MAKING RECONSTRUCTION FID PREDICTIVE OF DIFFUSION GENERATION FID

Anonymous authors
Paper under double-blind review

ABSTRACT

It is well known that the reconstruction FID (rFID) of a VAE correlates poorly with the generation FID (gFID) of a latent diffusion model. We propose **interpolated FID (iFID)**, a simple variant of rFID that exhibits a strong correlation with gFID. Specifically, for each data point, we retrieve its nearest neighbor in latent space, interpolate between the two latent representations, decode the interpolated latent, and compute the FID between the decoded samples and the original dataset. We provide theoretical intuition and examples for why iFID correlates well with gFID, drawing connections to recent results on diffusion generalization. Empirically, we construct a VAE benchmark comprising 22 pre-trained VAEs and 15 VAE metrics, along with 44 diffusion models trained for class-conditional image generation on ImageNet. Using this benchmark, we show that iFID correlates strongly with gFID, achieving Pearson and Spearman correlation coefficients of approximately 0.85. The source code is provided in supplementary materials.

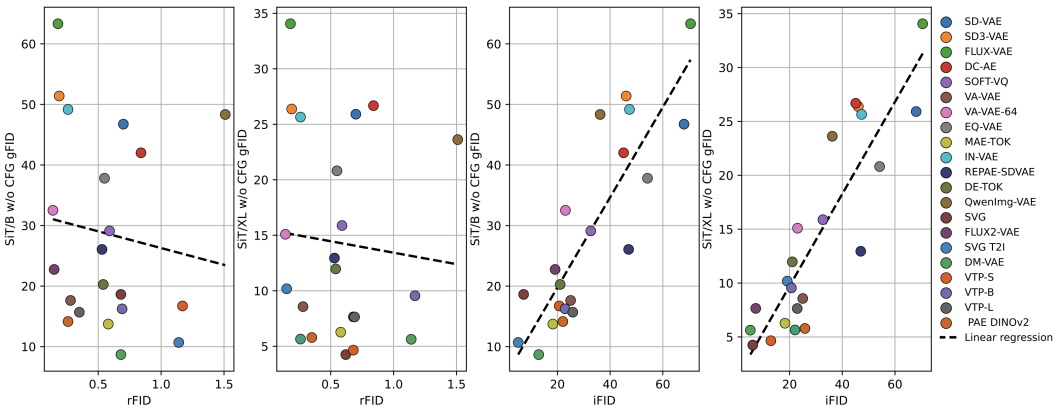

Figure 1: *Left two panels*: The rFID values of VAEs are uncorrelated with, or even negatively correlated with, the gFID values of diffusion models. *Right two panels*: The iFID metric exhibits a strong positive correlation with the gFID values of diffusion models.

1 INTRODUCTION

Variational autoencoders (VAEs) are foundational components of latent diffusion models (LDMs) (Rombach et al., 2022; Labs, 2024; Esser et al., 2024). In an LDM, a VAE first maps images into a latent space, where a diffusion model is subsequently trained to generate samples that are decoded back into pixel space. VAEs are typically optimized and evaluated in terms of reconstruction quality (Rombach et al., 2022; Esser et al., 2024), *e.g.*, reconstruction Fréchet Inception Distance (rFID) (Heusel et al., 2017). Intuitively, a VAE with better reconstruction quality should preserve more image details and therefore yield better generation performance. However, a phenomenon termed the "reconstruction–generation dilemma" has recently been observed in several studies (Yao & Wang, 2025; Kouzelis et al., 2025; Ye et al., 2025; Skorokhodov et al., 2025; Chen et al., 2025a): VAEs with excellent rFID can yield poor generation FID (gFID), whereas those with worse rFID can achieve better generation performance.

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.

In this work, we propose **interpolated FID (iFID)**, a simple variant of rFID that correlates strongly with gFID. Specifically, for each data point in the dataset, we identify its nearest neighbor in latent space. We then interpolate between their latent representations and decode the resulting latent into an image. Finally, we compute the FID between these decoded interpolated samples and the original dataset. Building on recent results on diffusion generalization, we provide an intuitive explanation for why iFID correlates well with diffusion sample quality. Empirically, we construct a benchmark comprising 22 VAEs and 15 VAE metrics, along with 44 diffusion models trained for class-conditional image generation on ImageNet. We find that iFID correlates strongly with diffusion gFID, achieving correlations of approximately 0.85.

In summary, our contributions are as follows:

- We propose interpolated FID (iFID), a simple variant of rFID based on nearest-neighbor interpolation in latent space. iFID is the first metric shown to strongly correlate with gFID across diverse diffusion VAEs, achieving high correlations across a wide range of VAEs.
- We provide an intuitive explanation for why iFID correlates well with generation performance by relating it to recent results on diffusion generalization.
- We construct a VAE benchmark comprising 22 VAE models and 15 VAE metrics, along with 44 diffusion models trained for class-conditional image generation on ImageNet. Using this benchmark, we show that iFID correlates strongly with diffusion gFID.

2 Preliminaries: Latent Diffusion Models

A latent diffusion model (Rombach et al., 2022) consists of a variational autoencoder (VAE) (Kingma & Welling, 2013), which projects images into a latent space, and a diffusion model (Ho et al., 2020), which generates samples in this latent space. Let the source image be $x \sim p(x)$ and its latent representation be $z \sim q(z|x)$, where $q(z|x)$ denotes the posterior distribution. The reconstructed image is given by $\hat{x} = g(z)$, where $g(\cdot)$ denotes the decoder. The diffusion model is trained in the latent space to approximate the latent distribution $p(z)$, and the generated latents are subsequently decoded using $g(\cdot)$. The forward process of the diffusion model constructs a Markov chain by incrementally adding noise over timesteps $t \in [0, T]$. More specifically, given diffusion parameters α_t and σ_t^2, the conditional forward kernel is

$$q(z_t|z) = \mathcal{N}(\alpha_t z, \sigma_t^2 I). \tag{1}$$

The reverse process of a diffusion model is described by a reverse-time stochastic differential equation (SDE) (Anderson, 1982). Sampling from the diffusion model involves simulating the reverse-time SDE from timestep T to timestep 0 (Song et al., 2020), which requires the score function $\nabla_{z_t} \log q(z_t)$. Accordingly, a diffusion model is commonly parameterized by a score estimator $s_\theta(z_t, t)$ that approximates $\nabla_{z_t} \log q(z_t)$ and is optimized by minimizing the denoising score-matching loss (Vincent, 2011):

$$\mathcal{L}\text{DSM} = \mathbb{E}[||s\theta(z_t, t) - \nabla_{z_t} \log q(z_t|z)||^2]. \tag{2}$$

3 Making Reconstruction FID Predictive of Generation FID

3.1 Reconstruction FID, Generation FID, and Interpolated FID

We denote the images in the validation dataset by $x^{1:N}$, their latent representations by $z^{1:N}$, the Fréchet Inception Distance (FID) between two sets of images by $d_{\text{FID}}(\cdot, \cdot)$, the decoder by $g(\cdot)$, the diffusion solver by $\Phi(\cdot, \cdot)$, and Gaussian noise by $\epsilon \sim \mathcal{N}(0, I)$. The reconstruction FID (rFID) and generation FID (gFID) are defined as

$$\text{rFID} = d_{\text{FID}}(x^{1:N}, g(z^{1:N})), \text{gFID} = d_{\text{FID}}(x^{1:N}, g(\Phi(\epsilon^{1:N}, T))). \tag{3}$$

It is well known that rFID does not correlate well with gFID (Kouzelis et al., 2025; Yao & Wang, 2025). This phenomenon is often referred to as the "reconstruction–generation trade-off". In this paper, we propose a simple variant of rFID that exhibits a strong correlation with gFID. Specifically,

rather than directly evaluating reconstructions from the original latents $z^{1:N}$, we linearly interpolate between each latent z_i and its nearest neighbor z^{i^*} in latent space:

$$\text{iFID} := d_{\text{FID}}(x^{1:N}, g(\hat{z}^{1:N})), \text{ where } \hat{z}^i = \frac{z^i + z^{i^*}}{2}, z^{i^*} := \arg\min_{j \neq i} ||z^j - z^i||. \tag{4}$$

We refer to this metric as **interpolated FID (iFID)** because it evaluates the quality of samples decoded from interpolated latent representations. As shown in Figure 1, iFID correlates strongly with gFID.

3.2 Why Interpolated FID Predicts Sample Quality

To understand why iFID is related to diffusion sample quality, we draw on a recent theoretical concept called the ridge set (He et al., 2026), which describes the manifold around which diffusion samples concentrate. More specifically, we show that the local ridge set lies close to the line connecting nearby training samples.

Intuition: diffusion models generalize through score smoothing and mode interpolation. First, we revisit the training objective of diffusion models. Specifically, for a fixed training dataset $z^{(1:M)}$, the score-matching objective admits a closed-form optimum known as the empirical score.

Lemma 1 (Denoising score matching has an optimal solution (Song et al., 2025))**.** *Given the forward diffusion process $z_t = \alpha_t z + \sigma_t \epsilon$ and a training set $\{z^{(1)}, \ldots, z^{(M)}\}$, denoising score matching in Eq. 2 has the following empirical optimum:*

$$s_*(z_t, t) = \frac{1}{\sigma_t^2}(-z_t + \alpha_t \sum_{i=1}^{M} w^i z^i), \text{ where } w^i = \frac{\exp\left(-||z_t - \alpha_t z^i||^2 / 2\sigma_t^2\right)}{\sum_{j=1}^{M} \exp\left(-||z_t - \alpha_t z^j||^2 / 2\sigma_t^2\right)}. \tag{5}$$

It is well established that when the score estimator $s_\theta(z_t, t)$ matches the empirical score $s_*(z_t, t)$, a diffusion model tends to replicate samples from the training dataset (Kamb & Ganguli, 2024; Bonnaire et al., 2025; Buchanan et al., 2025). Prior work suggests that diffusion models generate unseen samples because the learned score is a smoothed version of the empirical score (Kadkhodaie et al., 2023; Somepalli et al., 2022; Yoon et al., 2023; Buchanan et al., 2025; Song et al., 2025).

How does this score-smoothing effect influence generated samples? Prior work suggests that score smoothing leads to mode interpolation, meaning that generated samples can be interpreted as interpolations or compositions of training images (Okawa et al., 2023; Kamb & Ganguli, 2024; Somepalli et al., 2022; Aithal et al., 2024; Deschenaux et al., 2024; Chandran.C et al., 2025).

Theory: diffusion samples converge to the ridge-set manifold. To formalize this intuition, we provide a theoretical perspective on why iFID predicts sample quality by connecting it to ridge-set theory (He et al., 2026), which characterizes how the learned score deviates from the empirical score and how this deviation affects generated samples. Ridge-set theory shows that, for random feature neural network, the Jacobian of the learned score estimator $s_\theta(z_t, t)$ is a low-rank approximation to that of the empirical score $s_*(z_t, t)$. This discrepancy prevents diffusion models from replicating the training set and instead causes their samples to concentrate around a manifold known as the ridge set.

Lemma 2 (Diffusion samples concentrate around the ridge set determined by the Jacobian (He et al., 2026))**.** *Diffusion samples z_t concentrate around the ridge set*

$$\mathcal{R}(\beta, t) = \{z_t | V(z_t, t, \beta) V(z_t, t, \beta)^T s_*(z_t, t) = 0\}, \tag{6}$$

where β is a hyperparameter and $V(z_t, t, \beta)$ is the matrix whose columns are the $d - d^*$ eigenvectors of the Jacobian J corresponding to the smallest $d - d^*$ eigenvalues $-\beta \geq \lambda_{d^*+1} \geq \cdots \geq \lambda_d$. The Jacobian is defined as

$$J := \nabla s_*(z_t, t) = -\frac{1}{\sigma_t^2} I + \frac{\alpha_t^2}{\sigma_t^4} \sum_{i=1}^{M} w^i (z^i - \mu)(z^i - \mu)^T, \text{where } \mu = \sum_{j=1}^{M} w^j z^j. \tag{7}$$

Theory: The local ridge set is close to the line connecting nearby training samples. How is the ridge-set manifold related to interpolation between training samples? In fact, we can show that,

under mild assumptions, the local ridge set lies close to the line connecting nearby training samples. In other words, within certain local regions, samples generated by a diffusion model lie close to the affine line connecting nearby training samples. Because iFID measures the quality of samples decoded from the midpoints of nearby training samples, it correlates strongly with the sample quality of diffusion models.

Theorem 3 (Local ridge set is close to the scaled line connecting nearby training samples)**.** *Assume that, in a local region $\mathcal{U}$, two nearby training points z^i, z^j dominate the other samples in terms of their weights w^k, i.e., for every $z_t \in \mathcal{U}$, $\epsilon(z_t) = \sum_{k \neq i,j} w^k \leq \epsilon$. Let $L_t = \{\alpha_t(z^i + a(z^j - z^i)) : a \in \mathbb{R}\}$ be the scaled affine line through the pair. Assume that $\tilde{w}^i = \frac{w^i}{w^i + w^j}, \tilde{w}^j = \frac{w^j}{w^i + w^j}$ are non-negligible such that $b = \tilde{w}^i \tilde{w}^j \|z^i - z^j\|^2 > \gamma$, and $|z^k|, |z_t| \leq R$. Then, for every ridge point $z_t \in \mathcal{R}$ in the local region $\mathcal{U}$,*

$$\operatorname{dist}(z_t, L_t) \leq \frac{24R^3\epsilon}{\gamma} + 2R\epsilon.$$

For time t close to clean samples with $\alpha_t \approx 1$, the scaled line in Theorem 3 becomes the line connecting nearby training samples.

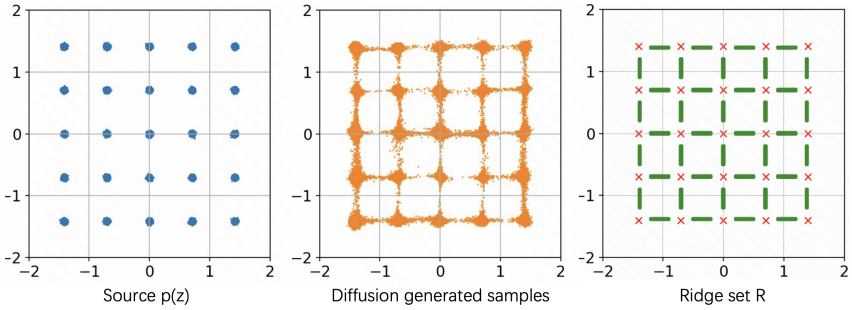

Figure 2: A 2D Gaussian mixture example. The samples generated by the diffusion model lie close to the ridge set $\mathcal{R}$ and the lines connecting nearby modes of the training distribution $p(z)$.

Example: 2D Gaussian mixture. To better illustrate the intuition and theory described above, we provide a toy example in Figure 2. We consider a 25-component Gaussian mixture arranged on a two-dimensional square grid. The samples generated by the diffusion model lie close to the ridge set $\mathcal{R}$. Moreover, the samples are close to the lines connecting nearby modes of the training distribution.

4 Diffusion VAE Benchmark

4.1 Benchmark Setup

Dataset and metrics. We conduct experiments on ImageNet at 256×256 resolution. All diffusion models are trained on the ImageNet training split, and all metrics are evaluated on the ImageNet validation split. To evaluate diffusion sample quality, we include gFID, IS, precision (prec.), and recall (rec.), which are standard metrics for evaluating diffusion models on ImageNet (Dhariwal & Nichol, 2021). We include 15 VAE metrics, as detailed in Table 1. For reconstruction metrics, we include PSNR, LPIPS, SSIM, and rFID, which are standard metrics for evaluating VAEs. For non-reconstruction VAE metrics, we include metrics proposed in recent work, including EQ Loss (Kouzelis et al., 2025), SE Loss (Skorokhodov et al., 2025), GMM Loss (Chen et al., 2025a), VIV, SEC (Zhong et al., 2026), LDS, CDS (Singh et al., 2025), VF Loss (Yao et al., 2025), Acc@1, Acc@5 (Yao et al., 2025), and our proposed iFID. We divide the VAE metrics into several categories, including reconstruction, equivariance, data rank, spatial smoothness, and semantic alignment. To quantify how well each metric predicts diffusion sample quality, we report the Pearson correlation coefficient (PCC), Spearman rank correlation coefficient (SRCC), and Kendall rank correlation coefficient (KRCC) between each VAE metric and each diffusion sample quality metric.

VAEs and diffusion models. We evaluate 22 representative VAEs with publicly available checkpoints; details are provided in Table 1. These VAEs span diverse latent dimensions, architectures, and training

Table 1: Our benchmark includes 22 VAEs with publicly available checkpoints and 15 VAE evaluation metrics. The VAEs vary in architecture, latent dimensionality, and alignment method.

VAE Models			
Name	Arch.	Latent Dim.	Align.
SD-VAE (Rombach et al., 2022)	UNet	4×32^2	-
SD3-VAE (Esser et al., 2024)	UNet	16×32^2	-
FLUX-VAE (Labs, 2024)	UNet	16×32^2	-
DC-AE (Chen et al., 2024)	UNet	32×8^2	-
SOFT-VQ (Chen et al., 2025b)	UNet	64×32	-
VA-VAE (Yao & Wang, 2025)	UNet	32×16^2	DINO
VA-VAE-64 (Yao & Wang, 2025)	UNet	64×16^2	DINO
EQ-VAE (Kouzelis et al., 2025)	UNet	4×32^2	EQ Loss
MAE-TOK (Chen et al., 2025a)	ViT	128×32	MAE Loss
IN-VAE (Leng et al., 2025)	UNet	32×16^2	-
REPAE (Leng et al., 2025)	UNet	4×32^2	REPA
DE-TOK (Yang et al., 2025)	UNet	128×32	MAE+Denoising
QW-VAE (Wu et al., 2025a)	UNet	16×32^2	-
RAE (Zheng et al., 2025)	ViT	768×16^2	DINO Encoder
SVG (Shi et al., 2025)	ViT	392×16^2	DINO Encoder
FLUX2-VAE (Labs, 2024)	ViT	128×16^2	REPA
SVG-T2I (Shi et al., 2025)	ViT	384×16^2	DINO Encoder
DM-VAE (Ye et al., 2025)	ViT	256×32	DINO
VTP-S (Yao et al., 2025)	ViT	64×16^2	DINO Loss
VTP-B (Yao et al., 2025)	ViT	64×16^2	DINO Loss
VTP-L (Yao et al., 2025)	ViT	64×16^2	DINO Loss
PAE (Yue et al., 2026)	UNet	32×16^2	DINO

VAE Metrics	
Name	Type
PSNR	Recon.
LPIPS (Zhang et al., 2018)	Recon.
SSIM (Wang et al., 2004)	Recon.
rFID (Heusel et al., 2017)	Recon.
EQ Loss (Kouzelis et al., 2025)	Equivar.
SE Loss (Skorokhodov et al., 2025)	Equivar.
GMM Loss (Chen et al., 2025a)	Data Rank
VIV (Zhong et al., 2026)	Data Rank
LDS (Singh et al., 2025)	Smoothness
CDS (Singh et al., 2025)	Smoothness
SEC (Zhong et al., 2026)	Smoothness
VF Loss (Yao & Wang, 2025)	Semantic
Acc@1 (Yao et al., 2025)	Semantic
Acc@5 (Yao et al., 2025)	Semantic
iFID	Proposed

Table 2: Diffusion sample quality metrics for 22 VAEs evaluated on the ImageNet 256×256 class-conditional image generation task. Inapplicable values are denoted by "–."

	SiT/B w/o CFG				SiT/B w/ CFG				SiT/XL w/o CFG				SiT/XL w/ CFG			
VAE	gFID ↓	IS ↑	Prec. ↑	Rec. ↑	gFID ↓	IS ↑	Prec. ↑	Rec. ↑	gFID ↓	IS ↑	Prec. ↑	Rec. ↑	gFID ↓	IS ↑	Prec. ↑	Rec. ↑
SD-VAE	46.74	30.43	0.47	0.63	9.91	162.90	0.84	0.32	25.91	53.90	0.60	0.62	6.33	178.64	0.84	0.43
SD3-VAE	51.39	27.61	0.43	0.60	11.89	151.90	0.82	0.30	26.38	50.39	0.59	0.62	7.16	189.95	0.87	0.36
FLUX-VAE	63.32	22.22	0.37	0.57	14.38	134.72	0.79	0.26	34.06	41.57	0.54	0.63	9.51	160.82	0.84	0.38
DC-AE	42.01	32.79	0.49	0.62	9.03	189.42	0.85	0.33	26.68	50.91	0.58	0.64	6.08	179.91	0.83	0.45
SOFT-VQ	29.11	47.19	0.58	0.59	7.48	217.76	0.83	0.36	15.88	76.09	0.67	0.60	5.14	240.19	0.86	0.44
VA-VAE	17.63	64.56	0.66	0.57	6.01	208.61	0.84	0.42	8.57	100.79	0.73	0.57	4.20	238.73	0.86	0.45
VA-VAE-64	32.51	39.13	0.56	0.59	7.76	208.07	0.84	0.32	15.09	73.06	0.68	0.58	4.95	183.03	0.83	0.47
EQ-VAE	37.81	37.24	0.54	0.62	9.32	177.31	0.88	0.32	20.81	63.14	0.65	0.60	6.24	201.39	0.88	0.39
MAE-TOK	13.73	92.70	0.67	0.58	5.69	237.12	0.82	0.45	6.27	132.73	0.75	0.60	3.74	199.88	0.81	0.54
IN-VAE	49.17	28.79	0.46	0.61	10.87	160.47	0.84	0.30	25.65	52.42	0.60	0.62	6.56	167.17	0.84	0.44
REPAE	26.06	52.01	0.61	0.61	6.46	176.37	0.85	0.43	12.95	85.11	0.70	0.60	4.17	211.95	0.86	0.47
DE-TOK	20.27	62.09	0.66	0.60	6.95	197.12	0.88	0.40	11.97	89.24	0.73	0.59	4.83	207.96	0.88	0.45
QW-VAE	48.34	29.10	0.44	0.59	10.97	148.86	0.82	0.30	23.62	55.86	0.61	0.61	7.25	180.18	0.85	0.41
RAE	–	–	–	–	–	–	–	–	4.25	192.71	0.85	0.51	3.50	256.41	0.88	0.45
SVG	18.63	78.27	0.68	0.51	8.66	233.68	0.81	0.33	7.65	127.75	0.76	0.56	4.79	247.56	0.85	0.46
FLUX2-VAE	22.75	57.02	0.62	0.62	5.92	185.76	0.84	0.45	10.18	96.68	0.72	0.61	4.10	229.93	0.87	0.49
SVG T2I	10.69	102.98	0.78	0.52	6.44	155.28	0.84	0.46	5.63	143.94	0.82	0.54	4.37	216.93	0.87	0.48
DM-VAE	8.69	106.01	0.76	0.57	4.89	174.05	0.84	0.50	4.65	20.00	0.80	0.56	3.39	220.37	0.87	0.51
VTP-S	16.72	69.04	0.70	0.58	6.07	213.12	0.87	0.43	9.55	98.19	0.75	0.59	4.58	223.93	0.87	0.47
VTP-B	16.24	76.37	0.70	0.59	5.56	220.22	0.85	0.44	7.64	119.41	0.76	0.59	4.02	245.77	0.86	0.48
VTP-L	15.69	84.72	0.69	0.57	5.50	232.28	0.85	0.43	5.79	140.61	0.78	0.59	3.48	211.17	0.83	0.53
PAE	14.15	86.02	0.69	0.61	4.58	250.76	0.86	0.46	5.65	135.82	0.77	0.60	2.85	214.77	0.83	0.56

objectives. Some, such as SD-VAE (Rombach et al., 2022), SD3-VAE (Esser et al., 2024), FLUX-VAE (Labs, 2024), DC-AE (Chen et al., 2024), IN-VAE (Leng et al., 2025), and QW-VAE (Wu et al., 2025a), are optimized solely for reconstruction. Other VAEs are optimized for diffusion modeling, including VA-VAE (Yao & Wang, 2025), SOFT-VQ (Chen et al., 2025b), MAE-TOK (Chen et al., 2025a), REPAE (Leng et al., 2025), DM-VAE (Yang et al., 2025), RAE (Zheng et al., 2025), SVG (Shi et al., 2025), FLUX2-VAE (Labs, 2024), DE-TOK (Ye et al., 2025), VTP (Yao et al., 2025), and PAE (Yue et al., 2026). For each VAE, we train two diffusion models of different sizes, SiT-B and SiT-XL, in the corresponding latent space. In total, we train 44 diffusion models. All diffusion models are trained for 40 epochs using the Adam optimizer. We report sample quality both with and without classifier-free guidance (CFG) (Ho & Salimans, 2022).

Table 3: Evaluation metrics for 22 VAEs. Inapplicable values are denoted by "–."

	Reconstruction				Equivariance		Data Rank		Spatial Smoothness			Semantic Alignment			Proposed	
VAE	PSNR ↑	LPIPS ↓	SSIM ↑	rFID ↓	EQ ↓	SE ↓	GMM ↓	VIV ↓	LDS ↓	CDS ↓	SEC ↓	VF ↓	Acc@1 ↑	Acc@5 ↑	iFID ↓	iIS ↑
SD-VAE	26.91	0.06	0.76	0.70	0.67	0.28	10.67	2.80	0.10	0.03	0.21	0.64	0.34	1.65	68.05	23.41
SD3-VAE	31.28	0.02	0.89	0.19	0.63	0.18	8.79	1.02	0.08	0.04	0.09	0.36	0.42	1.63	46.10	40.38
FLUX-VAE	32.87	0.01	0.92	0.18	0.69	0.17	18.06	1.10	0.12	0.04	0.13	0.41	1.24	4.25	70.52	21.46
DC-AE	24.55	0.08	0.65	0.84	0.89	0.35	11.50	1.36	0.12	0.03	0.24	0.35	2.31	7.72	45.12	28.33
SOFT-VQ	23.92	0.09	0.66	0.59	1.16	0.25	-25.73	–	–	0.09	0.11	0.18	0.14	0.68	32.66	43.00
VA-VAE	27.71	0.04	0.80	0.28	1.28	0.29	-16.20	0.57	0.29	0.09	0.05	0.08	14.33	34.14	25.01	55.45
VA-VAE-64	30.72	0.02	0.88	0.14	0.91	0.20	-6.21	0.60	0.27	0.08	0.03	0.09	14.58	35.12	23.04	51.53
EQ-VAE	26.16	0.06	0.74	0.55	0.32	0.22	12.52	2.13	0.27	0.06	0.01	0.59	0.56	2.33	54.18	27.02
MAE-TOK	24.57	0.09	0.68	0.58	0.27	0.26	-11.46	–	–	0.00	0.44	0.33	4.51	14.67	18.26	67.39
IN-VAE	28.92	0.03	0.83	0.26	0.78	0.23	9.02	1.22	0.10	0.04	0.20	0.40	2.78	8.23	47.35	31.87
REPAE	26.10	0.06	0.74	0.53	0.91	0.32	34.72	2.90	0.11	0.06	0.10	0.65	0.40	1.62	47.02	30.53
DE-TOK	25.24	0.06	0.71	0.54	0.67	0.19	15.11	–	–	0.01	0.24	0.46	3.19	10.17	21.05	46.18
QW-VAE	32.19	0.04	0.90	1.51	0.60	0.15	12.69	0.90	0.12	0.05	0.05	0.36	–	–	36.22	42.35
RAE	19.21	0.15	0.50	0.62	0.41	0.39	-36.55	0.78	0.11	0.04	0.11	0.01	75.45	93.63	6.05	142.81
SVG	23.35	0.10	0.65	0.68	0.78	0.32	-8.92	0.50	0.10	0.04	0.04	0.06	49.73	78.51	7.17	117.78
FLUX2-VAE	34.33	0.01	0.94	0.15	0.78	0.17	70.53	1.30	0.10	0.04	0.24	0.34	5.01	14.22	19.13	55.32
SVG T2I	18.90	0.16	0.49	1.14	0.74	0.44	-12.57	0.50	0.10	0.04	0.04	0.06	49.94	78.84	5.12	129.45
DM-VAE	22.46	0.11	0.63	0.68	0.66	0.29	19.34	–	–	0.05	0.18	0.22	15.70	32.66	12.91	77.00
VTP-S	24.17	0.10	0.67	1.17	0.78	0.37	-5.12	1.15	0.07	0.04	0.19	0.21	0.22	1.75	20.70	50.82
VTP-B	26.32	0.07	0.75	0.69	0.78	0.30	-0.31	1.15	0.06	0.04	0.20	0.22	0.30	1.60	22.92	52.41
VTP-L	26.85	0.05	0.77	0.35	0.80	0.26	7.70	1.11	0.06	0.03	0.21	0.22	0.92	2.74	25.85	54.02
PAE	27.28	0.05	0.79	0.26	0.93	0.28	-16.19	1.13	0.02	0.03	0.18	0.20	9.16	24.71	22.07	55.03

Table 4: Correlations between 15 VAE metrics and diffusion gFID. The correlations are computed using metrics obtained from SiT models trained with 22 pre-trained VAEs. EQ Loss and SE Loss are not evaluated for VAEs with 1D latents, and RAE is not evaluated with SiT-B. PCC denotes the Pearson correlation coefficient, SRCC denotes Spearman rank correlation coefficient, and KRCC denotes Kendall rank correlation coefficient. **Bold**: best.

Metrics	gFID SiT/B						gFID SiT/XL					
	w/o CFG			w/ CFG			w/o CFG			w/ CFG		
	PCC↑	SRCC↑	KRCC↑	PCC↑	SRCC↑	KRCC↑	PCC↑	SRCC↑	KRCC↑	PCC↑	SRCC↑	KRCC↑
Reconstruction												
-PSNR	-0.60	-0.58	-0.42	-0.47	-0.33	-0.24	-0.54	-0.54	-0.39	-0.51	-0.39	-0.31
LPIPS	-0.58	-0.60	-0.47	-0.43	-0.35	-0.27	-0.54	-0.56	-0.42	-0.46	-0.40	-0.32
-SSIM	-0.57	-0.55	-0.39	-0.45	-0.30	-0.21	-0.52	-0.50	-0.36	-0.48	-0.37	-0.28
rFID	-0.12	-0.24	-0.16	-0.05	-0.02	0.00	-0.08	-0.17	-0.11	0.01	0.00	0.01
Equivariance												
EQ Loss	-0.14	-0.12	-0.08	-0.22	-0.30	-0.22	-0.05	0.04	0.02	-0.13	-0.19	-0.13
SE Loss	-0.53	-0.55	-0.38	-0.46	-0.41	-0.27	-0.49	-0.52	-0.35	-0.49	-0.48	-0.32
Data Rank												
GMM Loss	0.23	0.36	0.22	0.13	0.25	0.15	0.29	0.43	0.30	0.22	0.28	0.20
VIV	0.24	0.22	0.20	0.06	0.02	0.04	0.32	0.41	0.32	0.12	0.04	0.05
Spatial Smoothness												
LDS	0.13	0.52	0.39	0.12	0.49	0.39	0.14	0.43	0.30	0.16	0.47	0.34
CDS	0.04	0.15	0.10	-0.03	0.07	0.03	0.01	0.01	0.00	-0.01	0.06	0.04
SEC	-0.20	-0.13	-0.10	-0.29	-0.29	-0.18	-0.13	-0.01	0.34	-0.24	-0.24	-0.18
Semantic Alignment												
VF Loss	0.51	0.57	0.37	0.38	0.46	0.28	0.58	0.65	0.47	0.44	0.47	0.33
-Acc@1	0.39	0.40	0.28	0.14	0.23	0.15	0.45	0.54	0.40	0.31	0.37	0.25
-Acc@5	0.43	0.42	0.29	0.20	0.23	0.16	0.49	0.54	0.41	0.34	0.35	0.25
Proposed												
iFID	**0.85**	**0.82**	**0.62**	**0.73**	**0.66**	**0.47**	**0.88**	**0.84**	**0.65**	**0.77**	**0.68**	**0.51**

4.2 Main Findings

The "reconstruction–generation dilemma" is observed only for certain reconstruction metrics. As shown in Table 4, image-to-image distortion metrics, including PSNR, LPIPS, and SSIM, are strongly negatively correlated with gFID. This observation supports the "reconstruction–generation dilemma" reported by Yao & Wang (2025): better reconstruction quality does not necessarily translate into better diffusion generation quality. However, in our benchmark of diverse VAEs, the distribution-level metric rFID appears to be uncorrelated with generation performance, as measured by gFID.

Existing metrics do not consistently perform well across our benchmark of diverse VAEs. Among the non-reconstruction metrics designed to predict diffusion sample quality, many show weak correlations in our diverse VAE setting. LDS, VIV (Zhong et al., 2026), GMM Loss (Chen et al.,

Table 5: Pearson correlations (PCCs) between metrics computed on generated samples and those computed on interpolated samples. All sample-quality-oriented metrics correlate strongly with their interpolated counterparts.

PCC w/ Metric	SiT/B w/o CFG				SiT/XL w/o CFG			
	-gFID↑	IS↑	Rrecision↑	Recall↑	-gFID↑	IS↑	Prec.↑	Rec.↑
Diffusion Sample Metrics								
-gFID	1.00	0.92	0.98	-0.40	1.00	0.72	0.97	-0.74
IS	-	1.00	0.94	-0.55	-	1.00	0.85	-0.64
Precision	-	-	1.00	-0.47	-	-	1.00	-0.83
Recall	-	-	-	1.00	-	-	-	1.00
Interpolated Metrics								
-iFID	**0.85**	0.61	0.75	-0.58	**0.88**	0.65	0.78	-0.76
iIS	0.65	**0.75**	0.70	-0.84	0.67	**0.70**	0.77	-0.89
iPrecision	0.80	0.63	**0.82**	-0.72	0.83	0.67	**0.87**	-0.86
iRecall	-0.51	-0.45	-0.47	**0.08**	-0.46	-0.43	-0.38	**0.08**

2025a), VF Loss (Yao et al., 2025), and ImageNet classification accuracy (Yao et al., 2025) retain moderate predictive correlations with gFID.

iFID exhibits a significantly stronger correlation with gFID. On the other hand, iFID is strongly correlated with gFID, achieving a Pearson correlation of of at least 0.85 without CFG. Unlike other metrics, iFID accounts for both the encoder and decoder of the VAE. This may explain why it stands out among non-reconstruction metrics when evaluated on diverse VAEs trained on different datasets.

Extension of iFID to IS, precision, and recall. The key idea of iFID is to compute metrics on samples decoded from interpolated latents $g(\hat{z}^i)$. Therefore, the interpolation-based approach is not limited to FID. In fact, other diffusion sample metrics, such as IS, precision, and recall, can also be computed on $g(\hat{z}^i)$. We refer to these metrics computed on interpolated images as iIS, iPrecision, and iRecall (interpolated IS, interpolated precision, and interpolated recall, respectively). In Table 5, we show the correlations between these interpolated metrics and diffusion sample quality metrics. iFID, iIS, and iPrecision. are highly correlated with the gFID, IS, and precision of diffusion samples, respectively. In contrast, iRecall is not correlated with Recall. This is because Recall primarily measures sample coverage rather than sample quality. Higher Recall often corresponds to worse sample quality (Ho & Salimans, 2022), and Recall is also strongly negatively correlated with the other generation metrics.

gFID with CFG is significantly harder to predict. Moreover, as shown in Table 4, the absolute values of all correlations with CFG are noticeably smaller than those without CFG. This indicates that both the "reconstruction–generation trade-off" and the ability of VAE metrics to predict gFID are weakened by the presence of CFG. This may be because CFG reduces the differences in gFID across latent spaces, making gFID significantly harder to predict.

4.3 Ablation Study and Analysis

Interpolation method and strength. In Table 6, we examine the effects of different implementation choices on the performance of iFID. First, we show that the choice of interpolation method between a latent z^i and its nearest neighbor z^{i^*} is important. Although linear interpolation is the most straightforward approach and is consistent with our theoretical results, it is outperformed by SLERP interpolation (Jang et al., 2024). This may be because Gaussian VAEs have a unit Gaussian prior, whose geometry is better preserved by SLERP interpolation, which is beneficial to diffusion training (Meral et al., 2026a).

By default, we set the interpolation strength of iFID to $\alpha = 0.5$, which corresponds to midpoint interpolation. When $\alpha = 0$, the interpolated latent is identical to the source latent z, and iFID reduces to rFID. Conversely, when $\alpha = 1$, the interpolated latent is identical to $\mathrm{NN}(z)$. As shown in Table 6, the correlation between iFID and gFID increases rapidly as α increases from 0.1 to 0.5.

Reference dataset size, runtime, and memory. The number of images used for nearest-neighbor search may affect the performance, runtime, and memory requirements of iFID. In Table 6, we show that increasing the reference dataset size from 50k to 200k improves the performance of iFID

by a large margin. However, further increasing the dataset size to 1M does not yield additional improvements. Moreover, using a reference set of 1M images doubles the GPU memory requirement. Therefore, we use a reference dataset of 200k images to balance performance and memory usage when computing iFID.

Robustness across condition types, models, datasets, and training durations. In Table 7, we show that iFID is robust across different condition types, models architectures, datasets, and training durations. Specifically, to test whether iFID generalizes to text-to-image generation, we directly use the text-to-image results from Diffusion-Bench (Leng et al., 2026). iFID maintains consistently high correlations with text-to-image evaluation metrics, including GenEval (Ghosh et al., 2023), DPG Bench (Hu et al., 2024) and GenAI Bench (Li et al., 2024). We further find that iFID generalizes across generation settings, model architectures, training budgets, and datasets: it remains effective for unconditional generation on FFHQ, for both SiT and U-Net architectures, and across training durations ranging from 20 to 160 epochs.

Visualization of nearest-neighbor and interpolated latents. Figure 3 visualizes the decoded latent z, its nearest-neighbor latent $\mathrm{NN}(z)$, and the interpolated latent $\hat{z}$ for different VAEs. For reconstruction-optimized VAEs, such as SD-VAE and FLUX-VAE (Rombach et al., 2022; Labs, 2024), $\mathrm{NN}(z)$ is often semantically unrelated to z, causing the decoded interpolated latent $\hat{z}$ to yield invalid images. In contrast, for diffusion-optimized VAEs, such as VA-VAE and RAE (Yao & Wang, 2025; Zheng et al., 2025), $\mathrm{NN}(z)$ is semantically similar to z, and the decoded interpolated latent $\hat{z}$ yields realistic images.

Table 6: Ablation studies on latent interpolation method (Interp.), interpolation strength (Interp. Strength), and number of images (# of Images) used for nearest-neighbor search. iFID is robust to these choices and consistently exhibits a strong correlation with gFID. Runtime and GPU memory usage are measured using SD3-VAE on 8 NVIDIA A800 GPUs.

Interp.	Interp. Strength	# of Images	gFID SiT/B w/o CFG			gFID SiT/XL w/o CFG			Cost (SD3-VAE 8×A800)	
			PCC↑	SRCC↑	KRCC↑	PCC↑	SRCC↑	KRCC↑	Time (min)↓	Mem. (GB)↓
Linear	0.1	200k	0.14	0.07	0.03	0.23	0.19	0.13	46.05	132.24
Linear	0.3	200k	0.33	0.34	0.24	0.60	0.65	0.51		
Linear	0.5	200k	0.69	0.74	0.55	0.75	0.79	0.61		
SLERP	0.1	200k	-0.14	-0.11	-0.12	0.14	0.15	0.11		
SLERP	0.3	200k	0.25	0.35	0.27	0.61	0.65	0.53		
SLERP	0.5	200k	0.85	0.82	0.62	0.88	0.84	0.65		
SLERP	0.5	50k	0.79	0.74	0.54	0.82	0.81	0.64	46.18	120.08
SLERP	0.5	1M	0.85	0.80	0.58	0.88	0.85	0.67	46.62	224.96

5 Related Work

Variational autoencoders for diffusion models. Early latent diffusion models typically employ VAEs trained solely for reconstruction (Rombach et al., 2022; Esser et al., 2024). Subsequent studies show that standard reconstruction-based optimization does not necessarily improve diffusion sampling, a phenomenon known as the "reconstruction–generation dilemma". Several works propose training VAEs with a diffusion prior to improve diffusion models (Wehenkel & Louppe, 2021; Vahdat et al., 2021; Heek et al., 2026). Other works analyze and regularize VAEs from a signal-processing perspective (Kouzelis et al., 2025; Skorokhodov et al., 2025; Liu et al., 2025; Fan et al., 2025). Additionally, many recent methods connect VAEs to image encoders trained through self-supervised learning, such as DINO and MAE (Oquab et al., 2023; He et al., 2022), to improve their suitability

Table 7: iFID is robust across condition types, datasets, model architectures, and training epochs.

	Text-to-Image			gFID ImageNet					gFID FFHQ
	Diffusion-Bench			SiT/XL w/o CFG				UNet w/o CFG	SiT/XL
	-GenEval	-DPG	-GenAI	20	40	80	160 epochs	40 epochs	40 epoch
PCC w/ iFID	0.79	0.58	0.88	0.88	0.88	0.87	0.87	0.72	0.72
SRCC w/ iFID	0.67	0.58	0.92	0.88	0.84	0.84	0.86	0.70	0.69
KRCC w/ iFID	0.55	0.45	0.83	0.73	0.65	0.68	0.71	0.52	0.51

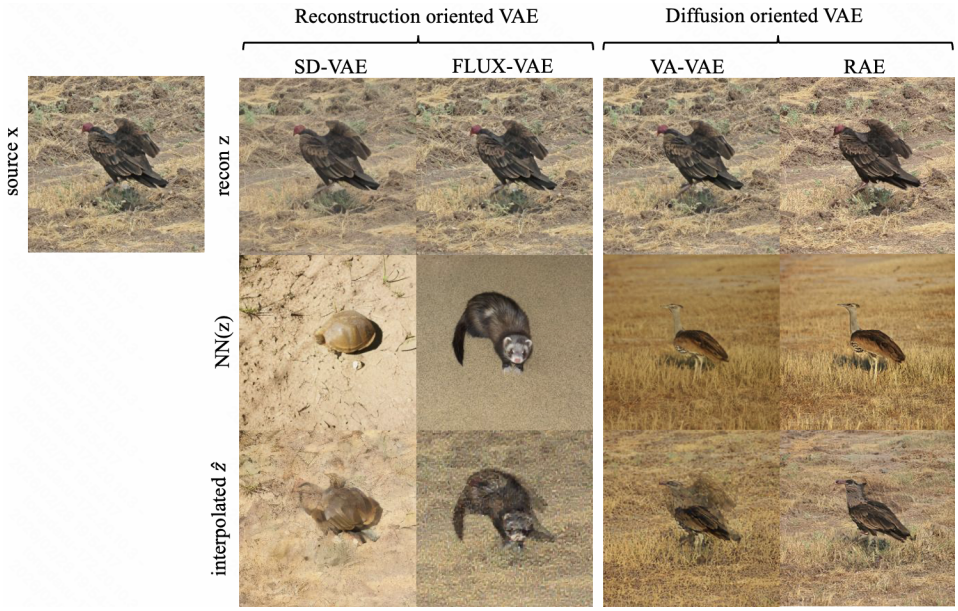

Figure 3: Visualization of nearest-neighbor latents $\mathrm{NN}(z)$ and interpolated latents $\hat{z}$. For reconstruction-oriented VAEs, $\mathrm{NN}(z)$ is often semantically different from z, causing the decoded interpolated latents to yield invalid images. In contrast, for diffusion-oriented VAEs, $\mathrm{NN}(z)$ is semantically similar to z, and the decoded interpolated latents yield realistic images.

for diffusion-based generation (Chen et al., 2025b;a; Zheng et al., 2025; Yao & Wang, 2025; Ye et al., 2025; Yang et al., 2025; Yao et al., 2025; Shi et al., 2025; Gao et al., 2025). Despite these advances, effective metrics for measuring VAE diffusability and explanations for the "reconstruction–generation dilemma" remain underexplored.

How diffusion models generalize. Understanding how diffusion models generalize to unseen samples is a fundamental question. In general, underfitting, underparameterization, and inductive biases in the score estimator are believed to cause the learned score to deviate from the empirical score (Scarvelis et al., 2023; Kadkhodaie et al., 2023; Kamb & Ganguli, 2024; Bonnaire et al., 2025; Song et al., 2025). These deviations prevent diffusion models from simply replicating the training data. Several works argue that diffusion models generate novel samples by composing and interpolating images from the training dataset (Kamb & Ganguli, 2024; Niedoba et al., 2024; Okawa et al., 2023; Deschenaux et al., 2024). Similarly, Zhang et al. (2025) demonstrate that diffusion models generalize through linear interpolation in the latent space of the score-matching network. Conversely, other works show that diffusion models can generate invalid samples, often referred to as hallucinations, by interpolating between nearby modes of the training data (Aithal et al., 2024; Chandran.C et al., 2025; Baptista et al., 2025).

6 Discussion and Conclusion

Although iFID is strongly correlated with gFID of diffusion models, there is currently no straightforward way to minimize it. Besides, currently the ridge set theory supports linear interpolation, but empirically SLERP works better. This might be related to shperical geometric preference of diffusion model (Meral et al., 2026b), which we leave for future works. In conclusion, we propose interpolated FID (iFID), a simple variant of rFID that is strongly correlated with diffusion generation FID. Our key findings are as follows. First, iFID provides a useful proxy for gFID because diffusion models often generate unseen samples by interpolating between training images, while iFID evaluates the quality of decoded latent interpolations. Theoretically, iFID evaluates decoded interpolations that align with an approximate ridge-set manifold around which diffusion samples concentrate. To our knowledge, iFID is the first metric shown to correlate strongly with diffusion gFID across diverse VAEs.

AI USE STATEMENT

In this work, we used generative AI tools for proof reading and polishing. We have not used generative AI tools for generating research ideas or generating code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sumukh K Aithal, Pratyush Maini, Zachary Chase Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *ArXiv*, abs/2406.09358, 2024. URL https://api.semanticscholar.org/CorpusID:270440527.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12:313–326, 1982. URL https://api.semanticscholar.org/CorpusID:3897405.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *arXiv preprint arXiv:2505.17638*, 2025.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models. *ArXiv*, abs/2508.17689, 2025. URL https://api.semanticscholar.org/CorpusID:280711597.

Barath Chandran.C, Srinivas Anumasa, and Dianbo Liu. Laplacian score sharpening for mitigating hallucination in diffusion models. *ArXiv*, abs/2511.07496, 2025. URL https://api.semanticscholar.org/CorpusID:282922289.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. *ArXiv*, abs/2502.03444, 2025a. URL https://api.semanticscholar.org/CorpusID:276116849.

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28358–28370, 2025b.

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*, abs/2410.10733, 2024. URL https://api.semanticscholar.org/CorpusID:273345623.

Justin Deschenaux, Igor Krawczuk, Grigorios G. Chrysos, and Volkan Cevher. Going beyond compositions, ddpms can produce zero-shot interpolations. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:270095412.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. URL https://api.semanticscholar.org/CorpusID:234357997.

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.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. *arXiv preprint arXiv:2512.19693*, 2025.

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *ArXiv*, abs/2512.07829, 2025. URL https://api.semanticscholar.org/CorpusID:283694483.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *ArXiv*, abs/2310.11513, 2023. URL https://api.semanticscholar.org/CorpusID:264288728.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Ye He, Yitong Qiu, and Molei Tao. Diffusion model's generalization can be characterized by inductive biases toward a data-dependent ridge manifold. *ArXiv*, abs/2602.06021, 2026. URL https://api.semanticscholar.org/CorpusID:285304188.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (ul): How to train your latents. 2026. URL https://api.semanticscholar.org/CorpusID:285787616.

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 *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:326772.

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

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

Xi-Wei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *ArXiv*, abs/2403.05135, 2024. URL https://api.semanticscholar.org/CorpusID:268296755.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wenkang Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. *ArXiv*, abs/2405.00571, 2024. URL https://api.semanticscholar.org/CorpusID:269484503.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and St'ephane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *ArXiv*, abs/2310.02557, 2023. URL https://api.semanticscholar.org/CorpusID:263620724.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *ArXiv*, abs/2412.20292, 2024. URL https://api.semanticscholar.org/CorpusID:275134200.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *ArXiv*, abs/2502.09509, 2025. URL https://api.semanticscholar.org/CorpusID:276317789.

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

Junho Lee, Jeongwoo Shin, Hyungwook Choi, and Joonseok Lee. Latent diffusion models with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17422–17431, 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*, abs/2504.10483, 2025. URL https://api.semanticscholar.org/CorpusID:277781071.

Xingjian Leng, Jaskirat Singh, Zhanhao Liang, Ethan Smith, Martin Bell, Aninda Saha, Yuhui Yuan, and Liang Zheng. Diffusionbench: On holistic evaluation of diffusion transformers. *ArXiv*, abs/2606.24888, 2026. URL https://api.semanticscholar.org/CorpusID:289625449.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *ArXiv*, abs/2406.13743, 2024. URL https://api.semanticscholar.org/CorpusID:270619531.

Shizhan Liu, Xinran Deng, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Jie Tang. Delving into latent spectral biasing of video vaes for superior diffusability. *arXiv preprint arXiv:2512.05394*, 2025.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas Matthew 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*, 2024. URL https://api.semanticscholar.org/CorpusID:267027717.

Tuna Meral, Kaan Oktay, Hidir Yesiltepe, Adil Kaan Akan, and Pinar Yanardag. Aligning latent geometry for spherical flow matching in image generation. *ArXiv*, abs/2605.15193, 2026a. URL https://api.semanticscholar.org/CorpusID:288254922.

Tuna Han Salih Meral, Kaan Oktay, Hidir Yesiltepe, Adil Kaan Akan, and Pinar Yanardag. Aligning latent geometry for spherical flow matching in image generation. *arXiv preprint arXiv:2605.15193*, 2026b.

Matthew Niedoba, Berend Zwartsenberg, Kevin Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization. *arXiv preprint arXiv:2411.19339*, 2024.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *ArXiv*, abs/2310.09336, 2023. URL https://api.semanticscholar.org/CorpusID:264146105.

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.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

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.

Christopher Scarvelis, Haitz S'aez de Oc'ariz Borde, and Justin Solomon. Closed-form diffusion models. *ArXiv*, abs/2310.12395, 2023. URL https://api.semanticscholar.org/CorpusID:264305821.

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*, abs/2510.15301, 2025. URL https://api.semanticscholar.org/CorpusID:282203316.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *ArXiv*, abs/2512.10794, 2025. URL https://api.semanticscholar.org/CorpusID:283737204.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *ArXiv*, abs/2502.14831, 2025. URL https://api.semanticscholar.org/CorpusID:276482178.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2022. URL https://api.semanticscholar.org/CorpusID:254366634.

Kiwhan Song, Jaeyeon Kim, Sitan Chen, Yilun Du, Sham M. Kakade, and Vincent Sitzmann. Selective underfitting in diffusion models. *ArXiv*, abs/2510.01378, 2025. URL https://api.semanticscholar.org/CorpusID:281724714.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. URL https://api.semanticscholar.org/CorpusID:227209335.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space, 2021. *URL https://arxiv. org/abs/2106.05931*, 2021.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23:1661–1674, 2011. URL https://api.semanticscholar.org/CorpusID:5560643.

Luting Wang, Yang Zhao, Zijian Zhang, Jiashi Feng, Si Liu, and Bingyi Kang. Image understanding makes for a good tokenizer for image generation. *Advances in Neural Information Processing Systems*, 37:31015–31035, 2024.

Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. URL https://api.semanticscholar.org/CorpusID:207761262.

Antoine Wehenkel and Gilles Louppe. Diffusion priors in variational autoencoders. *arXiv preprint arXiv:2106.15671*, 2021.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18770–18780, 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 and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025. URL https://api.semanticscholar.org/CorpusID:275211938.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards Scalable Pre-training of Visual Tokenizers for Generation. 12 2025. URL http://arxiv.org/abs/2512.13687.

Sen Ye, Jianning Pei, Mengde Xu, Shuyang Gu, Chunyu Wang, Liwei Wang, and Han Hu. Distribution matching variational autoencoder. *arXiv preprint arXiv:2512.07778*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. 5 2024. URL http://arxiv.org/abs/2405.14867.

Taeho Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion Probabilistic Models Generalize when They Fail to Memorize. Technical report, 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. *ArXiv*, abs/2410.06940, 2024. URL https://api.semanticscholar.org/CorpusID:273229262.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, and Yali Wang. What matters for diffusion-friendly latent manifold? prior-aligned autoencoders for latent diffusion. *ArXiv*, abs/2605.07915, 2026. URL https://api.semanticscholar.org/CorpusID:288147976.

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.

Zekai Zhang, Xiao Li, Xiang Li, Lianghe Shi, Meng Wu, Molei Tao, and Qing Qu. Generalization of diffusion models arises with a balanced representation space. *ArXiv*, abs/2512.20963, 2025. URL https://api.semanticscholar.org/CorpusID:284153760.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *ArXiv*, abs/2510.11690, 2025. URL https://api.semanticscholar.org/CorpusID:282057986.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. 2026. URL https://api.semanticscholar.org/CorpusID:288903703.

A Proof of Main Results

Theorem 3. (Local ridge set is close to the scaled line connecting nearby training samples) *Assume that, in a local region $\mathcal{U}$, two nearby training points z^i, z^j dominate the other samples in terms of their weights w^k, i.e., for every $z_t \in \mathcal{U}$, $\epsilon(z_t) = \sum_{k \neq i,j} w^k \leq \epsilon$. Let $L_t = \{\alpha_t(z^i + a(z^j - z^i)) : a \in \mathbb{R}\}$ be the scaled affine line through the pair. Assume that $\tilde{w}^i = \frac{w^i}{w^i + w^j}, \tilde{w}^j = \frac{w^j}{w^i + w^j}$ are non-negligible such that $b = \tilde{w}^i \tilde{w}^j \|z^i - z^j\|^2 > \gamma$, and $|z^k|, |z_t| \leq R$. Then, for every ridge point $z_t \in \mathcal{R}$ in the local region $\mathcal{U}$,*

$$\operatorname{dist}(z_t, L_t) \leq \frac{24R^3\epsilon}{\gamma} + 2R\epsilon.$$

Proof. Let

$$e = \frac{z^i - z^j}{\|z^i - z^j\|}$$

and denote by

$$P_2 = I - ee^T$$

the orthogonal projector onto the subspace normal to the line L. Since scaling a line by a nonzero scalar does not change its direction, P_2 is also the normal projector of L_t.

Therefore,

$$\begin{aligned} \operatorname{dist}(z_t, L_t) &= |P_2(z_t - \alpha_t z^i)| \\ &\leq |P_2(z_t - \alpha_t \mu)| + |\alpha_t||P_2(\mu - z^i)|, \end{aligned} \tag{8}$$

where

$$\mu = \sum_k w^k z^k.$$

We bound these two terms separately. First, the empirical score can be written as

$$s_*(z_t, t) = \frac{\alpha_t \mu - z_t}{\sigma_t^2}.$$

Since $|z_t|, |z^k| \leq R$, we have $|\mu| \leq R$ and hence

$$\|s_*(z_t, t)\| \leq \frac{R + |\alpha_t| R}{\sigma_t^2}.$$

Now consider the Jacobian induced by only the two dominant training samples. Define

$$\tilde{\mu} = \tilde{w}^i z^i + \tilde{w}^j z^j.$$

The covariance of these two points is

$$\Sigma_2 = \tilde{w}^i \tilde{w}^j (z^i - z^j)(z^i - z^j)^T = b\, ee^T.$$

Following Lemma 2, the corresponding two-point Jacobian is therefore

$$J_2 = -\frac{1}{\sigma_t^2} I + \frac{\alpha_t^2 b}{\sigma_t^4} ee^T.$$

Its eigenvalue in the direction e is

$$-\frac{1}{\sigma_t^2} + \frac{\alpha_t^2 b}{\sigma_t^4},$$

whereas its eigenvalue in every direction orthogonal to e is

$$-\frac{1}{\sigma_t^2}.$$

Thus, the spectral gap between the tangential and normal eigenspaces is

$$\delta = \frac{\alpha_t^2 b}{\sigma_t^4} \geq \frac{\alpha_t^2 \gamma}{\sigma_t^4}.$$

Next, we bound the difference between the full covariance Σ and its two-point approximation Σ_2. As in the original argument, the local dominance assumption gives

$$\|\mu - \tilde{\mu}\| \leq 2R\epsilon$$

and

$$\|\Sigma - \Sigma_2\| \leq 6R^2\epsilon.$$

Since

$$J = -\frac{1}{\sigma_t^2} I + \frac{\alpha_t^2}{\sigma_t^4}\Sigma,$$

we obtain

$$\|J - J_2\| \leq \frac{6\alpha_t^2 R^2 \epsilon}{\sigma_t^4}.$$

Let P denote the projector onto the normal eigenspace of J. By the Davis–Kahan theorem,

$$\begin{aligned} |P - P_2| &\leq 2\frac{|J - J_2|}{\delta} \\ &\leq 2\frac{6\alpha_t^2 R^2 \epsilon/\sigma_t^4}{\alpha_t^2 \gamma/\sigma_t^4} \\ &\leq \frac{12R^2\epsilon}{\gamma}. \end{aligned} \tag{9}$$

Notice that the factors α_t^2 cancel. Thus, the stability of the normal eigenspace is unchanged by the forward scaling.

By the definition of the ridge set $\mathcal{R}$,

$$P s_*(z_t, t) = 0.$$

Consequently,

$$\begin{aligned} |P_2 s_*(z_t, t)| &= |(P_2 - P)s_*(z_t, t)| \\ &\leq |P_2 - P||s_*(z_t, t)| \\ &\leq \frac{12R^2\epsilon}{\gamma}\frac{R + |\alpha_t|R}{\sigma_t^2}. \end{aligned} \tag{10}$$

Using

$$z_t - \alpha_t \mu = -\sigma_t^2 s_*(z_t, t),$$

we therefore obtain

$$\begin{aligned} |P_2(z_t - \alpha_t\mu)| &= \sigma_t^2 |P_2 s_*(z_t, t)| \\ &\leq \frac{12R^2(R + |\alpha_t|R)\epsilon}{\gamma}. \end{aligned} \tag{11}$$

For the second term in Eq. 8, since

$$P_2(z^j - z^i) = 0,$$

we have

$$\begin{aligned} |P_2(\mu - z^i)| &= \left|\sum_k w^k P_2(z^k - z^i)\right| \\ &= \left|\sum_{k\neq i,j} w^k P_2(z^k - z^i)\right| \\ &\leq \sum_{k\neq i,j} w^k |z^k - z^i| \\ &\leq 2R\epsilon. \end{aligned} \tag{12}$$

Hence,

$$|\alpha_t|\|P_2(\mu - z^i)\| \leq 2|\alpha_t|R\epsilon.$$

Combining the two bounds, with the convention that $\alpha_t \leq 1$, we have

$$\text{dist}(z_t, L_t) \leq \frac{24R^3\epsilon}{\gamma} + 2R\epsilon.$$

This completes the proof. □

B Additional Experimental Results

B.1 Pre-Trained VAEs

SD-VAE The Stable Diffusion VAE (SD-VAE) (Rombach et al., 2022) is among the first VAEs to be adopted in large-scale latent diffusion models. It is trained with a combination of KL-divergence regularization, mean-squared error, LPIPS loss, and adversarial loss. It has a 2D latent with shape $4 \times 32 \times 32$. We adopt the pre-trained checkpoint from `https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5/tree/main/vae`

IN-VAE ImageNet VAE (IN-VAE) (Yao & Wang, 2025) serves as the baseline VAE for REPA-E (Leng et al., 2025). It is trained with a mixture of KL divergence penalties, mean-squared error, LPIPS, and adversarial loss. It uses a 2D latent representation with shape $32 \times 16 \times 16$. We adopt the pre-trained checkpoint from `https://huggingface.co/REPA-E/invae/tree/main`.

FLUX-VAE FLUX-VAE (Labs, 2024) is the VAE for FLUX-1 models. It extends the channel dimension of SD-VAE to 32 and achieves higher reconstruction performance. It has a 2D latent representation with shape $16 \times 32 \times 32$. We adopt the pre-trained checkpoint from `https://huggingface.co/black-forest-labs/FLUX.1-dev/tree/main/vae`.

QW-VAE Qwen-Image VAE (QW-VAE) (Wu et al., 2025a) is the VAE for the Qwen-Image Model. It uses the same latent shape as FLUX-VAE, namely a 2D latent with shape $16 \times 32 \times 32$. We adopt the pre-trained checkpoint from `https://huggingface.co/Qwen/Qwen-Image/tree/main/vae`.

SD3-VAE Stable Diffusion 3 VAE (SD3-VAE) (Esser et al., 2024) is the VAE used in Stable Diffusion 3 models. It has the same latent shape as FLUX-VAE. It uses a 2D latent with shape $16 \times 32 \times 32$. We adopt the pre-trained checkpoint at `https://huggingface.co/black-forest-labs/FLUX.1-dev/tree/main/vae`.

EQ-VAE Equivariance VAE (EQ-VAE) (Kouzelis et al., 2025) is proposed improve diffusion training. In addition to the SD-VAE loss, EQ-VAE is regularized with an equivariance constraint that ensures that an affine transformation applied to the latent space has the same effect as the affine transform applied to the image. It has a 2D latent with shape $4 \times 32 \times 32$. We adopt the pre-trained checkpoint at `https://huggingface.co/zelaki/eq-vae/tree/main`.

VA-VAE Vision foundation model Aligned Variational AutoEncoder (VA-VAE) (Yao & Wang, 2025) is proposed to improve Gaussian VAEs for diffusion training. In addition to the standard VAE loss, VA-VAE additionally uses vision foundation models such as DINO (Oquab et al., 2023) and aligns the latent space linearly to the feature encoded by DINOv2. Such alignment empirically improves diffusion sample quality. We evaluate two variants of VA-VAE, with 2D latents of shape $32 \times 16 \times 16$ and $64 \times 16 \times 16$, respectively. We adopt the pre-trained checkpoint in `https://huggingface.co/hustvl/va-vae-imagenet256-experimental-variants/tree/main`.

SOFT-VQ Soft VQ-VAE (SOFT-VQ) (Chen et al., 2025b) is a variational autoencoder with a 1D latent space and a transformer architecture. It also adopts vision foundation models, such as DINOv2, for alignment. It has a 1D latent with shape 64×32. We adopt the pre-trained checkpoint from `https://huggingface.co/SoftVQVAE/softvq-l-64`.

MAE-TOK Masked Autoencoders (MAE-TOK) (Chen et al., 2025a) is an extension SOFT-VQ, which additionally introduces mask training following MAE (He et al., 2022) and uses auxiliary decoder loss functions aligned with DINOv2 features. It has a 1D latent with shape 128×32. We adopt the pre-trained checkpoint from `https://huggingface.co/MAETok/maetok-b-128`.

DE-TOK Denoising Tokenizer (DE-TOK) is a combination of a denoising autoencoder (Vincent, 2011) and MAE (He et al., 2022). During training, it combines token corruption with masked training. It has a 1D latent with shape 128×32. We adopt the pre-trained checkpoint from `https://huggingface.co/jjiaweiyang/l-DeTok`.

DM-VAE Distribution Matching VAE (DM-VAE) adopts Distribution Matching Distillation (DMD) (Yin et al., 2024) as the prior for the VAE. We find that alignment to the DINO prior derived from ten ImageNet classes is most beneficial for generation. It has a 1D latent with shape 256×32. We adopt the pre-trained checkpoint in `https://huggingface.co/sen-ye/dmvae/tree/main`.

REPAE-VAE End-to-end representation alignment VAE (REPAE-VAE) (Leng et al., 2025) is obtained by end-to-end optimization of a VAE with REPA-based alignment (Yu et al., 2024). We adopt the REPAE-VAE that starts from SD-VAE. It has a 2D latent representation with shape $4 \times 32 \times 32$. We adopt the pre-trained checkpoint in `https://huggingface.co/REPA-E/e2e-sdvae`.

RAE Representation-aligned autoencoder (RAE) (Zheng et al., 2025) removes the reconstruction loss in the encoding stage. It directly adopts DINOv2 (Oquab et al., 2023) as the image encoder and trains a decoder for reconstruction. It has a 2D latent representation with shape $768 \times 16 \times 16$. We adopt the pre-trained checkpoint in `https://huggingface.co/nyu-visionx/RAE-collections/tree/main`

Training We train all SiT models (Ma et al., 2024) on the ImageNet training split for 40 epochs, following Leng et al. (2025). For VAEs with 2D latents, we adopt patch sizes of 1 or 2 to produce a consistent sequence length of 256. We employ 1D/2D sine-cosine positional embeddings for VAEs with 1D/2D latents. We additionally apply the time-shifting strategy introduced by (Esser et al., 2024; Zheng et al., 2025) to eliminate the effect of different total latent dimensions on a diffusion model.

Sampling We adopt the Euler SDE solver implemented in (Ma et al., 2024) with 250 steps. For results with CFG, we search the classifier-free guidance scale in the range [1.0, 6.0] with a step size of 0.25 and report the best gFID.

B.2 METRICS

PSNR, SSIM, LPIPS and rFID The peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS), together with reconstruction FID (rFID) (Wang et al., 2004; Zhang et al., 2018; Heusel et al., 2017), are among the most commonly used reconstruction metrics for training and evaluating VAEs. PSNR, SSIM, and LPIPS are similar in that they are computed between individual source images and their reconstructions. However, they are strongly negatively correlated with the quality of samples produced by diffusion models. rFID measures the distributional discrepancy between the original images and their reconstructions. Additionally, its correlation with gFID is weak.

EQ Loss (Kouzelis et al., 2025) The Equivariance Loss (EQ Loss) evaluates whether a spatial transformation applied to the latent representation is preserved after decoding. Given an input image x and its latent representation z, and letting $g(\cdot)$ denote the decoder, the transformed reconstruction term is

$$\mathcal{L}_{\mathrm{EQ}}(x;\tau) = \Delta(\tau \circ x,\ g(\tau \circ z))\,. \tag{13}$$

where $\Delta(\cdot,\cdot)$ denotes the reconstruction distortion. The transformation τ is sampled from spatial transformations composed of scaling and 90°-multiple rotations. The final score is averaged over the evaluation set. Lower EQ Loss indicates better equivariance under such spatial transformations.

VF Loss (Yao & Wang, 2025) The Vision Foundation model alignment loss (VF loss) aligns the VAE latent features with the features of a frozen vision foundation model. Given an input x, let z denote the latent representation produced by the encoder and let f denote the feature representation extracted by the foundation model. A linear projection W is first used to match the latent dimension to the foundation-model feature dimension:

$$z' = Wz. \tag{14}$$

Let d denote the number of feature positions after flattening all non-channel dimensions. We use z'_i and f_i to denote the feature vectors at position $i \in \{1, \ldots, d\}$ in z' and f, respectively. The VF loss

consists of two terms. The first term is a marginal cosine similarity term:

$$\mathcal{L}_{\text{mcos}} = \frac{1}{d}\sum_{i=1}^{d}\text{ReLU}\left(1 - m_1 - \frac{z_i' \cdot f_i}{\|z_i'\|\,\|f_i\|}\right). \tag{15}$$

The second term is a marginal distance matrix similarity term, where $i, j \in \{1, \ldots, d\}$ index two feature positions:

$$\mathcal{L}_{\text{mdms}} = \frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\text{ReLU}\left(\left|\frac{z_i' \cdot z_j'}{\|z_i'\|\,\|z_j'\|} - \frac{f_i \cdot f_j}{\|f_i\|\,\|f_j\|}\right| - m_2\right). \tag{16}$$

Here, m_1 and m_2 are margin hyperparameters. We define the overall VF loss as

$$\mathcal{L}_{\text{VF}} = \mathcal{L}_{\text{mcos}} + \mathcal{L}_{\text{mdms}}. \tag{17}$$

SE Loss (Skorokhodov et al., 2025) The Scale-Equivariance Loss (SE Loss) regularizes reconstruction by enforcing consistency between downsampled RGB signals and reconstructions from downsampled latents. Given an input image x and its latent representation z, let $s(\cdot)$ denote a downsampling operator, and define $\tilde{x} = s(x)$ and $\tilde{z} = s(z)$. The SE objective is

$$\mathcal{L}_{\text{SE}}(x) = \Delta(x, g(z)) + \alpha\,\Delta(\tilde{x}, g(\tilde{z})), \tag{18}$$

where $\Delta(\cdot, \cdot)$ denotes the reconstruction distortion and α is the scale-equivariance regularization weight. Lower SE Loss indicates better scale equivariance.

GMM Loss (Chen et al., 2025a) The Gaussian Mixture Model Loss (GMM Loss) quantifies the structural complexity of the latent space by fitting a Gaussian mixture model with a fixed number of components. Given latent representations $z^{(1:N)}$, we fit a K-component Gaussian mixture model (GMM) to the latent space and evaluate its negative log-likelihood (NLL):

$$\mathcal{L}_{\text{GMM}} - \frac{1}{N}\sum_{n=1}^{N}\log p_{\text{GMM},K}(z^{(n)}), \tag{19}$$

where $p_{\text{GMM},K}$ denotes the density of the fitted K-component GMM. A lower GMM Loss indicates that the latent distribution can be better approximated by a K-component GMM, suggesting a simpler latent structure.

In (Chen et al., 2025a), the authors theoretically demonstrate that a latent space with lower GMM Loss also exhibits lower diffusion loss, and therefore propose using GMM Loss as an indicator of diffusion sample quality. In Figure 6, we show that GMM Loss exhibits a strong correlation with diffusion loss. However, diffusion loss does not necessarily correlate highly with sample quality when the latent spaces differ.

B.3 The Relationship Between Metrics and gFID

Additionally, we provide comprehensive plots illustrating the relationship between gFID and all metrics, as shown in Figures 4-9.

B.4 Visualization

In Figure 10, we provide additional visualizations of the source image x, the reconstructed image $g(z)$, the nearest neighbour image $g(\text{NN}(z))$, and the interpolated image $g(\hat{z})$. For VAEs optimized for reconstruction, the nearest neighbours $\text{NN}(z)$ are semantically different from z, and the interpolations $\hat{z}$ are not realistic. In contrast, for VAEs optimized with diffusion, the nearest neighbours $\text{NN}(z)$ are semantically similar to z, and the interpolations $\hat{z}$ are more realistic.

C Additional Discussions

C.1 Explanation on Diffusibility of Previous VAEs

Pure Reconstruction VAEs VAEs trained solely with a reconstruction metric exhibit the worst generative performance (SD-VAE, IN-VAE, FLUX-VAE, SD3-VAE, QW-VAE) (Yang et al., 2025;

Table 8: Correlation between 15 VAE metrics and diffusion gFID. Correlations are computed using metrics obtained from SiT models trained with 22 pre-trained VAEs. EQ Loss and SE Loss are not evaluated for VAEs with 1D latents and RAE is not evaluated with SiT-B. KRCC denotes Kendall rank correlation coefficient, and Top-K denotes Top-K overlap ratio. **Bold**: best.

Metrics	gFID SiT/B				gFID SiT/XL			
	w/o CFG		w/ CFG		w/o CFG		w/ CFG	
	KRCC↑	Top-K↑	KRCC↑	Top-K↑	KRCC↑	Top-K↑	KRCC↑	Top-K↑
Reconstruction								
-PSNR	0.42	0.6	0.24	0.5	0.39	0.6	0.31	0.5
LPIPS	-0.47	0.3	-0.27	0.4	-0.42	0.3	-0.32	0.4
-SSIM	0.37	0.6	0.18	0.5	0.32	0.6	0.25	0.4
rFID	-0.16	0.4	0.00	0.5	-0.11	0.3	0.01	0.5
Equivariance								
EQ Loss	-0.08	0.4	-0.22	0.3	0.02	0.4	-0.13	0.4
SE Loss	-0.38	0.2	-0.27	0.2	-0.35	0.1	-0.32	0.2
Data Rank								
GMM Loss	0.22	0.8	0.15	**0.7**	0.30	**0.8**	0.20	0.6
VIV	0.20	0.7	0.04	**0.7**	0.32	0.6	0.05	0.6
Spatial Smoothness								
LDS	0.39	0.7	0.39	**0.7**	0.30	0.7	0.34	**0.7**
CDS	0.10	0.7	0.03	0.5	0.00	0.6	0.04	0.5
SEC	-0.10	0.3	-0.18	0.3	0.34	0.4	-0.18	0.4
Semantic Alignment								
VF Loss	0.37	**0.8**	0.28	**0.7**	0.47	**0.8**	0.33	0.6
-Acc@1	0.28	0.7	0.15	0.6	0.40	0.7	0.25	**0.7**
-Acc@5	0.29	0.7	0.16	0.6	0.41	0.7	0.25	0.6
Proposed								
iFID	**0.62**	**0.8**	**0.47**	**0.7**	**0.65**	**0.8**	**0.51**	**0.7**

Labs, 2024; Esser et al., 2024; Wu et al., 2025a; Yao & Wang, 2025). According to our findings, reconstruction loss results in a more separable latent space, thereby leading to degraded generative performance.

VAEs with Signal Processing Regularization Several previous VAEs incorporate signal processing regularization to constrain the shape of the latent space by suppressing high-frequency information (EQ-VAE, FT-SE, SSVAE) (Kouzelis et al., 2025; Skorokhodov et al., 2025; Liu et al., 2025). Our findings indicate that the performance of diffusion models is primarily determined by the interpolability of the latent space. Suppressing high-frequency components in the latent space enhances latent smoothness, thereby improving interpolability.

VAEs with Vision Foundation Models Many previous VAEs incorporate vision foundation models as alignment mechanisms or encoders (e.g., VA-VAE, SOFT-VQ, MAE-TOK, DE-TOK, DM-VAE, REPAE-VAE, RAE, VMAE) (Yao & Wang, 2025; Chen et al., 2025b;a; Ye et al., 2025; Leng et al., 2025; Yu et al., 2024; Yang et al., 2025; Lee et al., 2025). According to our findings, the performance of diffusion models is primarily determined by the interpolability of the latent space. The latent space of these vision foundation models, such as DINOv2 (Oquab et al., 2023), MAE (He et al., 2022), and Denoising Auto-encoder (Vincent, 2011), is semantically continuous and interpolable. This explains why adopting these vision foundation models improves generation quality.

C.2 Practical Guidance to Diffusion VAEs

Although iFID cannot currently be optimized directly, our findings suggest that reconstruction objectives may harm generation performance by encouraging more separable and isolated latent representations. A practical implication is therefore to reduce or avoid reconstruction-based supervision of the encoder when generation quality is the primary objective. This observation is also consistent with the performance of RAE (Zheng et al., 2025), which achieves the best gFID without CFG in our benchmark.

C.3 Relationship to Discrete Tokenizers

As early as 2022, researchers found that using vision foundation models as the encoder for vector quantized variational autoencoders (VQ-VAE) improves autoregressive image generation (Peng et al., 2022). Subsequently, the superiority of vision foundation model encoders over reconstruction-based VQ-VAEs has been verified in numerous studies (Wang et al., 2024; Wu et al., 2025b; Xiong et al., 2025; Geng et al., 2025). However, similar to VAEs in diffusion models, metrics on VQ-VAE for predicting the generation performance of autoregressive models remain under-explored. It is possible to extend iFID to VQ-VAE and evaluate its relationship to the generation performance of autoregressive models. Nevertheless, we leave this topic to future work.

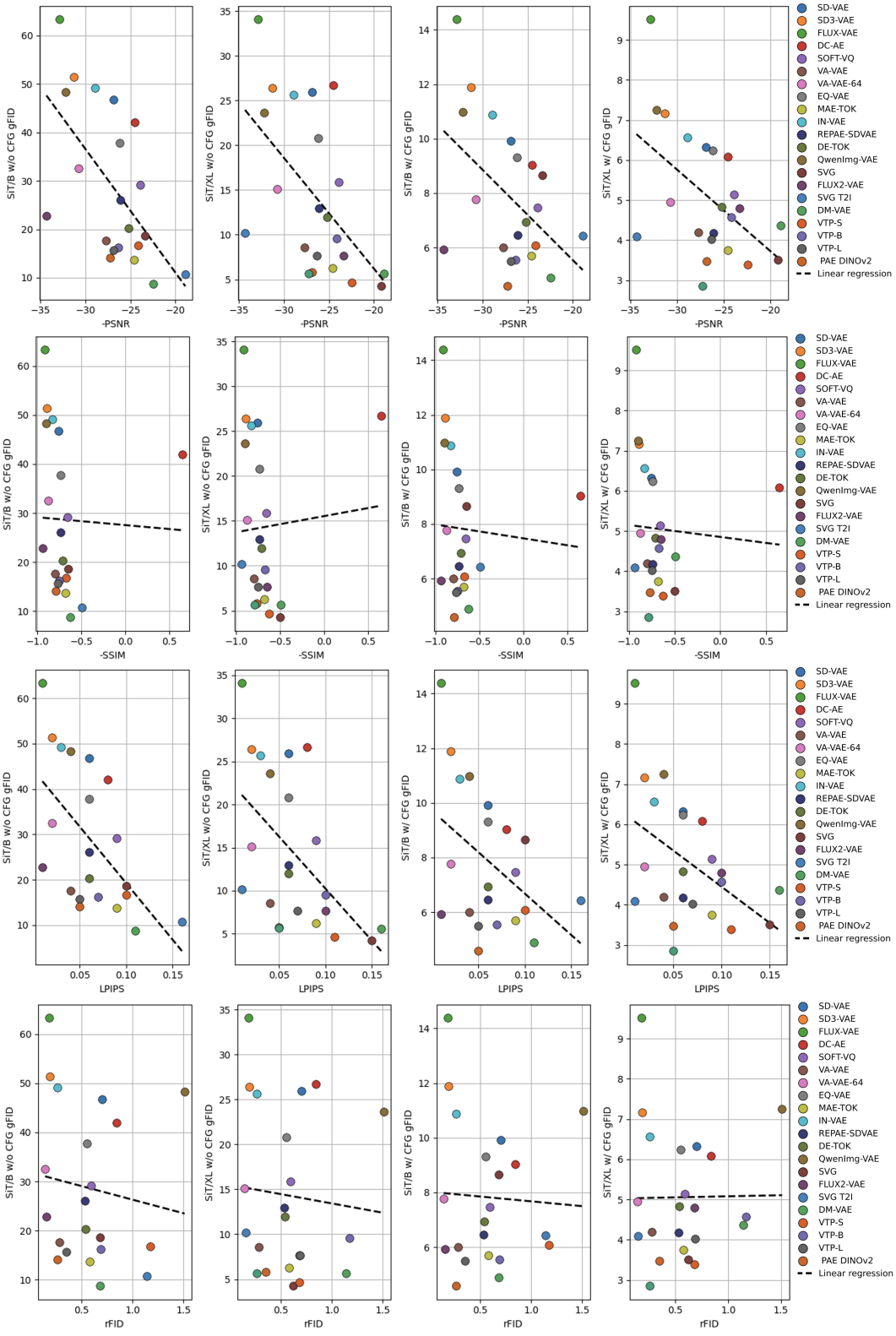

Figure 4: The relationship between reconstruction metrics and gFID.

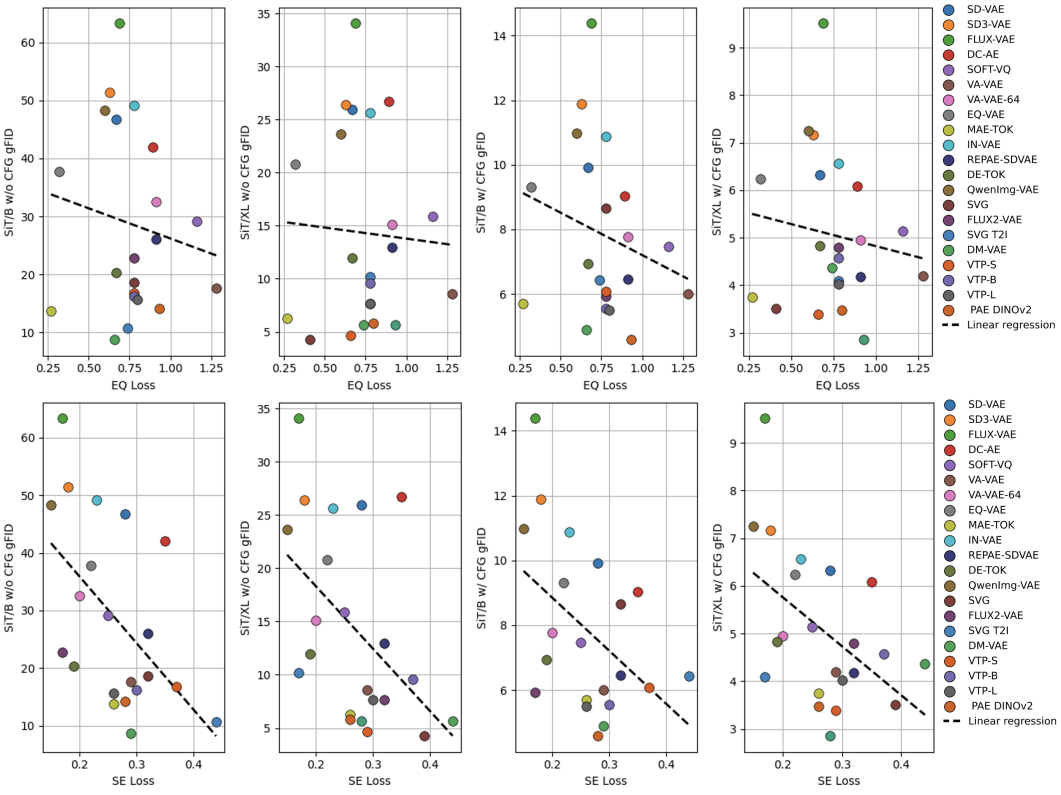

Figure 5: The relationship between equivariance metrics and gFID.

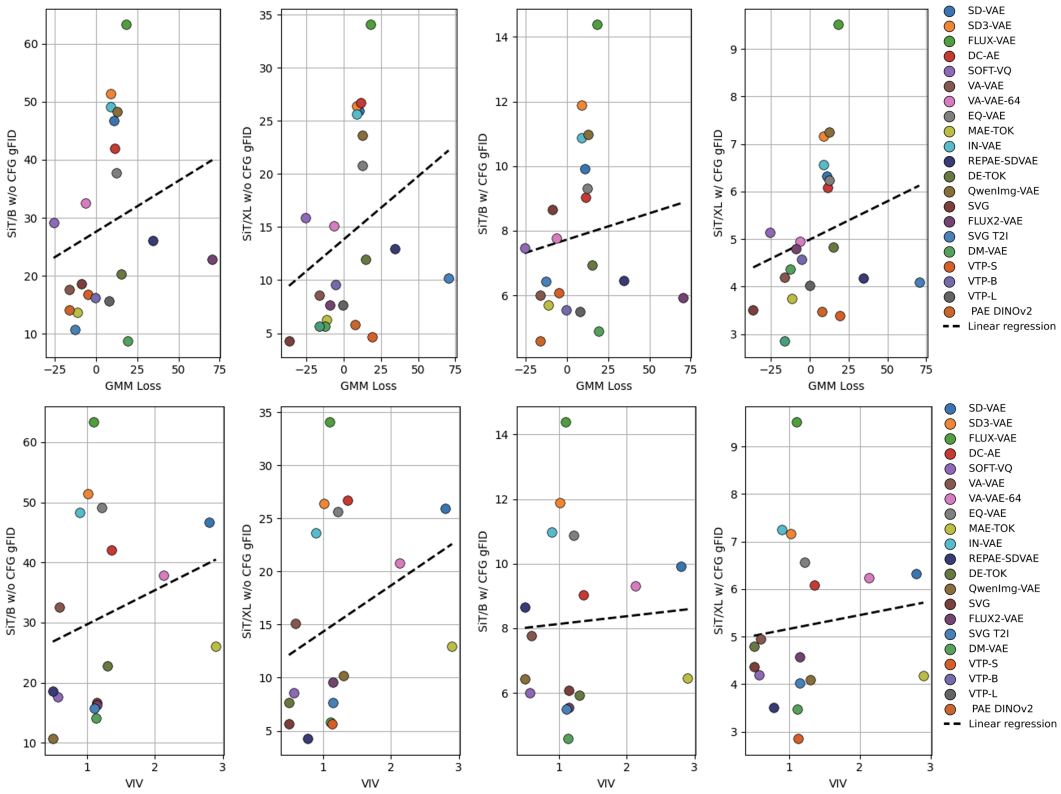

Figure 6: The relationship between data rank metrics and gFID.

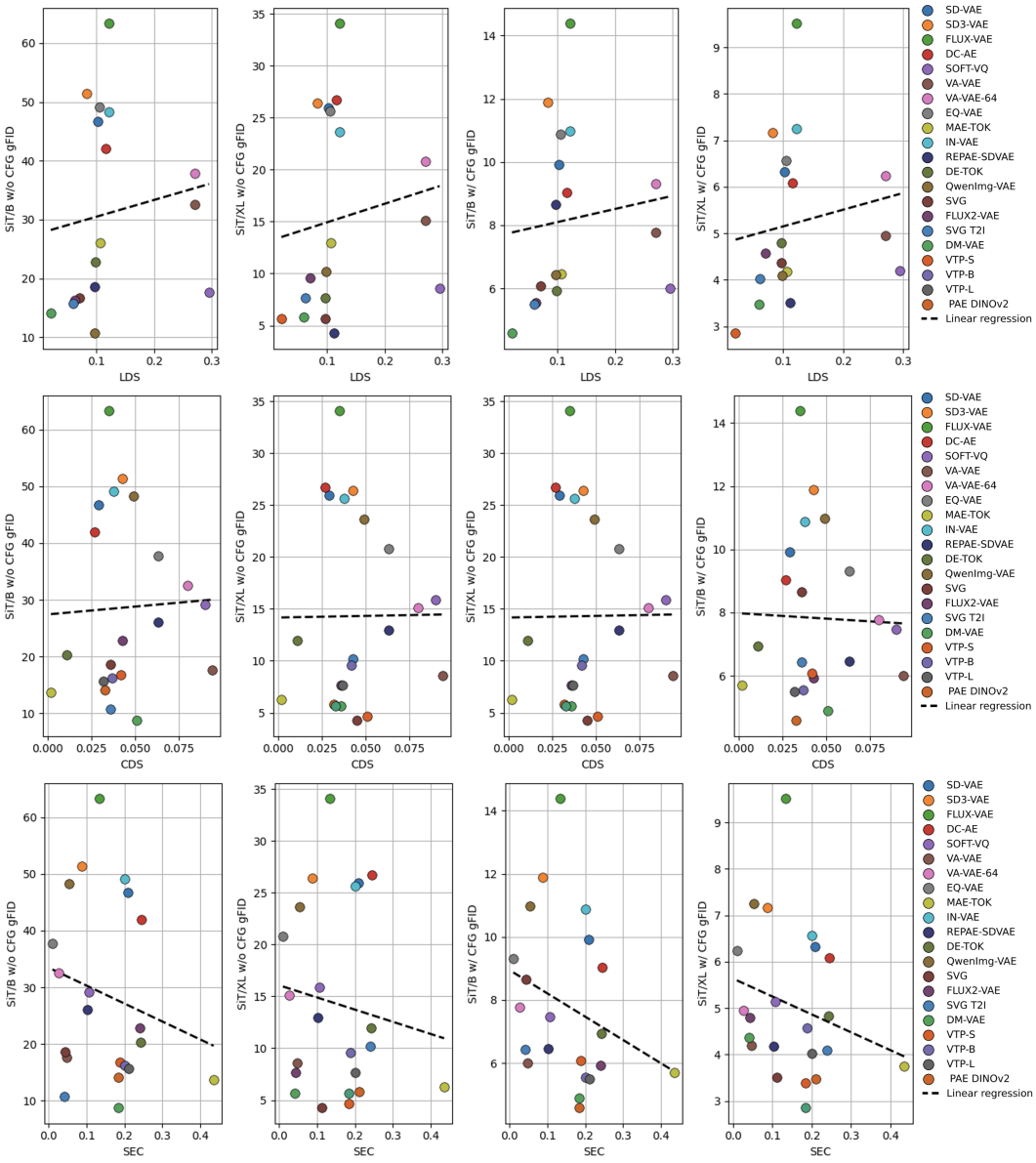

Figure 7: The relationship between spatial smoothness metrics and gFID.

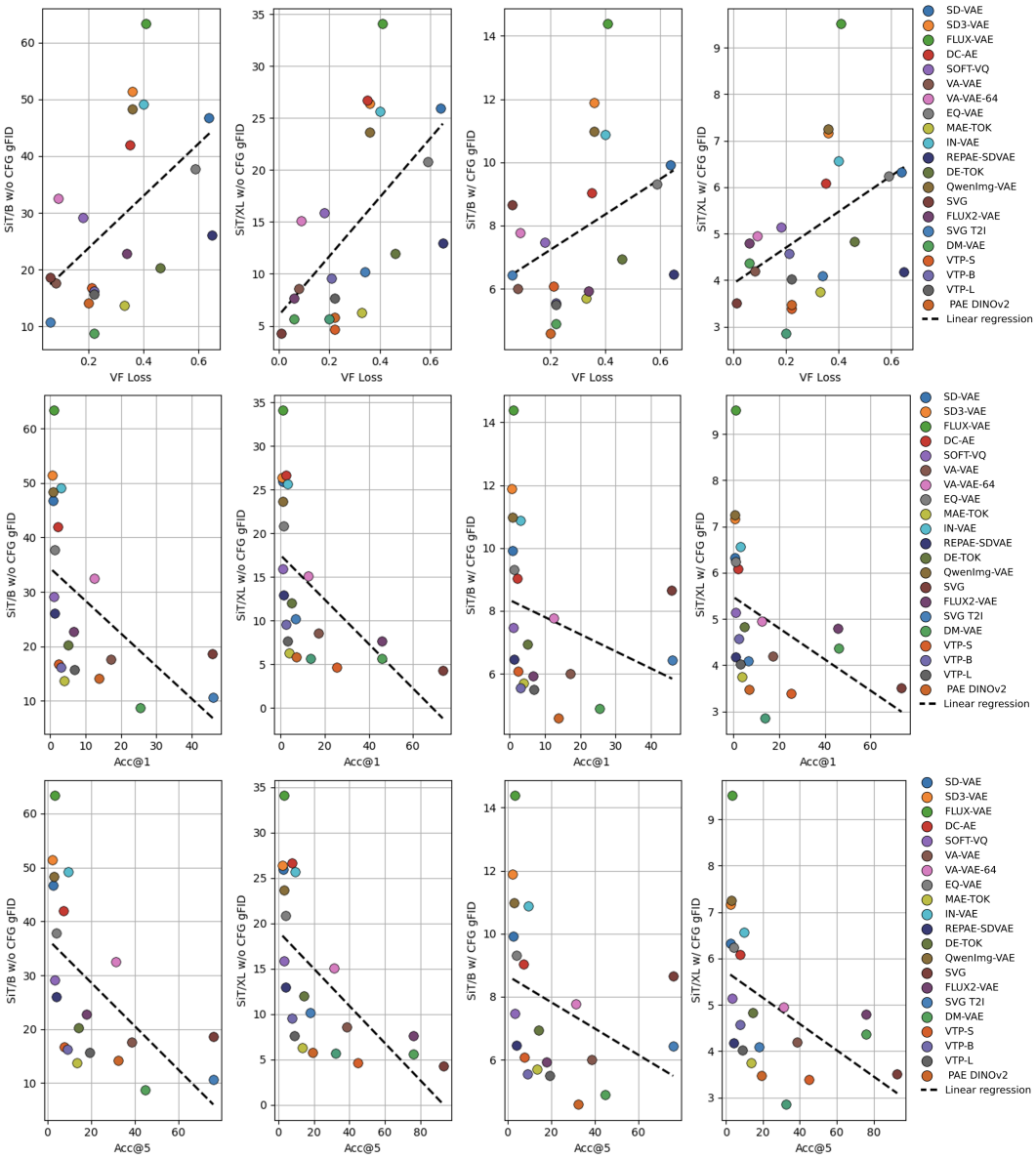

Figure 8: The relationship between semantic alignment metrics and gFID.

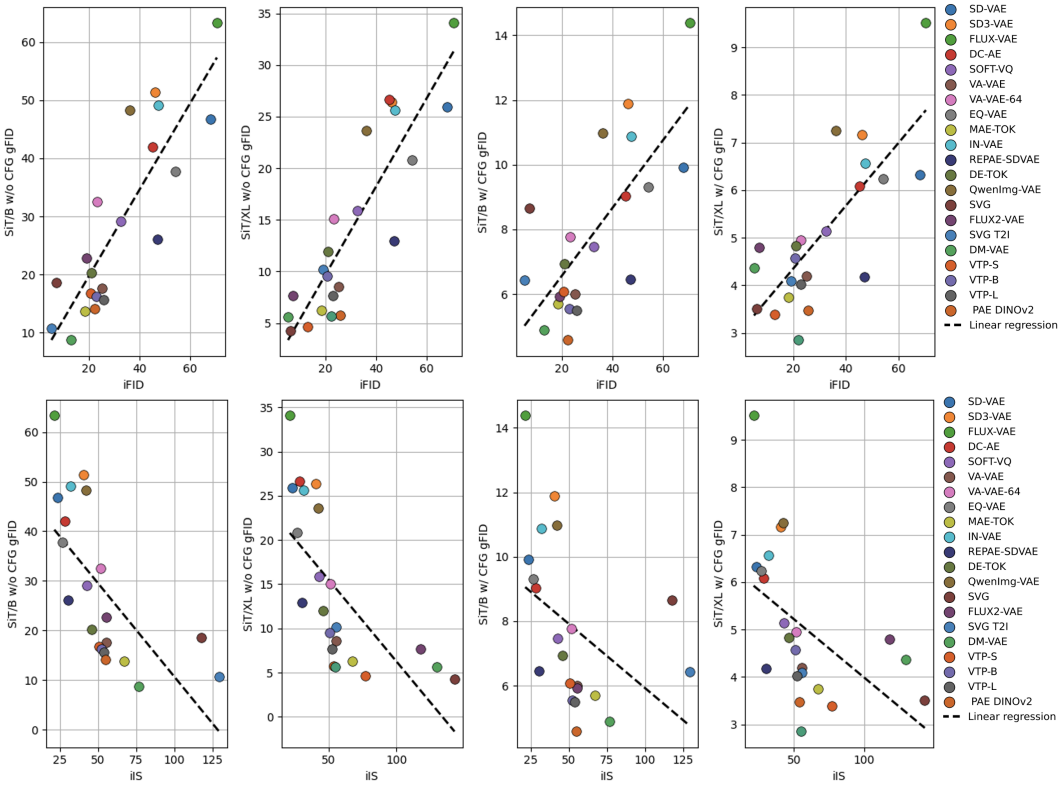

Figure 9: The relationship between proposed metrics and gFID.

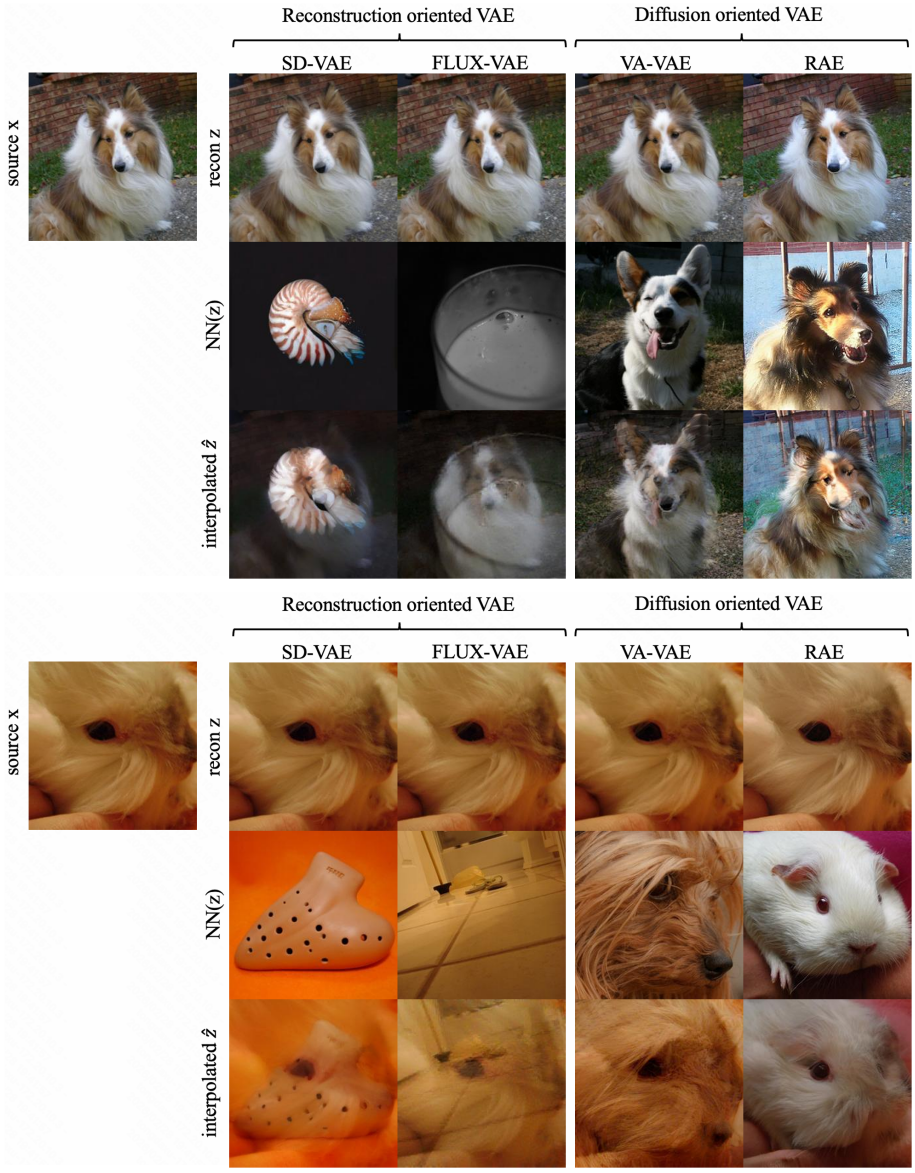

Figure 10: Visualization of the decoded nearest neighbour latent NN(z) and the interpolated latent $\hat{z}$. For reconstruction-oriented VAEs, NN(z) is semantically different from z, and the interpolated $\hat{z}$ are invalid images. In contrast, for diffusion-oriented VAEs, NN(z) is semantically similar to z, and the interpolated $\hat{z}$ are realistic images.