\rangle$ denotes spatial mean. The loss is the weighted mean over bands:

$$\mathcal{L}_{\text{dist}} = \frac{\sum_k \omega_k\, r_k}{\sum_k \omega_k}, \tag{2}$$

r_k is formulated as a relative metric because coefficient values in low-frequency bands are significantly larger than those in high-frequency bands, which would otherwise cause low frequencies to dominate the gradient. The normalization in Eq. 1 ensures critical bands are not underrepresented in the optimization objective. The weight (ω_k) of each band can be explicitly set or tuned (more in Sec. 5.1). The overall training objective now becomes

$$\mathcal{L} \;=\; \lambda_{\text{rec}}\,\mathcal{L}_{\text{L1}} + \lambda_{\text{lp}}\,\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{dist}}\,\mathcal{L}_{\text{dist}}, \tag{3}$$

with the first two terms are computed against the ground truth input of the student. As illustrated in Figure 4, both the teacher and the student models share the same pretrained autoencoder, differing only in the resolution of the input image. Consequently, their respective latents maintain identical channel counts and downsampling factors, differing only in spatial resolutions. To account for the resolution mismatch between the teacher and student outputs, we upsample $\hat{y}_S$ via bicubic interpolation[1] to match the teacher's resolution.

[1] we explored other interpolation techniques like PixelShuffle (Shi et al., 2016) but it did not make a difference in the VAE reconstruction quality.

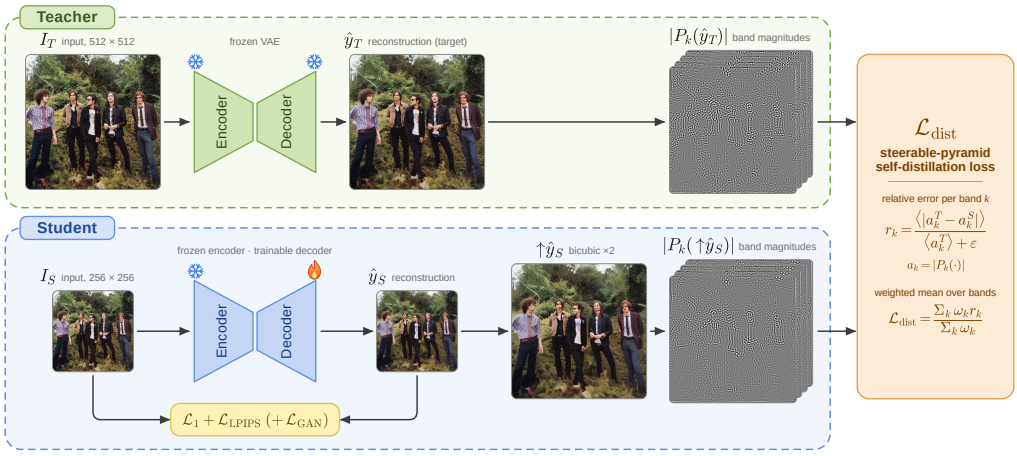

Figure 4: **Overview of `FS-VAE`:** The teacher (top) is a frozen VAE which takes a higher-resolution image and it's reconstruction is the target for our frequency loss. The student (bottom) encodes the same image at a lower resolution with the same frozen encoder, and only its decoder is trained. Steerable Pyramid operator is applied on both the student and teacher reconstructions and the difference between their magnitude maps is penalized.

Why fine-tune only the decoder? Following (Stability AI, 2022), freezing the encoder preserves the underlying latent representation, prevents latent space drift, and eliminates the need to retrain downstream diffusion models bound to the original latent space (Rombach et al., 2022). Restricting fine-tuning to the decoder also allows effective adaptation to domain-specific datasets without degrading cross-domain generalizability(Ni et al., 2023). See Appendix A.5 for generation results.

4.2 Stage 2: Adversarial Post-Tuning

While Stage 1 successfully recovers the missing high frequency details, we empirically observe that in lower capacity (i.e., 4-channel) VAEs, L_{dist} introduces a subtle periodic grain artifact across the reconstructed image (Fig. 5 (a)). To diagnose the origin of this pattern, we analyze the 2D Fast Fourier Transform (FFT) of the reconstruction residual ($I_S - \hat{y}_S$). Specifically, we compute the squared magnitude of the residual FFT, average it across 200 images, and visualize the result on a logarithmic scale with zero frequency at the center. Notice from Fig. 5 (b) how, compared to the baseline, the Stage 1 residual magnitude spectrum reveals prominent, localized bright spots at higher frequency regions. We hypothesize that this behavior arises particularly in low-capacity VAEs because forcing the decoder to match the teacher's high-resolution energy profile constrains the total power per frequency band, but does not impose any constraint on where within the band that energy should be allocated. This manifests as concentrated localized energy generating periodic spatial artifacts. To eliminate these artifacts, we briefly post-tune the decoder using a compound objective comprising the standard ℓ_1 reconstruction loss (Rombach et al., 2022), LPIPS perceptual loss (Zhang et al., 2018), and a PatchGAN adversarial loss (Isola et al., 2017). The PatchGAN term, particularly, enforces realistic local texture statistics by penalizing non-natural spectral distributions, effectively suppressing high-norm spectral spikes and restoring clean spatial reconstructions.

5 Experiments

5.1 Training Setup

Models: For image experiments, we train on the Stable Diffusion autoencoder (KL-f8) with the improved `ft-MSE` weights (Stability AI, 2022) which yields compression of $1\times8\times8$ and 4-channel latents. We specifically chose this model as our baseline as it is finetuned for improving facial reconstructions and serves as a direct comparison. We also report results on SDXL-VAE (Podell et al., 2024). For video experiments, we use Wan 2.1 VAE (Wan et al., 2025), a causal 3D autoencoder with $4\times8\times8$ compression and a 16-channel latent. Additionally, to highlight the plug and play na-

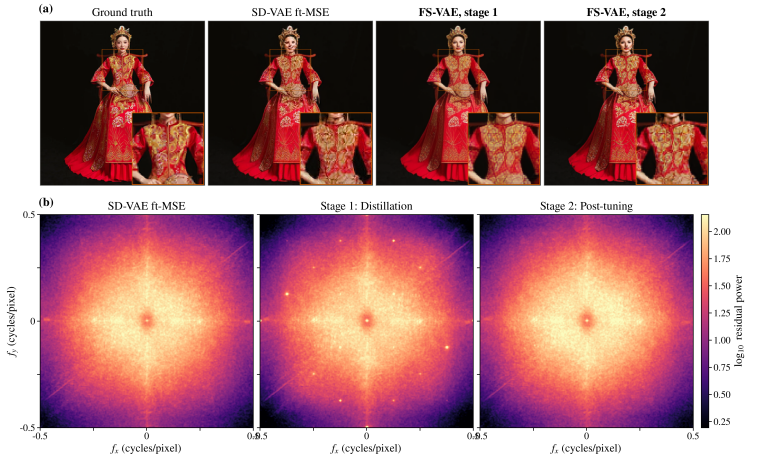

Figure 5: We illustrate **(a) high-frequency artifacts** that Stage 1 introduces in 4 channel VAEs, this motivates our adversarial post-tuning step, described in Sec. 4.2. The GAN loss removes the artifact and enhances the texture. **(b) Power spectrum of the residual:** We compute residual maps by subtracting the ground truth and reconstructed images, then applying Fast Fourier Transform on them. We calculate the power at each spatial frequency averaged over 200 images. Notice the bright spots after Stage 1 of training, which are eliminated after Stage-2 of training. f_x and f_y are the horizontal and vertical frequencies in cycles per pixel.

ture of our method, we also apply `FS-VAE` to RefDecoder (Fan et al., 2026), a reference-conditioned decoder for Wan 2.1 latent space, by adding our objective on top of its training loss.

Data: For the image experiments with SD-VAE ft-MSE, we train on a subset of $\sim 1M$ images from the 50M LAION-Face (Zheng et al., 2022) collection. Each image is then resized to the two resolutions used during training (512^2 and 256^2), for teacher and student, respectively. SDXL-VAE training details are in Appendix A.3. For video VAEs, we train on HumanVid (Wang et al., 2024) a 20K human-centric videos, keeping the human-centric domain. The teacher and student resolutions during the training: $832\times480\times5$ and $416\times240\times5$. More details in Appendix A.3. While Section 3 showed that mid-to-high frequency bands are most critical, empirical results demonstrate that maintaining non-zero loss weights (ω_k) across all subbands yields perceptually smoother reconstructions.

Metrics: We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), rFID (Heusel et al., 2017), face similarity score (F-Sim) against the ground-truth images (and videos) on TokBench Wu et al. (2025) (Table 1). During evaluation on TokBench images, we take the largest square edge, crop and resize to 256^2. In the video experiments, we also report rFVD (Unterthiner et al., 2019) to measure temporal consistency. We report on ImageNet-1k validation set for images and the HumanVid (Wang et al., 2024) validation set in Appendix A.3. We evaluate all videos at 416×240. We emphasize that these metrics evaluate global, high-level quality rather than the subtle, fine-grained details that `FS-VAE` is explicitly designed to restore.

5.2 Results

Table 2: **Face similarity (F-Sim) by face size**, with the relative gain over the teacher for Stage 1 shown as a green subscript.

	Face size (relative scale s)			
	$s < 0.1$	$0.1-0.2$	$0.2-0.3$	$s > 0.3$
instances	3,196	6,429	1,579	1,513
SD-VAE ft-MSE	0.2090	0.2368	0.4054	0.7204
FS-VAE, stage 1	**0.2271**$_{+8.7\%}$	**0.2610**$_{+10.2\%}$	**0.4258**$_{+5.0\%}$	**0.7318**$_{+1.6\%}$

From Table 1, note that by matching the per-band energy of the 512^2 teacher, Stage 1 (Sec. 4.1) restores high-frequency details missing from the 256^2 baseline. This leads to clear improvements in PSNR, SSIM, and F-Sim ($+4.12\%$ (from $0.45 \rightarrow 0.47$ on TokBench (Wu et al., 2025)). In Stage 2, the adversarial loss introduces a standard perception–distortion tradeoff (Esser et al., 2021), which leads to a slight dip in the pixel-level quality metrics, with an improvement of rFID. For video models, we apply only the Stage 1 objective (Sec. 4.1) to the decoder of the Wan 2.1 VAE. On TokBench (Wu et al., 2025) (Table 1) the finetuned decoder stays on par with the teacher, with PSNR ($30.55 \rightarrow 30.61$) and F-Sim slightly improved.

To show that the objective generalizes beyond a single decoder, we also apply it on top of RefDecoder (Fan et al., 2026), a reference-conditioned Wan 2.1 VAE decoder, by adding the steerable

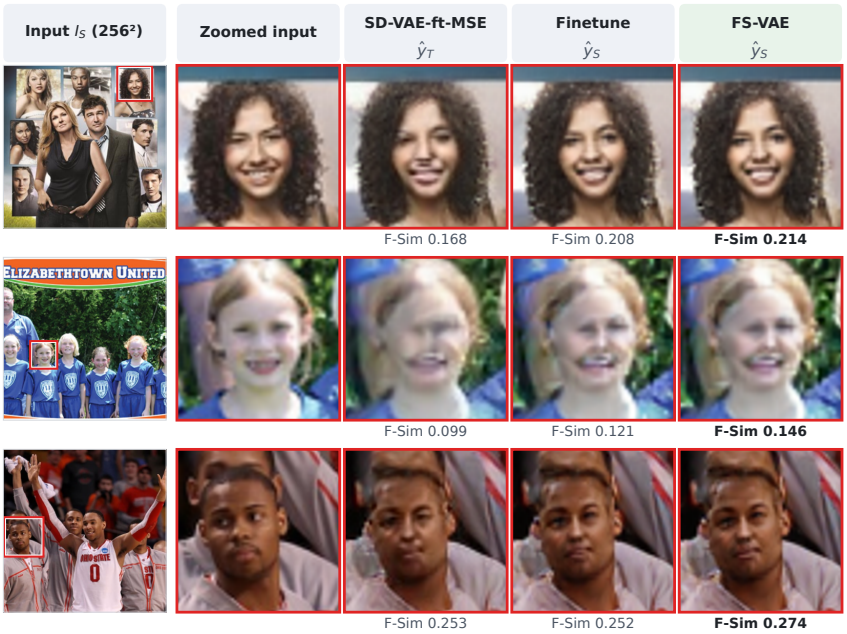

Figure 6: **Our results on LAION Face (Zheng et al., 2022) at 256^2:** These images were not used in the training. Baseline: SD-VAE-ft-MSE (Stability AI, 2022), Finetuned: Baseline finetuned with $el1$, LPIPS and GAN loss on training data, `FS-VAE`: two variants where we introduce all the steerable pyramid bands at once and at 5 every epoch interval starting from finest band to the coarsest band Table 6. Notice how `FS-VAE` shows clear improvement over the baseline and the finetuned versions even on small faces, especially in retaining the fine-grained details like shape of the eye, around the mouth, and nose.

pyramid term to its original training loss. We compare against RefDecoder fine-tuned on the same data with its original loss only, so the pyramid term is the only difference between the two models. With the pyramid term, PSNR, SSIM, LPIPS, rFID and rFVD all improve on TokBench (rFID $9.70 \rightarrow 9.11$, rFVD $9.65 \rightarrow 8.11$), and F-Sim increases by 2.73% over the released RefDecoder. This model gives the best PSNR, SSIM, LPIPS, rFID and F-Sim among all 16-channel video VAEs we evaluate on TokBench. The same trend holds on the validation split (Table 4): LPIPS, rFID and rFVD improve ($3.84 \rightarrow 3.40$), while PSNR stays close to the fine-tuned RefDecoder (34.69 vs. 34.77). Unlike in the image setting, video reconstructions do not exhibit high-frequency checkerboard artifacts, likely due to Wan 2.1's 16-channel latent space and higher input resolutions. Consequently, we omit Stage 2 adversarial post-tuning for video models and report Stage 1 results.

Performance Across Face Sizes: We evaluate `FS-VAE` across face scales $s = \max(w, h)/256$, defined following TokBench as the longer bounding box side relative to the cropped image dimension. Instances are categorized into TokBench bins: < 0.1, $[0.1, 0.2)$, $[0.2, 0.3)$, and $[0.3, 1.0]$. As shown in Table 2, substantial gains occur on small faces, i.e., for $s < 0.2$: $+8.7\%$ and $+10.2\%$. These improvements taper off as face scale increases, demonstrating that while the baseline effectively reconstructs prominent faces, `FS-VAE` successfully addresses its weakness on small-scale facial features.

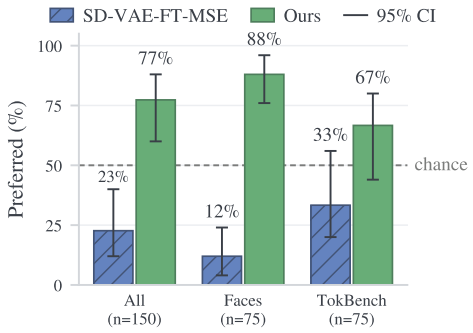

Figure 7: **User study:** `FS-VAE` vs. the baseline (Stability AI, 2022), 3 raters $\times$ 50 images.

User study: To fully understand the subtle, fine-grained details that `FS-VAE` is explicitly designed to restore, we conduct a user study. Three participants performed pairwise preference evaluations on 50 image pairs from LAION-Face (Zheng et al.,

Table 1: **Reconstruction quality on TokBench across image and video autoencoders.** Highlighted in green are FS−VAE while cream correspond to the baseline we train on. Italic: best per latent-channel group; **bold**: best within each decoder-swap block (baseline, finetune, ours).

Tokenizer	PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓	rFVD ↓	F-Sim ↑
Image VAEs						
32-channel latent						
DC-AE (Chen et al., 2025)	23.92	0.6884	0.1865	8.00	–	0.3020
VA-VAE (Yao et al., 2025)	27.08	0.8090	0.1293	6.81	–	0.5739
FLUX.2-VAE (klein-4B) (Black Forest Labs, 2026)	31.45	0.9066	0.0936	5.81	–	0.8569
16-channel latent						
SD-3.5-VAE (Stability AI, 2024)	30.39	0.8896	0.1017	6.26	–	0.7983
FLUX.1-VAE (dev) (Black Forest Labs, 2024a)	31.40	0.9079	0.0941	5.89	–	0.8470
4-channel latent						
SD-1.5-VAE (Rombach et al., 2022)	25.11	0.7433	0.1664	7.64	–	0.4288
EQ-VAE (Kouzelis et al., 2025)	25.22	0.7467	0.1638	7.56	–	0.4268
SDXL-VAE (Podell et al., 2024)	26.53	0.7904	0.1471	**7.10**	–	0.5011
+ **FS−VAE**, stage 1	**26.78**	**0.7981**	0.1467	8.37	–	**0.5116**
+ **FS−VAE**, stage 1+2	26.76	0.7974	**0.1456**	7.58	–	0.5106
SD-VAE-ft-MSE (Stability AI, 2022)	26.15	0.7703	0.1536	7.86	–	0.4542
+ finetune ($\lambda_{\text{rec}}\,\mathcal{L}_{\text{L1}} + \lambda_{\text{lp}}\,\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{gan}}\,\mathcal{L}_{\text{GAN}}$)	25.86	0.7672	0.1515	**7.38**	–	0.4595
+ **FS−VAE**, stage 1	**26.36**	**0.7804**	0.1525	9.04	–	**0.4729**
+ **FS−VAE**, stage 1+2	26.15	0.7760	**0.1503**	7.76	–	0.4667
Video VAEs						
32-channel latent						
WF-VAE-M (Li et al., 2025)	30.30	0.8950	0.1319	10.61	11.50	0.5201
16-channel latent						
Cosmos-Tokenizer (NVIDIA, 2024)	30.73	0.9055	0.1194	10.25	10.89	0.5349
CogVideoX-VAE (Yang et al., 2025)	30.63	0.9040	0.1266	11.32	11.99	0.5201
EasyAnimate-VAE v5.1 (Xu et al., 2026)	30.27	0.8944	0.1238	10.31	10.53	0.5094
CV-VAE (Zhao et al., 2024)	29.97	0.8924	0.1354	12.28	12.06	0.4807
IV-VAE (Wu et al., 2024)	30.87	0.9068	0.1158	9.88	9.20	0.5293
Reducio-VAE (T/2) (Tian et al., 2025)	29.59	0.8795	0.1490	11.80	34.07	0.5518
Wan 2.1 VAE (Wan et al., 2025)	30.55	**0.9056**	0.1169	**9.13**	**8.30**	0.5215
+ **FS−VAE**, stage 1	**30.61**	0.9050	**0.1168**	9.54	8.80	**0.5219**
RefDecoder (Fan et al., 2026)	30.53	0.9062	0.1203	10.37	10.68	0.5533
+ finetune ($\lambda_{\text{rec}}\,\mathcal{L}_{\text{L1}} + \lambda_{\text{lp}}\,\mathcal{L}_{\text{LPIPS}}$)	30.99	0.9108	0.1151	9.70	9.65	0.5644
+ **FS−VAE**, stage 1	**31.13**	**0.9130**	**0.1121**	**9.11**	**8.11**	**0.5684**

2022) and TokBench (Wu et al., 2025). Reconstructions from the baseline (Stability AI, 2022) and FS−VAE were displayed side-by-side in a randomized order and position, with human subjects selecting the image exhibiting more realistic fine detail. As shown in Fig. 7, FS−VAE was preferred in 77% of overall comparisons, receiving 88% preference on LAION-Face and 67% on TokBench.

6 Conclusion

In this work, we demonstrate that information loss in low-resolution VAEs is non-uniformly distributed across the frequency spectrum. In particular, 4-channel VAEs exhibit localized degradation primarily within mid-to-high frequency steerable pyramid subbands. Leveraging this insight, we design a frequency-targeted self-distillation loss that transfers high-frequency details from a high-resolution teacher model. By fine-tuning solely the decoder, FS−VAE preserves full compatibility with downstream diffusion pipelines. Empirically, FS−VAE significantly enhances facial identity preservation across varying scale ranges, particularly on small faces, where conventional VAEs systematically fail. Furthermore, FS−VAE achieves lower FVD scores across VideoVAEs, demonstrating enhanced temporal consistency and identity stability over time.

Limitations and Future Work: Our findings highlight a broader need for specialized evaluation tools: while metrics like F-Sim effectively measure facial identity preservation, they inevitably inherit the domain-specific biases of their underlying feature extractors. Developing metrics that assess localized, fine-grained frequency discrepancies directly will be crucial for diagnosing subtle, yet crucial reconstruction artifacts in future generative modeling research. We also plan to extend FS−VAE to additional domains where VAE reconstruction degradation particularly impact high-frequency information, such as fine text rendering and complex fabric textures, as well as higher-capacity 16- and 32-channel architectures.

AI USE STATEMENT

We used generative AI tools (Anthropic, 2026) to only polish the writing of this paper (rephrasing sentences, improving clarity and grammar, and assisting with LaTeX formatting of tables), to assist with code development, and to produce plotting scripts for figures based on our experimental results. We did not use generative AI tools for research ideation, methodology design, data collection, and no other task requiring disclosure applies to this work. All AI-assisted text was reviewed and edited by the authors, all AI-assisted code was verified and tested for correctness by the authors, all figures were checked against the underlying logged numbers, and every claim was checked against our own experimental results and the cited literature. We take full responsibility for the final content of this work, including all text, claims, code, figures, and artifacts.

ETHICAL STATEMENT

We train our image models on a subset of LAION-Face (Zheng et al., 2022), collected by scraping publicly available images from the web. The dataset automatically inherits the biases in LAION (Schuhmann et al., 2022). We use the data only to fine-tune a VAE decoder for reconstruction. Our video experiments use HumanVid, (Wang et al., 2024), which is built from Pexels (Pexels GmbH, 2026) videos. We will do a thorough analysis of bias and safety and retrain only on a filtered subset before releasing of the checkpoints. The people in the figures are images from the above mentioned datasets and TokBench (Wu et al., 2025), they are only added to show the reconstruction quality.

REPRODUCIBILITY STATEMENT

Training and evaluation details are given in Sec. 5 and Appendix A.3. The implementation is provided in the supplementary material.

REFERENCES

Anthropic. Claude. https://claude.ai, 2026. Large language model. Accessed: 2026-09-26.

Sara Björk, Jonas Nordhaug Myhre, and Thomas Haugland Johansen. Simpler is better: Spectral regularization and up-sampling techniques for variational autoencoders. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3778–3782, 2022. doi: 10.1109/ICASSP43922.2022.9746027.

Black Forest Labs. FLUX.1 [dev]. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024a. 12B rectified flow transformer, guidance-distilled; FLUX.1 [dev] Non-Commercial License. Code: https://github.com/black-forest-labs/flux.

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

Black Forest Labs. FLUX.2 [klein] 4B. https://huggingface.co/black-forest-labs/FLUX.2-klein-4B, 2026. Model weights, Apache 2.0 license. Blog: https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence.

Xin Cai, Zhiyuan You, Zhoutong Zhang, and Tianfan Xue. Da-vae: Plug-in latent compression for diffusion via detail alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18703–18713, June 2026.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=ajnBafpqmE.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models, 2025. URL https://arxiv.org/abs/2410.10733.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=89ia77nZ8u.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis . In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12868–12878, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01268. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01268.

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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024. arXiv:2403.03206.

Xiang Fan, Yuheng Wang, Bohan Fang, Zhongzheng Ren, and Ranjay Krishna. Refdecoder: Enhancing visual generation with conditional video decoding, 2026. URL https://arxiv.org/abs/2605.15196.

W.T. Freeman and E.H. Adelson. The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(9):891–906, 1991. doi: 10.1109/34.93808.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 5111–5120, January 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13919–13929, October 2021.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. 2026. doi: 10.48550/arXiv.2606.03748. URL https://arxiv.org/abs/2606.03748.

Andrew Kiruluta. Wavelet-based variational autoencoders for high-resolution image generation, 2025. URL https://arxiv.org/abs/2504.13214.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 31648–31666. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kouzelis25a.html.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Bolin Lai, XuDong Wang, Saketh Rambhatla, James M. Rehg, Zsolt Kira, Rohit Girdhar, and Ishan Misra. Toward diffusible high-dimensional latent spaces: A frequency perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 43450–43460, June 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 18262–18272, October 2025.

Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, and Li Yuan. Wf-vae: Enhancing video vae by wavelet-driven energy flow for latent video diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17778–17788, June 2025.

Xinmiao Lin, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Yu Kong. Catch missing details: Image reconstruction with frequency augmented variational autoencoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1736–1745, June 2023.

Tejaswini Medi, Hsien-Yi Wang, Arianna Rampini, and Margret Keuper. Decoupling high and low frequencies for faithful image generation with fine details. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), Computer Vision – ECCV 2026, pp. 119–138, Cham, 2026. Springer Nature Switzerland.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18444–18455, 2023.

Mang Ning, Mingxiao Li, Le Zhang, Lanmiao Liu, Matthew B. Blaschko, Albert Ali Salah, and Itir Onal Ertugrul. Spectrum matching: a unified perspective for superior diffusability in latent diffusion, 2026. URL https://arxiv.org/abs/2603.14645.

NVIDIA. Cosmos tokenizer: A suite of image and video neural tokenizers. https://github.com/NVIDIA/Cosmos-Tokenizer, 2024. Archived Feb. 2025; superseded by https://github.com/NVIDIA/Cosmos.

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

Pexels GmbH. Pexels: Free stock photos, royalty free images & videos. https://www.pexels.com, 2026. Accessed: 2026-09-04.

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 The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision, 40(1):49–70, 2000. doi: 10.1023/A:1026553619983.

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 (CVPR), pp. 10684–10695, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Litevae: Lightweight and efficient variational autoencoders for latent diffusion models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 3907–3936. Curran Associates, Inc., 2024. doi: 10.52202/079017-0128. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0735ab358bf9bf0f30af6c871b666ec8-Paper-Conference.pdf.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In Advances in Neural Information Processing Systems, volume 35, pp. 25278–25294, December 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1874–1883, 2016.

E.P. Simoncelli and W.T. Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In Proceedings., International Conference on Image Processing, volume 3, pp. 444–447 vol.3, 1995. doi: 10.1109/ICIP.1995.537667.

E.P. Simoncelli, W.T. Freeman, E.H. Adelson, and D.J. Heeger. Shiftable multiscale transforms. IEEE Transactions on Information Theory, 38(2):587–607, 1992. doi: 10.1109/18.119725.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 55876–55905. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/skorokhodov25a.html.

Stability AI. sd-vae-ft-mse: Fine-tuned vae decoder for stable diffusion. https://huggingface.co/stabilityai/sd-vae-ft-mse, 2022.

Stability AI. Introducing stable diffusion 3.5. https://stability.ai/news-updates/introducing-stable-diffusion-3-5, October 2024. Accessed: 2026-09-18.

Anni Tang, Tianyu He, Junliang Guo, Xinle Cheng, Li Song, and Jiang Bian. Vidtok: A versatile and open-source video tokenizer, 2024. URL https://arxiv.org/abs/2412.13061.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5644–5659, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.310. URL https://aclanthology.org/2023.acl-long.310/.

Rui Tian, Qi Dai, Jianmin Bao, Kai Qiu, Yifan Yang, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Reducio! generating 1k video within 16 seconds using extremely compressed motion latents, 2025. URL https://arxiv.org/abs/2411.13552.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric challenges, 2019. URL https://arxiv.org/abs/1812.01717.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. Humanvid: Demystifying training data for camera-controllable human image animation, 2024. URL https://arxiv.org/abs/2407.17438.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Junfeng Wu, Dongliang Luo, Weizhi Zhao, Zhihao Xie, Yuanhao Wang, Junyi Li, Xudong Xie, Yuliang Liu, and Xiang Bai. Tokbench: Evaluating your visual tokenizer before visual generation, 2025. URL https://arxiv.org/abs/2505.18142.

Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Improved video vae for latent video diffusion model, 2024. URL https://arxiv.org/abs/2411.06449.

Jiaqi Xu, Kunzhe Huang, Xinyi Zou, Yunkuo Chen, Bo Liu, MengLi Cheng, Jun Huang, and Xing Shi. Easyanimate: High-performance video generation framework with hybrid windows attention and reward backpropagation, 2026. URL https://arxiv.org/abs/2405.18991.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15703–15712, June 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models, 2024. URL https://arxiv.org/abs/2405.20279.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18697–18709, June 2022.

A Appendix

A.1 Additional Related Work

Cross-resolution regularization: Prior work treats resolution in autoencoders in one of two ways. The first treats it as a consistency constraint: scale-consistency methods (Skorokhodov et al., 2025) regularize the latent so that representations agree across resolutions. SR-VAE (Skorokhodov et al., 2025), for instance, requires that decoding a downscaled latent reproduce the correspondingly downscaled image. The second treats it as a capacity limitation, assuming that detail lost at lower resolution cannot be represented by a smaller latent and must be addressed by enlarging it; DA-VAE (Cai et al., 2026) follows this route, increasing latent capacity to capture details that emerge at higher resolutions. Unlike these methods, we neither regularize nor enlarge the latent, but distill detail from higher-resolution reconstructions into the decoder alone.

Video VAEs: Modern latent video diffusion models rely on causal 3D autoencoders that compress a video both spatially and temporally, such as the Cosmos tokenizer (NVIDIA, 2024), VidTok (Tang et al., 2024), CogVideoX (Yang et al., 2025), Wan 2.1 (Wan et al., 2025) and EasyAnimate (Xu et al., 2026). In addition, CV-VAE (Zhao et al., 2024) regularizes the video latent towards the latent of a pretrained image VAE so that image models can be reused for video, IV-VAE (Wu et al., 2024) splits temporal compression into a keyframe branch and group causal convolutions, WF-VAE (Li et al., 2025) routes low-frequency energy through a wavelet branch, and Reducio (Tian et al., 2025) compresses motion into an small latent conditioned on a content frame.

All of these retrain the full autoencoder and therefore define a new latent space. In contrast, RefDecoder (Fan et al., 2026) leaves the encoder and latent untouched and replaces only the decoder, injecting a high-fidelity reference frame into the decoding process through reference attention; designed for decoding in image-to-video generation.

A.2 Design choices

Why not downsample the teacher's reconstruction? We upsample the student's reconstruction to resolve the resolution mismatch between teacher and student. Downsampling the teacher's reconstruction instead would irreversibly discard the high-frequency detail we aim to recover from the low-resolution latents in the first place.

A.3 Implementation Details

Implementation: The hyperparameters for Stage 1 finetuning are in Table 3. λ_{sp} is the weight of the steerable-pyramid term relative to ℓ_1 and LPIPS, which both have unit weight. The band weights are from $L0$ to $L4$: 0.12, 0.18, 0.27, 0.60, 1.50, along with 0.1 and 0.05 for high and low pass respectively. Stage 2 (Sec. 4.2) is only done for low channel VAEs where we observe the visible periodic grain artifacts. Starting from the Stage 1 weights, we post-tune for 5 epochs with the standard LDM autoencoder objective (Rombach et al., 2022): ℓ_1, LPIPS (Zhang et al., 2018), and a PatchGAN adversarial term (Isola et al., 2017) with an adaptive weight of 0.25 on a subset of 100k samples from the same dataset as Stage 1. All other hyperparameters are unchanged. We initialize the discriminator from the pretrained PatchGAN used in SD-VAE (Rombach et al., 2022). For SDXL-VAE, we train only on a randomly sampled subset of 200k images from our training data, with 15 epochs for Stage 1 and 2 epochs for Stage 2, due to compute constraints. This was done to show the general applicability of `FS-VAE` to other VAEs.

Data Preprocessing: For the image experiments, we train on a subset of approximately 1M images from the 50M LAION-Face (Zheng et al., 2022) collection. To ensure sufficient high-frequency detail at both training resolutions, we retain only images whose shorter side exceeds 512 pixels. Each image is then resized to the two resolutions used during training (512^2 and 256^2).

For video VAEs, we train on HumanVid (Wang et al., 2024) a 20K human-centric videos, keeping the the main video experiments in the same human-centric domain as the image ones. The original HumanVid data is a mix of landscape and portrait videos. Prior to training, we center every video on the person using a pose detector (Jocher et al., 2026), crop every frame to the size 832×480, yielding $18,455$ videos, used at two resolutions during the training: $832\times480\times5$ and $416\times240\times5$.

Table 3: **Stage 1 training configuration** (Sec. 4.1).

Setting	Image	Video
Setup		
Student resolution	256^2	416×240
Teacher resolution	512^2	832×480
Video length	–	5 frames
Steerable pyramid		
Scales $\times$ orientations	5×4	4×4
Subbands (incl. residuals)	22	18
Weight, highpass (H)	0.10	0.10
Weight, scales ($L0 \rightarrow L4$)	0.12–1.50	0.12–0.60
Weight, lowpass (Lo)	0.05	0.05
λ_{sp}	1.0	0.5
Optimization		
Epochs	25	80
Learning rate	10^{-5}	10^{-4}
Batch size	192	384
EMA decay	0.9999	0.9999

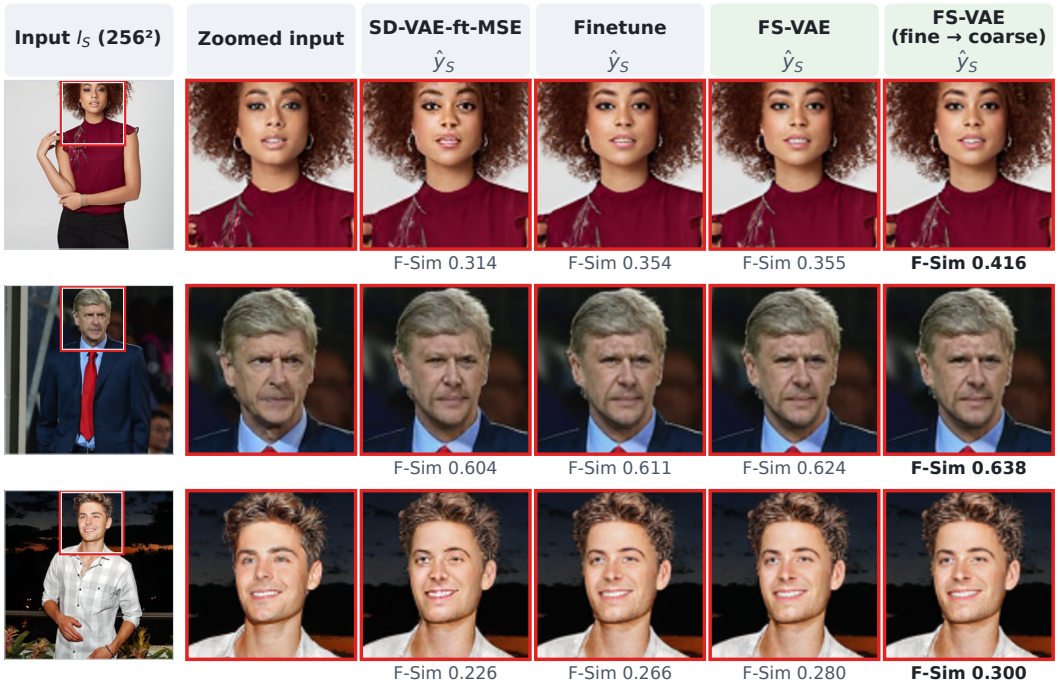

Figure 8: Additional qualitative results comparing FS−VAE with the baseline (Stability AI, 2022), showing the recovery of fine facial detail at 256^2 and the corresponding increase in F-Sim. We also show results for a different band introduction schedule that introduces bands progressively from fine to coarse every 5 epochs (Sec. A.6.2).

A.4 Additional Results

To test the generalizability beyond faces, we also evaluate FS−VAE on the Imagenet-1K split (Deng et al., 2009). Stage 1 (Sec. 4.1) improves PSNR, SSIM, and LPIPS over the baseline (Stability AI, 2022), nearly matching SDXL-VAE (Podell et al., 2024), but it also increases rFID. Stage 2 (Sec. 4.2) brings rFID below the baseline with a small drop in PSNR and SSIM. Longer Stage 2

Table 4: **Reconstruction quality on the ImageNet-1k validation set (images) and the HumanVid (Wang et al., 2024) validation set (videos).** Highlighted in green are FS−VAE while cream correspond to the baseline we train on. Italic: best per latent-channel group; **bold**: best within each decoder-swap block (baseline, finetune, ours).

Tokenizer	PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓	rFVD ↓
Image VAEs					
32-channel latent					
DC-AE (Chen et al., 2025)	25.000	0.677	0.157	0.65	–
VA-VAE (Yao et al., 2025)	27.705	0.787	0.097	0.27	–
FLUX.2-VAE (klein-4B) (Black Forest Labs, 2026)	33.084	0.918	0.035	0.10	–
16-channel latent					
SD-3.5-VAE (Stability AI, 2024)	31.287	0.880	0.060	0.19	–
FLUX.1-VAE (dev) (Black Forest Labs, 2024a)	32.866	0.911	0.044	0.175	–
4-channel latent					
SD-1.5-VAE (Rombach et al., 2022)	25.830	0.721	0.145	0.86	–
SD-XL-VAE (Podell et al., 2024)	27.376	0.763	0.121	0.67	–
EQ-VAE (Kouzelis et al., 2025)	25.967	0.723	0.141	0.82	–
SD-VAE ft-MSE (Stability AI, 2022)	26.913	0.741	0.130	**0.70**	–
+ **FS−VAE**, stage 1	**27.352**	**0.755**	**0.121**	1.91	–
+ **FS−VAE**, stage 1+2	27.041	0.748	0.123	1.18	–
Video VAEs					
16-channel latent					
WF-VAE-L (Li et al., 2025)	34.06	0.9383	0.1148	16.11	3.16
Cosmos-Tokenizer (NVIDIA, 2024)	34.36	0.9417	0.1084	15.96	4.71
CogVideoX-VAE (Yang et al., 2025)	34.25	0.9420	0.1131	18.28	5.01
EasyAnimate-VAE v5.1 (Xu et al., 2026)	33.60	0.9344	0.1101	16.22	4.66
CV-VAE (Zhao et al., 2024)	33.67	0.9344	0.1175	19.32	5.94
IV-VAE (Wu et al., 2024)	34.45	0.9409	0.1057	15.67	4.01
Reducio-VAE (T/2) (Tian et al., 2025)	33.44	0.9397	0.1119	16.17	8.15
Wan 2.1 VAE (Wan et al., 2025)	34.11	0.9403	0.1059	**14.83**	3.57
+ **FS−VAE**, stage 1	**34.28**	**0.9409**	**0.1042**	15.75	**3.48**
RefDecoder (Fan et al., 2026)	34.33	0.9434	0.1076	16.49	5.59
+ finetune ($\mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{LPIPS}}$)	**34.77**	0.9464	0.1033	15.19	3.84
+ **FS−VAE**, stage 1	34.69	**0.9466**	**0.1011**	**14.81**	**3.40**

training further improves rFID. On HumanVid (Wang et al., 2024) validation split (Table 4, notice that Wan 2.1 (Wan et al., 2025) with FS−VAE improves PSNR (34.11 → 34.28), SSIM, LPIPS and rFVD (3.57 → 3.48) over the teacher. We also see similar improvement with using FS−VAE with RefDecoder (Fan et al., 2026). This shows our method is complementary to the architectural changes. The improvement in rFVD suggests it improves temporal consistency. We show additional qualitative results in Figures 8, 11, and 9.

A.5 Compatibility with Diffusion Models

Our key advantage of FS−VAE is its compatibility with many diffusion models built on the SD-VAE (Rombach et al., 2022) latent space. This is primarily because our encoder is frozen and only the decoder is finetuned with the frequency steering loss, thereby a drop-in replacement in the existing pipelines. We show the results on DiT-XL/2 (Peebles & Xie, 2023), sampling with 250 DDPM steps and learned variance, a classifier-free guidance scale of 1.5, and a fixed seed. As shown in Figure 10, decoding the same latents with FS−VAE yields noticeably better face generation than the baseline (Stability AI, 2022).

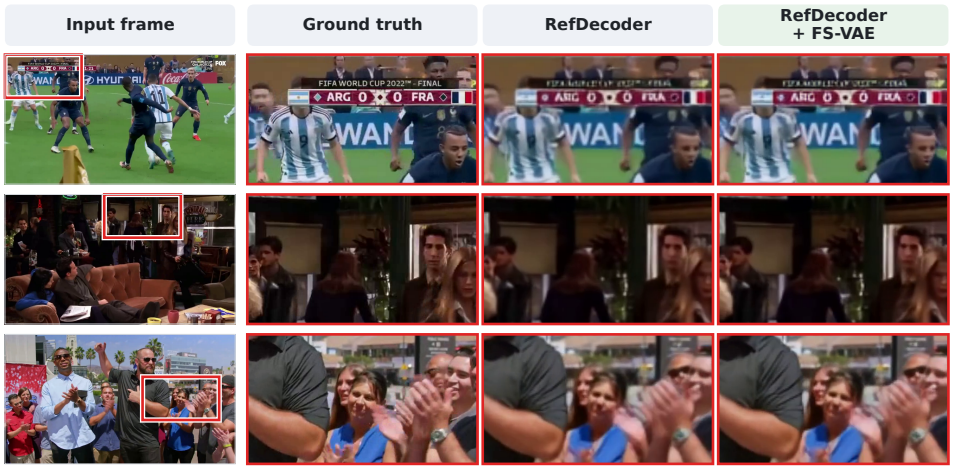

Figure 9: **Video Results on TokBench (Wu et al., 2025)**, with FS−VAE applied to RefDecoder (Fan et al., 2026).



Figure 10: **Compatibility with diffusion models:** Generations from DiT-XL/2 (Peebles & Xie, 2023) decoded with the baseline decoder (Stability AI, 2022) and with FS−VAE. Since the encoder is frozen, both decode the same sampled latents, so any improvement comes from the decoder alone. FS−VAE recovers noticeably more facial detail, particularly around the eyes, mouth, and nose.

A.6 Additional Ablations

A.6.1 Distillation Target

We ablate two targets for our Stage 1 (Sec. 4.1): the teacher reconstructions at $512^2$and the ground truth images at 512^2. Using the teacher gives better image quality metrics and a higher TokBench F-Sim. While ground truth has slightly better rFID in Stage 1, their values are pretty close in Stage 2. We attribute this to that the teacher is produced by the same VAE at a higher resolution, it serves a soft target and only has details that the model can represent at an increased latent size. We show the results in Table 5.

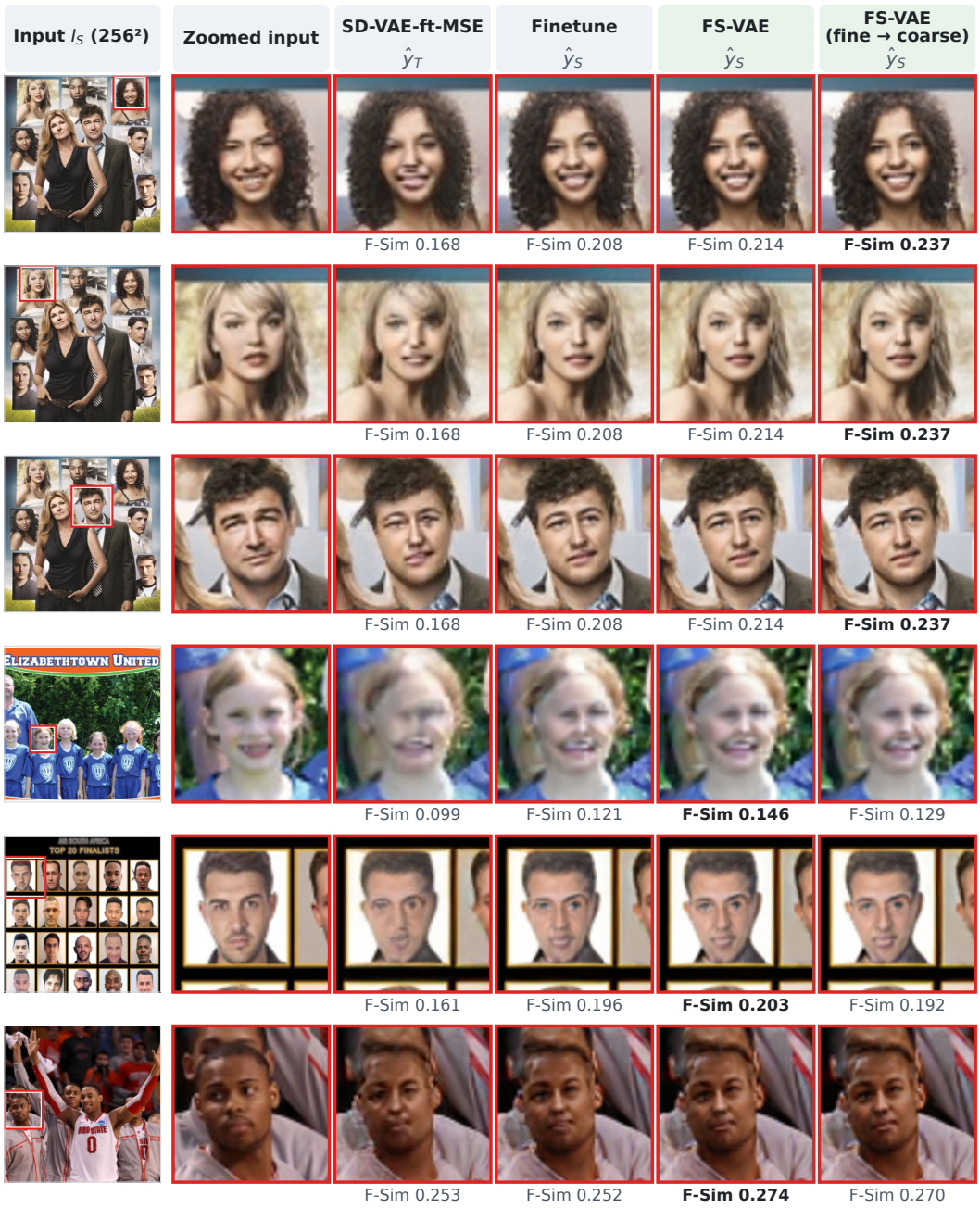

Figure 11: Additional qualitative results comparing FS−VAE with the baseline (Stability AI, 2022), showing the fine facial detail recovered at 256^2.

A.6.2 Band schedule

In this analysis, we explore when to activate the steerable bands in two settings. For the first one, we introduce all the bands from the start of the training. In second setting, we introduce every level of bands at a 5 epoch interval, starting with $L0$ at 0, $L1$ at 5 and so on. We combine residual high pass and low pass bands with $L0$ and $L4$ respectively. We also increase the weight of the newly introduced bands to its full weight over a ramp of 2 epochs. Introducing the bands from fine to coarse keeps the finer bands from being dominated by the coarser ones early in training, and gives

Table 5: **Teacher reconstruction vs. ground truth:** We compare two targets for our distillation objective, the teacher reconstruction at 512^2 and the ground truth at 512^2. The teacher target gives better image quality and a higher F-Sim, while the ground-truth target is slightly better on rFID. Evaluated on TokBench faces.

Distillation target	PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓	F-Sim ↑
SD-VAE ft-MSE	26.15	0.7703	0.1536	7.86	0.4542
Stage 1: Distillation					
teacher reconstruction @512^2	**26.36**	**0.7804**	0.1525	9.04	**0.4729**
ground truth @512^2	26.18	0.7766	**0.1519**	**8.58**	0.4669
Stage 1 + 2: + adversarial post-tuning					
from teacher reconstruction target	26.15	**0.7760**	**0.1503**	7.77	**0.4667**
from ground truth target	26.15	0.7749	0.1510	**7.73**	0.4638

them more training time, which in turn leads slightly better fine-grained details as seen in Figure 11 (rightmost column). The two approaches yielded very similar results, with the all band introduction at the start, giving a better increase in F-Sim score and better image quality metrics, while fine to coarse band introduction gave better rFID as in shown in Table 6.

Table 6: **Effect of the band schedule:** We analyze when to introduce the steerable pyramid bands and study two settings: introducing them all at once at the start, and introducing them at a 5-epoch interval starting from the finest band. Evaluated on TokBench faces.

Band schedule	PSNR ↑	SSIM ↑	LPIPS ↓	rFID ↓	F-Sim ↑
SD-VAE ft-MSE	26.15	0.7703	0.1536	7.86	0.4542
Stage 1: Self-distillation					
All bands from epoch 0	**26.36**	**0.7804**	0.1525	**9.04**	0.4729
Fine to coarse band introduction	26.33	0.7796	**0.1523**	9.06	**0.4734**
Stage 1 + 2: + adversarial post-tuning					
All bands at epoch 0 stage 1	**26.15**	**0.7760**	**0.1503**	7.77	**0.4667**
Fine to coarse band introduction stage 1	25.97	0.7718	**0.1503**	**7.40**	0.4623