A LATENT DISTRIBUTION PERSPECTIVE ON EVALUATING AND IMPROVING LATENT GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In latent generative models, reconstruction quality is often assumed to correlate with generative performance. However, reconstruction FID (rFID) can exhibit weak or even negative correlation with generation FID (gFID). We attribute this misalignment to a latent distribution mismatch: reconstruction evaluates the decoder on encoder-induced latents, whereas generation uses the same decoder on latents produced by the generative model. To characterize this shift, we introduce generation-aware reconstruction (GAR), which constructs a continuous trajectory from standard reconstruction toward generation by perturbing encoder latents with noise and denoising them through the generative model before decoding. GAR probes the decoder behavior along this trajectory, making the transition from encoder to generation-time latent distributions observable and diagnosable. The resulting trajectory-based diagnostic, GAR-FID, exhibits strong empirical correlation with gFID across diverse tokenizers and scales. Importantly, intermediate GAR latents become more generation-aware while preserving correspondence with their source images, thereby retaining paired supervision that is absent for fully generated latents. This correspondence enables decoder adaptation on intermediate GAR latents, consistently improving generative quality across model scales. Overall, latent distribution mismatch provides a useful perspective for evaluating and improving latent generative models.

1 INTRODUCTION

Latent generative modeling has become a dominant paradigm for high-quality image synthesis, typically using a visual tokenizer to map images into a latent space where generation is modeled by diffusion or flow matching (Rombach et al., 2022; Preechakul et al., 2022; Peebles & Xie, 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023; Geng et al., 2025). Reconstruction quality is commonly used to evaluate the tokenizer and is often implicitly assumed to correlate with downstream generative performance: a decoder that better reconstructs images from encoder latents might also be expected to produce better samples during generation (van den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022; Preechakul et al., 2022). However, recent empirical studies (Yao & Wang, 2025; Kouzelis et al., 2025; Ye et al., 2025; Skorokhodov et al., 2025; Chen et al., 2025b; Xu et al., 2026) have challenged this assumption. In practice, reconstruction FID (rFID) can exhibit weak or even negative correlation with generation FID (gFID), suggesting a possible misalignment between reconstruction quality and generative performance.

In this paper, we investigate the causes of misalignment between reconstruction quality and generation performance in latent generative models.We hypothesize that this misalignment can be explained by structural differences between the latent distributions encountered during reconstruction and generation. That is, the fact that reconstruction evaluates the decoder on encoder latents $z_e \sim \mathcal{P}_e$, whereas generation uses the same decoder on generative latents $z_g \sim \mathcal{P}_g$, as illustrated in Figure 1(a). When $\mathcal{P}_e$ and $\mathcal{P}_g$ differ, decoder behavior under encoder latents need not reflect its behavior at generation time, potentially explaining why reconstruction quality may not closely track generative performance. Moreover, comparing only the two endpoints leaves unobserved how decoder behavior changes as its inputs shift from encoder to generation-time latent distributions.

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.

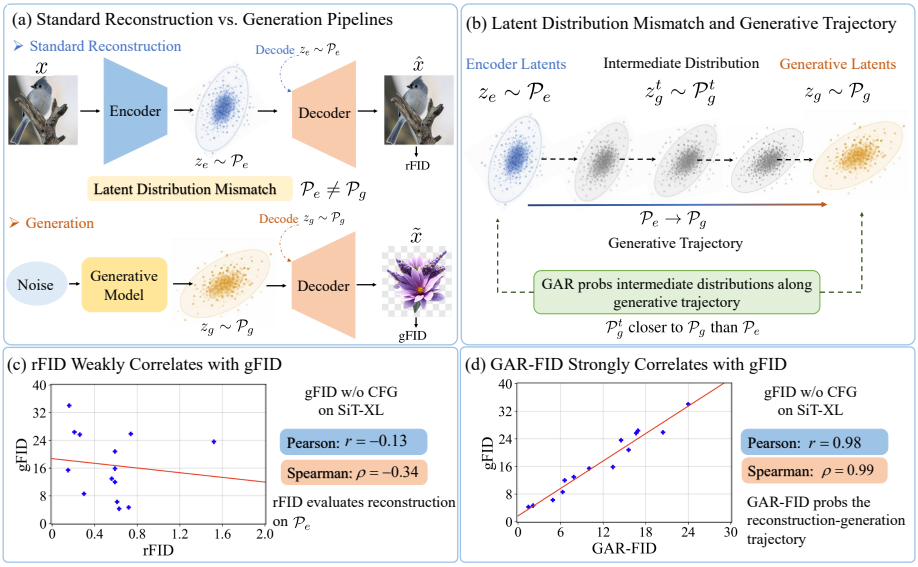


Figure 1: **Generation-aware reconstruction exposes the reconstruction-generation latent transition.** (a) Reconstruction and generation use the same decoder on different latent distributions: reconstruction decodes encoder latents $z_e \sim \mathcal{P}_e$, whereas generation decodes generator-produced latents $z_g \sim \mathcal{P}_g$. (b) GAR uses the generative process to construct a trajectory of intermediate latent distributions $\mathcal{P}_g^t$ from reconstruction toward generation. (c) In this setting, rFID is weakly correlated with gFID, with Spearman rank correlation $\rho = -0.34$. (d) GAR-FID at $\eta_t = 0.8$ exhibits strong correlation with gFID, achieving $\rho = 0.99$.

To expose this transition, we introduce *generation-aware reconstruction* (GAR), which constructs a continuous trajectory from standard reconstruction toward generation, as illustrated in Figure 1(b). Specifically, GAR perturbs encoder latents with noise and denoises them through the generative model before decoding, producing intermediate latent distributions between the reconstruction and generation endpoints. By probing the decoder along this trajectory, GAR makes the reconstruction-generation transition traceable. The resulting trajectory-based diagnostic, GAR-FID, exhibits high correlation with gFID across diverse tokenizers, generator architectures, and model scales—in contrast to the weak association observed for rFID (Figure 1(c,d)). As such, GAR complements endpoint measurements by revealing how decoder behavior evolves between reconstruction and generation.

Finally, the GAR trajectory also provides a natural opportunity for improving generation. At intermediate noise levels, GAR latents move toward generation-time distributions while retaining correspondence with their source images, thereby providing generation-aware paired supervision. Based on this insight, we propose a simple *decoder adaptation* approach that freezes the encoder and generative model and fine-tunes only the decoder on intermediate GAR latents using standard tokenizer objectives. This generation-oriented adaptation exposes the decoder to latent distributions closer to those encountered during generation while preserving meaningful image-level supervision, leading to consistent improvements in generative quality across model scales. Overall, our results establish latent distribution mismatch as a useful perspective for evaluating and improving latent generative models.

Our contributions are summarized as follows:

1. We identify a **latent distribution mismatch** between reconstruction and generation: the same decoder is evaluated on encoder latents during reconstruction but operates on generator-produced latents during generation. This provides a distributional explanation for why reconstruction quality need not track generative performance.
2. We introduce **generation-aware reconstruction (GAR)**, which constructs and probes the reconstruction-generation latent trajectory. GAR-FID quantifies decoder behavior along this trajectory and provides a reliable evaluation proxy for generative performance, exhibiting strong correlation with gFID across diverse tokenizer and generator configurations.
3. We show that this trajectory provides a principled route from **diagnosis to adaptation**: intermediate GAR latents move toward generation-time distributions while retaining source correspondence, providing paired, generation-aware supervision for decoder adaptation and consistently improving generative quality across model scales.

2 Revisiting Reconstruction and Generation

We revisit the reconstruction and generation pipelines from a latent distribution perspective. Despite sharing the same decoder, they operate on different latent distributions, creating a reconstruction-generation mismatch that cannot be understood from either endpoint alone. We therefore introduce generation-aware reconstruction (GAR) to characterize the transition between them.

2.1 Latent Distribution Mismatch Between Reconstruction and Generation

Latent visual generative models (Rombach et al., 2022) are typically composed of two separately trained components: (i) a continuous visual tokenizer (Chen et al., 2025c;b; Yang et al., 2026b; Kouzelis et al., 2025), and (ii) a latent diffusion or flow-matching model (Preechakul et al., 2022; Peebles & Xie, 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023; Geng et al., 2025). This decomposition gives rise to two distinct pathways: reconstruction and generation.

As illustrated in Figure 1(a), standard reconstruction maps an input image x to an encoder latent $z_e = \mathcal{E}_\theta(x) \sim \mathcal{P}_e$. The decoder $\mathcal{D}_\phi$ reconstructs the image as $\hat{x} = \mathcal{D}_\phi(z_e)$, and reconstruction quality is commonly measured by Fréchet Inception Distance (FID; Heusel et al., 2017), referred to as rFID. Generation, in contrast, starts from noise $\epsilon \sim \mathcal{N}(0, I)$. A generative model $\mathcal{G}_\varphi$ produces a latent $z_g = \mathcal{G}_\varphi(\epsilon) \sim \mathcal{P}_g$, which is decoded as $\tilde{x} = \mathcal{D}_\phi(z_g)$. Generative quality is measured by FID between real and generated images, referred to as gFID.

The key distinction is therefore distributional: reconstruction evaluates the decoder under $\mathcal{P}_e$, whereas generation uses it under $\mathcal{P}_g$. When these distributions differ, good decoder behavior under encoder latents does not necessarily imply good behavior on generator-produced latents. Accordingly, rFID characterizes reconstruction quality under $\mathcal{P}_e$, while gFID reflects the generative system under $\mathcal{P}_g$. This explains why reconstruction quality need not track generative performance. At the same time, $\mathcal{P}_e$ and $\mathcal{P}_g$ describe only the two endpoints and do not reveal how decoder behavior evolves as its input distribution shifts from reconstruction to generation.

2.2 Generation-Aware Reconstruction

To characterize this transition, we introduce *generation-aware reconstruction* (GAR), which constructs a controlled trajectory from standard reconstruction toward generation. As illustrated in Figure 2, given an encoder latent $z_e \sim \mathcal{P}_e$, GAR perturbs it with noise, yielding $z^t_{\text{noisy}} = a_t z_e + b_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and a_t, b_t are the noise-schedule coefficients at timestep t. We define the relative noise level as $\eta_t = b_t/(a_t + b_t)$. The generative model then denoises z^t_{noisy} from timestep t, producing $z^t_g = \mathcal{G}_\varphi(z^t_{\text{noisy}}, t) \sim \mathcal{P}^t_g$, decoded as $\bar{x} = \mathcal{D}_\phi(z^t_g)$. The noise level η_t determines the position along the reconstruction-generation trajectory. At $\eta_t = 0$, no perturbation or denoising is applied, so $z^t_g = z_e$ and $\mathcal{P}^t_g = \mathcal{P}_e$, recovering standard reconstruction. At $\eta_t = 1$, the encoder latent is fully replaced by noise and the full generative process is applied, so $z^t_g = z_g$ and $\mathcal{P}^t_g = \mathcal{P}_g$, recovering the generation endpoint. As η_t increases, the intermediate distributions $\mathcal{P}^t_g$ progressively move from $\mathcal{P}_e$ toward $\mathcal{P}_g$, while source correspondence gradually weakens. GAR thus connects the two endpoints through a controlled latent-distribution trajectory.

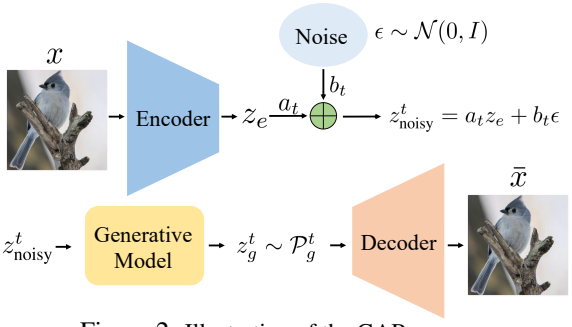


Figure 2: Illustration of the GAR process.

We define GAR-FID as the FID between source images and their GAR outputs $\bar{x}$ at a specified noise level η_t. By varying η_t, GAR-FID tracks how decoder behavior changes as its inputs move along the generator-induced trajectory from $\mathcal{P}_e$ toward $\mathcal{P}_g$: rFID characterizes the reconstruction endpoint, GAR-FID probes the intermediate distributions $\mathcal{P}^t_g$, and gFID characterizes the generation endpoint. As shown in Section 4, GAR-FID becomes strongly associated with gFID as the trajectory approaches the generation-time regime across diverse tokenizer and generator configurations.

At intermediate noise levels, GAR latents can become more generation-aware while retaining source correspondence. This paired structure allows the same trajectory used for diagnosis to also provide a training distribution for decoder adaptation, as described in Section 6.

3 Characterizing the Reconstruction-Generation Latent Shift

We now empirically examine the latent distribution perspective developed in Section 2. We first quantify the mismatch between $\mathcal{P}_e$ and $\mathcal{P}_g$. We then verify that GAR induces the expected shift from $\mathcal{P}_e$ toward $\mathcal{P}_g$ as the noise level increases. Finally, we study how the endpoint mismatch is associated with generative performance. Additional analyses of structural and perceptual reconstruction quality along the GAR trajectory are provided in Appendix C.

3.1 Latent Distribution Mismatch

To characterize the distribution mismatch between encoder latents z_e and generative latents z_g, we conduct a systematic empirical analysis using the SD-VAE (Rombach et al., 2022) encoder and iMF (Geng et al., 2026) as representative models. Since each latent is high-dimensional ($z_e \in \mathbb{R}^{32\times32\times4}$, i.e., 4096 dimensions), conventional low-dimensional visualization techniques such as 2D PCA were not sufficiently informative in our analyses. To overcome this limitation, we adopt Linear Discriminant Analysis (LDA) to identify the direction that maximizes separation between the two distributions, and visualize the resulting one-dimensional projections as Gaussian-smoothed densities, as shown in Figure 3.

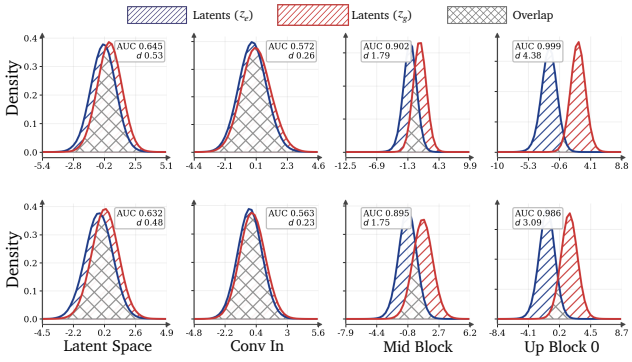


Figure 3: LDA projections of z_e and z_g for iMF-B/2 (top) and iMF-XL/2 (bottom). The vertical axis shows the Gaussian-smoothed, normalized histogram density of the one-dimensional LDA scores.

Our results reveal that the distribution gap between generative and encoder latents is smaller for iMF-XL/2 compared to iMF-B/2. This trend is further corroborated by a high-dimensional latent Fréchet distance (Latent-FD, see Appendix A), which yields a lower distance for iMF-XL/2 (91.55) than for iMF-B/2 (103.64), consistent with lower LDA AUC and reduced differences in distributional means. These results show that, within this controlled iMF comparison, the larger generative model exhibits a smaller measured mismatch. Furthermore, separability increases systematically in progressively deeper decoder feature spaces. At Decoder Up Block 0, the LDA AUC reaches 0.999, indicating near-complete separation between encoder- and generator-induced representations. We refer to this phenomenon as the *decoder amplification effect*: differences that are moderate in latent space become substantially more pronounced in downstream decoder representations. This observation suggests that the decoder can amplify latent distribution mismatch as representations propagate toward image space.

3.2 Distribution Shift Along the Reconstruction-Generation Trajectory

We next verify the latent distribution shift induced by GAR by measuring the discrepancy between $\mathcal{P}_g^t$ and the two endpoint distributions $\mathcal{P}_e$ and $\mathcal{P}_g$. Using the SD-VAE encoder (Rombach et al., 2022) and iMF-B/2 (Geng et al., 2026), we extract GAR latents on ImageNet (Deng et al., 2009) and compute the corresponding Latent-FD.

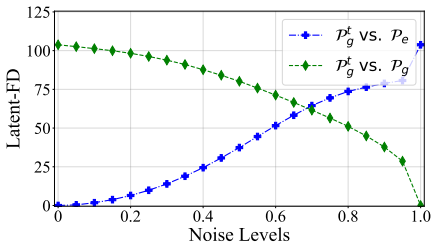


Figure 4: Latent-FD from intermediate GAR distributions $\mathcal{P}_g^t$ to the reconstruction endpoint $\mathcal{P}_e$ and generation endpoint $\mathcal{P}_g$.

As shown in Figure 4, increasing the noise level produces a consistent distributional shift: the Latent-FD between $\mathcal{P}_g^t$ and $\mathcal{P}_e$ increases, while that between $\mathcal{P}_g^t$ and $\mathcal{P}_g$ decreases. This confirms that GAR progressively moves decoder inputs away from the reconstruction regime and

toward the generation-time regime, consistent with the reconstruction-generation trajectory defined in Section 2.2. As source correspondence weakens along this trajectory, structural and perceptual reconstruction quality also gradually deteriorates; detailed results are provided in Appendix C.

3.3 Latent Distribution Mismatch and Generative Performance

We finally examine whether the endpoint mismatch between $\mathcal{P}_e$ and $\mathcal{P}_g$ is associated with generative performance. We measure the mismatch using Latent-FD and compare it with gFID, both with and without classifier-free guidance (CFG). Our evaluation uses the SD-VAE tokenizer (Rombach et al., 2022) with iMF-B/2, iMF-M/2, and iMF-XL/2 (Geng et al., 2026) on ImageNet (Deng et al., 2009).

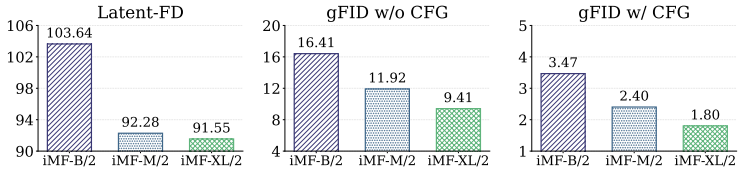


Figure 5: Latent distribution mismatch and generative performance across iMF model scales. Smaller Latent-FD between $\mathcal{P}_e$ and $\mathcal{P}_g$ coincides with lower gFID, both without and with CFG.

As shown in Figure 5, models with smaller Latent-FD consistently achieve lower gFID across the three model scales, both with and without CFG. This trend suggests a close association between latent distribution mismatch and generative performance. One possible explanation is that a larger mismatch places the decoder on generation-time latents farther from the distribution encountered during reconstruction, potentially degrading sample quality. Motivated by this empirical relationship, we next provide a theoretical perspective showing how latent distribution mismatch can contribute to an upper bound on generative error.

Proposition 1 (A Lipschitz Bound Linking Latent-FD and gFID). *Let $\psi : \mathcal{X} \to \mathbb{R}^{2048}$ denote the Inception feature extractor used for FID, and define the decoder-feature map $T = \psi \circ \mathcal{D}_\phi$. Suppose that $\mathcal{P}_e$ and $\mathcal{P}_g$ are Gaussian latent distributions. Assume that the decoder $\mathcal{D}_\phi$ and feature extractor ψ are Lipschitz with constants $L_\mathcal{D}$ and L_ψ, respectively. Then we have*

$$\sqrt{\text{gFID}} \le \sqrt{\text{rFID}} + L_\psi L_\mathcal{D} \sqrt{\text{Latent-FD}(\mathcal{P}_e, \mathcal{P}_g)}. \tag{1}$$

The proof is provided in Appendix B. Proposition 1 provides a theoretical lens for the empirical trend in Figure 5. Under the stated Gaussian and Lipschitz assumptions, gFID is upper-bounded by the reconstruction term together with a contribution that grows with latent distribution mismatch and is scaled by the decoder-feature Lipschitz constant. The bound does not imply that reducing Latent-FD must reduce gFID; rather, it identifies latent mismatch as one factor that can enlarge the gap between reconstruction and generation, consistent with the empirical association above.

4 Tracking Decoder Behavior from Reconstruction to Generation

We now examine how decoder behavior evolves along the reconstruction-generation trajectory and how GAR-FID captures this evolution. We first use a controlled setting with a fixed tokenizer to isolate variation induced by the generative model and connect GAR-FID to the underlying latent distribution shift. We then examine whether the relationship between GAR-FID and generative performance remains stable when both tokenizers and generators vary.

4.1 Controlled Study With a Fixed Tokenizer

We begin with a controlled study that fixes the SD-VAE tokenizer (Rombach et al., 2022) while varying the generative model across three scales: iMF-B/2, iMF-M/2, and iMF-XL/2 (Geng et al., 2026). Fixing the tokenizer places all latent distributions in a shared latent space, making $\text{Latent-FD}(\mathcal{P}_e, \mathcal{P}_g)$ directly comparable across generators

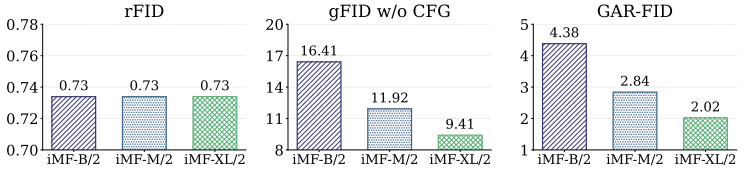


Figure 6: Metric behavior under a fixed tokenizer (SD-VAE) as generator scale increases from iMF-B/2 to iMF-XL/2. rFID remains unchanged by construction, while gFID and GAR-FID vary with the generative model.

and isolating variation induced by the generative model. Figure 6 shows that gFID decreases substantially with increasing generator scale, while rFID remains unchanged because the tokenizer is fixed. This invariance is expected: rFID characterizes reconstruction under $\mathcal{P}_e$ and does not depend on the generative model. GAR-FID, in contrast, varies consistently with generator scale because its intermediate distributions $\mathcal{P}_g^t$ depend on the corresponding generator. This controlled setting therefore highlights their distinct roles: rFID characterizes tokenizer reconstruction, whereas GAR-FID reflects generator-induced variation along the reconstruction-generation trajectory.

Connecting GAR-FID to Latent Distribution Mismatch. Section 3.3 showed that, within this fixed-tokenizer setting, a larger endpoint mismatch between $\mathcal{P}_e$ and $\mathcal{P}_g$ is empirically associated with worse generative performance. We summarize this observed relationship as $\text{Latent-FD}(\mathcal{P}_e, \mathcal{P}_g) \uparrow \rightsquigarrow \text{gFID} \uparrow$. Hereafter, $\rightsquigarrow$ denotes an empirical relationship rather than a causal implication.

We next examine how GAR-FID connects to this endpoint relationship through the intermediate distributions $\mathcal{P}_g^t$. As shown in Figure 7, GAR-FID exhibits near-perfect rank correlation with $\text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e)$ across the evaluated noise levels (Spearman's $\rho = 1.0$, $p < 10^{-8}$). Thus, as GAR progressively shifts decoder inputs away from $\mathcal{P}_e$, GAR-FID changes systematically with the underlying latent shift: $\text{GAR-FID} \uparrow \rightsquigarrow \text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e) \uparrow$. Moreover, under the fixed-tokenizer setting, $\text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e)$ preserves the model ordering of the endpoint mismatch $\text{Latent-FD}(\mathcal{P}_g, \mathcal{P}_e)$ across noise levels (Appendix D). We summarize this ordering relationship as $\text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e) \uparrow \rightsquigarrow \text{Latent-FD}(\mathcal{P}_g, \mathcal{P}_e) \uparrow$.

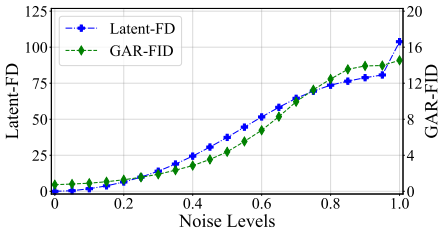


Figure 7: GAR-FID (w/o CFG) closely tracks $\text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e)$ across noise levels.

Combining these observations yields the following empirical association chain:

$$\text{GAR-FID} \uparrow \rightsquigarrow \text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e) \uparrow \rightsquigarrow \text{Latent-FD}(\mathcal{P}_g, \mathcal{P}_e) \uparrow \rightsquigarrow \text{gFID} \uparrow . \tag{2}$$

The three links respectively capture the correlation between GAR-FID and intermediate latent mismatch, the preservation of generator ordering from intermediate to endpoint mismatch, and the empirical association between endpoint mismatch and generative performance. This chain provides an empirical explanation for why GAR-FID becomes increasingly associated with gFID as the trajectory approaches the generation-time regime.

Beyond the Controlled Setting. Because the tokenizer is fixed, all latent distributions share the same representation space and Latent-FD can be compared directly across generators. When tokenizers vary, their latent spaces differ and Latent-FD is no longer directly comparable. We therefore turn to metric correlations to assess whether the GAR-FID–gFID relationship remains stable across tokenizer and generator variations.

4.2 Robustness Across Tokenizers and Generators

We next examine whether the relationship between GAR-FID and gFID remains stable when both the tokenizer and generative model vary. Following iFID (Xu et al., 2026), we evaluate 13 tokenizers with two latent generative models, SiT-B and SiT-XL (Ma et al., 2024), and compute rFID, GAR-FID, and gFID for each tokenizer-generator configuration. All metrics are reported in Table 4 and Table 5, with detailed experimental settings provided in Appendix H.

As shown in Table 1, rFID exhibits negative correlation with gFID across SiT-B, SiT-XL, and mixed-generator configurations, consistent with the reconstruction-generation discrepancy observed in prior work (Yao & Wang, 2025; Kouzelis et al., 2025; Ye et al., 2025; Skorokhodov et al., 2025; Chen et al., 2025b; Xu et al., 2026). By contrast, the correlation between GAR-FID and gFID strengthens as η_t increases. At $\eta_t = 0.8$, GAR-FID already exhibits consistently high correlation across SiT-B, SiT-XL, and their mixed configurations, with SRCC ranging from 0.94 to 0.99.

The mixed SiT-B/XL setting further tests whether the GAR-FID-gFID relationship remains stable across generator scales. Correlations computed within a single generative model, such as SiT-B or SiT-XL (Ma et al., 2024), primarily capture tokenizer variation under a fixed generator scale. By contrast, the mixed setting jointly varies tokenizer choice and generator scale, providing a more

Table 1: Correlation between FID-based metrics and generative performance (gFID). Correlations are computed across models obtained by training SiT with 13 pre-trained VAEs. We report Pearson correlation (PCC) and Spearman rank correlation (SRCC), with and without CFG, for SiT-B, SiT-XL, and their mixed configurations. The correlation between GAR-FID and gFID strengthens as the noise level η_t increases, while rFID remains negatively correlated with gFID across settings. **Bold** indicates the best results. † denotes results from (Xu et al., 2026), and * indicates our reproduced results. Full results are provided in Table 7 in Appendix H.

Metrics	gFID SiT/B				gFID SiT/XL				gFID SiT/mixed			
	w/o CFG		w/ CFG		w/o CFG		w/ CFG		w/o CFG		w/ CFG	
	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)
rFID†	-0.04	-0.31	-0.07	-0.31	-0.06	-0.21	-0.15	-0.31	–	–	–	–
iFID†	0.85	0.86	0.82	0.84	0.89	0.91	0.88	0.92	–	–	–	–
rFID*	-0.11	-0.36	-0.11	-0.30	-0.13	-0.34	-0.15	-0.39	-0.10	-0.30	-0.10	-0.28
iFID*	0.86	0.83	0.80	0.79	0.90	0.88	0.88	0.90	0.72	0.77	0.64	0.70
GAR-FID ($\eta_t = 0.6$)	0.81	0.79	0.77	0.78	0.79	0.84	0.80	0.79	0.86	0.89	0.84	0.85
GAR-FID ($\eta_t = 0.7$)	0.93	0.90	0.89	0.87	0.91	0.90	0.91	0.86	0.95	0.96	0.93	0.91
GAR-FID ($\eta_t = 0.8$)	0.98	0.96	0.96	0.94	0.98	0.99	0.97	0.96	0.99	0.99	0.97	0.95
GAR-FID ($\eta_t = 0.9$)	**1.00**	**0.99**	**0.98**	0.98	**1.00**	**1.00**	**0.98**	**0.96**	**1.00**	**1.00**	**0.98**	**0.97**
GAR-FID ($\eta_t = 1.0$)	**1.00**	**0.99**	**0.98**	**0.99**	**1.00**	**1.00**	**0.98**	**0.96**	**1.00**	**1.00**	**0.98**	**0.97**

demanding test that better reflects cross-system comparisons in latent generative modeling. In this setting, the correlation of iFID with gFID decreases, whereas GAR-FID maintains consistently strong correlation; for example, at $\eta_t = 0.8$, GAR-FID achieves SRCCs of 0.99 and 0.95 without and with CFG, respectively, compared with 0.77 and 0.70 for iFID. These results indicate that the GAR-FID–gFID relationship remains robust across heterogeneous tokenizer-generator configurations.

Leave-one-out robustness across tokenizers. To test whether the GAR-FID–gFID correlation is driven by any individual tokenizer, we perform a leave-one-out analysis over the 14 SiT-XL tokenizer configurations, recomputing SRCC after omitting each tokenizer. Correlations are more sensitive to individual tokenizers at low noise levels but become uniformly high at $\eta_t \geq 0.8$. At $\eta_t = 0.8$, the leave-one-out SRCC is 0.99 for every omission without CFG and ranges from 0.96 to 0.97 with CFG, showing that the strong correlation is not driven by any single tokenizer. Full results are provided in Tables 8 and 9 in Appendix H.

Computational cost. For multi-step generators, intermediate GAR-FID can additionally be more computationally efficient than endpoint gFID, as it requires only a partial generative trajectory. Detailed runtime comparisons are provided in Table 3 in Appendix F.

5 A Unified Perspective on Latent Distribution Mismatch

As discussed in Section 4, latent distribution mismatch provides a useful explanation for the gap between reconstruction and generation: the same decoder operates under different latent distributions in the two regimes. GAR makes this shift observable by exposing decoder behavior along the intermediate distributions $\mathcal{P}_g^t$ between $\mathcal{P}_e$ and $\mathcal{P}_g$. Beyond this diagnostic role, GAR reveals a broader insight: *generation-oriented evaluation and training should account for the latent distributions encountered during generation*.

This perspective naturally leads to two complementary ways to address the latent distribution mismatch. The first is to improve the latent generative process, such that the resulting distribution $\mathcal{P}_g$ more faithfully approximates the encoder-induced distribution $\mathcal{P}_e$ that the generator is trained to model. The second is to adapt the decoder, such that it reliably operates on samples drawn from $\mathcal{P}_g$ or intermediate generative distributions $\mathcal{P}_g^t$. These correspond to reducing the mismatch on the generator side, or adapting the decoder to the latent distributions encountered during generation.

A large body of prior work has focused on the first direction, improving the latent generative process through multiple complementary avenues, including stronger architectures and scaling strategies, improved objectives and transport formulations, enhanced multimodal conditioning, and efficient sampling or distillation techniques (Peebles & Xie, 2023; Esser et al., 2024; Mei et al., 2024; Tong et al., 2026; Chen et al., 2024; Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023; Geng et al., 2025; Wu et al., 2025; Labs, 2024; Xie et al., 2025; Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Song & Dhariwal, 2024; Sauer et al., 2024). Additional approaches leverage reinforcement learning, preference-based optimization, or richer latent representations to further improve generative quality (Yu et al., 2025; Liu et al., 2025; Xue et al., 2025; Lee et al., 2023; Fan et al., 2023; Xu et al., 2023; Wallace et al., 2023; Zheng et al., 2026; Leng et al., 2025; Yang et al., 2026b; Chen et al., 2025b;c; Ye et al., 2025). Although these methods do not necessarily

explicitly optimize the discrepancy between $\mathcal{P}_e$ and $\mathcal{P}_g$, they broadly improve the latent representation being modeled. In the standard two-stage setting, where the latent generator is trained to model the encoder-induced distribution $\mathcal{P}_e$, improving its modeling fidelity can be viewed as reducing the generator-side mismatch between $\mathcal{P}_g$ and $\mathcal{P}_e$. Consistent with this perspective, our controlled study shows that stronger generators exhibit a smaller measured discrepancy between $\mathcal{P}_e$ and $\mathcal{P}_g$.

Our results motivate the complementary decoder-side perspective. Even as the generative model improves, generation still relies on a decoder trained under $\mathcal{P}_e$ but operating on latents drawn from $\mathcal{P}_g$. Improving the generator therefore does not directly optimize decoder behavior under generation-time latent distributions. Moreover, as evidenced by the *decoder amplification effect*, discrepancies in latent space can become substantially more pronounced in downstream decoder representations. Consequently, generative performance depends not only on the mismatch between $\mathcal{P}_e$ and $\mathcal{P}_g$, but also on how reliably $\mathcal{D}_\phi$ operates as its input distribution shifts toward the generation-time regime.

This motivates aligning decoder training with generation-relevant latent distributions. Intermediate GAR latents provide a practical way to do so: they move toward the generation-time regime while retaining correspondence with their source images, thereby preserving paired supervision that is unavailable for fully generated latents. We instantiate this idea through *decoder adaptation* (Section 6), which fine-tunes the decoder on intermediate GAR latents while keeping the encoder and generative model fixed. In this way, the same GAR trajectory used to diagnose the reconstruction-generation shift also provides a natural training distribution for improving downstream generative performance.

6 Decoder Adaptation via Latent Distribution Alignment

We instantiate the decoder-side perspective developed above through *decoder adaptation* (DA), which keeps the encoder $\mathcal{E}_\theta$ and generative model $\mathcal{G}_\varphi$ fixed and updates only the decoder $\mathcal{D}_\phi$ on intermediate GAR latents. Given an image x, we first obtain its encoder latent $z_e = \mathcal{E}_\theta(x) \sim \mathcal{P}_e$. Following Section 2.2, we perturb z_e as $z^t_{\text{noisy}} = a_t z_e + b_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, and denoise it through the fixed generative model to obtain $z^t_g = \mathcal{G}_\varphi(z^t_{\text{noisy}}, t) \sim \mathcal{P}^t_g$, which is decoded as $\bar{x}(\phi) = \mathcal{D}_\phi(z^t_g)$. Only the decoder parameters ϕ are optimized during adaptation.

The choice of intermediate GAR latents is central to DA. Standard decoder training operates on encoder latents $z_e \sim \mathcal{P}_e$, which retain source correspondence but do not expose the decoder to generation-relevant latent distributions. Fully generated latents $z_g \sim \mathcal{P}_g$, in contrast, reflect the generation-time distribution but have no paired target images. Intermediate GAR latents bridge these regimes: as η_t increases, $\mathcal{P}^t_g$ moves toward $\mathcal{P}_g$, while intermediate noise levels retain correspondence with the source image x. DA therefore trains the decoder on generation-relevant latents while preserving paired supervision. In this sense, latent distribution alignment is achieved by changing the decoder's training distribution rather than introducing an explicit distribution-matching objective. We use noise levels $\eta_t \in [0.25, 0.45]$ for adaptation, with sensitivity studied in Appendix E.

To optimize the decoder, we adopt standard objectives commonly used in modern visual tokenizers (Tian et al., 2024; Fang et al., 2026; Yu et al., 2026), combining pixel-level, perceptual, adversarial, and feature-statistics supervision:

$$\mathcal{L}_{\text{DA}}(\phi) = \lambda_1 \|\bar{x}(\phi) - x\|_2^2 + \lambda_2 \mathcal{L}_{\text{LPIPS}} + \lambda_3 \mathcal{L}_{\text{GAN}} + \lambda_4 \mathcal{L}_{\text{Gram}}. \tag{3}$$

The pixel, LPIPS, and Gram terms exploit the correspondence between the GAR output $\bar{x}(\phi)$ and its source image x. Specifically, $\mathcal{L}_{\text{LPIPS}}$ measures perceptual similarity using deep feature representations (Zhang et al., 2018; Simonyan & Zisserman, 2015), while $\mathcal{L}_{\text{Gram}}$ matches feature covariance statistics through Gram matrices (Gatys et al., 2016; Lu et al., 2025). The adversarial term $\mathcal{L}_{\text{GAN}}$ encourages realistic decoded outputs and is implemented using a hinge-based adversarial objective (Isola et al., 2017; Lim & Ye, 2017; Karras et al., 2019; 2020). We set $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.7$, and $\lambda_4 = 10.0$ across all experiments. Importantly, these objectives are standard; the key change in DA is the training distribution presented to the decoder, replacing encoder-only latents with paired, generation-aware GAR latents.

7 Experiments

Experimental Setup. To evaluate decoder adaptation, we conduct experiments on four iMF model scales, iMF-B/2, iMF-M/2, iMF-L/2, and iMF-XL/2 (Geng et al., 2026), all sharing the SD-VAE

Table 2: **Effect of decoder adaptation (DA) on generative performance.** We report gFID without and with CFG across iMF model scales under two adaptation settings: DA w/o CFG, trained for 10 epochs using GAR latents generated without CFG, and CFG-aware DA, trained for one epoch using GAR latents generated with CFG. †: results reported in iMF (Geng et al., 2026). ‡: our reproduction using the official iMF protocol. ⋆: our reproduction using the OpenAI protocol. See Appendix J for protocol details.

Setting	gFID w/o CFG (↓)				gFID w/ CFG (↓)			
	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2
w/o DA†	-	-	-	-	3.39	2.27	1.86	1.72
w/o DA‡	16.58	11.92	9.28	9.78	3.37	2.27	1.86	1.73
DA w/o CFG (10 ep.)‡	12.21 $\uparrow_{4.37}$	9.52 $\uparrow_{2.40}$	7.34 $\uparrow_{1.94}$	7.47 $\uparrow_{2.31}$	2.90 $\uparrow_{0.47}$	2.12 $\uparrow_{0.15}$	1.65 $\uparrow_{0.21}$	1.56 $\uparrow_{0.17}$
CFG-aware DA(1 ep.)‡	-	-	-	-	2.81 $\uparrow_{0.56}$	1.98 $\uparrow_{0.29}$	1.58 $\uparrow_{0.28}$	1.50 $\uparrow_{0.23}$
w/o DA⋆	16.41	11.92	9.26	9.41	3.47	2.40	1.90	1.80
DA w/o CFG (10 ep.)⋆	12.41 $\uparrow_{4.00}$	9.66 $\uparrow_{2.26}$	7.41 $\uparrow_{1.85}$	7.61 $\uparrow_{1.80}$	3.12 $\uparrow_{0.35}$	2.30 $\uparrow_{0.10}$	1.80 $\uparrow_{0.10}$	1.73 $\uparrow_{0.07}$
CFG-aware DA (1 ep.)⋆	-	-	-	-	3.01 $\uparrow_{0.46}$	2.19 $\uparrow_{0.21}$	1.70 $\uparrow_{0.20}$	1.61 $\uparrow_{0.19}$

tokenizer (Rombach et al., 2022). During adaptation, the tokenizer encoder and latent generative model are frozen, and only the decoder is updated. We consider two settings: DA w/o CFG, which uses GAR latents generated without CFG and trains for 10 epochs, and CFG-aware DA, which uses CFG during GAR construction and trains for one epoch. Experiments are conducted on ImageNet-1k at 256×256 resolution (Deng et al., 2009). Additional implementation details are provided in Appendix K. Code and checkpoints will be publicly released.

Main Results. As shown in Table 2, DA w/o CFG consistently improves generative performance across all model scales, with larger gFID gains observed without CFG and for smaller models. CFG-aware DA further improves guided generation despite using only one epoch of adaptation, outperforming or matching the 10-epoch DA w/o CFG setting across all reported gFID evaluations with CFG. The gains remain consistent across both the official iMF and OpenAI evaluation protocols, suggesting that the improvement is not tied to a particular FID implementation. Importantly, neither adaptation setting introduces additional inference cost, as only the decoder weights are updated while the inference pipeline and generative model remain unchanged.

Additional Analyses. We provide further analyses of the 10-epoch DA w/o CFG setting in Appendix L. First, repeated evaluation over three random seeds shows consistently small variance and stable gFID improvements across model scales (Table 12). Second, after adaptation, rFID varies more consistently with gFID across generator scales, indicating a closer correspondence between reconstruction and generation behavior (Figure 12). Third, the noise-level sensitivity analysis in Table 13 reveals a clear trade-off between generation awareness and source correspondence, with $\eta_t = 0.5$ achieving the best gFID without CFG across all scales under fixed noise. Fourth, reconstruction results in Table 14 show that DA trades standard reconstruction fidelity for improved behavior as GAR approaches the generation-time regime. Finally, qualitative comparisons show that DA reduces visual artifacts and yields richer details and more semantically consistent samples (Figures 13-23).

Generalization Study. Beyond iMF (Geng et al., 2026), we further apply the same one-epoch CFG-aware decoder adaptation to MeanFlow-XL/2 (Geng et al., 2025) and DMF-XL/2+ (Lee et al., 2026). Under the official iMF evaluation protocol, decoder adaptation consistently improves gFID from 3.43 to 2.96 on MeanFlow-XL/2 and from 2.16 to 1.90 on DMF-XL/2+. These results show that the benefits of decoder adaptation extend beyond iMF and remain effective across both from-scratch and distillation-based generative models, supporting its generalization across different generator training paradigms. A broader system-level comparison on class-conditional ImageNet 256×256 is provided in Table 15 of Appendix L.

8 Conclusion

This work presents a latent distribution perspective on the reconstruction-generation discrepancy in latent generative models. We identify that reconstruction and generation evaluate the same decoder under fundamentally different latent distributions, and introduce GAR to make this shift observable through a controlled trajectory. GAR-FID provides a reliable proxy for generative performance, and decoder adaptation on intermediate GAR latents consistently improves generation quality across scales. We hope this perspective encourages future work to account for generation-time latent distributions in tokenizer evaluation and training.

9 AI USE STATEMENT

Generative AI tools were used in this work solely to assist with language refinement and editorial improvements. They did not contribute to the development of the research ideas, methodological design, theoretical derivations, experimental setup, data preparation, implementation, or interpretation of the scientific results. Any AI-assisted wording was carefully checked and edited by the authors. The authors remain fully responsible for the accuracy, integrity, and final presentation of the manuscript.

10 ETHICS STATEMENT

This research does not involve human-subject studies, user studies, crowdsourcing, participant interaction, or the collection of new personal or sensitive data. Therefore, Institutional Review Board (IRB) approval or an equivalent ethics review was not required.

Our experiments are conducted using existing publicly available datasets and model architectures, including ImageNet. The manuscript presents a limited number of images generated by generative models trained on ImageNet for qualitative visualization and analysis of the experimental results. We do not collect additional information about the depicted content, attempt to identify individuals, or conduct analyses related to personal attributes.

This work does not introduce a new dataset or collect new human-subject data. We are not aware of additional ethical risks introduced by the proposed methodology beyond those associated with the publicly available datasets and models used in our experiments. All resources are used for research purposes in accordance with their applicable terms and conditions.

11 REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility and independent verification of the empirical results reported in this work. The supplementary material includes the implementation code used in our experiments, together with instructions for reproducing the main results.

For the analytical studies in Sections 3 and 4, we describe the experimental setups and evaluation procedures in detail, and provide additional implementation details for extracting intermediate decoder representations in Appendix I. For the decoder adaptation experiments, Appendix K documents the model configurations, training and optimization settings, computational resources, and inference and evaluation protocols. We also report the classifier-free guidance (CFG) parameters used for gFID evaluation in Table 11. In addition, Appendix J details the differences between the official iMF evaluation protocol and the commonly used OpenAI gFID protocol, including implementation choices that can affect reported FID values.

Together, these materials provide the information needed to reproduce the main experimental results and verify the reported findings. Upon publication, we will publicly release the complete codebase, model checkpoints, configuration files, and execution instructions to further facilitate reproducibility and future research.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.

Hansheng Chen, Kai Zhang, Hao Tan, Leonidas J. Guibas, Gordon Wetzstein, and Sai Bi. pi-flow: Policy-based few-step generation via imitation distillation. *ArXiv*, 2025a.

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. In *ICML*, 2025b.

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 *CVPR*, 2025c.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

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 *ICML*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, P. Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 2023.

Xianghong Fang, Yuan Yuan, Dehan Kong, and Tim G. J. Rudner. VQ-transplant: Efficient plug-and-play vq-module integration for pre-trained visual tokenizers. In *ICLR*, 2026.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *NeurIPS*, 2025.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *CVPR*, 2026.

Jiatao Gu, Tianrong Chen, David Berthelot, Huangjie Zheng, Yuyang Wang, Ruixiang Zhang, Laurent Dinh, Miguel Angel Bautista, Joshua M. Susskind, and Shuangfei Zhai. Starflow: Scaling latent normalizing flows for high-resolution image synthesis. In *NeurIPS*, 2025.

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 *NeurIPS*, 2017.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, 2023.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. In *CVPR*, 2025.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

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

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

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, P. Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *ArXiv*, 2023.

Kyungmin Lee, Sihyun Yu, and Jinwoo Shin. Decoupled meanflow: Turning flow models into flow maps for accelerated sampling. In *ICLR*, 2026.

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*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. In *ICLR*, 2025.

Jae Hyun Lim and J. C. Ye. Geometric gan. *ArXiv*, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *ArXiv*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. *ArXiv*, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *ArXiv*, 2023.

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 *ECCV*, 2024.

Kangfu Mei, Zhengzhong Tu, Mauricio Delbracio, Hossein Talebi, Vishal M. Patel, and Peyman Milanfar. Bigger is not always better: Scaling properties of latent diffusion models. *TMLR*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

William S. Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Yansong Peng, Kai Zhu, Yu Liu, Ping Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. Flow-anchored consistency models. In *ICLR*, 2026.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *ICCV*, 2025.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, A. Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *ICML*, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *ArXiv*, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *ArXiv*, 2026.

Michael Tschannen, André Susano Pinto, and Alexander Kolesnikov. Jetformer: An autoregressive generative model of raw images and text. In *ICLR*, 2025.

Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *CVPR*, 2021.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq R. Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2023.

Shuai Wang, Zhichao Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *ArXiv*, 2025a.

Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective. *ArXiv*, 2025b.

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*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

Tongda Xu, Min He, Shady Abu-Hussein, José Miguel Hernández-Lobato, Haotian Zhang, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid. *ArXiv*, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation. *ArXiv*, 2025.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yong-Hong Tian, and Yue Wang. Representation fréchet loss for visual generation. *ArXiv*, 2026a.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers. In *ICLR*, 2026b.

Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

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

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *ICCV*, 2024.

Qihang Yu, Qihao Liu, Ju He, Xinyang Zhang, Yang Liu, Liang-Chieh Chen, and Xi Chen. Autoregressive image generation with masked bit modeling. *ArXiv*, 2026.

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 *ICLR*, 2025.

Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. *ArXiv*, 2019.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. Alphaflow: Understanding and improving meanflow models. *ArXiv*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *ICML*, 2025.

A Latent Fréchet Distance

To measure the distributional discrepancy in latent space, we adopt the Fréchet distance between two sets of latent variables, $z_i^a \sim \mathcal{P}_A$ and $z_i^b \sim \mathcal{P}_B$. Specifically, we estimate the empirical mean and covariance of the two latent distributions, denoted by (μ_1, Σ_1) and (μ_2, Σ_2), respectively. Based on these statistics, we define the latent Fréchet distance (Latent-FD) as

$$\mathrm{FD}(\mathcal{P}_A, \mathcal{P}_B) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\Big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\Big), \tag{4}$$

This metric provides a well-established measure of discrepancy between two distributions by capturing both first-order (mean) and second-order (covariance) statistics, and is closely related to the widely used Fréchet Inception Distance (FID). Both metrics compute the Fréchet distance under a Gaussian assumption, differing only in the representation space used to estimate these statistics. Specifically, FID operates in a learned semantic feature space, where each image is mapped to a 2048-dimensional representation using a pretrained InceptionV3 (Szegedy et al., 2016) network. In contrast, our Latent-FD is computed directly in the model latent space, using latents from the encoder or generative process. This distinction allows Latent-FD to directly quantify distributional discrepancies relevant to the generative pipeline, while retaining the same underlying geometric interpretation as FID.

B Proof of Proposition 1

Let $\mathcal{P}_{\mathrm{data}}$ denote the real image distribution, let ψ be the Inception feature extractor used for FID, and define $T = \psi \circ \mathcal{D}_\phi$. We write

$$\mathcal{Q}_{\mathrm{data}} = \psi_{\#}\mathcal{P}_{\mathrm{data}}, \qquad \mathcal{Q}_e = T_{\#}\mathcal{P}_e, \qquad \mathcal{Q}_g = T_{\#}\mathcal{P}_g.$$

For any distribution $\mathcal{Q}$ with finite second moment, let $\Gamma(\mathcal{Q}) = \mathcal{N}(\mu_{\mathcal{Q}}, \Sigma_{\mathcal{Q}})$ be the Gaussian distribution with the same mean and covariance as $\mathcal{Q}$. The FID between two feature distributions $\mathcal{Q}$ and $\mathcal{R}$ is the squared 2-Wasserstein distance between their Gaussian approximations:

$$\mathrm{FID}(\mathcal{Q}, \mathcal{R}) = W_2^2(\Gamma(\mathcal{Q}), \Gamma(\mathcal{R}))\,.$$

Therefore, using the triangle inequality for W_2,

$$\begin{aligned}\sqrt{\mathrm{gFID}} &= W_2(\Gamma(\mathcal{Q}_{\mathrm{data}}), \Gamma(\mathcal{Q}_g)) \\ &\le W_2(\Gamma(\mathcal{Q}_{\mathrm{data}}), \Gamma(\mathcal{Q}_e)) + W_2(\Gamma(\mathcal{Q}_e), \Gamma(\mathcal{Q}_g)) \\ &= \sqrt{\mathrm{rFID}} + \sqrt{\mathrm{FID}(\mathcal{Q}_e, \mathcal{Q}_g)}.\end{aligned}$$

The Gaussian Fréchet distance between the moment-matched Gaussians is upper bounded by the true Wasserstein distance between the original distributions, since

$$\mathrm{FID}(\mathcal{Q}_e, \mathcal{Q}_g) \le W_2^2(\mathcal{Q}_e, \mathcal{Q}_g).$$

Moreover, because T is L-Lipschitz, pushing any coupling of $\mathcal{P}_e$ and $\mathcal{P}_g$ through T gives

$$W_2(\mathcal{Q}_e, \mathcal{Q}_g) = W_2(T_{\#}\mathcal{P}_e, T_{\#}\mathcal{P}_g) \le L\, W_2(\mathcal{P}_e, \mathcal{P}_g).$$

Finally, since $\mathcal{P}_e$ and $\mathcal{P}_g$ are Gaussian, the latent Fréchet distance equals the squared Wasserstein distance between them:

$$\text{Latent-FD}(\mathcal{P}_e, \mathcal{P}_g) = W_2^2(\mathcal{P}_e, \mathcal{P}_g).$$

Combining the preceding inequalities yields

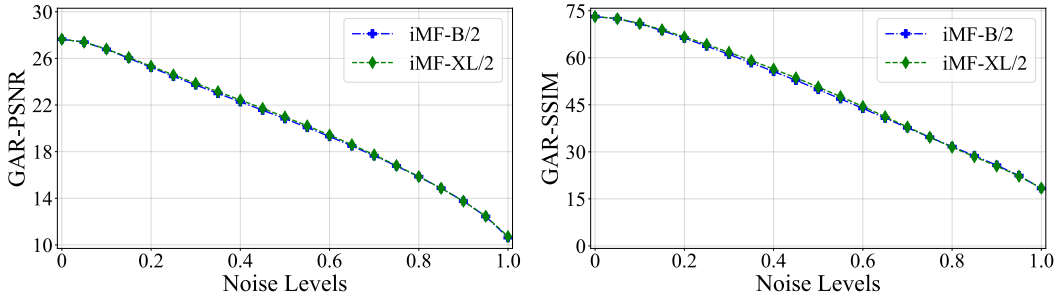
$$\sqrt{\mathrm{gFID}} \le \sqrt{\mathrm{rFID}} + L_\psi L_{\mathcal{D}} \sqrt{\text{Latent-FD}(\mathcal{P}_e, \mathcal{P}_g)}.$$

C Structural and Perceptual Quality under GAR

We provide additional analysis of reconstruction quality under GAR using both structural and perceptual metrics. This analysis complements the main text by illustrating how latent distribution shift manifests in reconstruction quality. For structural fidelity, we consider peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). For perceptual quality, we use Fréchet Inception Distance (FID; Heusel et al., 2017) and learned perceptual image patch similarity (LPIPS; Zhang et al., 2018). When computed under the GAR framework, we refer to these metrics as GAR-PSNR, GAR-SSIM, GAR-FID, and GAR-LPIPS, respectively. All experiments are conducted on the ImageNet validation set (Deng et al., 2009), using the SD-VAE tokenizer (Rombach et al., 2022) and two representative flow-matching models, iMF-B/2 and iMF-XL/2 (Geng et al., 2026).

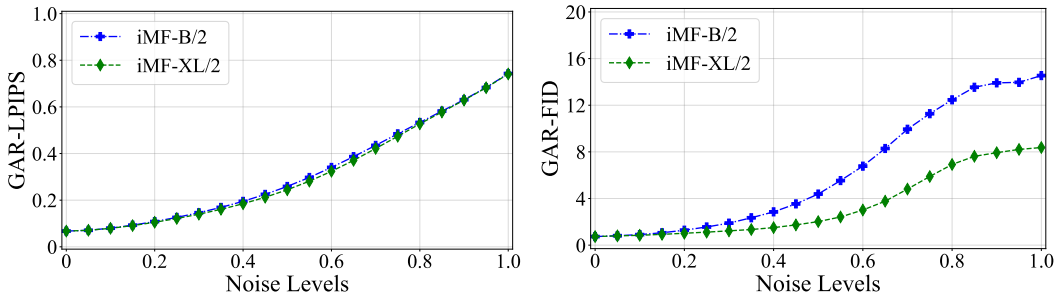


Figure 8: Illustration of the progressive loss of source structure along the GAR trajectory. Each row shows an original input image (leftmost column) followed by GAR reconstructions at noise levels $\eta_t \in \{0.0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ from left to right. As the noise level increases, reconstructions progressively deviate from the original inputs: while coarse semantic content may remain loosely recognizable at intermediate noise levels, fine-grained structural details and local textures are gradually degraded. At high noise levels (e.g., $\eta_t \geq 0.6$), the reconstructed images no longer preserve the original structure, resulting in semantically plausible but structurally inconsistent outputs. This progression illustrates how source-level structural fidelity weakens as GAR moves toward the generation endpoint.



(a) PSNR vs. noise level. (b) SSIM vs. noise level.

Figure 9: Structural quality under GAR across noise levels. PSNR (left) and SSIM (right) degrade consistently as noise increases. Increasing generative model capacity (iMF-B/2 vs. iMF-XL/2) does not improve structural quality, as both follow nearly identical trends.



(a) LPIPS vs. noise level. (b) FID vs. noise level.

Figure 10: Perceptual quality under GAR across noise levels. LPIPS (left) and FID without CFG (right) both degrade as noise increases. Increasing model capacity (iMF-B/2 vs. iMF-XL/2) has little effect on LPIPS but significantly improves FID across noise levels.

Structural Quality. As depicted in Figure 9, both structural metrics, PSNR and SSIM, decline consistently as the noise level increases. At sufficiently large noise levels (e.g., $\eta_t \geq 0.6$), GAR outputs deviate substantially from the original inputs: coarse semantic content may remain partially recognizable, while fine-grained structure and local textures are increasingly altered, as illustrated in Figure 8. This behavior is expected as GAR moves toward the generation endpoint and source correspondence progressively weakens. Increasing generative model capacity (iMF-B/2 vs. iMF-XL/2) has little effect on these structural metrics, with both models exhibiting nearly identical trends across noise levels.

Perceptual Quality. As shown in Figure 10, both perceptual metrics, LPIPS and FID without CFG, increase consistently as the noise level grows (higher values indicate worse quality). Notably,

increasing generative model capacity (iMF-B/2 vs. iMF-XL/2) has little effect on LPIPS, as the curves nearly overlap, but leads to a substantial improvement in FID across all noise levels. This indicates that increasing generative model capacity has little effect on input-level perceptual similarity as measured by LPIPS, while yielding better distribution-level FID.

Latent Distribution Shift Perspective. The trends above are consistent with the latent distribution shift induced along the GAR trajectory. As shown in Figure 4, increasing the noise level moves $\mathcal{P}_g^t$ progressively away from $\mathcal{P}_e$ and toward $\mathcal{P}_g$. At the same time, source correspondence weakens, leading GAR outputs to become less similar to their corresponding input images. This behavior is reflected by the decreasing PSNR and SSIM and the increasing LPIPS and FID observed along the trajectory. Together, these results illustrate how the transition from reconstruction toward generation is accompanied by a progressive loss of source-level structural and perceptual fidelity.

D Intermediate Latent-FD as an Indicator of Final Latent Mismatch

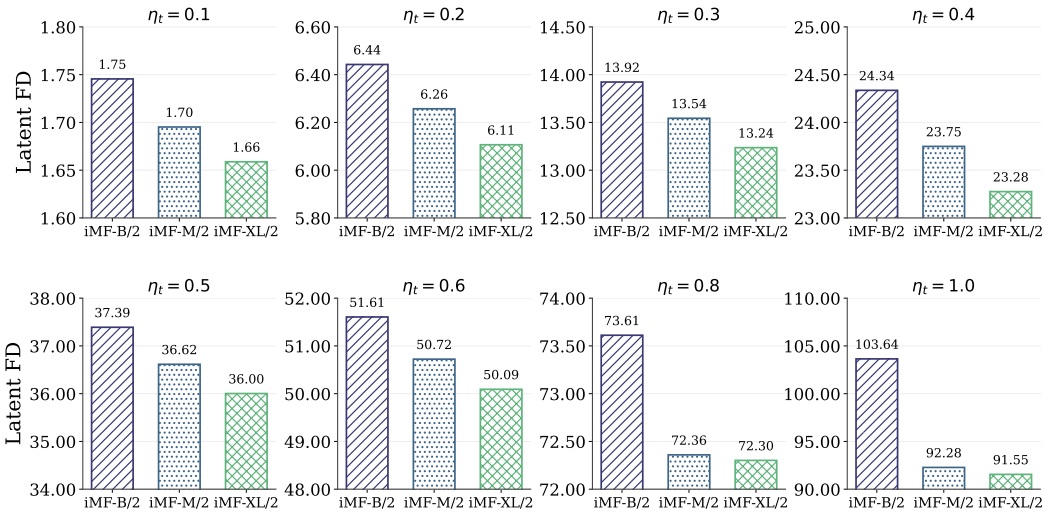


Figure 11: Latent-FD($\mathcal{P}_g^t, \mathcal{P}_e$) across noise levels for different generative models, with $\eta_t = 1.0$ corresponding to Latent-FD($\mathcal{P}_g, \mathcal{P}_e$). The model ordering is preserved across noise levels.

Figure 11 compares Latent-FD($\mathcal{P}_g^t, \mathcal{P}_e$) across noise levels for different generative models, where $\eta_t = 1.0$ corresponds to Latent-FD($\mathcal{P}_g, \mathcal{P}_e$). We observe that the model ordering is preserved across all evaluated noise levels. This consistent ordering indicates that larger intermediate latent mismatch is associated with larger final mismatch at $\eta_t = 1.0$. We summarize this empirical relationship as

$$\text{Latent-FD}(\mathcal{P}_g^t, \mathcal{P}_e) \uparrow \rightsquigarrow \text{Latent-FD}(\mathcal{P}_g, \mathcal{P}_e) \uparrow .$$

Under the fixed-tokenizer setting, this ordering consistency supports using intermediate latent mismatch to characterize the relative final mismatch across generative models.

E Noise Level Choice in Decoder Adaptation

A key practical consideration in decoder adaptation is the choice of noise level. Intermediate GAR latents can in principle be constructed using any $\eta_t \in (0, 1)$, but the noise level controls the trade-off between generation awareness and source correspondence. At large noise levels, GAR latents move closer to the generation-time regime while progressively losing correspondence with the source image x, making paired reconstruction supervision less reliable. At small noise levels, the latents remain close to the encoder distribution $\mathcal{P}_e$, providing limited exposure to the latent distributions encountered during generation. We therefore sample intermediate noise levels as $\eta_t \sim \text{Unif}[\eta_{\min}, \eta_{\max}]$ to balance these two effects. In our experiments, we use $\eta_{\min} = 0.25$ and $\eta_{\max} = 0.45$, which provides effective decoder adaptation across the evaluated model scales.

F COMPUTATIONAL COST OF RFID, GAR-FID, AND GFID

We compare the computational cost of rFID, GAR-FID, and gFID using REPAE-VAE, whose full generation process requires 250 function evaluations (NFEs). The three metrics correspond to different positions along the reconstruction-generation trajectory. rFID evaluates the reconstruction endpoint and requires no generative-model evaluation, whereas gFID evaluates the generation endpoint and requires the full 250-step generative process. GAR-FID operates between these two endpoints: at noise level η_t, it starts from a partially noised encoder latent and applies only the remaining generative trajectory before decoding. Its computational cost therefore increases with η_t and approaches that of gFID as $\eta_t \to 1$. As shown in Table 3, rFID is the least expensive, requiring only 0.095 hours in this setting. GAR-FID exhibits a gradual increase in computational cost along the trajectory. For example, GAR-FID requires 2.936 hours at $\eta_t = 0.5$ and 4.545 hours at $\eta_t = 0.8$, corresponding to approximately 53% and 82% of the runtime of gFID, respectively. At $\eta_t = 1.0$, GAR reaches the generation endpoint and its runtime (5.572 hours) is effectively identical to that of gFID (5.570 hours).

Table 3: Computational cost of rFID, GAR-FID, and gFID on REPAE-VAE. rFID requires no generative-model evaluations, GAR-FID traverses a partial generation trajectory determined by η_t, and gFID uses the full 250-step generation process.

Metric	Noise Level η_t	NFE	Time (hours)
rFID	0.0	0	0.095
GAR-FID	0.1	32	0.799
GAR-FID	0.2	56	1.327
GAR-FID	0.3	81	1.879
GAR-FID	0.4	105	2.407
GAR-FID	0.5	129	2.936
GAR-FID	0.6	153	3.465
GAR-FID	0.7	177	3.992
GAR-FID	0.8	202	4.545
GAR-FID	0.9	226	5.075
GAR-FID	1.0	250	5.572
gFID	1.0	250	**5.570**

G LIMITATIONS

Our work has several limitations.

First, our empirical study focuses on latent generative models based on diffusion and flow matching. While latent distribution mismatch may extend more broadly, we do not empirically investigate autoregressive latent generators or other generation frameworks.

Second, our explanation for the relationship between GAR-FID and generative performance is primarily empirical. Although we observe consistent associations among GAR-FID, intermediate latent mismatch, final latent mismatch, and gFID, we do not provide formal theoretical guarantees for these relationships. Proposition 1 provides a complementary perspective under Gaussian and Lipschitz assumptions.

Third, our latent distribution analysis relies on comparisons within a shared latent space. The relationship between Latent-FD and gFID is most directly examined in fixed-tokenizer settings, where latent distributions are comparable across generators. While broader experiments show that the GAR-FID–gFID relationship remains strong when both tokenizers and generators vary, direct distribution-level comparisons across tokenizer latent spaces remain challenging.

Fourth, decoder adaptation is explicitly generation-oriented and need not simultaneously improve reconstruction quality. Adapting the decoder toward generation-relevant latent distributions can trade reconstruction performance under $\mathcal{P}_e$ for improved behavior under generation-time latents.

Finally, GAR and decoder adaptation introduce additional computation by passing intermediate latents through the generative model during evaluation or training. GAR cost depends on the underlying generative process and noise level: for multi-step generators, intermediate GAR evaluations can require fewer function evaluations than full generation, although this advantage is not universal. Decoder adaptation also depends on the intermediate noise range, reflecting a trade-off between generation awareness and source correspondence that may require tuning across model families.

GAR-FID should therefore be viewed as a trajectory-based diagnostic rather than a replacement for direct generative evaluation. It reveals how decoder behavior evolves from reconstruction toward generation and correlates strongly with gFID, but gFID remains necessary when final sample quality is the primary evaluation target.

Table 4: Quantitative results for 13 tokenizers. For each tokenizer, we evaluate rFID, gFID (w/o and w/ CFG), and GAR-FID without CFG at noise levels η_t ranging from 0.1 to 1.0. All results are obtained using SiT/B models trained on the corresponding tokenizer (Xu et al., 2026).

Tokenizers	rFID (↓)	gFID (↓)		GAR-FID (η_t) (↓)									
		w/o CFG	w CFG	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SD-VAE (Rombach et al., 2022)	0.73	46.64	9.91	1.16	2.28	4.85	9.63	17.10	25.94	33.98	38.91	41.35	42.03
IN-VAE (Leng et al., 2025)	0.26	48.28	10.87	0.46	0.88	1.86	3.86	7.68	14.53	24.37	35.13	41.78	43.36
FLUX-VAE (Labs, 2024)	0.16	62.92	14.38	1.05	2.32	5.42	11.09	18.97	29.14	39.71	49.25	55.71	58.33
QwenImage-VAE (Wu et al., 2025)	1.52	48.03	10.97	2.14	2.94	4.54	7.14	11.17	16.84	24.20	32.57	40.49	43.72
SD3-VAE (Esser et al., 2024)	0.21	51.58	11.89	0.59	1.06	2.25	4.76	9.34	16.36	26.03	35.92	43.70	46.82
EQ-VAE (Kouzelis et al., 2025)	0.59	37.49	9.32	3.37	6.54	10.11	13.91	17.79	21.83	26.12	29.70	32.06	33.26
VA-VAE-16 (Yao & Wang, 2025)	0.30	18.88	6.01	2.52	3.71	4.91	6.21	7.46	8.85	10.48	12.57	14.90	15.78
VA-VAE-64 (Yao & Wang, 2025)	0.15	33.56	7.76	1.90	3.10	5.02	7.44	10.32	13.48	17.07	21.10	25.98	29.50
SOFT-VQ (Chen et al., 2025c)	0.59	29.48	7.48	1.64	5.63	10.52	15.19	18.97	21.93	24.13	25.38	25.93	26.07
MAE-TOK (Chen et al., 2025b)	0.61	13.75	5.69	1.66	3.42	4.95	6.34	7.55	8.64	9.93	10.89	11.56	11.64
DE-TOK (Yang et al., 2026b)	0.59	20.17	6.95	0.63	0.72	0.91	1.30	2.03	3.48	6.74	12.05	16.16	16.87
DM-VAE (Ye et al., 2025)	0.72	8.38	4.89	0.81	0.87	0.95	1.05	1.20	1.47	2.10	3.60	5.75	6.76
REPAE-VAE (Leng et al., 2025)	0.56	25.75	6.46	0.94	1.78	3.10	4.91	7.08	9.86	13.12	17.18	21.04	22.14

Table 5: Quantitative results for 13 tokenizers. For each tokenizer, we evaluate rFID, gFID (w/o and w/ CFG), and GAR-FID without CFG at noise levels η_t ranging from 0.1 to 1.0. All results are obtained using SiT/XL models trained on the corresponding tokenizer (Xu et al., 2026).

Tokenizers	rFID (↓)	gFID (↓)		GAR-FID (η_t) (↓)									
		w/o CFG	w CFG	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SD-VAE (Rombach et al., 2022)	0.73	25.91	6.33	1.11	1.76	2.91	5.03	8.27	12.82	17.31	20.43	21.85	22.30
IN-VAE (Leng et al., 2025)	0.26	25.65	6.56	0.45	0.69	1.15	1.97	3.48	6.18	10.75	16.65	20.58	21.51
FLUX-VAE (Labs, 2024)	0.16	34.06	7.82	0.96	1.66	3.13	5.54	8.87	13.45	18.62	24.00	27.85	29.65
QwenImage-VAE (Wu et al., 2025)	1.52	23.62	6.19	2.13	2.61	3.33	4.35	5.82	7.88	10.76	14.52	18.29	19.92
SD3-VAE (Esser et al., 2024)	0.21	26.38	6.31	0.59	0.90	1.51	2.59	4.50	7.38	11.62	16.91	21.27	22.89
EQ-VAE (Kouzelis et al., 2025)	0.59	20.81	6.24	2.53	4.14	5.79	7.39	9.29	11.32	13.37	15.59	16.99	17.53
VA-VAE-16 (Yao & Wang, 2025)	0.30	8.57	4.20	2.11	2.75	3.30	3.85	4.32	4.88	5.49	6.30	7.32	7.80
VA-VAE-64 (Yao & Wang, 2025)	0.15	15.40	5.36	1.67	2.35	3.23	4.28	5.51	6.82	8.25	10.03	12.05	13.33
SOFT-VQ (Chen et al., 2025c)	0.59	15.88	5.14	1.29	3.39	5.83	8.01	9.92	11.56	12.65	13.37	13.62	13.62
MAE-TOK (Chen et al., 2025b)	0.61	6.27	3.74	1.23	1.99	2.59	3.08	3.54	3.94	4.44	4.91	5.19	5.21
DE-TOK (Yang et al., 2026b)	0.59	11.97	4.83	0.65	0.73	0.88	1.13	1.52	2.26	3.77	6.58	9.14	9.59
DM-VAE (Ye et al., 2025)	0.72	4.65	3.39	0.81	0.86	0.92	1.00	1.09	1.24	1.50	2.12	3.17	3.86
REPAE-VAE (Leng et al., 2025)	0.56	12.95	4.17	0.90	1.36	1.94	2.72	3.57	4.64	6.00	7.85	9.89	10.60
RAE (Zheng et al., 2026)	0.63	4.25	3.50	0.64	0.68	0.82	0.89	0.99	1.10	1.25	1.47	1.92	4.38

H Additional Results for Cross-Tokenizer and Cross-Model Generalization

To examine whether the relationship observed in the controlled setting (Section 4.1) remains stable when both the tokenizer and generative model vary, we follow iFID (Xu et al., 2026) and consider 14 VAEs with publicly available checkpoints. Among them, 13 are evaluated with both SiT-B and SiT-XL, while RAE is evaluated with SiT-XL only. The details of these models are provided in Table 6. These VAEs cover a diverse set of design choices. Some models, such as SD-VAE (Rombach et al., 2022), IN-VAE (Leng et al., 2025), QW-VAE (Wu et al., 2025), FLUX-VAE (Labs, 2024), and SD3-VAE (Esser et al., 2024), are optimized primarily for reconstruction. Others incorporate additional inductive biases, such as equivariance regularization in EQ-VAE (Kouzelis et al., 2025). A third group leverages contrastive learning-based image encoders, including VA-VAE (Yao & Wang, 2025), SOFT-VQ (Chen et al., 2025c), MAE-TOK (Chen et al., 2025b), DE-TOK (Chen et al., 2025b), DM-VAE (Ye et al., 2025), REPAE-VAE (Leng et al., 2025), and RAE (Zheng et al., 2026).

For each VAE, we use the publicly available SiT-B and SiT-XL (Ma et al., 2024) generative models released by iFID (Xu et al., 2026), which are trained on the 256×256 ImageNet dataset (Deng et al., 2009) in the corresponding latent space. We directly use their provided checkpoints without additional retraining. For each tokenizer-generator configuration, we compute rFID, GAR-FID (without CFG) across noise levels $\eta_t \in [0.1, 1.0]$, and gFID both with and without classifier-free guidance (CFG). The results are summarized in Table 4 for SiT-B and Table 5 for SiT-XL.

Note on FID references. As shown in Table 4 and Table 5, GAR-FID (without CFG) at $\eta_t = 1.0$ is not equal to gFID without CFG. This discrepancy arises because the two metrics are evaluated with different reference image statistics. Specifically, gFID is computed using the precomputed ImageNet statistics (`VIRTUAL_imagenet256_labeled.npz`) (Deng et al., 2009), whereas GAR-FID is computed against the 50K input-image PNG reference set used in the paired reconstruction protocol. As a result, GAR-FID at $\eta_t = 1.0$ is not expected to exactly match gFID without CFG.

Table 6: List of VAEs included in our study. We follow iFID (Xu et al., 2026) and use publicly available generative-model checkpoints trained in their corresponding latent spaces.

VAE Name	Latent Dim.	Arch.	Training
SD-VAE (Rombach et al., 2022)	$4 \times 32 \times 32$	UNet	Recon.
IN-VAE (Leng et al., 2025)	$16 \times 16 \times 16$	UNet	Recon.
FLUX-VAE (Labs, 2024)	$16 \times 32 \times 32$	UNet	Recon.
QwenImage-VAE (Wu et al., 2025)	$16 \times 32 \times 32$	UNet	Recon.
SD3-VAE (Esser et al., 2024)	$16 \times 32 \times 32$	UNet	Recon.
EQ-VAE (Kouzelis et al., 2025)	$4 \times 32 \times 32$	UNet	Recon. + Equivariance
VA-VAE-16 (Yao & Wang, 2025)	$16 \times 16 \times 16$	UNet	Recon. + DINO alignment
VA-VAE-64 (Yao & Wang, 2025)	$64 \times 16 \times 16$	UNet	Recon. + DINO alignment
SOFT-VQ (Chen et al., 2025c)	64×32	ViT	Recon. + DINO alignment
MAE-TOK (Chen et al., 2025b)	128×32	ViT	Recon. + Mask + DINO alignment
DE-TOK (Yang et al., 2026b)	128×32	ViT	Recon. + Mask + latent denoising
DM-VAE (Ye et al., 2025)	256×32	ViT	Recon. + Distribution Matching
REPAE-VAE (Leng et al., 2025)	$4 \times 32 \times 32$	UNet	Recon. + REPA Loss
RAE (Zheng et al., 2026)	$768 \times 16 \times 16$	ViT	DINO Encoder + Recon. Decoder

Table 7: Correlation between FID-based metrics and generative performance (gFID). Correlations are computed across models obtained by training SiT with 13 pre-trained VAEs. We report both Pearson correlation (PCC) and Spearman rank correlation (SRCC) under different settings (with and without CFG) for SiT-B, SiT-XL, and their mixed configurations. GAR-FID without CFG consistently exhibits stronger correlation with gFID across noise levels η_t, whereas the standard reconstruction metric (rFID) shows negative correlation. **Bold** indicates the best results. † denotes results from (Xu et al., 2026), and * indicates our reproduced results.

Metrics	gFID SiT/B				gFID SiT/XL				gFID SiT/mixed			
	w/o CFG		w/ CFG		w/o CFG		w/ CFG		w/o CFG		w/ CFG	
	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)	PCC (↑)	SRCC (↑)
rFID†	-0.04	-0.31	-0.07	-0.31	-0.06	-0.21	-0.15	-0.31	–	–	–	–
iFID†	0.85	0.86	0.82	0.84	0.89	0.91	0.88	0.92	–	–	–	–
rFID*	-0.11	-0.36	-0.11	-0.30	-0.13	-0.34	-0.15	-0.39	-0.10	-0.30	-0.10	-0.28
iFID*	0.86	0.83	0.80	0.79	0.90	0.88	0.88	0.90	0.72	0.77	0.64	0.70
GAR-FID ($\eta_t = 0.1$)	-0.10	-0.19	-0.13	-0.14	0.01	0.00	0.10	0.01	0.04	0.03	0.05	0.01
GAR-FID ($\eta_t = 0.2$)	-0.04	-0.04	-0.08	-0.04	0.12	0.22	0.19	0.20	0.16	0.25	0.16	0.22
GAR-FID ($\eta_t = 0.3$)	0.14	0.15	0.09	0.14	0.25	0.38	0.32	0.35	0.34	0.42	0.34	0.39
GAR-FID ($\eta_t = 0.4$)	0.38	0.33	0.33	0.34	0.45	0.54	0.49	0.51	0.55	0.64	0.54	0.60
GAR-FID ($\eta_t = 0.5$)	0.63	0.68	0.57	0.69	0.61	0.66	0.65	0.61	0.74	0.80	0.72	0.76
GAR-FID ($\eta_t = 0.6$)	0.81	0.79	0.77	0.78	0.79	0.84	0.80	0.79	0.86	0.89	0.84	0.85
GAR-FID ($\eta_t = 0.7$)	0.93	0.90	0.89	0.87	0.91	0.90	0.91	0.86	0.95	0.96	0.93	0.91
GAR-FID ($\eta_t = 0.8$)	0.98	0.96	0.96	0.94	0.98	0.99	0.97	0.96	0.99	0.99	0.97	0.95
GAR-FID ($\eta_t = 0.9$)	**1.00**	**0.99**	**0.98**	0.98	**1.00**	**1.00**	**0.98**	**0.96**	**1.00**	**1.00**	**0.98**	**0.97**
GAR-FID ($\eta_t = 1.0$)	**1.00**	**0.99**	**0.98**	**0.99**	**1.00**	**1.00**	**0.98**	**0.96**	**1.00**	**1.00**	**0.98**	**0.97**

To quantify the relationship between these metrics and generative performance, we follow iFID (Xu et al., 2026) and compute the Pearson correlation coefficient (PCC) and Spearman rank correlation coefficient (SRCC) between each metric and gFID, both with and without CFG. As reported in Table 7, rFID exhibits consistently negative correlation with gFID across SiT-B, SiT-XL, and their mixed configurations. This indicates that better reconstruction FID can correspond to worse generative FID, consistent with the reconstruction-generation dilemma observed in prior work (Yao & Wang, 2025; Kouzelis et al., 2025; Ye et al., 2025; Skorokhodov et al., 2025; Chen et al., 2025b; Xu et al., 2026). In contrast, the correlation between GAR-FID and gFID strengthens substantially as the noise level η_t increases. At $\eta_t = 0.8$, GAR-FID already achieves consistently high correlation across settings, with SRCC ranging from 0.94 to 0.99. When $\eta_t \geq 0.8$, GAR-FID also outperforms iFID (Xu et al., 2026) on both SiT-B and SiT-XL across the reported correlation settings, with the correlation approaching saturation at higher noise levels.

The mixed SiT-B/XL setting further tests whether the GAR-FID-gFID relationship remains stable across generator scales. Correlations computed within a single generative model, such as SiT-B or SiT-XL (Ma et al., 2024), primarily capture tokenizer variation under a fixed generator scale. By contrast, the mixed setting jointly varies tokenizer choice and generator scale, providing a more demanding test of cross-system comparisons. In this setting, the correlation of iFID with gFID decreases, whereas GAR-FID maintains consistently strong correlation at higher noise levels. These results indicate that the GAR-FID–gFID relationship remains robust across heterogeneous tokenizer-generator configurations.

Leave-one-out robustness across tokenizers. To test whether the strong correlation between GAR-FID and gFID is driven by any individual tokenizer, we perform a leave-one-out analysis using all 14 tokenizer configurations in Table 5. Specifically, we omit each tokenizer in turn and recompute the SRCC over the remaining 13 configurations. Tables 8 and 9 report the complete SRCC results without and with CFG, respectively.

Table 8: Leave-one-out SRCC between GAR-FID and gFID without CFG. Each column omits the indicated tokenizer when computing the correlation.

Omitted tokenizer	SD-VAE	IN-VAE	FLUX-VAE	QwenImage-VAE	SD3-VAE	EQ-VAE	VA-VAE-16	VA-VAE-64	SOFT-VQ	MAE-TOK	DE-TOK	DM-VAE	REPAE-VAE	RAE
GAR-FID ($\eta_t = 0.1$)	0.03	0.18	0.00	-0.03	0.21	-0.03	0.05	-0.01	-0.01	0.02	-0.05	-0.11	-0.07	-0.13
GAR-FID ($\eta_t = 0.2$)	0.22	0.31	0.21	0.20	0.25	0.16	0.26	0.17	0.16	0.21	0.10	0.03	0.12	-0.04
GAR-FID ($\eta_t = 0.3$)	0.38	0.45	0.37	0.37	0.47	0.37	0.44	0.37	0.37	0.32	0.27	0.21	0.32	0.19
GAR-FID ($\eta_t = 0.4$)	0.52	0.65	0.48	0.55	0.68	0.58	0.55	0.51	0.58	0.55	0.47	0.42	0.54	0.42
GAR-FID ($\eta_t = 0.5$)	0.66	0.79	0.64	0.69	0.70	0.73	0.65	0.65	0.73	0.64	0.62	0.58	0.63	0.58
GAR-FID ($\eta_t = 0.6$)	0.85	0.90	0.82	0.89	0.90	0.89	0.85	0.85	0.89	0.84	0.85	0.82	0.84	0.82
GAR-FID ($\eta_t = 0.7$)	0.90	0.92	0.88	0.92	0.93	0.93	0.89	0.88	0.93	0.89	0.90	0.88	0.88	0.88
GAR-FID ($\eta_t = 0.8$)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GAR-FID ($\eta_t = 0.9$)	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GAR-FID ($\eta_t = 1.0$)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	1.00

Table 9: Leave-one-out SRCC between GAR-FID and gFID with CFG. Each column omits the indicated tokenizer when computing the correlation.

Omitted tokenizer	SD-VAE	IN-VAE	FLUX-VAE	QwenImage-VAE	SD3-VAE	EQ-VAE	VA-VAE-16	VA-VAE-64	SOFT-VQ	MAE-TOK	DE-TOK	DM-VAE	REPAE-VAE	RAE
GAR-FID ($\eta_t = 0.1$)	0.04	0.23	0.01	-0.03	0.20	-0.03	0.03	-0.01	-0.01	0.02	-0.02	-0.08	-0.08	-0.10
GAR-FID ($\eta_t = 0.2$)	0.20	0.34	0.17	0.15	0.23	0.12	0.19	0.15	0.14	0.17	0.10	0.03	0.07	-0.05
GAR-FID ($\eta_t = 0.3$)	0.35	0.45	0.33	0.34	0.43	0.34	0.38	0.36	0.36	0.29	0.27	0.19	0.29	0.17
GAR-FID ($\eta_t = 0.4$)	0.49	0.66	0.45	0.51	0.63	0.54	0.51	0.51	0.58	0.51	0.46	0.40	0.51	0.40
GAR-FID ($\eta_t = 0.5$)	0.61	0.79	0.57	0.61	0.64	0.64	0.58	0.61	0.68	0.58	0.56	0.52	0.58	0.52
GAR-FID ($\eta_t = 0.6$)	0.81	0.89	0.77	0.84	0.85	0.84	0.79	0.83	0.86	0.79	0.81	0.78	0.79	0.78
GAR-FID ($\eta_t = 0.7$)	0.85	0.91	0.82	0.87	0.88	0.86	0.85	0.85	0.90	0.84	0.86	0.83	0.85	0.83
GAR-FID ($\eta_t = 0.8$)	0.97	0.97	0.96	0.96	0.97	0.96	0.97	0.96	0.96	0.96	0.97	0.96	0.97	0.96
GAR-FID ($\eta_t = 0.9$)	0.96	0.97	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.96
GAR-FID ($\eta_t = 1.0$)	0.97	0.97	0.95	0.96	0.97	0.96	0.96	0.96	0.96	0.95	0.96	0.95	0.97	0.95

At low noise levels, particularly when $\eta_t < 0.5$, the correlations are relatively weak and more sensitive to the removal of individual tokenizers. They increase substantially as η_t grows and become uniformly high at $\eta_t \geq 0.8$. In particular, at $\eta_t = 0.8$, the leave-one-out SRCC is 0.99 for every omitted tokenizer without CFG and ranges from 0.96 to 0.97 with CFG. This demonstrates that the strong GAR-FID–gFID relationship at $\eta_t = 0.8$ is not driven by any single tokenizer.

I Decoder Feature Spaces for Latent Distribution Analysis

In Section 3.1, we analyze intermediate representations from the SD-VAE decoder (Rombach et al., 2022). Specifically, we extract features from three representative decoder stages: (i) *Conv In*, the initial convolutional projection of the latent input; (ii) *Mid Block*, the bottleneck stage at the lowest spatial resolution; and (iii) *Up Block 0*, the first upsampling stage of the decoder.

These feature spaces correspond to progressively deeper stages of the decoding process and are identified from the official implementation. We collect activations using forward hooks at `decoder.conv_in`, `decoder.mid_block`, and `decoder.up_blocks[0]` for both encoder and generator latents. For visualization, activations from each feature space are adaptively pooled to a fixed 4×4 spatial resolution, flattened, and projected onto a one-dimensional LDA direction fitted to distinguish encoder-induced from generator-induced representations.

J gFID Evaluation Protocol and Implementation Details

Table 10: gFID is sensitive to the evaluation protocol. We compare two protocols on iMF-B/2: the official iMF evaluation (Geng et al., 2026) and the OpenAI protocol. They differ in computational platform (TPU vs. GPU), Inception network implementation and weights (PyTorch vs. TensorFlow), and reference image statistics.

Setting	Platform	Inception	Reference Image Statistics	FID	IS
Official iMF	TPU	PyTorch	`jit_in256_stats.npz`	3.39	255.3
Official iMF (our reproduction)	GPU	PyTorch	`jit_in256_stats.npz`	3.37	255.7
OpenAI Setup	GPU	TensorFlow	`VIRTUAL_imagenet256_labeled.npz`	3.47	255.2

The official iMF-B/2 (Geng et al., 2026) results are obtained under a different setup from the commonly used OpenAI evaluation protocol. Specifically, the two protocols differ in three aspects: (i) computational platform (TPU vs. GPU), (ii) Inception network implementation and weights (PyTorch vs. TensorFlow), and (iii) reference image statistics.

Table 10 summarizes these differences and their impact on reported FID scores. We observe that replacing the TPU platform with GPU, while keeping the other components unchanged, yields slightly improved FID in our reproduction. However, further adopting the OpenAI evaluation setup by switching to the TensorFlow Inception implementation and using different reference image statistics leads to noticeably worse FID. This indicates that the commonly used OpenAI evaluation protocol tends to produce systematically higher (worse) FID scores. This observation is also consistent with FD-loss (Yang et al., 2026a), who similarly report higher FID scores under the OpenAI evaluation protocol.

Unless otherwise specified, we adopt the commonly used OpenAI evaluation protocol (TensorFlow Inception network with `VIRTUAL_imagenet256_labeled.npz` statistics) for all reported gFID results, to ensure consistency and comparability across different models and prior works.

K Experimental Details in Section 7

Data Preprocessing. All experiments are conducted on ImageNet-1k (Deng et al., 2009). Following LlamaGen (Sun et al., 2024), we apply iterative box downsampling to resize all images to a resolution of 256×256.

Training Details. We perform decoder adaptation on four latent generative models, iMF-B/2, iMF-M/2, iMF-L/2, and iMF-XL/2 (Geng et al., 2026), which share a common tokenizer, SD-VAE (Rombach et al., 2022). During adaptation, the tokenizer encoder and latent generative model are kept frozen, and only the tokenizer decoder parameters are updated. We consider two adaptation settings. In the standard setting, intermediate GAR latents are generated without classifier-free guidance (CFG), and the decoder is adapted for 10 epochs. We additionally evaluate a *CFG-aware decoder adaptation* setting, in which intermediate GAR latents are generated with CFG to better match the guided generation regime used at inference time. CFG-aware adaptation is performed for one epoch, while all other training settings remain unchanged. To further evaluate the generality of decoder adaptation beyond iMF, we apply the same one-epoch CFG-aware adaptation setting to MeanFlow-XL/2 (Geng et al., 2025) and the distillation-based DMF-XL/2+ (Lee et al., 2026). All experiments are conducted on 8 NVIDIA H100 GPUs (80GB memory each) using the AdamW optimizer (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We use a fixed learning rate of 10^{-5}, a weight decay of 10^{-4}, and a batch size of 32 per GPU. Under the standard 10-epoch setting, training typically completes within approximately 12 hours for iMF-B/2 and 14 hours for iMF-XL/2.

Adversarial Training. For adversarial supervision used in Section 6, we follow established visual-tokenizer training practices (Tian et al., 2024; Chen et al., 2025c; Fang et al., 2026; Li et al., 2025) and employ a discriminator built on a frozen DINO-S backbone (Caron et al., 2021; Oquab et al., 2024). We further use differentiable augmentation (DiffAug) (Zhao et al., 2020), consistency regularization (Zhang et al., 2019), and LeCAM regularization (Tseng et al., 2021) for adversarial training. These training choices and loss weights are kept fixed across model scales.

Loss Weights and Noise Level. We fix $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 0.7$, and $\lambda_4 = 10.0$ across all experiments. For noise level selection in decoder adaptation, we sample $\eta_t \sim \mathrm{Uniform}[\eta_{\min}, \eta_{\max}]$, where $\eta_{\min} = 0.25$ and $\eta_{\max} = 0.45$, as discussed in Appendix E.

CFG parameters. The official iMF implementation employs three classifier-free guidance (CFG) parameters: $t_{\min}$, $t_{\max}$, and ω. In the original iMF repository, these parameters are selected via a fine-grained grid search. Since decoder adaptation is performed without CFG during training, the original CFG hyperparameters are no longer optimal after updating the decoder. We therefore conduct a new grid search to identify a new set of CFG parameters for evaluation. The updated CFG parameters are reported in Table 11. We will release the corresponding checkpoints to further facilitate reproducibility by the community.

Table 11: CFG parameters used for gFID evaluation across different model scales. Left: official parameters from the iMF repository, obtained via fine-grained grid search without decoder adaptation. Right: re-optimized parameters after decoder adaptation, obtained via a new grid search to account for decoder changes.

CFG Parameters	w/o Decoder Adaptation				w/ Decoder Adaptation			
	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2
$t_{\min}$	0.40	0.40	0.40	0.42	0.50	0.55	0.44	0.48
$t_{\max}$	0.65	0.60	0.60	0.62	0.65	0.625	0.60	0.65
ω	8.0	10.5	10.5	8.0	9.0	10.5	8.0	5.25

L Additional Experimental Results in Section 7

We provide additional experimental results in this appendix due to space constraints in the main paper. Specifically, we report robustness analyses across random seeds, examine how decoder adaptation

Table 12: Repeated-run evaluation of decoder adaptation across three random seeds. We report mean ± standard deviation of gFID under the OpenAI evaluation protocol for four iMF model scales, with and without CFG, before and after decoder adaptation (DA).

Setting	w/o CFG				w/ CFG			
	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2
w/o DA	16.36 ± 0.12	11.79 ± 0.04	9.28 ± 0.04	9.52 ± 0.17	3.49 ± 0.01	2.39 ± 0.02	1.91 ± 0.01	1.81 ± 0.03
w/ DA	12.30 ± 0.07	9.51 ± 0.01	7.33 ± 0.05	7.48 ± 0.09	3.14 ± 0.01	2.27 ± 0.03	1.81 ± 0.02	1.73 ± 0.03

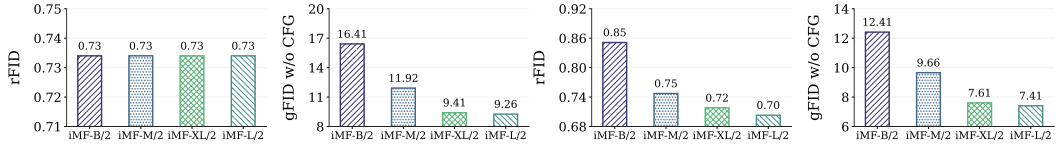


Figure 12: rFID vs. gFID (w/o CFG) across model scales, before (left two) and after (right two) decoder adaptation. rFID is uninformative before adaptation but becomes aligned with gFID after adaptation.

Table 13: **Effect of fixed noise levels in decoder adaptation.** We report gFID without CFG, evaluated using the OpenAI protocol, and rFID across iMF model scales for decoder adaptation with different fixed noise levels η_t. See Appendix J for protocol details.

Noise Level	gFID w/o CFG (↓)				rFID (↓)			
	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2	iMF-B/2	iMF-M/2	iMF-L/2	iMF-XL/2
$\eta_t = 0.0^{\ddagger}$	16.41	11.92	9.26	9.41	0.73	0.73	0.73	0.73
$\eta_t = 0.0$	15.27	11.30	8.38	8.82	**0.58**	0.62	0.59	0.62
$\eta_t = 0.1$	14.76	11.03	8.19	8.64	0.62	**0.61**	**0.58**	**0.60**
$\eta_t = 0.2$	13.50	10.15	7.89	8.03	0.71	0.67	0.62	0.64
$\eta_t = 0.3$	12.86	9.99	7.57	7.84	0.77	0.70	0.67	0.67
$\eta_t = 0.4$	12.08	9.52	7.51	7.24	0.94	0.79	0.75	0.74
$\eta_t = 0.5$	**11.58**	**9.18**	**7.36**	**7.19**	1.20	1.01	0.94	0.89
$\eta_t = 0.6$	11.88	9.34	7.41	7.40	1.81	1.42	1.26	1.21

affects the relationship between reconstruction and generation, study sensitivity to adaptation noise levels, evaluate reconstruction performance, provide system-level comparisons with a broad range of generative models, and present additional qualitative visual results.

Robustness across random seeds. To assess the robustness of decoder adaptation to random-seed variation, we repeat the gFID evaluation with three random seeds under the OpenAI evaluation protocol. As shown in Table 12, the standard deviation is consistently small across all model scales, both with and without CFG. Moreover, DA improves the mean gFID in every evaluated setting. Without CFG, the improvements are substantial across all four model scales, while with CFG the gains are smaller but remain consistent relative to the observed run-to-run variation. For example, on iMF-XL/2 with CFG, DA improves gFID from 1.81 ± 0.03 to 1.73 ± 0.03. These results indicate that the improvements from DA are stable across random seeds rather than being driven by a favorable evaluation run.

Effect of Decoder Adaptation on the rFID-gFID Relationship. Figure 12 shows the relationship between rFID and gFID (w/o CFG) across model scales, before and after decoder adaptation. Before adaptation, rFID remains constant across models because the tokenizer is fixed, while gFID varies substantially with generator scale. After adaptation, rFID decreases together with gFID across model scales, indicating a closer correspondence between reconstruction and generation behavior. This trend is consistent with our latent distribution perspective: adapting the decoder on generation-relevant GAR latents changes its behavior under $\mathcal{P}_e$ in a generator-dependent manner, bringing reconstruction performance into closer agreement with downstream generative performance.

Sensitivity Analyses on Noise Levels. To study the effect of the noise level used for decoder adaptation, we conduct a sensitivity analysis by fixing $\eta_t \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ during adaptation. As shown in Table 13, $\eta_t = 0.5$ achieves the best gFID without CFG across all iMF model scales in this fixed-noise setting. When η_t is too small, the intermediate GAR latents $z_g^t \sim \mathcal{P}_g^t$ remain relatively close to the encoder distribution $\mathcal{P}_e$, providing limited exposure to generation-relevant latent distributions and resulting in smaller gFID improvements. Increasing η_t moves the decoder inputs further toward the generation-time regime, but also progressively weakens their

correspondence with the source image x, making paired reconstruction supervision less reliable at large noise levels. Consistently, rFID increases monotonically with η_t across all iMF model scales, reflecting the growing shift away from the reconstruction regime. These results highlight a trade-off between generation awareness and source correspondence and motivate using intermediate noise levels for decoder adaptation.

Reconstruction Results. As shown in Table 14, we compare reconstruction performance before and after decoder adaptation on the iMF-B/2 generative model, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), Fréchet Inception Distance (FID; Heusel et al., 2017), perceptual similarity (LPIPS; Zhang et al., 2018), and inception score (IS; Salimans et al., 2016). All results are reported without CFG. Standard reconstruction (SR, $\eta_t = 0$) degrades consistently after adaptation across all metrics, indicating that the adapted decoder is no longer well aligned with encoder latents z_e.

For generation-aware reconstruction (GAR), we observe a systematic trade-off in the effect of decoder adaptation as the noise level η_t increases. For structural fidelity, measured by PSNR and SSIM, decoder adaptation consistently degrades performance; however, the magnitude of degradation gradually diminishes with increasing η_t, indicating that the mismatch between the adapted decoder and the input latents becomes less severe as the latents move closer to the generative distribution. In contrast, perceptual quality, measured by LPIPS, IS, and FID, exhibits an opposite trend. At low noise levels, decoder adaptation slightly worsens performance, while at higher noise levels it yields consistent improvements. In particular, FID shows substantial gains when $\eta_t \geq 0.3$, suggesting that decoder adaptation improves perceptual alignment under generation-consistent latent distributions.

System-Level Comparison of Generative Performance. We present a system-level comparison on class-conditional ImageNet 256×256 across diffusion/flow, GAN, and autoregressive or masked generative models, with inference cost measured by the number of function evaluations (NFE). For diffusion and flow models trained from scratch, we include 1-NFE methods MeanFlow-XL/2 (Geng et al., 2025), TiM-XL/2 (Wang et al., 2025b), α-Flow-XL/2+ (Zhang et al., 2025), and iMF (Geng et al., 2026) at multiple scales (B/2, M/2, L/2, XL/2), together with 2-NFE methods IMM-XL/2 (Zhou et al., 2025), MeanFlow-XL/2+ (Geng et al., 2025), α-Flow-XL/2+ (Zhang et al., 2025), and iMF-XL/2. We further include distillation-based 1-NFE methods π-Flow-XL/2 (Chen et al., 2025a), DMF-XL/2+ (Lee et al., 2026), and FACM-XL/2 (Peng et al., 2026), as well as multi-NFE diffusion/flow models ADM-G (Dhariwal & Nichol, 2021), LDM-4-G (Rombach et al., 2022), SimDiff (Hoogeboom et al., 2023), DiT-XL/2 (Peebles & Xie, 2023), SiT-XL/2 (Ma et al., 2024), SiT-XL/2+REPA (Yu et al., 2025), SiD2 (Hoogeboom et al., 2025), LightningDiT-XL/2 (Yao & Wang, 2025), DDT-XL/2 (Wang et al., 2025a), and RAE+DiT^{DH}-XL (Zheng et al., 2026). For broader reference, we also compare with GAN-based methods BigGAN (Brock et al., 2019), GigaGAN (Kang et al., 2023), and StyleGAN-XL (Sauer et al., 2022), and autoregressive or masked models JetFormer-L (Tschannen et al., 2025), MaskGIT (Chang et al., 2022), RQ-Transformer (Lee et al., 2022), STARFlow (Gu et al., 2025), LLamaGen-3B (Sun et al., 2024), VAR-$d30$ (Tian et al., 2024), MAR-H (Li et al., 2024), RAR-XXL (Yu et al., 2024), and xAR-H (Ren et al., 2025).

As shown in Table 15, under the most constrained $\text{NFE} = 1$ setting, decoder adaptation consistently improves generative performance across all evaluated iMF scales. Importantly, the gains extend beyond iMF: applying the same decoder adaptation procedure to MeanFlow-XL/2 and the distillation-based DMF-XL/2+ also yields consistent improvements, indicating that the benefit is not specific to a particular generator architecture or training paradigm. Among the evaluated 1-NFE models, iMF-XL/2 + DA achieves the lowest gFID, while the adapted MeanFlow and DMF models also improve over their corresponding baselines. These results highlight decoder adaptation as a lightweight yet effective mechanism for improving generative quality without introducing additional inference cost.

Visual Comparison of Generated Samples. We provide additional qualitative comparisons for decoder adaptation (before and after) on iMF-XL/2 across 11 different classes, as shown in Figures 13-23. For each class, we present two sets of eight samples: the top row shows results before decoder adaptation, while the bottom row shows results after adaptation. Samples in the same column are generated from identical generative latents z_g, with the only difference being the decoder parameters. Across all classes, decoder adaptation consistently produces samples with richer fine-grained details,

Table 14: Reconstruction performance on the iMF-B/2 generative model. We compare reconstruction results without and with decoder adaptation. SR denotes standard reconstruction ($\eta_t = 0$), while GAR corresponds to generation-aware reconstruction with noise levels $\eta_t \in [0.1, 0.5]$.

Pipelines	w/o Decoder Adaptation					w/ Decoder Adaptation				
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	IS (↑)	FID (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	IS (↑)	FID (↓)
SR ($\eta_t = 0.0$)	27.64	73.1	0.067	208.53	0.73	26.72	70.2	0.077	200.00	0.85
GAR ($\eta_t = 0.1$)	26.77	70.8	0.080	204.92	0.90	26.17	68.7	0.084	197.29	0.93
GAR ($\eta_t = 0.2$)	25.23	66.3	0.108	198.00	1.28	24.96	65.2	0.107	193.30	1.10
GAR ($\eta_t = 0.3$)	23.72	61.1	0.146	187.94	1.88	23.59	60.6	0.141	185.83	1.44
GAR ($\eta_t = 0.4$)	22.28	55.6	0.195	172.56	2.90	22.18	55.2	0.187	173.58	2.04
GAR ($\eta_t = 0.5$)	20.83	49.9	0.259	154.66	4.38	20.73	49.3	0.251	155.10	3.10

Table 15: **System-level comparison on class-conditional ImageNet 256×256**. **NFE** denotes the number of function evaluations during inference. All numbers are with CFG when applicable, and ×2 in NFE indicates that the CFG computation doubles NFEs at inference time. †: results cited from iMF (Geng et al., 2026).

Method	# Params	NFE	FID
1-NFE diffusion/flow from scratch			
MeanFlow-XL/2†	676M	1	3.43
TiM-XL/2†	664M	1	3.26
α-Flow-XL/2+†	676M	1	2.58
MeanFlow-XL/2 + DA	676M	1	2.96
iMF-B/2†	89M	1	3.39
iMF-M/2†	174M	1	2.27
iMF-L/2†	409M	1	1.86
iMF-XL/2†	610M	1	1.72
iMF-B/2 + DA (Our)	89M	1	2.81
iMF-M/2+ DA (Our)	174M	1	1.98
iMF-L/2+ DA (Our)	409M	1	1.58
iMF-XL/2+DA (Our)	610M	1	**1.50**
2-NFE diffusion/flow from scratch			
IMM-XL/2†	675M	1×2	7.77
MeanFlow-XL/2+†	676M	2	2.20
α-Flow-XL/2+†	676M	2	1.95

Method	# Params	NFE	FID
2-NFE diffusion/flow from scratch			
iMF-XL/2†	610M	2	1.54
1-NFE diffusion/flow (distillation)			
π-Flow-XL/2†	675M	1	2.85
DMF-XL/2+†	675M	1	2.16
FACM-XL/2†	675M	1	1.76
DMF-XL/2+ + DA (Our)	676M	1	1.90
Multi-NFE diffusion/flow			
ADM-G†	554M	250×2	4.59
LDM-4-G†	400M	250×2	3.60
SimDiff†	2B	1000×2	2.77
DiT-XL/2†	675M	250×2	2.27
SiT-XL/2†	675M	250×2	2.06
SiT-XL/2 + REPA†	675M	250×2	1.42
SiD2†	–	512×2	1.38
LightningDiT-XL/2†	675M	250×2	1.35
DDT-XL/2†	675M	250×2	1.26

Method	# Params	NFE	FID
Multi-NFE diffusion/flow			
RAE + DiT^{DH}-XL†	839M	250×2	1.13
GANs			
BigGAN†	112M	1	6.95
GigaGAN†	569M	1	3.45
StyleGAN-XL†	166M	1	2.30
autoregressive/masking			
JetFormer-L†	2.75B	256×2	6.64
MaskGIT†	227M	8	6.18
RQ-Transformer†	3.8B	256×2	3.80
STARFlow†	1.4B	1024×2	2.40
LLamaGen-3B†	3.1B	256×2	2.18
VAR-$d30$†	2B	10×2	1.92
MAR-H†	943M	256×2	1.55
RAR-XXL†	1.5B	256×2	1.48
xAR-H†	1.1B	50×2	1.24

whereas samples before adaptation tend to appear slightly smoother. In addition, decoder adaptation reduces visual artifacts and yields more semantically consistent samples.

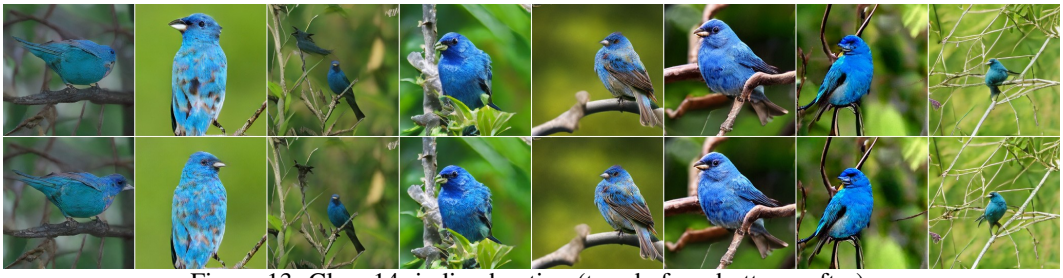


Figure 13: Class 14: indigo bunting (top: before, bottom: after).



Figure 14: Class 42: agama (top: before, bottom: after).

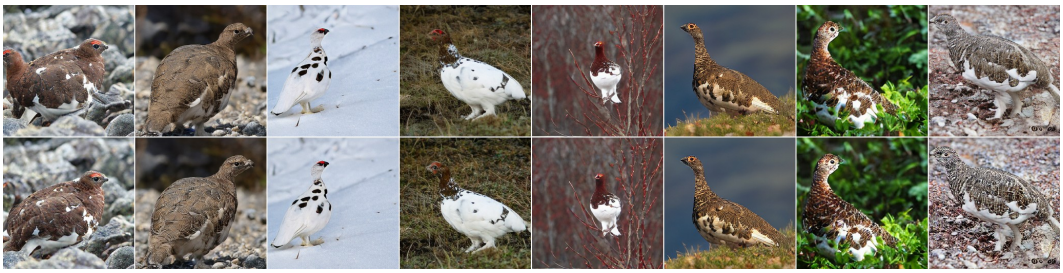


Figure 15: Class 81: ptarmigan (top: before, bottom: after).

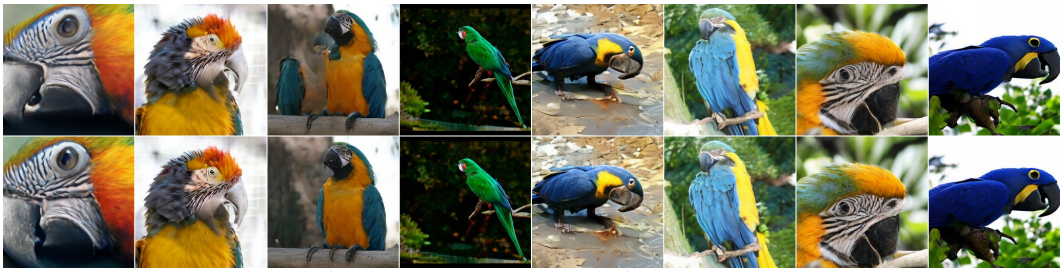


Figure 16: Class 88: macaw (top: before, bottom: after).

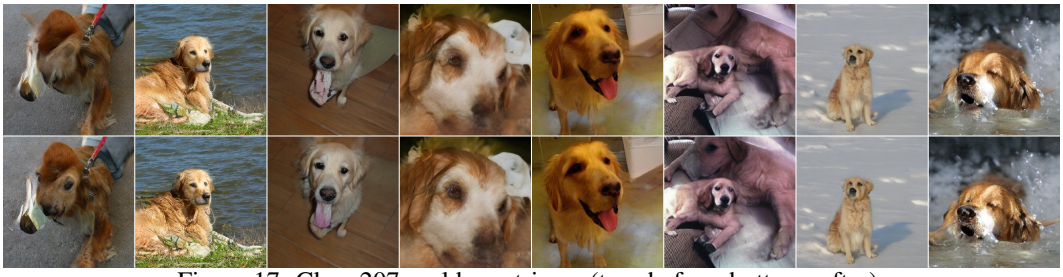


Figure 17: Class 207: golden retriever (top: before, bottom: after).

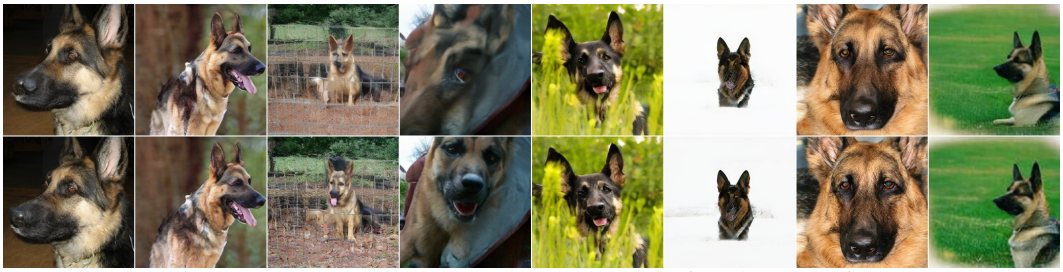


Figure 18: Class 235: German shepherd (top: before, bottom: after).

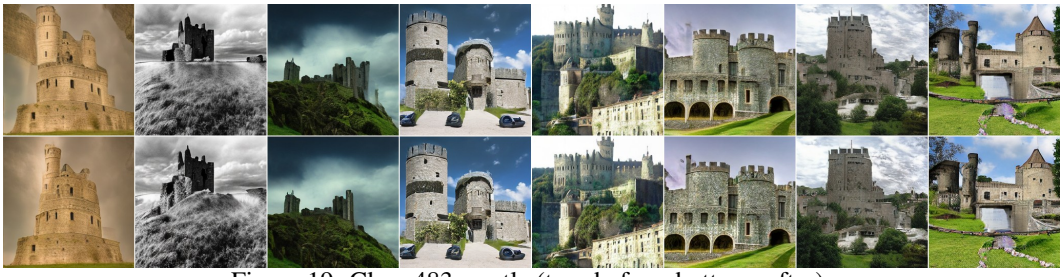


Figure 19: Class 483: castle (top: before, bottom: after).

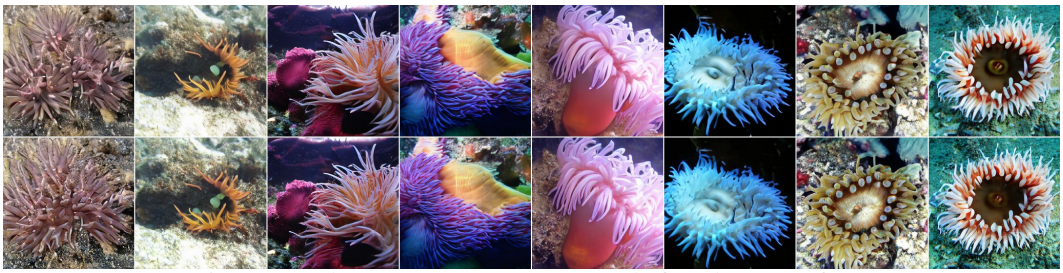


Figure 20: Class 108: sea anemone (top: before, bottom: after).

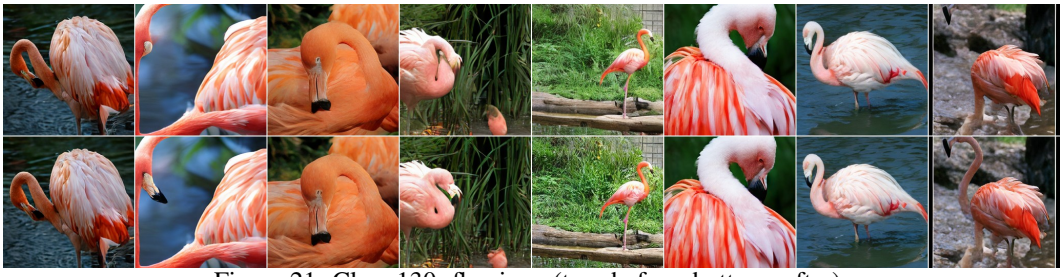


Figure 21: Class 130: flamingo (top: before, bottom: after).



Figure 22: Class 22: bald eagle (top: before, bottom: after).



Figure 23: Class 817: sports car (top: before, bottom: after).