IMPROVING RECONSTRUCTION OF REPRESENTATION AUTOENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work leverages vision foundation models as image encoders for latent diffusion models (LDMs), as their semantic representations are favorable for generative modeling. However, these representations often lack low-level visual information (*e.g.*, color and texture), limiting reconstruction fidelity and posing a bottleneck to further scaling LDMs. To address this limitation, we propose LV-RAE, a representation autoencoder that augments pretrained semantic features with complementary low-level information, enabling high-fidelity reconstruction while preserving semantic structure. Beyond reconstruction, we identify a decoder-side source of the reconstruction–generation trade-off: sensitivity to fine-grained latent variations supports faithful reconstruction but can also amplify errors in generated latents and degrade generation quality. Our analysis suggests that excessive decoder responses along off-manifold directions contribute to this issue in high-dimensional latent spaces. We improve decoder robustness through noise-augmented fine-tuning and further enhance generation quality through controlled noise injection at inference. Moreover, replacing the Base decoder with a larger XL configuration, while keeping the encoder, latent representation, and trained diffusion model fixed, simultaneously improves reconstruction and generation, increasing PSNR from 30.91 to 32.07 dB and reducing unguided gFID from 2.42 to 1.86. These results demonstrate that LV-RAE supports both high-fidelity reconstruction and strong generation, while highlighting the decoder as an important factor in the observed reconstruction–generation balance.

1 INTRODUCTION

Variational Autoencoders (VAEs) (Kingma & Welling, 2013) serve as the foundation of latent diffusion models (LDMs) (Rombach et al., 2022), enabling diffusion processes to operate in a compact latent space while preserving visual fidelity. This paradigm has been instrumental in scaling diffusion models to high-resolution image synthesis.

Building on this paradigm, recent studies have further shown that leveraging Vision Foundation Models (VFMs) as feature encoders can substantially improve the generative performance of diffusion models, as their semantic features are linearly separable and easy for diffusion models to learn. However, as shown in Fig. 1(b), images reconstructed from VFM semantic features often suffer from degraded visual fidelity due to missing low-level information, limiting their applicability in tasks such as low-level vision (Liu et al., 2025; Chang et al., 2025) and precise image editing (Zhang et al., 2023; Duan et al., 2025). More importantly, recent studies (Labs et al., 2025; Team et al., 2025) indicate that reconstruction fidelity has emerged as a primary bottleneck when further scaling LDMs.

The fundamental issue stems from a mismatch between what VFMs are trained to capture and what features are required for faithful reconstruction. Based on the manifold hypothesis (Carlsson, 2009), as shown in Fig. 2, we consider the real data manifold as a low-dimensional manifold decomposed into a smooth base manifold for global semantics and local variations for low-level information. Since semantic features focus on high-level understanding, they capture the smooth base but ignore local variations, explaining why VFM-based reconstructions often lack visual fidelity.

To bridge this gap, a straightforward solution is to fine-tune the VFM with a reconstruction loss to fit local variations, alongside an alignment loss to preserve the base manifold structure. However, this strategy can be sub-optimal. The two losses introduce competing objectives, causing the base

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.

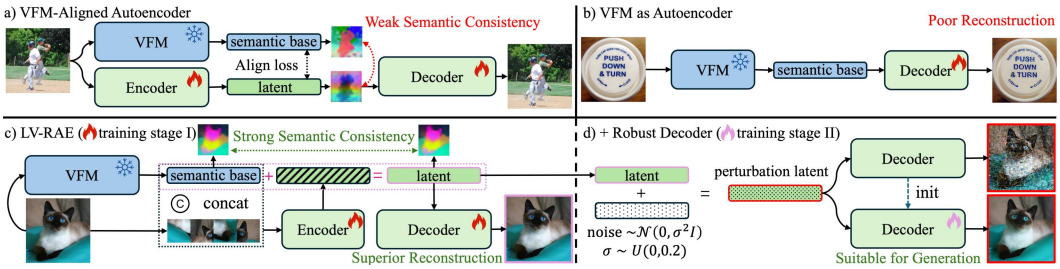


Figure 1: **Overview of previous methods and LV-RAE.** a) Training an autoencoder to align with VFMs fails to adequately preserve semantic consistency. b) Directly using a VFM as an autoencoder suffers from severely degraded reconstruction quality. c) The proposed LV-RAE significantly enhances reconstruction fidelity while effectively maintaining semantic representations. d) Fine-tune the LV-RAE decoder to improve robustness of latent perturbations, making it suitable for generation.

manifold itself to continuously drift during training and depriving the local variations of a stable reference. As a result (section 4.2), the learning of local variations becomes unstable and fails to converge to a consistent solution.

To address this limitation, we propose the **L**ocal-**V**ariations Augmented **R**epresentation **A**uto**e**ncoder (LV-RAE). As shown in Fig. 1(c), rather than fine-tuning the VFM to produce the latent directly, LV-RAE treats the semantic features as a fixed base manifold, and employs a shallow encoder to learn the missing low-level information (i.e., local variations) not captured by the VFM. Specifically, the encoder takes both the input image and its corresponding semantic features as inputs. The decoder then reconstructs the image by adding the encoder outputs to the semantic features. Notably, we find that only minimal adjustments to the semantic features are required to achieve high-fidelity image reconstruction (PSNR$\sim$32.32) while remaining highly aligned with semantics (CKNNA$\sim$0.99).

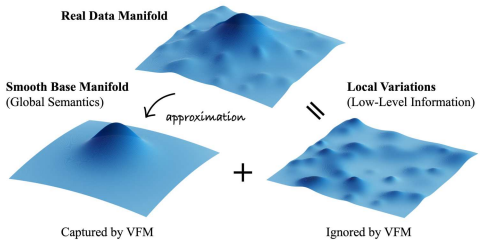


Figure 2: **Conceptual decomposition of the real data manifold.** We hypothesize that the real data manifold can be decomposed into two distinct components: a smooth base manifold representing global semantics captured by VFMs, and local variations representing low-level information ignored by VFMs.

Furthermore, as shown in Fig. 1(d), we observe pronounced sensitivity to latent perturbations in the high-dimensional LV-RAE representation. A controlled toy experiment illustrates how off-manifold decoder responses can amplify small latent deviations into pronounced visual artifacts, providing a possible explanation for this behavior. This observation highlights a decoder-side source of the reconstruction–generation trade-off: faithful reconstruction requires the decoder to exploit fine-grained latent variations, but this responsiveness can also amplify errors in generated latents. The observed trade-off therefore depends not only on the difficulty of modeling information-rich latent distributions, but also on how the decoder translates residual modeling errors into image-space distortions. To mitigate this issue, we adopt a two-stage strategy that addresses decoder behavior during both training and inference. First, we fine-tune the decoder with stochastic latent noise to regularize its local response and improve robustness to latent perturbations. Second, after diffusion sampling, we inject a controlled amount of Gaussian noise into the final latent before decoding, which provides an empirical control over decoding behavior and improves generation quality within an appropriate noise range. Beyond robustness, we find that replacing the Base decoder with a larger XL configuration, while keeping the encoder, latent representation, and trained diffusion model fixed, simultaneously improves reconstruction fidelity and generation quality. Together, these findings suggest that the observed reconstruction–generation balance is not determined by the latent representation alone: decoder-side changes can improve both reconstruction and generation without retraining the diffusion prior.

Our contributions are summarized as follows:

- We propose LV-RAE, a representation autoencoder that augments pretrained semantic features with complementary low-level information, enabling high-fidelity reconstruction while preserving strong semantic alignment.
- We observe substantial decoder sensitivity to latent perturbations, and use a controlled toy experiment to illustrate how off-manifold decoder responses can amplify errors in generated latents. A noise-augmented decoder training recipe and controlled inference-time noise improve the reported robustness and generation results.
- We show that replacing the Base decoder with a larger XL configuration improves both reconstruction and generation while keeping the encoder, latent representation, and diffusion prior fixed. Reconstruction PSNR increases from 30.91 to 32.07 dB and unguided gFID decreases from 2.42 to 1.86, highlighting the decoder as an important factor in the observed reconstruction–generation balance.

2 Related Work

Here, we discuss previous work on the line of representation autoencoders and noise augmentation. We present a more detailed related work discussion in Section C.

Representation autoencoders. Instead of aligning a compact latent with semantic features from VFMs (Yao et al., 2025; Leng et al., 2025; Xiong et al., 2025), a growing body of work adopts VFMs (Oquab et al., 2023; Siméoni et al., 2025) directly as encoders (Zheng et al., 2025; Shi et al., 2025; Chen et al., 2025a; Bi et al., 2025; Gao et al., 2025). Since diffusion models are traditionally thought to struggle with high-dimensional latent distributions, several works distill semantic features into lower-dimensional latents (Chen et al., 2025a; Bi et al., 2025; Gao et al., 2025), whereas RAE (Zheng et al., 2025) shows that diffusion models can be trained directly on high-dimensional semantic features with minimal architectural modifications. A key limitation remains: VFM features are optimized for high-level tasks and lack the low-level information that reconstruction requires, so decoded images exhibit noticeable visual degradation. LV-RAE targets this bottleneck.

Noise augmentation. Perturbing inputs to improve robustness is a classical strategy (Matsuoka, 1992; Bishop, 1995; An, 1996) that underlies denoising autoencoders (Vincent et al., 2008; 2010; Irobe et al., 2024) and, more recently, tokenizer learning for generative models (Qiu et al., 2025; Zheng et al., 2025; Zhai et al., 2024), where it primarily serves to bridge the gap between encoded and generated latent distributions. Building on this line of work, we study decoder behavior in an information-rich semantic latent space, focusing on how decoder robustness and decoder-side upgrades affect the reconstruction–generation balance when the latent representation and diffusion prior are held fixed.

3 Method

In this section, we first introduce the proposed LV-RAE (section 3.1). We then analyze the decoder's sensitivity to latent perturbations via a toy experiment (section 3.2), and propose an augmentation strategy to enhance decoder robustness (section 3.3). The framework is illustrated in Fig. 1(c) and (d).

3.1 Improving Representation Autoencoder

Motivation As illustrated in Fig. 1, our goal is to address the poor reconstruction fidelity when directly using VFMs as encoders. VFM features typically reside in extremely high-dimensional spaces, *i.e.*, with channel dimensions far exceeding those of conventional VAE latents. We argue that in such high-dimensional semantic spaces, even a minimal change can lead to substantial improvements in reconstruction fidelity.

A straightforward approach is to finetune the VFM using both a reconstruction loss and an alignment loss. However, we find that this strategy is sub-optimal. Such a training paradigm implicitly forces the encoder to simultaneously fit the base manifold and the local variations defined on top of it. Since the base manifold itself keeps evolving during training, the corresponding local variations

lack a stable reference and therefore cannot converge consistently toward a well-defined direction, ultimately hindering effective optimization.

To overcome this limitation, we propose the Local-Variations Augmented Representation Autoencoder (LV-RAE). Rather than directly fine-tuning the VFM, LV-RAE treats the VFM semantic features as a fixed base manifold. Instead of forcing the encoder to extract both semantic and low-level information into a shared latent, the encoder is tasked solely with learning the low-level information that is missing from the semantic features, leading to more stable and effective optimization. The alignment objective is retained to regularize the deviation of the enhanced representation from the semantic reference.

Architecture Both the residual encoder and decoder adopt Transformer architectures equipped with RoPE positional embeddings, using a patch size of 16×16. The residual encoder comprises six Transformer layers and is designed to capture low-level information complementary to the pretrained VFM features. The decoder is instantiated in two configurations, Base and XL, which let us compare decoder upgrades under a fixed latent representation. Detailed architectural configurations are provided in Appendix F.

Training Let Φ denote a VFM, *i.e.*, DINOv3 (Siméoni et al., 2025), and let $E(\cdot)$ and $D(\cdot)$ denote the encoder and decoder, respectively. Given an image $X \in \mathbb{R}^{H\times W\times 3}$, the VFM Φ extracts a pre-normalization feature $u \in \mathbb{R}^{N\times D}$, taken before the final LayerNorm LN_Φ of Φ. We write $\bar{u} = \mathrm{LN}_\Phi(u)$ for the corresponding semantic feature, which serves as the alignment target below.

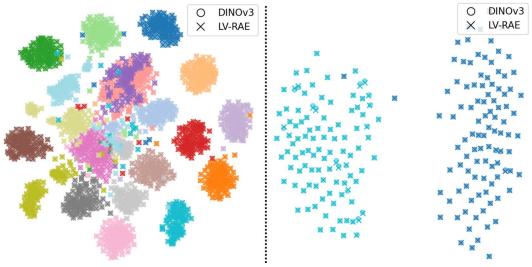


Figure 3: **t-SNE visualization of DINOv3 semantic features and LV-RAE latents. Left**: 20-class setting. **Right**: 2-class setting.

The image X is first projected into a patch-level representation $x_{\text{in}} \in \mathbb{R}^{N\times D}$ via a 16×16 convolution, where $N = \frac{H}{16} \times \frac{W}{16}$. The projected features are then concatenated with the pre-normalization feature u along the token dimension, forming

$$x = [x_{\text{in}}; u] \in \mathbb{R}^{2N\times D}, \tag{1}$$

which serves as the input to the encoder E. Let $r = E(x)[:N, :] \in \mathbb{R}^{N\times D}$ denote the first N tokens of the encoder output. We interpret these tokens as encoding the low-level information missing from the semantic features. The complete latent $z \in \mathbb{R}^{N\times D}$ is then obtained by element-wise addition of r and u, followed by a trainable LayerNorm LN_θ:

$$z = \mathrm{LN}_\theta(r + u). \tag{2}$$

Notably, we initialize the final linear layer of the encoder with all-zero weights, and initialize LN_θ with the parameters of LN_Φ. At initialization we therefore have $r = 0$ and $z = \bar{u}$, so the semantic feature is left unchanged.

The decoder D then reconstructs the image from z, producing $\bar{X} \in \mathbb{R}^{H\times W\times 3}$. For image reconstruction, we employ a combination of the pixel-wise loss $\mathcal{L}_1$ and the perceptual loss $\mathcal{L}_{lpips}$:

$$\mathcal{L}_{rec} = \alpha\mathcal{L}_1(X, \bar{X}) + \beta\mathcal{L}_{Lpips}(X, \bar{X}), \tag{3}$$

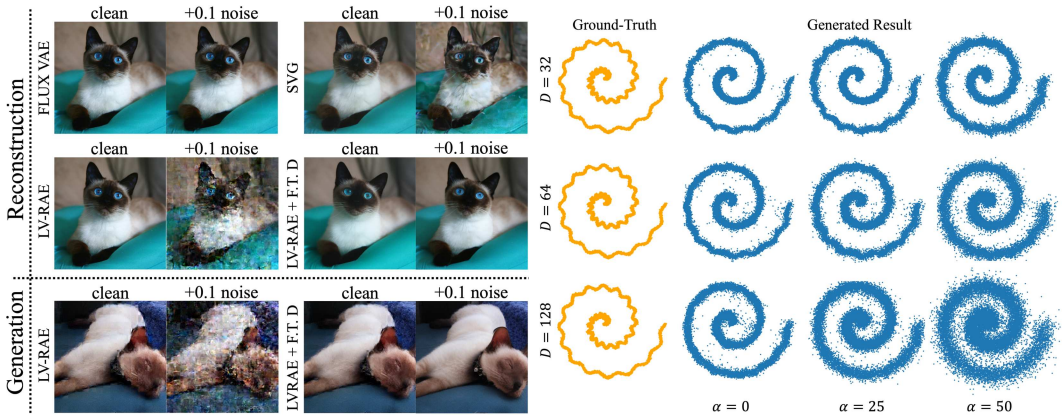
where $\alpha = \beta = 1$ are weighting coefficients. Simultaneously, to ensure the learned latent space remains grounded in the semantic manifold, we explicitly align the latent representation z with the semantic feature $\bar{u}$ via an $\mathcal{L}_2$ loss:

$$\mathcal{L}_{\text{align}} = \|z - \bar{u}\|_2^2. \tag{4}$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \eta \cdot \mathcal{L}_{\text{align}}, \tag{5}$$

where $\eta = 5$ is the weighting coefficient. As shown qualitatively in Fig. 3, the latents produced by LV-RAE overlap closely with the corresponding VFM semantic features; we quantify this alignment using CKNNA in section 4.



(a) Decoder sensitivity to latent perturbations.

(b) Toy experiment.

Figure 4: **Decoder sensitivity analysis and toy experiment. (a) Top:** Small latent perturbations can cause pronounced reconstruction artifacts. **Bottom:** Noise-augmented decoder fine-tuning and inference-time noise injection reduce generation artifacts. **(b)** Two-dimensional data are embedded into a D-dimensional latent space for diffusion modeling. The parameter α controls the decoder's sensitivity to off-manifold directions. Despite identical, exact on-manifold reconstruction, larger α leads to greater distortions in decoded samples, particularly at $D = 128$.

3.2 Toy Experiment

Despite its high reconstruction fidelity, LV-RAE exhibits substantial decoder sensitivity to latent perturbations, as shown in Fig. 4(a). Under the manifold assumption, the latent data distribution is concentrated on a low-dimensional manifold embedded in a higher-dimensional space. Excessive decoder responses along directions orthogonal to this manifold can amplify small latent deviations into pronounced visual artifacts. We illustrate this possible mechanism through a controlled toy experiment, in which the decoder response along off-manifold directions can be varied explicitly.

We begin by sampling points $\hat{x} \in \mathbb{R}^2$ from a simple distribution and embedding them into a D-dimensional latent space using a random matrix $P \in \mathbb{R}^{D\times 2}$ with orthonormal columns, satisfying $P^\top P = I_2$. The resulting latents are $z = P\hat{x} \in \mathbb{R}^D$. We train a diffusion model, implemented as a five-layer ReLU MLP with 512-dimensional hidden layers, to model this latent distribution.

To isolate the effect of decoder behavior at a fixed latent dimensionality, we construct a decoder that responds to directions both parallel and orthogonal to the data manifold. Let $U \in \mathbb{R}^{D\times(D-2)}$ be an orthonormal basis for the orthogonal complement of the column space of P, satisfying $U^\top U = I_{D-2}$ and $P^\top U = 0$. We define the decoder as

$$D(z) = P^\top z + \alpha W^\top \sin(\beta U^\top z), \tag{6}$$

where α and β are hyperparameters, the sine function is applied element-wise, and $W \in \mathbb{R}^{(D-2)\times 2}$ has entries independently sampled from $\mathcal{N}(0, 1/(D-2))$. Importantly, for any latent on the data manifold, $U^\top z = 0$ and therefore $D(P\hat{x}) = \hat{x}$, regardless of α and β. Thus, these decoders achieve identical, exact reconstruction on the data manifold while differing in their responses to off-manifold perturbations. The decoder Jacobian is

$$J_D(z) = \frac{\partial D(z)}{\partial z} = P^\top + \alpha\beta W^\top \operatorname{Diag}\big(\cos(\beta U^\top z)\big)\, U^\top. \tag{7}$$

The first term recovers variations within the data manifold, whereas the second governs the response to off-manifold perturbations. For simplicity, we set $\beta = 1$ and vary α to control the strength of the off-manifold response.

We evaluate different latent dimensionalities and values of α, with results shown in Fig. 4(b). When $\alpha = 0$, the decoder reduces to a projection onto the data-manifold directions and discards all off-manifold components. In this setting, the decoded distributions remain largely stable across the tested dimensionalities. When $\alpha > 0$, off-manifold errors in generated latents also contribute to

the decoded outputs. In the evaluated synthetic settings, higher-dimensional latents exhibit more pronounced decoded distortions when the decoder responds to off-manifold errors, particularly for larger α, leading to substantial loss of structure. These observations illustrate how decoder sensitivity can magnify residual latent modeling errors and degrade image-level generation quality.

Crucially, in this synthetic setting these differences in generation quality arise despite identical reconstruction on the data manifold, showing that reconstruction fidelity alone does not determine the reliability of decoding generated latents. While the geometry of real LV-RAE latents is not characterized here, this analysis offers a possible explanation for the perturbation sensitivity in Fig. 4(a), and motivates preserving responsiveness to fine-grained variations while suppressing excessive responses to off-manifold errors. Decoder robustness therefore matters alongside latent modeling difficulty when interpreting and mitigating the reconstruction–generation trade-off.

3.3 Decoder with Noise Augmentation

Motivation To mitigate the decoder-side reconstruction–generation trade-off, we fine-tune the decoder with stochastic latent noise. Reconstruction supervision on perturbed latents encourages stable local responses, providing a practical means of regularizing excessive decoder sensitivity. The goal is to improve robustness to latent errors while preserving responsiveness to the fine-grained variations required for faithful reconstruction. Because the accuracy of latent modeling varies across diffusion priors, the appropriate balance between robustness and reconstruction fidelity may also differ. We therefore adopt a decoupled framework that combines noise-augmented decoder training with adjustable noise injection at inference. During training, randomly sampled noise amplitudes expose the decoder to varying levels of latent corruption. At inference, controlled noise injection into the final generated latents provides an empirical means of adjusting decoding behavior without modifying the diffusion model.

Training Starting from the LV-RAE trained in the previous stage, we freeze the encoder and fine-tune the decoder using Gaussian-perturbed latents:

$$\tilde{z} = z + \sigma\epsilon, \quad \sigma \sim \mathcal{U}(0, \tau), \quad \epsilon \sim \mathcal{N}(0, I), \tag{8}$$

where $\tau = 0.2$ controls the maximum noise amplitude. The decoder reconstructs the original image X from the perturbed latent, producing $\bar{X} = D(\tilde{z})$. In addition to the reconstruction loss $\mathcal{L}_{\text{rec}}$, we employ an adversarial loss (Goodfellow et al., 2014) to encourage perceptual realism. To balance the two objectives, we compute an adaptive weight based on their gradient norms with respect to the parameters θ_{last} of the decoder's final layer:

$$w_{\text{gan}} = \frac{\|\nabla_{\theta_{\text{last}}} \mathcal{L}_{\text{rec}}\|_2}{\|\nabla_{\theta_{\text{last}}} \mathcal{L}_{\text{gan}}\|_2}. \tag{9}$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \kappa\, w_{\text{gan}}\, \mathcal{L}_{\text{gan}}, \tag{10}$$

where $\kappa = 0.75$ controls the relative strength of adversarial supervision.

Sampling After diffusion sampling, we inject controlled Gaussian noise into the final generated latent z_0 before decoding:

$$\tilde{z}_0 = z_0 + \bar{\sigma}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{11}$$

where $\bar{\sigma}$ specifies the inference-time noise amplitude. For the noise-augmented decoder, this parameter provides an empirical control over decoding behavior, and an appropriate noise level improves generation quality, as shown in section 4.2. Qualitatively, larger $\bar{\sigma}$ yields visually smoother outputs. The final image is obtained as $X = D(\tilde{z}_0)$.

4 Experiments

4.1 Experiments Setup

Training We train all models on the ImageNet-1K (Russakovsky et al., 2015) 256×256 dataset. For generation, we use LightningDiT (Yao et al., 2025) and DiT$^{\text{DH}}$ (Zheng et al., 2025) as backbones

Table 1: **Reconstruction fidelity and semantic alignment of LV-RAE versus VFM fine-tuning.** Init. D. indicates whether the decoder is pretrained with a frozen VFM before joint fine-tuning.

Method	Target VFM	Init. D.	PSNR↑	SSIM↑	CKNNA↑
F.T. VFM	DINOv3-B	✗	29.94	0.899	0.925
		✓	30.88	0.920	0.925
LV-RAE	DINOv3-B	✗	**32.32**	**0.941**	**0.987**
F.T. VFM	DINOv2-B	✗	25.58	0.770	0.873
LV-RAE	DINOv2-B	✗	**29.30**	**0.892**	**0.982**
F.T. VFM	SigLip2	✗	23.48	0.683	0.802
LV-RAE	SigLip2	✗	**27.66**	**0.840**	**0.989**

Table 2: **Quantitative comparison of different decoders under increasing levels of latent corruption.**

Decoder	Latent	PSNR↑	SSIM↑	LPIPS↓
original	z	32.32	0.941	0.015
	$z + 0.1\epsilon$	17.72	0.437	0.466
	$z + 0.2\epsilon$	13.68	0.228	0.621
F.T. w/ $\sigma = 0.1$	z	25.45	0.810	0.092
	$z + 0.1\epsilon$	23.56	0.735	0.093
	$z + 0.2\epsilon$	19.98	0.581	0.182
F.T. w/ $\sigma \sim \mathcal{U}(0, 0.2)$	z	30.91	0.930	0.018
	$z + 0.1\epsilon$	23.78	0.735	0.100
	$z + 0.2\epsilon$	21.61	0.637	0.152

of diffusion model. We apply QK-Norm (Henry et al., 2020) to stabilize training and adopt the time-shift training strategy proposed in RAE (Zheng et al., 2025). More implementation details can be found in the Appendix F.

Evaluation For reconstruction, we evaluate on the ImageNet-1k (Russakovsky et al., 2015) validation set and the COCO2017 (Lin et al., 2014) validation set at 256×256 resolution, reporting PSNR, SSIM, and LPIPS (Zhang et al., 2018). For generation, we report the Frechet Inception Distance(gFID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), as well as Precision and Recall. These metrics are computed using 50K generated images, except for ablation studies, where 10K samples are used. Additionally, we report the Frechet Distance computed on DINOv2 (Oquab et al., 2023) features (FDD) to complement FID, as FDD is a more reliable metric (Stein et al., 2023; Skorokhodov et al., 2025).

4.2 Ablations

Effect of LV-RAE compared with fine-tuning the VFM. We compare LV-RAE with directly fine-tuning the VFM as the encoder (F.T. VFM) across different backbones. Reconstruction is evaluated using PSNR and SSIM on the COCO2017 validation set, and semantic alignment using CKNNA on the first 5,000 ImageNet validation images. As shown in Tab. 1, LV-RAE consistently outperforms direct fine-tuning in both reconstruction fidelity and semantic alignment. We adopt DINOv3-B as the default backbone, as it achieves the best reconstruction fidelity among the tested VFMs while maintaining strong semantic alignment. Beyond CKNNA, a linear classifier trained on the original semantic features attains the same reported top-1 accuracy when evaluated on the LV-RAE latents without adaptation (Appendix H.3). Further analysis of training dynamics is provided in Appendix H.4.

Effect of noise augmentation for reconstruction We investigate the reconstruction ability of different decoders under varying levels of latent corruption, with quantitative results reported in Tab. 2. When evaluated on clean latent z, the decoder without fine-tuning achieves the best reconstruction performance. However, its performance degrades sharply even under mild noise perturbations ($z + 0.1\epsilon$), indicating limited robustness to latent corruption. In contrast, the fine-tuned decoder exhibits a substantial improvement in robustness, maintaining significantly better reconstruction quality as the noise level increases. We further compare fine-tuning with a fixed noise magnitude ($\sigma = 0.1$) and fine-tuning with randomly sampled noise amplitudes ($\sigma \sim \mathcal{U}(0, 0.2)$). The results show that fine-tuning with a fixed noise level severely harms reconstruction performance on clean latent. We attribute this degradation to an implicit high-frequency truncation effect, where the decoder no longer attends to fine-grained latent information. In comparison, fine-tuning with random noise amplitudes enhances robustness while largely preserving reconstruction quality on clean latent inputs. We therefore adopt LV-RAE equipped with a fine-tuned decoder using $\sigma \sim \mathcal{U}(0, 0.2)$ as the default setting.

Effect of noise augmentation for generation After diffusion sampling, we inject a controlled amount of Gaussian noise into the final latent before decoding. For the noise-augmented decoder, this provides an empirical control over decoding behavior at inference time.

Table 3: **Quantitative comparison of reconstruction performance across different autoencoders.** Our proposed LV-RAE achieves state-of-the-art reconstruction quality among all autoencoders.

Tokenizer	Config.	ImageNet-1K				COCO 2017			
		PSNR↑	SSIM↑	LPIPS↓	rFDD↓	PSNR↑	SSIM↑	LPIPS↓	rFDD↓
SD-VAE (Rombach et al., 2022)	f8d4	25.65	0.748	0.068	13.78	25.38	0.760	0.065	43.72
FLUX-VAE (Labs et al., 2025)	f8d16	31.01	0.914	**0.017**	2.88	30.89	0.921	**0.015**	8.86
VA-VAE (Yao et al., 2025)	f16d32	26.26	0.786	0.046	4.18	26.07	0.799	0.044	18.76
SVG (Shi et al., 2025)	f16d392	21.87	0.632	0.114	24.92	21.71	0.645	0.112	60.66
RAE (Zheng et al., 2025)	f16d768	18.92	0.522	0.148	13.39	18.41	0.516	0.153	45.37
LV-RAE (ours)	f16d768	**32.50**	**0.938**	**0.017**	5.51	**32.32**	**0.941**	**0.015**	13.20
+ Noise Augmentation (ours)	f16d768	31.11	0.927	0.019	**2.49**	30.91	0.930	0.018	**8.74**
+ XL decoder (ours)	f16d768	32.19	0.935	0.018	2.82	32.07	0.939	0.016	8.77



Figure 5: **Qualitative comparison of reconstruction fidelity with SVG.** LV-RAE better preserves fine-grained structures, such as text and faces. Red boxes highlight key differences.

Fig. 6 visualizes the effect of different noise levels on the final generation quality. As the noise magnitude increases, the generation quality improves consistently: both FID and FDD decrease with increasing noise strength and begin to converge at around $\bar{\sigma} \approx 0.08$. We further compare models trained with different numbers of epochs (320 vs. 640) and observe that longer training has a limited impact on the overall trend, mainly resulting in a vertical shift of the curves. In contrast, the choice of model architecture has a much more significant effect. Specifically, when the noise level is low, larger-capacity models (e.g., DiT$^{\text{DH}}$-XL) achieve substantially better performance than their smaller counterparts (e.g., DiT-XL).

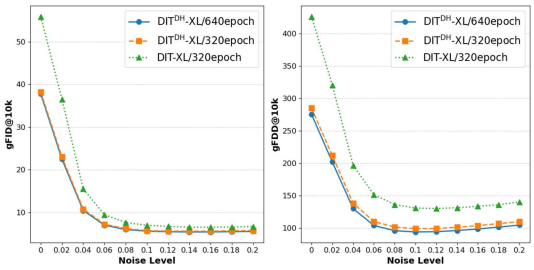


Figure 6: **Generation quality trends under different levels of noise injection at inference time.**

Decoder upgrades with fixed representations The experiments above ask how robust a fixed decoder is to latent perturbations. We now ask the complementary question: how much of the image-level generation quality is determined by the decoder itself? To this end, we keep the LV-RAE encoder, the latent representation, the trained diffusion model, and the sampling procedure unchanged, replace the default Base decoder with the larger XL configuration, and decode the *same* generated latents with both. As reported in Tab. 4, the XL decoder improves the unguided gFID of DiT$^{\text{DH}}$-XL from 2.42 to 1.86 and its guided gFID from 1.82 to 1.42, while reconstruction fidelity also improves (PSNR from 30.91 to 32.07). Applying the same substitution to DiT-XL improves the guided gFID from 2.61 to 1.98, indicating that the effect is not specific to a single diffusion model or training budget in our experiments. Since both configurations decode the same generated latents, these improvements arise from changes on the decoder side rather than from changes in latent modeling. The Base and XL decoders follow different initialization and training schedules, as detailed in Appendix F.2; this comparison therefore evaluates a decoder-only upgrade rather than a training-matched ablation of capacity alone. A more detailed analysis, including a comparison against a decoder trained directly on the semantic features, is provided in Appendix H.2.

Table 4: **Class-conditional generation performance on ImageNet 256×256.**

Method	#Params	Epochs	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Latent Diffusion										
DiT (Peebles & Xie, 2022)	675M	1400	9.62	121.5	0.67	0.67	2.27	278.2	**0.83**	0.57
SiT (Ma et al., 2024)	675M	1400	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
REPA (Yu et al., 2025)	675M	800	5.90	157.8	0.70	**0.69**	1.42	305.7	0.80	0.65
VA-VAE (Yao et al., 2025)	675M	800	2.17	205.6	0.77	0.65	1.35	295.3	0.79	0.65
DDT (Wang et al., 2025b)	675M	400	6.27	154.7	0.68	**0.69**	**1.26**	**310.6**	0.79	0.65
SVG (Shi et al., 2025)	675M	1400	3.36	181.2	-	-	1.92	264.9	-	-
RAE (Zheng et al., 2025)	839M	800	**1.60**	**242.7**	-	-	1.28	262.9	-	-
Latent Diffusion with LV-RAE + 0.1 noise (Ours)										
DiT-XL	675M	400	3.77	185.4	0.76	0.60	2.61	208.7	0.76	0.64
+ XL decoder	675M	400	2.94	195.0	0.78	0.61	1.98	219.2	0.77	0.64
DiT^{DH}-XL	839M	800	2.42	223.8	0.77	0.64	1.82	249.7	0.75	0.67
+ XL decoder	839M	800	1.86	232.3	**0.78**	0.64	1.42	257.7	0.76	**0.68**

4.3 Reconstruction Performance Comparison

Reconstruction results are summarized in Tab. 3. Compared with other tokenizers, the latents produced by LV-RAE preserve richer fine-grained details, leading to state-of-the-art reconstruction performance. After noise-augmented decoder fine-tuning, reconstruction performance decreases but remains better than other tokenizers. Qualitative results are shown in Fig. 5. Previous representation autoencoders (e.g., SVG) suffer from severe structural distortions that significantly degrade visual quality. In contrast, LV-RAE handles these cases effectively, producing structurally coherent and visually faithful reconstructions.

4.4 Generation Performance Comparison

Tab. 4 summarizes the class-conditional generation results on ImageNet 256×256. With the Base decoder, LV-RAE achieves an unguided gFID of 3.77 using DiT-XL trained for 400 epochs and 2.42 using DiT^{DH}-XL trained for 800 epochs. For the latter model, replacing the Base decoder with XL further improves unguided and guided gFID to 1.86 and 1.42, respectively, without changing the latent representation or the trained diffusion model (section 4.2). The resulting unguided gFID is close to RAE (Zheng et al., 2025) (1.86 vs. 1.60), while LV-RAE achieves over 13 dB higher reconstruction PSNR on COCO 2017. LV-RAE thus supports high-fidelity reconstruction while maintaining competitive generation performance.

Table 5: **Additional generation comparison.** LV-RAE achieves the best gFDD without guidance. † denotes results reproduced using official open-source weights.

Method	Epochs	Steps	gFID↓	gFDD↓
SVG† (Shi et al., 2025)	1400	250	3.47	130.4
VA-VAE† (Yao et al., 2025)	800	250	2.23	74.3
LV-RAE (ours)	800	250	2.42	58.2
+ XL decoder (ours)	800	250	**1.86**	**35.1**

To complement Inception-based gFID, we also evaluate generation quality with DINOv2-based gFDD (Tab. 5). With the Base decoder, LV-RAE obtains a gFDD of 58.2, outperforming the reproduced SVG and VA-VAE baselines. The XL decoder further reduces gFDD to 35.1, the best on both metrics among the compared methods. Additional comparisons and discussions are provided in Appendix H.

5 Conclusion

In this paper, we propose LV-RAE, which augments pretrained semantic features with complementary low-level information, enabling faithful reconstruction while preserving semantic alignment. We observe that the LV-RAE decoder is sensitive to latent perturbations, and a controlled toy experiment illustrates how off-manifold decoder responses can amplify generation errors, providing a possible explanation for this behavior. We mitigate this issue through noise-augmented decoder fine-tuning and controlled noise injection at inference. Moreover, replacing the Base decoder with a larger XL configuration improves both reconstruction and generation while keeping the encoder, latent representation, and trained diffusion model fixed. Together, these findings highlight the decoder as an important factor in the observed reconstruction–generation balance.

REFERENCES

Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models. *arXiv preprint arXiv:2510.18457*, 2025.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7 (1):108–116, 1995.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Zewei Chang, Zheng-Peng Duan, Jianxing Zhang, Chun-Le Guo, Siyu Liu, Hyungju Chun, Hyunhee Park, Zikun Liu, and Chongyi Li. Pertouch: Vlm-driven agent for personalized and semantic image retouching. *arXiv preprint arXiv:2511.12998*, 2025.

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. *arXiv preprint arXiv:2509.25162*, 2025a.

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. *arXiv preprint arXiv:2410.10733*, 2024.

Junyu Chen, Dongyun Zou, Wenkun He, Junsong Chen, Enze Xie, Song Han, and Han Cai. Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. *arXiv preprint arXiv:2508.00413*, 2025b.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025c.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Zheng-Peng Duan, Jiawei Zhang, Siyu Liu, Zheng Lin, Chun-Le Guo, Dongqing Zou, Jimmy Ren, and Chongyi Li. A diffusion-based framework for occluded object movement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2816–2824, 2025.

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.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer, 2023.

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *arXiv preprint arXiv:2512.07829*, 2025.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

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. *Advances in neural information processing systems*, 30, 2017.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Hiroo Irobe, Wataru Aoki, Kimihiro Yamazaki, Yuhui Zhang, Takumi Nakagawa, Hiroki Waida, Yuichiro Wada, and Takafumi Kanamori. Robust vaes via generating process of noise augmented data. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 587–592. IEEE, 2024.

Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. *arXiv preprint arXiv:2212.11972*, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Siyu Liu, Zheng-Peng Duan, Jia OuYang, Jiayi Fu, Hyunhee Park, Zikun Liu, Chun-Le Guo, and Chongyi Li. Faceme: Robust blind face restoration with personal identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5567–5575, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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*, pp. 23–40. Springer, 2024.

Kiyotoshi Matsuoka. Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):436–440, 1992.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Kai Qiu, Xiang Li, Hao Chen, Jason Kuen, Xiaohao Xu, Jiuxiang Gu, Yinyi Luo, Bhiksha Raj, Zhe Lin, and Marios Savvides. Image tokenizer needs post-training. *arXiv preprint arXiv:2509.12474*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025b.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv preprint arXiv:2504.08736*, 2025.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers. *arXiv preprint arXiv:2507.15856*, 2025.

Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in Neural Information Processing Systems*, 37:56166–56189, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models. *arXiv preprint arXiv:2412.06329*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

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*, pp. 586–595, 2018.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in neural information processing systems*, 33:7559–7570, 2020.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. In *Transactions on Machine Learning Research (TMLR)*, 2024.

A AI ASSISTANCE

We used large language models (LLMs) as assistive tools to polish the manuscript, search for and identify related papers, and generate some non-critical code for our experiments. We did not use LLMs to generate synthetic datasets or prove mathematical claims. All LLM-assisted work, including the retrieved references and generated code, was reviewed and checked by the authors. The authors take full responsibility for the final content of this paper, including its text, references, methodology, results, and conclusions.

B IMPACT STATEMENT

This paper improves the autoencoders used in latent diffusion models, focusing on the simultaneous enhancement of reconstruction quality and semantic distribution. These advancements support diverse applications in creative content generation and generative AI research. Our method does not expand the functional domain of existing models; therefore, it inherits the standard ethical considerations and societal challenges associated with large-scale generative systems.

C EXTENDED RELATED WORK

This section expands the condensed discussion of section 2 in the main text.

C.1 AUTOENCODERS FOR LATENT DIFFUSION MODELS

Autoencoders are a core component of LDMs, as their design directly determines the quality and efficiency of the models. Prior studies have primarily focused on improving the compression ratio (Chen et al., 2024; 2025b) and reconstruction fidelity (Labs et al., 2025; Esser et al., 2024). Recent studies suggest that the structure of the latent space, *i.e.*, *diffusability* (Skorokhodov et al., 2025), plays a critical role in diffusion model training. To improve diffusability, some studies (Kouzelis et al., 2025; Skorokhodov et al., 2025) have observed that latent representations often contain excessively high-frequency components and have proposed simple regularization strategies to suppress these components. In parallel, a line of research (Yao et al., 2025; Leng et al., 2025; Xiong et al., 2025) has explored aligning latent with semantic features extracted from VFMs (Oquab et al., 2023; Siméoni et al., 2025), demonstrating substantially faster convergence and improved generative performance. Despite their success, these methods often suffer from an information bottleneck when aligning compact latents with high-dimensional VFM features. This constraint makes it challenging for the latent space to simultaneously accommodate high-level semantic priors and fine-grained low-level information, thereby both degrading reconstruction fidelity and hindering further performance gains.

C.2 REPRESENTATION AUTOENCODERS

To further enrich the semantic density of the latent space, recent studies (Zheng et al., 2025; Shi et al., 2025; Chen et al., 2025a; Bi et al., 2025; Gao et al., 2025) have explored the direct use of VFMs as encoders. Traditionally, diffusion models are thought to struggle with learning high-dimensional latent distributions. Consequently, several works (Chen et al., 2025a; Bi et al., 2025; Gao et al., 2025) have attempted to distill high-dimensional semantic features into lower-dimensional latents while preserving semantic information. In contrast, RAE (Zheng et al., 2025) demonstrates that diffusion models can be trained directly on high-dimensional semantic features and achieve strong generative performance with only minimal architectural modifications. Nevertheless, a key limitation remains: VFM features are primarily optimized for high-level tasks and lack the low-level information required for image reconstruction. As a result, image reconstruction based on such semantic features often suffers from noticeable visual degradation.

C.3 NOISE AUGMENTATION

Injecting noise into inputs to improve neural network robustness is a classical strategy that has been widely studied and applied (Matsuoka, 1992; Bishop, 1995; An, 1996). Although these methods often

employ similar input perturbations, their modeling objectives, underlying mechanisms, and practical benefits can differ substantially across tasks, leading to several distinct research directions. Early representative works, such as denoising autoencoders (Vincent et al., 2008; 2010; Irobe et al., 2024), focused on training models to reconstruct clean images from corrupted inputs, thereby encouraging the learning of feature representations that are invariant to local perturbations. More recently, this paradigm has been extended to tokenizer learning for generative models (Qiu et al., 2025; Zheng et al., 2025; Zhai et al., 2024), where the primary goal is to bridge the gap between encoded and generated latent distributions. Building on noise-augmented decoder training, we study decoder behavior in information-rich semantic latent spaces. Our focus is on how decoder robustness and decoder-side upgrades affect the reconstruction–generation balance when the latent representation and diffusion prior are held fixed.

C.4 Robust Decoders for Latent Diffusion Models

The decoder plays a crucial role in generation. Recent works (Yang et al., 2025; Qiu et al., 2025; Zheng et al., 2025) have shown that fine-tuning the decoder with perturbed latent inputs can significantly enhance decoder robustness, leading to substantial improvements in generative performance. Our observations are consistent with these findings. Building on them, we study decoder behavior in an information-rich semantic latent space, and show in Appendix H.2 that decoder-side changes can shift the reconstruction–generation balance while the latent representation and diffusion prior are held fixed.

D Background

Flow Matching Flow matching (Liu et al., 2022; Lipman et al., 2022; Esser et al., 2024) learns a deterministic probability flow that continuously transports a known prior distribution $p_1 = \mathcal{N}(0, \mathrm{I})$ to the target data distribution p_0. To estimate a time-dependent velocity field v_t that induces a probability path between p_0 and p_1, we can define a forward interpolating process between $x_0 \sim p_0$ and $\epsilon \sim p_1$:

$$x_t = \alpha_t x_0 + \beta_t \epsilon, \quad 0 \le t \le 1, \tag{12}$$

where the coefficients satisfy the boundary conditions $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$.

The distribution of this stochastic interpolant coincides with that of the solution to a deterministic ODE whose drift is the corresponding conditional expectation. Let $v_t(X) = \mathbb{E}\left[\frac{dx_t}{dt} \,\middle|\, x_t = X\right]$, then the probability flow ODE

$$dX_t = v_t(X_t)dt, \quad X_1 \sim p_1 \tag{13}$$

satisfy $\mathrm{Law}(X_t) = \mathrm{Law}(x_t)$ for all $t \in [0, 1]$.

To learn the velocity field v_t, we can use empirical risk minimization to approximate this conditional expectation in practice. This is achieved by optimizing the square loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \epsilon}\left[\|v_\theta(x_t, t) - \dot{x}_t\|_2^2\right], \tag{14}$$

where $\dot{x}_t = \frac{dx_t}{dt}$ is analytically available.

E Limitations and Future Work

Despite the promising results, several limitations of our work remain to be addressed:

Domain Specificity: Our framework is built upon a Vision Foundation Model (VFM) pre-trained predominantly on natural images. Consequently, its generalization to specialized domains—such as medical imaging or remote sensing—may be constrained by the discrepancy in both underlying representations and the domain-specific statistics of local variations.

Sensitivity to Hyperparameters: The generation quality is sensitive to the noise level, which currently requires empirical tuning. Since different generative models capture the latent distribution to varying degrees, the optimal noise intensity is context-dependent. Suboptimal calibration often leads to a trade-off: excessive noise may cause over-smoothing of fine details, while insufficient noise fails to provide the necessary robustness against latent perturbations.

Dimensionality and Guidance Efficacy: We observe a diminishing return of Classifier-Free Guidance (CFG) in high-dimensional latent spaces. Compared to its effectiveness in low-dimensional settings, the guidance signals in higher dimensions appear less potent at steering the generative process, suggesting that CFG's scaling behavior is sensitive to the latent manifold's dimensionality.

We leave the exploration of these challenges for future work.

F Implementation Details

F.1 Diffusion Model

DiT and DiT$^{\text{DH}}$ model details. Our DiT and DiT$^{\text{DH}}$ models are built upon the RAE (Zheng et al., 2025) configuration. We adopt SwiGLU (Shazeer, 2020) activations and RMSNorm (Zhang & Sennrich, 2019), together with a Gaussian Fourier embedding (Song et al., 2020) layer for timestep encoding. For positional encoding, we apply Absolute Positional Embeddings to the input tokens in addition to RoPE (Su et al., 2024). To improve training stability, we incorporate QK-Norm (Henry et al., 2020); while QK-Norm slightly affects final generation performance, it provides more stable and reliable optimization during training (Yao et al., 2025; Chen et al., 2025a).

We follow prior work (Peebles & Xie, 2022; Zheng et al., 2025) and use the same model sizes for DiT-XL and DiT$^{\text{DH}}$-XL. We use DiT$^{\text{DH}}$-S as the guiding model for AutoGuidance(Karras et al., 2024):

- DiT-XL: hidden dimensionality of 1152, 28 transformer blocks, and 16 attention heads.
- DiT-S: hidden dimensionality of 384, 12 transformer blocks, and 6 attention heads.
- DiT$^{\text{DH}}$-XL: DiT-XL augmented with a DDT head (consisting of a hidden dimensionality of 2048, 2 transformer blocks, and 16 attention heads.)
- DiT$^{\text{DH}}$-S: DiT-S augmented with a DDT head (consisting of a hidden dimensionality of 2048, 2 transformer blocks, and 16 attention heads.)

DiT and DiT$^{\text{DH}}$ training details. All DiT and DiT$^{\text{DH}}$ models are trained on NVIDIA A100 40GB GPUs using mixed-precision training with bfloat16. We use the AdamW(Loshchilov & Hutter, 2017) optimizer with a learning rate of 2.0×10^{-4} for the first 40 epochs, which is then decayed to 2.0×10^{-5} until epoch 800. The total batch size is set to 1024. Following Zheng et al. (2025), we apply EMA of model weights with a decay rate of 0.9995.

Time shifting. We adopt the time-shifting schedule from RAE (Zheng et al., 2025) for both training and sampling:

$$t' = \frac{\alpha \cdot t}{1 + (\alpha - 1) \cdot t}, \quad \alpha = \sqrt{\frac{c \cdot h \cdot w}{4096}}. \tag{15}$$

Here, t denotes the diffusion timestep, ranging from 1 (full noise) to 0 (clean image), and α is the shifting coefficient. For a 256×256 input image, our autoencoder encodes it into a latent representation with $h = 16$, $w = 16$, and $c = 768$, resulting in $\alpha \approx 6.93$.

Sampling. We use standard ODE-based sampling with the Euler sampler, employing 250 sampling steps by default. Following Zheng et al. (2025), we adopt AutoGuidance as our primary guidance method. We train the smallest variant, DiT$^{\text{DH}}$-S, for 16 epochs as the guiding model and use a guidance scale of 1.4.

F.2 Autoencoders

LV-RAE model details. LV-RAE is implemented as a Transformer-based autoencoder equipped with RoPE (Su et al., 2024) and T5-MLP(Raffel et al., 2020), using a patch size of 16×16 and an MLP ratio of 4. The model configuration is as follows:

- Encoder: hidden dimensionality of 768, 6 transformer blocks, and 12 attention heads.

- Decoder (Base): hidden dimensionality of 768, 12 transformer blocks, and 12 attention heads.
- Decoder (XL): hidden dimensionality of 1152, 28 transformer blocks, and 16 attention heads.

The Base and XL variants share the same encoder, so the latent representation is identical in both cases and the two decoders can be compared in isolation.

LV-RAE training details. During the local-variations augmented stage (training stage I), each autoencoder is trained using the AdamW optimizer (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. We use a constant learning rate of 1.0×10^{-4} and a total batch size of 512. Each autoencoder is trained for 80k steps.

In the noise augmentation stage (training stage II), the encoder is frozen and the decoder is fine-tuned using the same optimization settings as in the previous stage. We first train the decoder using only the reconstruction loss for 10k steps, and then introduce the GAN loss for an additional 90k training steps. Following RAE (Zheng et al., 2025), we use DINO-S/8 as the discriminator. We additionally apply differentiable augmentations (Zhao et al., 2020), followed by a random crop to 224×224 before feeding samples into the discriminator.

The XL decoder is randomly initialized and trained on top of the same frozen LV-RAE encoder as the Base configuration, so that the latent representation is identical in both settings. It uses the same optimization settings and noise augmentation scheme, with 50k steps supervised by the reconstruction loss only, followed by 100k steps with the GAN loss.

G Evaluation Details

We evaluate our method from both reconstruction quality and generative performance perspectives, using standard metrics widely adopted in prior work.

G.1 Standard Feature Alignment Metrics

We use **CKNNA**(Centered Kernel Nearest-Neighbor Alignment)(Huh et al., 2024) to evaluate the alignment between the latent representations learned by LV-RAE and the semantic features extracted from VFM. CKNNA is a relaxed version of Centered Kernel Alignment(Kornblith et al., 2019) that measures representation alignment by comparing local neighborhood structures. We follow the evaluation protocol of the original work[1] for its computation.

G.2 Standard Reconstruction Metrics

To assess reconstruction fidelity, we report **PSNR**, **SSIM**, and **LPIPS**.

- **PSNR** measures pixel-wise reconstruction accuracy and is sensitive to low-level differences.
- **SSIM** evaluates structural similarity between reconstructed images and ground truth, emphasizing luminance, contrast, and structural consistency.
- **LPIPS** measures perceptual similarity using deep features from pretrained networks and correlates well with human perception.

G.3 Standard Generative Metrics

To evaluate generative performance, we report **FID**, **IS**, **Precision** and **Recall**. We strictly follow the setup and use the same reference batches of ADM[2](Dhariwal & Nichol, 2021) for evaluation.

- **FID** measures the distance between the distributions of generated and real images in the feature space of the Inception-v3 network (Szegedy et al., 2016).

[1] https://github.com/minyoungg/platonic-rep
[2] https://github.com/openai/guided-diffusion

Table 6: **Class-conditional generation performance on ImageNet 256×256.**

Method	#Params	Epochs	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Autoregressive										
MaskGiT (Chang et al., 2022)	227M	555	6.18	182.1	0.80	0.51	-	-	-	-
MagViT-v2 (Yu et al., 2023)	307M	1080	3.65	200.5	-	-	1.78	319.4	-	-
VAR(Tian et al., 2024)	2.0B	350	1.92	323.1	0.82	0.59	1.73	350.2	0.82	0.60
MAR(Li et al., 2024)	943M	800	2.35	227.8	0.79	0.62	1.55	303.7	0.81	0.62
Pixel Diffusion										
ADM (Dhariwal & Nichol, 2021)	554M	400	10.94	101.0	0.69	0.63	3.94	215.8	0.83	0.53
RIN (Jabri et al., 2022)	410M	480	3.42	182.0	-	-	-	-	-	-
PixelFlow (Chen et al., 2025c)	677M	320	-	-	-	-	1.98	282.1	0.81	0.60
PixNerd (Wang et al., 2025a)	700M	160	-	-	-	-	2.15	297.0	0.79	0.59
Latent Diffusion										
MaskDiT (Zheng et al., 2024)	675M	1600	5.69	177.9	0.74	0.60	2.28	276.6	0.80	0.61
DiT (Peebles & Xie, 2022)	675M	1400	9.62	121.5	0.67	0.67	2.27	278.2	0.83	0.57
SiT (Ma et al., 2024)	675M	1400	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
FasterDiT (Yao et al., 2024)	675M	400	7.91	131.3	0.67	0.69	2.03	264.0	0.81	0.60
MDTv2 (Gao et al., 2023)	675M	1080	-	-	-	-	1.58	314.7	0.79	0.65
REPA (Yu et al., 2025)	675M	800	5.90	157.8	0.70	0.69	1.42	305.7	0.80	0.65
VA-VAE (Yao et al., 2025)	675M	800	2.17	205.6	0.77	0.65	1.35	295.3	0.79	0.65
DDT (Wang et al., 2025b)	675M	400	6.27	154.7	0.68	0.69	1.26	310.6	0.79	0.65
SVG (Shi et al., 2025)	675M	1400	3.36	181.2	-	-	1.92	264.9	-	-
RAE (Zheng et al., 2025)	839M	800	1.60	242.7	-	-	1.28	262.9	-	-
Latent Diffusion with LV-RAE + 0.1 noise (Ours)										
DiT-XL	675M	400	3.77	185.4	0.76	0.60	2.61	208.7	0.76	0.64
+ XL decoder	675M	400	2.94	195.0	0.78	0.61	1.98	219.2	0.77	0.64
DiT^{DH}-XL	839M	800	2.42	223.8	0.77	0.64	1.82	249.7	0.75	0.67
+ XL decoder	839M	800	1.86	232.3	0.78	0.64	1.42	257.7	0.76	0.68

- **IS** evaluates both sample quality and diversity by measuring the confidence and entropy of class predictions.
- **Precision** and **Recall** explicitly disentangle fidelity and diversity: precision measures the quality of generated samples, while recall reflects coverage of the real data distribution.

G.4 Frechet Distance computed on top of DINOv2 features (FDD)

In addition to standard FID, we report FDD, a feature distribution distance computed using DINOv2 features. Specifically, we replace Inception features with class token extracted from a pretrained DINOv2 model[3] and compute the Fréchet distance in this semantic feature space.

Compared to Inception-based FID, DINO-based FID correlates better with human perceptual judgments(Stein et al., 2023), and thus provides a more appropriate measure for evaluating image distributions.

H Additional Quantitative Results and Discussion

H.1 Complete Generation Results

We provide the complete quantitative comparison of class-conditional image generation on ImageNet 256×256 in Tab. 6, covering autoregressive models, pixel-space diffusion models, and latent diffusion models under both unguided and guided settings. DiT-XL trained with LV-RAE reaches an unguided gFID of 3.77 and an IS of 185.4 after 400 epochs; scaling to DiT^{DH}-XL and training for 800 epochs improves these to 2.42 and 223.8. Substituting the XL decoder improves the unguided gFID of the two models further, to 2.94 and 1.86 respectively, without any change to the encoder, the latent representation, or the diffusion model. Although these results are not uniformly superior to all prior methods, they remain competitive with existing latent diffusion models of comparable scale.

[3] https://huggingface.co/facebook/dinov2-large

We note that LV-RAE is not intended to claim state-of-the-art generation performance across all metrics. Instead, our goal is to show that the proposed representation autoencoder can substantially improve reconstruction fidelity and semantic alignment while still maintaining strong generation quality.

H.2 Decoder-Side Reconstruction–Generation Trade-off

We further analyze the remaining gap in gFID.

First, we emphasize that most competing methods in Tab. 6 are built upon SD-VAE, whose reconstruction PSNR is around 26 and is therefore substantially lower than that of LV-RAE. This suggests that the comparison reflects not only generation quality, but also a broader trade-off between reconstruction fidelity and generation quality.

Our claim in this section is that the gap in gFID should not be interpreted solely as evidence of worse latent distribution modeling. Since gFID is computed after decoding generated latents into images, the decoder mediates between latent-space error and image-level statistics, and this mediation is itself a design axis.

H.2.1 Local Sensitivity of the Decoder

For a reference encoded latent z and a nearby point $z + \delta$, a first-order expansion of the decoder around z gives

$$D(z + \delta) \approx D(z) + J_D(z)\,\delta, \tag{16}$$

where $J_D(z)$ is the decoder Jacobian introduced in section 3.2. This expansion illustrates how decoder responses translate latent deviations into output changes: δ measures how far the point lies from the reference latent, whereas $J_D(z)$ determines how strongly such a deviation is expressed in image space. Two factors therefore influence image-level generation quality:

- the latent-space error δ left by the generative model;
- the decoder response J_D, which the noise augmentation of section 3.3 is designed to regularize.

Only the first is a property of the latent representation. The two comparisons below instead vary the decoder while holding the generated latents fixed.

H.2.2 Comparison 1: A Decoder Trained on Semantic Features

We first train a decoder directly on the frozen semantic feature u using reconstruction and GAN loss, i.e., $\min \mathcal{L}(X, D(u))$. We denote this decoder as the DINO decoder, which can be viewed as an RAE-style decoder. It adopts the same Base architecture as the LV-RAE decoder, but receives a different input latent. No noise perturbation is used for the DINO decoder during either training or inference, whereas the LV-RAE decoders use the settings reported in section 3.3. Its reconstruction performance is reported in Tab. 7.

As shown in Tab. 7, the DINO decoder exhibits substantially worse reconstruction quality than the LV-RAE decoder. Meanwhile, adding r, which can be regarded as a perturbation to u, does not severely degrade the decoding results of the DINO decoder, indicating that this decoder is comparatively insensitive to such perturbations.

We then decode the **same generated latents** with different decoders and report the generation results in Tab. 8. The results show that the choice of decoder alone can lead to a substantial difference in gFID. Using the DINO decoder improves unguided FID from 2.42 to 1.86 and guided FID from 1.82 to 1.41, despite its much worse reconstruction fidelity.

As further illustrated in Fig. 7, the images decoded from the same latent are structurally consistent across decoders, while the visible differences lie mainly in fine-grained texture.

H.2.3 Comparison 2: A Larger Decoder Configuration

The analysis above relates part of the remaining gFID gap to the limited expressiveness of our decoder. This suggests a concrete prediction: upgrading the decoder, while leaving the latent representation

Table 7: **Reconstruction performance of different decoders.** The DINO decoder trained on the semantic feature u achieves substantially worse reconstruction quality than LV-RAE, whereas the larger XL decoder further improves reconstruction fidelity.

Latent	Decoder	Dataset	PSNR↑	SSIM↑	LPIPS↓
u	DINO decoder	COCO17	20.24	0.586	0.129
$u + r$	DINO decoder	COCO17	19.83	0.564	0.142
$u + r$	LV-RAE decoder (Base)	COCO17	30.91	0.930	0.018
$u + r$	LV-RAE decoder (XL)	COCO17	32.07	0.939	0.016

Table 8: **Effect of the decoder configuration on generation performance.** We decode the same generated latents using different decoders. Both the DINO decoder and the larger XL decoder substantially improve gFID relative to the Base LV-RAE decoder, showing that image-level generation metrics depend strongly on the decoder rather than on the latent representation alone.

Method	Decoder	Epoch	gFID w/o guidance↓	IS w/o guidance↑	gFID w/ guidance↓	IS w/ guidance↑
DiT^{DH}-XL	LV-RAE decoder (Base)	800	2.42	223.8	1.82	249.7
DiT^{DH}-XL	LV-RAE decoder (XL)	800	1.86	232.3	1.42	257.7
DiT^{DH}-XL	DINO decoder	800	1.86	228.3	1.41	257.7



Figure 7: **Qualitative comparison of decoded samples.**

and the generative model untouched, should improve image-level generation quality. We test this directly.

We replace the default LV-RAE decoder (Base; 768 wide, 12 blocks, 12 heads) with a larger configuration (XL; 1152 wide, 28 blocks, 16 heads, matching the DiT-XL configuration), trained with the same objective on the same frozen encoder. The encoder, the latent representation, the trained diffusion model, and the sampling procedure are all held fixed, and we decode the **same generated latents** used in Comparison 1. As detailed in Appendix F.2, the two decoders differ in initialization and training schedule in addition to size, so this comparison evaluates a decoder-only upgrade rather than a training-matched ablation of capacity alone.

As reported in Tab. 8, the XL decoder improves the unguided gFID from 2.42 to 1.86 and the guided gFID from 1.82 to 1.42. Because both configurations decode the same generated latents, these improvements arise from changes on the decoder side rather than from changes in latent modeling. In terms of eq. (16), the error δ is held fixed while the decoder is varied.

In contrast to Comparison 1, the XL decoder does not pay for this gain with reconstruction quality. As shown in Tab. 7, its PSNR instead improves from 30.91 to 32.07.

Taken together, the decoder configurations achieve different combinations of reconstruction fidelity and generation quality, even when the underlying diffusion samples are fixed. In particular, similar gFID values can accompany substantially different reconstruction fidelity: the DINO decoder and the XL decoder both reach an unguided gFID of 1.86, yet differ by 12.2 dB in PSNR. These observations show that image-level generation metrics do not, by themselves, isolate the quality of latent distribution modeling.

H.3 Analysis of the Residual Latent Formulation

We further analyze the residual latent formulation used in LV-RAE. Given a semantic feature u extracted by a pretrained vision foundation model, our goal is to improve reconstruction fidelity while preserving the semantic structure of u. Since u contains rich semantic information but lacks sufficient low-level details for high-fidelity reconstruction, LV-RAE introduces an additional branch to compensate for the missing reconstruction-relevant information.

A straightforward design, such as SVG, represents the latent as two parallel components, (u, r), where the decoder receives u and r as separate latent streams. While this design increases reconstruction capacity, it may also allow r to behave as an independent latent, making the decoder less dependent on the original semantic feature u. In contrast, LV-RAE adopts a residual formulation and represents the final latent as $u + r$. In this form, r serves as a correction to u rather than an independent latent stream. The enhanced latent keeps the same dimensionality and representation form as u, which helps anchor it to the pretrained semantic feature space.

To examine this property, we conduct a linear-probing experiment on ImageNet. We first train a linear classifier using only the original semantic feature u. After training, we keep the classifier fixed and evaluate it with either u or $u + r$ as input. This directly tests whether a classifier trained on u can still operate on the enhanced latent $u + r$ without adaptation.

Table 9: **Linear-probing results on ImageNet.** The classifier is trained using only u and evaluated with either u or $u + r$ as input.

Latent	Acc@1
u	80.8
$u + r$	80.8

As shown in Tab. 9, the classifier achieves the same Acc@1 when evaluated on u and $u + r$. This indicates that the residual correction does not disrupt the semantic decision boundary learned from u. If r had significantly shifted the representation away from the original semantic space, the classifier trained only on u would likely suffer a performance drop on $u + r$. The unchanged accuracy therefore suggests that $u + r$ remains semantically compatible with u.

Overall, the residual latent formulation allows LV-RAE to supplement u with reconstruction-relevant details while preserving compatibility with the pretrained semantic representation. This differs from a parallel latent formulation such as (u, r), where the auxiliary component can be treated as an independent latent channel.

H.4 Training Dynamics

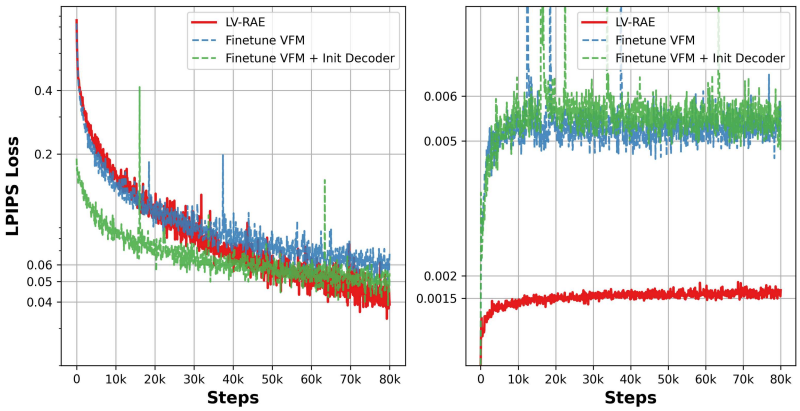


Figure 8: **Training dynamics comparison. Left:** LPIPS loss over training steps. **Right:** Semantic alignment loss over training steps. LV-RAE achieves lower and more stable alignment loss and converges to lower LPIPS loss than directly fine-tuning the VFM.

Fig. 8 compares the evolution of the training losses of LV-RAE and direct VFM fine-tuning. Fine-tuning the VFM continuously alters the semantic space against which the alignment loss is measured, making the alignment objective a moving target and leading to visibly unstable optimization. LV-RAE instead keeps the semantic space fixed and learns only the residual, so its alignment loss is both lower and markedly more stable throughout training, while the LPIPS loss converges to a lower value. This is consistent with the ablation in Tab. 1, where LV-RAE outperforms direct VFM fine-tuning in both reconstruction fidelity and semantic alignment.

I Additional Qualitative Results

I.1 PCA Visualizations

Fig. 9 shows the PCA (Hotelling, 1933) visualizations of the features extracted by LV-RAE and DINOv3.

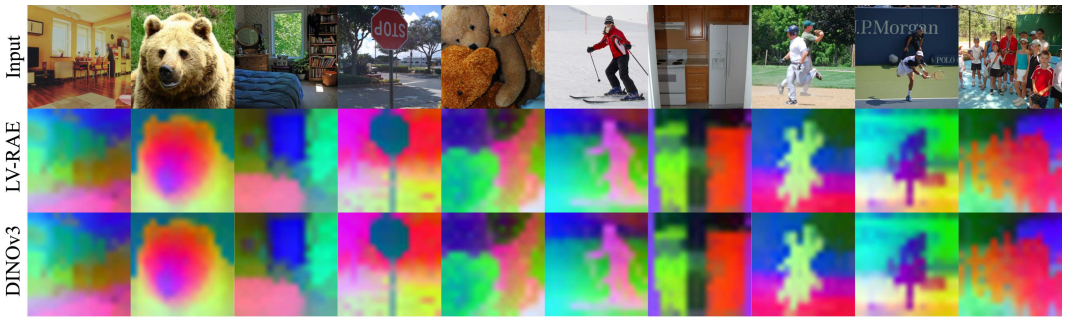


Figure 9: **PCA visualizations of LV-RAE latent and DINOv3 semantic features.** The two visualizations exhibit strong consistency, providing a qualitative view of the alignment quantified by CKNNA in Tab. 1 and by the fixed linear probe in Tab. 9.

I.2 Effect of the LV-RAE Encoder

To verify that the LV-RAE encoder output $\mathbf{r}$ captures the low-level information missing from the semantic features $\mathbf{u}$, we visualize reconstructions produced by the LV-RAE decoder under different latent inputs. As shown in figs. 10 to 12, when the encoder output $\mathbf{r}$ is not included, the reconstructed images exhibit noticeable color shifts and lack correct fine-grained textures. This observation indicates that the LV-RAE encoder effectively learns complementary low-level details that are critical for accurate image reconstruction.

I.3 Effect of Noise Level on Generation Results

To improve generation quality, we inject noise into the generated latent and provide visualizations illustrating the effect of different noise strengths, as shown in Figs. 13 and 14.



Figure 10: **Visualization of LV-RAE reconstructions under different latent inputs.** Excluding the encoder output $\mathbf{r}$ leads to color distortions and missing texture details.



Figure 11: **Visualization of LV-RAE reconstructions under different latent inputs.** Excluding the encoder output $\mathbf{r}$ leads to color distortions and missing texture details.



Figure 12: **Visualization of LV-RAE reconstructions under different latent inputs.** Excluding the encoder output $\mathbf{r}$ leads to color distortions and missing texture details.

I.4 Uncurated Generation Visual Results

We provide uncurated generation results for specific classes in figs. 15 to 26.



Figure 13: **Visualization of the effect of latent noise injection.**

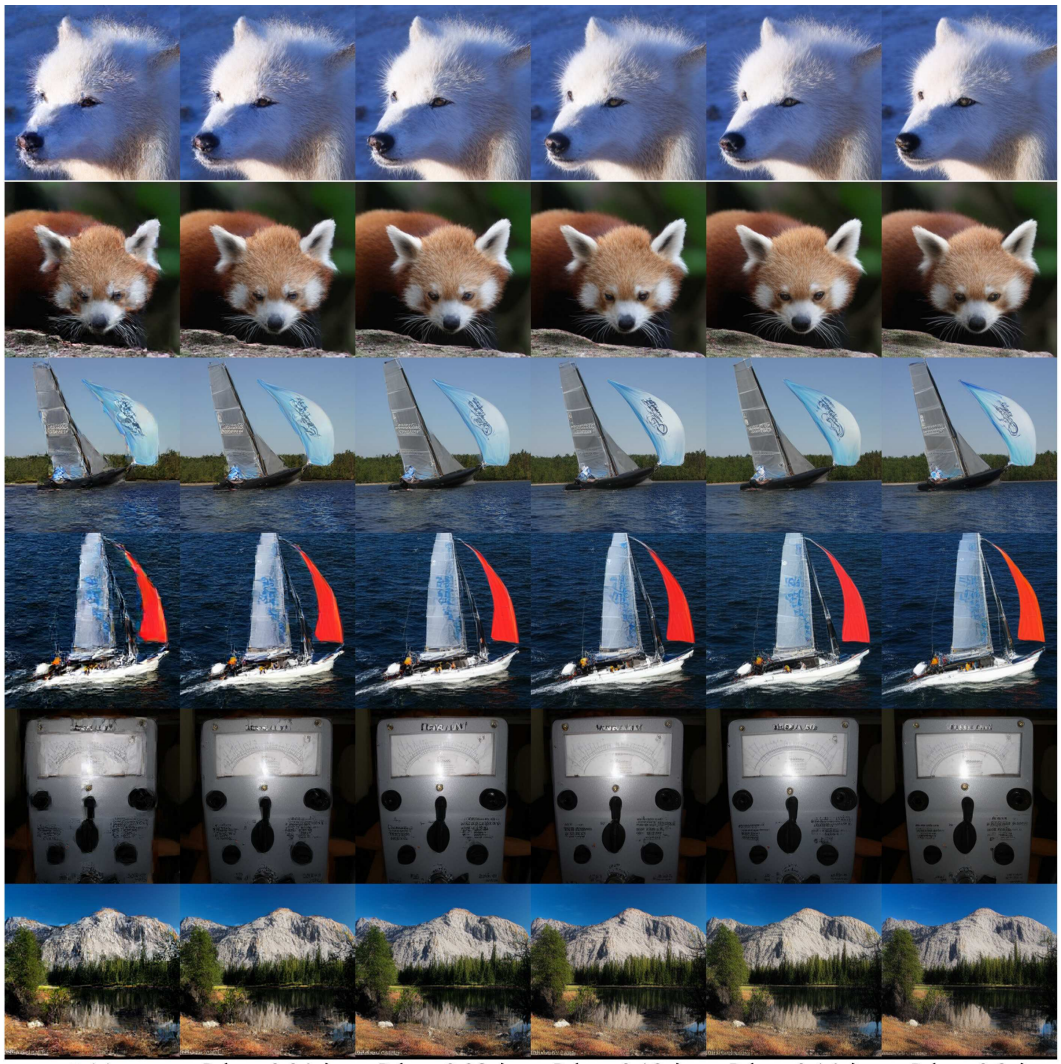


Figure 14: **Visualization of the effect of latent noise injection.**



Figure 15: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=15



Figure 16: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=20



Figure 17: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=88



Figure 18: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=146



Figure 19: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=207



Figure 20: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=250



Figure 21: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=270



Figure 22: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=484



Figure 23: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=688



Figure 24: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=974



Figure 25: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=979



Figure 26: **Uncurated** 256×256 **DiT^{DH}-XL samples.** AutoGuidance Scale=1.4, Class label=980