CoAE: A Visual Tokenizer with Complementary Latent Subspaces for Image Generation

Anonymous authors
Paper under double-blind review

Abstract

Vision foundation models (VFMs) have emerged as effective encoders for image generation due to their rich semantic information. However, VFM features often lack low-level details, leading to inferior reconstruction quality. Existing methods attempt to address this issue by fusing semantic and detail information into a homogeneous latent space. This forces the two types of information to compete within the same representation, leading to a reconstruction–generation tradeoff. In this work, we propose Complementary AutoEncoder (CoAE), which organizes the original VFM features and detail features as distinct subspaces within a heterogeneous latent space, allowing them to preserve their own structures and serve complementary roles. We further introduce a diffusion training strategy tailored to this latent organization. Ablation studies show that homogeneous latent designs rely on alignment to improve generation quality at the cost of reconstruction, while CoAE performs best without alignment. On class-conditional ImageNet 256×256, CoAE achieves strong performance in both reconstruction and generation, with an rFID of 0.21 and a gFID of 2.16 without guidance after only 80 epochs. CoAE also transfers effectively to text-to-image generation and outperforms other latent spaces.

1 Introduction

Diffusion models have made progress in image generation (Peebles & Xie, 2023). With the advent of latent diffusion models (LDMs) (Rombach et al., 2022), the visual tokenizer has become a critical component, as its latent space determines both reconstruction and generation quality. Recent works show that vision foundation models (VFMs) provide diffusion-friendly feature spaces with rich semantic information, making them suited as latent representations for generation (Yu et al., 2025; Zheng et al., 2026; 2025; Tong et al., 2026). This has motivated a series of VFM-based tokenizers.

Despite these advances, reconstruction and generation have long been competing objectives in LDMs, as reconstruction favors low-level details while generation favors semantic information. VFM features excel at the latter but lack the former, leaving VFM-based tokenizers with a reconstruction gap. Closing this gap by introducing detail information is not difficult, the challenge is doing so while preserving the original diffusion-friendly manifold structure. The prevailing approach is to fuse detail information into the VFM space while aligning the resulting features with the original VFM features, hoping to obtain a latent that is friendly to both tasks. Yet this goal is difficult to achieve, because a representation is fundamentally asked to serve two conflicting objectives.

We argue that how information is organized in the latent space is the key factor. Our method, Complementary AutoEncoder (CoAE), organizes VFM features and detail features as distinct subspaces rather than fusing them into a new homogeneous space. This design allows each subspace to preserve its own structure and serve complementary roles. Rather than using fusion to construct a new latent that balances semantics and details, CoAE focuses on effectively modeling and utilizing the two representations already available. The detail subspace improves reconstruction, but generation also requires the diffusion model to predict it accurately. However, the two subspaces have basically different distributions and dimensions, and training a diffusion model on their concatenation leads to gradient imbalance and training instability. We address these issues with a group-wise MSE objective and split prediction heads. Experiments show that jointly denoising both subspaces allows

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.

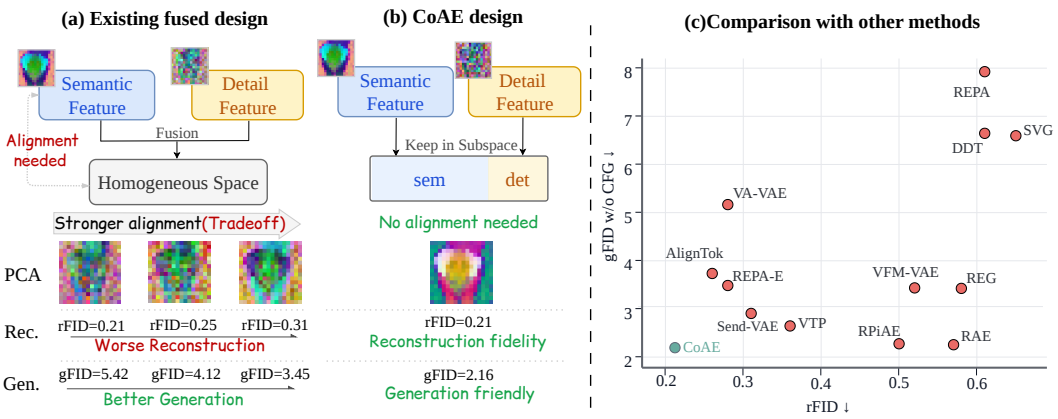

Figure 1: (a) Existing fused design forms a homogeneous latent and causes a tradeoff. (b) CoAE keeps semantic and detail features in distinct subspaces. (c) Comparison on ImageNet 256×256 at 80 epochs; $*$ denotes 64-epoch results.

detail prediction to benefit from the generation-friendly semantic prior provided by the VFM target, leading to better generation performance.

To study how latent organization affects reconstruction and generation, we conduct systematic alignment experiments across multiple fusion strategies. In fused latent variants, stronger VFM alignment improves generation but hurts reconstruction, revealing the classic tradeoff. In contrast, for CoAE, forcing alignment makes both reconstruction and generation worse, because it undermines the distinct roles that each subspace is designed to serve. This contrast suggests that the key to balancing reconstruction and generation is not stronger alignment, but clearer distinction.

CoAE achieves strong performance on both reconstruction and generation. On ImageNet 256×256, it reaches an rFID of 0.21 and a gFID of 2.16 at 80 epochs. It also obtains an FD_r^6 of 3.05 at 80 epochs, better than RAE's FD_r^6 of 3.26 at 800 epochs. With longer training, CoAE reaches a gFID of 1.02. When extended to text-to-image generation, CoAE also shows better generation quality than other latent spaces.

Our contributions are summarized as follows:

- We propose CoAE, which organizes semantic and detail information into complementary subspaces and introduces a diffusion training strategy tailored to this organization.
- Through ablation studies of different latent organizations and VFM alignment strengths, we show that keeping semantic and detail features distinct is more important than enforcing stronger alignment.
- By preserving the strengths of both semantic and detail features, CoAE achieves strong reconstruction and generation performance on class-conditional ImageNet 256×256 and transfers effectively to text-to-image generation.

2 Related Work

2.1 Visual Tokenizers for Image Generation

Latent diffusion models generate images in a compressed latent space learned by a visual tokenizer (Rombach et al., 2022). Early tokenizers based on VQ-VAE (van den Oord et al., 2018), VQGAN (Esser et al., 2021), or continuous VAEs (Rombach et al., 2022; Peebles & Xie, 2023) are optimized for reconstruction fidelity, producing latent spaces that are not inherently generation-friendly and typically require long diffusion training (Peebles & Xie, 2023; Ma et al., 2024). These observations have motivated tokenizer designs that consider not only reconstruction fidelity, but also whether the latent space contains high-level representation structure useful for generation (Yao et al., 2025; Yue et al., 2026; Yao et al., 2026).

2.2 VFM Representations for Image Generation

Recent work has found that pretrained vision foundation models provide feature spaces with rich semantic structure that benefits image generation. One line of methods uses VFM features as external alignment signals during diffusion training (Yu et al., 2025; Singh et al., 2026a; Wang et al., 2026b), injecting semantic priors to improve optimization efficiency and sample quality. Another line introduces VFM features into the tokenizer training process. Some methods align the VAE latent space to VFM representations (Yao et al., 2025; Liu et al., 2026a; Page et al., 2026), while others discard the VAE entirely and directly use frozen VFM features as the latent space (Zheng et al., 2026; Gao et al., 2025). However, high-level VFM features are not designed to preserve fine visual details, so these tokenizers still face a reconstruction gap. This has motivated a series of VFM-based tokenizer designs that seek to recover missing details while preserving the semantic structure.

2.3 Advances in VFM-Based Tokenizers

VFM-based tokenizers face the same core challenge: how to recover the missing visual details, while preserving the semantic structure that makes them generation-friendly. Existing methods pursue this goal through different paths. Some fine-tune the VFM encoder on reconstruction tasks (Tang et al., 2026; Chang et al., 2026), which improves detail but risks disrupting the pretrained representation structure and thus requires alignment to limit the drift. Others introduce extra trainable branches to supply reconstruction-oriented details, together with additional alignment or regularization objectives (Shi et al., 2026; Yue et al., 2026; Liu et al., 2026b; Fan et al., 2025). Since the branch output is typically merged with the VFM feature into a single latent, these constraints are needed to prevent the combined representation from drifting too far. Still others exploit the layer-wise structure of VFMs, fusing shallow features that retain more detail with deeper semantic features (Singh et al., 2026b; Zhu et al., 2026; Jia et al., 2026), or design multi-stage training strategies to balance reconstruction and representation preservation (Chen et al., 2026; Gong et al., 2026).

Although these methods differ in architecture and training strategy, they generally combine semantic and detail information into a homogeneous latent. Since the two types of information compete within a shared representation structure, alignment or regularization is needed to preserve generation-friendly semantics, at the cost of reconstruction. We view this tradeoff as a consequence of latent organization rather than the particular mechanism used to inject details. CoAE instead preserves semantic and detail information as distinct subspaces within a heterogeneous latent, allowing them to serve complementary roles without explicit alignment.

3 Method

CoAE is based on a simple idea: semantic and detail information are complementary, but they should not be fused into a single homogeneous latent. Instead, they should be preserved as distinct subspaces, and the generative model should be adapted to the resulting heterogeneous latent space. Accordingly, our method consists of three parts: latent organization, the CoAE tokenizer, and diffusion training adaptation.

3.1 Latent Organization

Let x denote an input image and let N be the number of spatial tokens. A VFM-based tokenizer naturally provides a semantic latent space $\mathbf{z}^{\mathrm{sem}} \in \mathbb{R}^{N \times C_s}$, which preserves the pretrained visual structure that makes VFM representations effective for diffusion generation. Faithful reconstruction also requires low-level information, represented as a detail feature $\mathbf{z}^{\mathrm{det}} \in \mathbb{R}^{N \times C_d}$. The key design question is how these two features should be organized in the latent space.

Homogeneous Fusion. A common choice is to merge semantic and detail information into a homogeneous C_s-dimensional latent space. We consider three forms in our experiments: Additive,

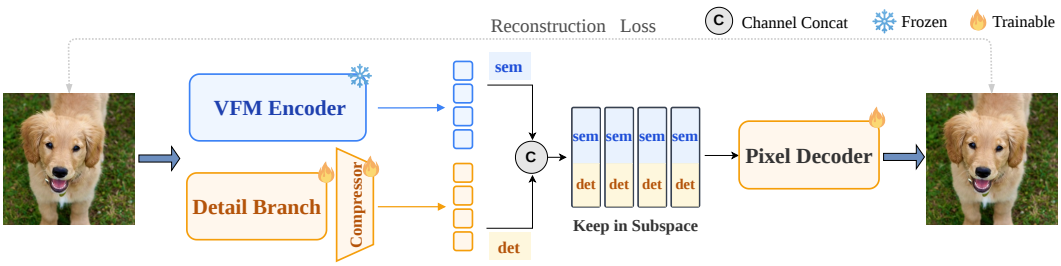

Figure 2: Tokenizer design of CoAE. A frozen VFM encoder preserves the semantic subspace, while a trainable detail branch provides the detail subspace.

Projection, and Modulation.

$$\mathbf{z}_{\text{add}} = \mathbf{z}^{\text{sem}} + W_a \mathbf{z}^{\text{det}}, \tag{1}$$

$$\mathbf{z}_{\text{proj}} = W_m[\mathbf{z}^{\text{sem}}; \mathbf{z}^{\text{det}}], \tag{2}$$

$$\mathbf{z}_{\text{mod}} = \mathbf{z}^{\text{sem}} \odot (1 + \gamma(\mathbf{z}^{\text{det}})) + \beta(\mathbf{z}^{\text{det}}). \tag{3}$$

Here W_a and W_m are linear projections, while $\gamma(\cdot)$ and $\beta(\cdot)$ predict channel-wise modulation parameters. Although implemented differently, all three variants collapse semantic and detail information into a homogeneous latent, modifying the original VFM representation.

Distinct Subspaces. An alternative is to preserve semantic and detail information in distinct subspaces. The simplest implementation is channel-wise concatenation:

$$\mathbf{z} = [\mathbf{z}^{\text{sem}}; \mathbf{z}^{\text{det}}] \in \mathbb{R}^{N \times (C_s + C_d)}. \tag{4}$$

In this form, the two feature types occupy separate channel groups while preserving their respective structures. We compare this organization with the fusion variants in Section 4.2.

3.2 CoAE Tokenizer

Following this organization, CoAE extracts semantic features with a frozen VFM encoder and complementary details with a trainable branch, and jointly decodes them to reconstruct the input image.

Semantic subspace. Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, a frozen VFM encoder E_v produces semantic features $\mathbf{z}^{\text{sem}} = E_v(x) \in \mathbb{R}^{N \times C_s}$, which form the semantic part of the final latent.

Detail subspace. Complementary detail information can be obtained from different sources: shallow VFM features (Singh et al., 2026b; Zhu et al., 2026), or an additional detail branch (Shi et al., 2026). We use the latter by default and show in Section 4.3 that CoAE does not depend on a specific detail source. In our main implementation, an independent trainable branch and a compressor produce the detail subspace: $\mathbf{z}^{\text{det}} = P_d(E_d(x)) \in \mathbb{R}^{N \times C_d}$. Here E_d is a lightweight ViT, and P_d is composed of an attention block and a convolutional projection to obtain a low-dimensional detail subspace for diffusion learning.

We then concatenate the two subspaces along the channel dimension to obtain the heterogeneous CoAE latent $\mathbf{z} = [\mathbf{z}^{\text{sem}}; \mathbf{z}^{\text{det}}] \in \mathbb{R}^{N \times (C_s + C_d)}$. A pixel decoder D reconstructs the image as $\hat{x} = D(\mathbf{z})$, and the tokenizer is trained with standard reconstruction losses, including pixel-level L1 loss, perceptual LPIPS loss, and adversarial loss:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{L1}}(x, \hat{x}) + \lambda_p \mathcal{L}_{\text{LPIPS}}(x, \hat{x}) + \lambda_g \mathcal{L}_{\text{GAN}}(x, \hat{x}). \tag{5}$$

During training, the VFM encoder remains frozen.

Unlike fusion-based tokenizers, CoAE does not require an additional alignment loss between semantic and detail features. Since the two types of information are designed to preserve their own structure, the effect of alignment is studied experimentally in Section 4.2.

3.3 Diffusion Training on Heterogeneous Latents

After training the tokenizer, we learn a diffusion model directly on the CoAE latent using flow matching. Given a clean latent $\mathbf{z}_0 = [\mathbf{z}_0^{\text{sem}}; \mathbf{z}_0^{\text{det}}] \in \mathbb{R}^{N\times(C_s+C_d)}$ and Gaussian noise $\boldsymbol{\epsilon}$ of the same shape, we define

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\boldsymbol{\epsilon}, \qquad \mathbf{v}_t = \boldsymbol{\epsilon} - \mathbf{z}_0,$$

and train the model to predict the complete velocity $\hat{\mathbf{v}}_t \in \mathbb{R}^{N\times(C_s+C_d)}$ from $\mathbf{z}_t$, timestep t, and conditioning signal c.

A standard element-wise MSE implicitly weights the two subspaces in proportion to their channel counts, causing optimization to be dominated by the semantic subspace.

We therefore use a group-wise MSE. To compute this loss, we split the predicted and target velocities along the channel dimension as $\mathbf{v}_t = [\mathbf{v}_t^{\text{sem}}; \mathbf{v}_t^{\text{det}}]$ and $\hat{\mathbf{v}}_t = [\hat{\mathbf{v}}_t^{\text{sem}}; \hat{\mathbf{v}}_t^{\text{det}}]$, where $\mathbf{v}_t^{\text{sem}}, \hat{\mathbf{v}}_t^{\text{sem}} \in \mathbb{R}^{N\times C_s}$ contain the first C_s semantic channels, and $\mathbf{v}_t^{\text{det}}, \hat{\mathbf{v}}_t^{\text{det}} \in \mathbb{R}^{N\times C_d}$ contain the remaining C_d detail channels. We then compute the MSE separately for the two channel groups:

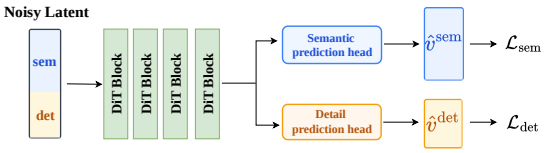

Figure 3: Diffusion training on the heterogeneous CoAE latent. A shared backbone jointly models the complete latent, while subspace-specific output heads and a group-wise loss operate on the semantic and detail channel groups.

$$\begin{aligned}\mathcal{L}_{\text{sem}} &= \mathbb{E}_{x,t,\boldsymbol{\epsilon}}\left[\frac{1}{NC_s}\left\|\hat{\mathbf{v}}_t^{\text{sem}} - \mathbf{v}_t^{\text{sem}}\right\|_2^2\right], \\ \mathcal{L}_{\text{det}} &= \mathbb{E}_{x,t,\boldsymbol{\epsilon}}\left[\frac{1}{NC_d}\left\|\hat{\mathbf{v}}_t^{\text{det}} - \mathbf{v}_t^{\text{det}}\right\|_2^2\right].\end{aligned} \tag{6}$$

We optimize $\mathcal{L}_{\text{FM}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{det}}$. This objective effectively increases the relative weight of the detail subspace, preventing it from being overwhelmed by the semantic channels, without introducing additional parameters.

The semantic and detail velocity targets also differ in distribution. Yet a standard diffusion model predicts all velocity channels through a single prediction head. We instead use two independent prediction heads to predict $\hat{\mathbf{v}}_t^{\text{sem}} \in \mathbb{R}^{N\times C_s}$ and $\hat{\mathbf{v}}_t^{\text{det}} \in \mathbb{R}^{N\times C_d}$ respectively. This change further improves generation quality while adding only 2M parameters.

4 Experiments

4.1 Class-Conditional Generation

Datasets and metrics. We evaluate CoAE for class-conditional image generation on ImageNet-1K (Deng et al., 2009) at 256×256 resolution. For reconstruction, we report PSNR and reconstruction FID (rFID) on the ImageNet validation set. For generation, we report generative Fréchet Inception Distance (gFID), Inception Score (IS), Precision, and Recall using the ADM evaluation suite (Dhariwal & Nichol, 2021). We additionally report FD_r^6 (Yang et al., 2026), which averages Fréchet distances across six representation spaces. All generation metrics are evaluated on 50K samples generated with EMA weights and a 50-step Euler ODE sampler.

Implementation details. In our default setting, we use DINOv2-B/14 (Oquab et al., 2024) as the frozen VFM encoder, with channel dimension $C_s = 768$. To extract detail features, we use an independent trainable 6-layer ViT branch (43M), followed by a compressor that reduces the detail dimension to $C_d = 32$. The full CoAE latent has size $16 \times 16 \times 800$. We use a ViT-XL decoder (415M) and train the tokenizer for 16 epochs using the reconstruction objective in Eq. (5). For downstream generation, CoAE preserves the complete VFM feature without channel compression, so we use DiT^{DH}-XL (839M) to model the high-dimensional latent. We train the diffusion model with the strategy described in Sec. 3.3, which adds only 2M parameters, while keeping other training settings the same as RAE (Zheng et al., 2026). More details are provided in Appendix A.

Table 1: Comparison on class-conditional ImageNet 256×256.

	Reconstruction		Generator		Generation w/o Guidance				Generation w/ Guidance			
Method	**rFID↓**	**PSNR↑**	**Params**	**Epochs**	**gFID↓**	**IS↑**	**Prec.↑**	**Rec.↑**	**gFID↓**	**IS↑**	**Prec.↑**	**Rec.↑**
REPA (Yu et al., 2025)	0.61	26.90	675M	80	7.90	122.60	0.70	0.65	1.42	305.70	0.80	0.65
VA-VAE (Yao et al., 2025)	0.28	**27.96**	675M	64	5.14	130.20	0.76	0.62	2.11	252.30	0.81	0.58
REG (Wu et al., 2025)	0.58	–	675M	80	3.40	184.10	–	–	1.86	**321.40**	0.76	**0.66**
REPA-E (Leng et al., 2025)	0.28	26.25	675M	80	3.46	159.80	0.77	0.63	1.67	266.30	0.80	0.63
DDT (Wang et al., 2026a)	0.61	26.90	675M	80	6.62	135.20	0.69	**0.67**	1.52	263.70	0.78	0.63
SVG (Shi et al., 2026)	0.65	23.89	675M	80	6.57	137.90	–	–	3.54	207.60	–	–
DINO-SAE (Chang et al., 2026)	0.37	26.20	839M	80	3.07	209.70	0.80	0.63	–	–	–	–
VFM-VAE (Bi et al., 2026)	0.52	22.99	675M	80	3.41	160.40	–	–	2.16	232.80	**0.82**	0.58
AlignTok (Chen et al., 2026)	0.26	25.83	675M	64	3.71	148.90	0.77	0.62	1.90	260.90	0.81	0.61
Send-VAE (Page et al., 2026)	0.31	27.62	675M	80	2.88	175.30	0.78	0.62	**1.41**	301.70	0.79	0.65
RPiAE (Gong et al., 2026)	0.50	21.30	675M	80	2.25	208.70	0.81	0.60	1.51	225.90	0.79	0.65
VTP (Yao et al., 2026)	0.36	–	675M	80	2.62	197.80	0.79	0.62	1.44	238.20	0.80	0.63
RAE (Zheng et al., 2026)	0.57	18.93	839M	80	**2.16**	**214.80**	**0.82**	0.59	–	–	–	–
DRoRAE (Zhu et al., 2026)	0.29	24.32	839M	80	2.68	197.80	0.80	0.59	1.65	230.60	0.81	0.60
CoAE	**0.21**	26.79	841M	80	**2.16**	207.08	0.79	0.61	**1.41**	236.90	0.81	0.62
SD-VAE (Rombach et al., 2022)	0.61	26.90	675M	1400	9.62	121.50	0.67	0.67	2.27	278.20	**0.83**	0.57
REPA (Yu et al., 2025)	0.61	26.90	675M	800	5.78	158.30	0.70	**0.68**	1.29	**306.30**	0.79	0.64
REPA-E (Leng et al., 2025)	0.28	26.25	675M	800	1.70	217.30	0.77	0.66	1.15	304.00	0.79	0.66
VA-VAE (Yao et al., 2025)	0.28	**27.96**	675M	800	2.17	205.60	0.77	0.65	1.35	295.30	0.79	0.65
SVG (Shi et al., 2026)	0.65	23.89	675M	1400	3.36	181.20	–	–	1.92	264.90	–	–
VTP (Yao et al., 2026)	0.36	–	675M	600	1.85	232.30	**0.79**	0.63	1.11	279.50	0.79	**0.67**
RAE (Zheng et al., 2026)	0.57	18.93	839M	800	**1.51**	**242.90**	**0.79**	0.63	1.13	262.60	0.78	**0.67**
CoAE	**0.21**	26.79	841M	640	1.83	231.00	0.77	0.65	**1.02**	262.10	0.79	0.66

Table 2: Representation Fréchet distance (FD_r) comparison on ImageNet 256×256. The Incep., ConvNeXt, DINOv2, MAE, SigLIP, and CLIP columns report FD_r computed in the corresponding representation space. FD_r^6 averages distances over six feature spaces.

Method	**Epochs**	**Incep.↓**	**ConvNeXt↓**	**DINOv2↓**	**MAE↓**	**SigLIP↓**	**CLIP↓**	**FD_r^6 ↓**
SiT (Ma et al., 2024)	800	1.26	2.02	7.89	5.62	16.14	17.69	8.44
DDT (Wang et al., 2026a)	800	0.75	**1.02**	4.26	4.11	10.16	13.86	5.70
REPA (Yu et al., 2025)	800	0.85	1.22	4.27	3.85	9.87	12.65	5.45
LightningDiT (Yao et al., 2025)	800	0.85	1.09	3.76	3.02	8.47	10.21	4.57
REPA-E (Leng et al., 2025)	800	0.70	1.28	2.44	2.52	5.04	6.28	3.04
RAE (Zheng et al., 2026)	800	0.69	1.79	2.11	3.30	3.79	7.87	3.26
CoAE	80	0.85	1.45	2.13	2.83	5.17	5.87	3.05
CoAE	640	**0.63**	1.22	**1.60**	**2.29**	**3.38**	**5.02**	**2.36**

Main Results. Table 1 compares CoAE with other methods on class-conditional ImageNet generation. CoAE achieves a strong balance between reconstruction and generation. For reconstruction, CoAE obtains an rFID of 0.21, substantially improving over the RAE baseline. Meanwhile, CoAE reaches a gFID of 2.16 at only 80 epochs, comparable to RAE, showing fast convergence efficiency. With longer training, CoAE reaches a gFID of 1.02 with AutoGuidance (Karras et al., 2024) at 640 epochs. Table 2 further reports FD_r^6. CoAE reaches 3.05 at only 80 epochs, outperforming RAE trained for 800 epochs (3.26). Figure 4 shows class-conditional samples generated by CoAE.

4.2 Analysis

Latent Organization. We analyze the latent organization choices defined in Section 3.1: three fused variants, *Additive*, *Projection*, and *Modulation*, and the distinct-subspace organization, *Concat*. All tokenizers are trained under the same setting and differ only in the fusion strategy. For diffusion training, the fused variants use the standard objective, whereas *Concat* uses the adapted strategy proposed in our method. We evaluate both latent properties and downstream generation quality. For latent properties, CKA similarity (Kornblith et al., 2019) measures representation similarity to the original VFM and detail features, denoted as CKA_v and CKA_d, respectively, and linear probing accuracy (LP@1) assesses semantic quality. The original VFM feature (DINOv2-B) is included as a reference. More details are provided in Appendix B.1.



Figure 4: Class-conditional ImageNet samples generated by CoAE at 80 epochs and 256×256 resolution.

As shown in Table 3, without alignment, *Concat* closely preserves the latent properties of the original VFM. In contrast, fused variants drift away from the VFM feature. For example, *Modulation* shows low similarity to the VFM feature ($\text{CKA}_v = 0.16$) but high similarity to the detail feature ($\text{CKA}_d = 0.97$). Its LP@1 also drops from 81.01 to 71.54. Thus, fused latents often require additional alignment to avoid drifting too far from the original VFM representation. We further add an ℓ_2 alignment loss to the tokenizer objective in Eq. (5). This loss aligns the detail-branch feature $E_d(x) \in \mathbb{R}^{N \times C_s}$ with the frozen semantic feature $\mathbf{z}^{\text{sem}} \in \mathbb{R}^{N \times C_s}$:

$$\mathcal{L}_{\text{align}} = \|E_d(x) - \mathbf{z}^{\text{sem}}\|_2^2, \qquad \mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{rec}} + \lambda_a \mathcal{L}_{\text{align}}.$$

We set $\lambda_a \in \{0, 0.05, 1.0\}$ for each organization strategy.

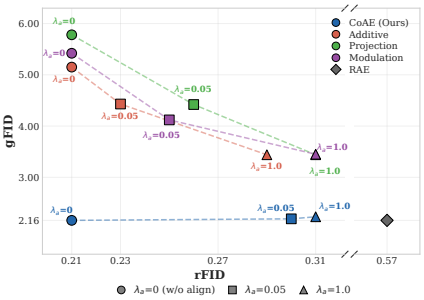

Figure 5: Reconstruction–generation tradeoff under different organization strategies and alignment strengths ($\lambda_a \in \{0, 0.05, 1.0\}$). Fused variants trace a clear tradeoff trajectory as λ_a increases.

Table 3 and Figure 5 present the full alignment sweep. Fused variants use one homogeneous latent to serve two competing tasks and show a clear reconstruction–generation tradeoff. Although stronger alignment improves gFID, rFID degrades at the same time. In contrast, *Concat* achieves the best balance without alignment. Adding alignment to *Concat* degrades both reconstruction and generation (rFID $0.21 \rightarrow 0.31$, gFID $2.16 \rightarrow 2.29$). This shows that CoAE benefits from organizing semantic and detail information into complementary subspaces, while forcing alignment makes their representations more homogeneous and undermines this design.

Why CoAE Works. We conduct a stage-wise comparison to understand the source of CoAE's effectiveness. The first row of Table 4 corresponds to the RAE baseline, which uses the VFM feature as both the tokenizer latent and the diffusion target, achieving strong generation (gFID 2.16) but limited reconstruction (rFID 0.57). We then train a detail-only tokenizer by removing the VFM branch from CoAE while keeping the remaining architecture unchanged, so the decoder reconstructs images only from the learned detail feature. Keeping this detail-only tokenizer fixed, we compare three diffusion training strategies: (i) directly training diffusion on the detail feature, (ii) retaining the same diffusion target but applying REPA alignment (Yu et al., 2025), which uses the VFM feature as auxiliary supervision for diffusion hidden states, and (iii) concatenating the frozen VFM feature with the detail feature and training diffusion on the joint $[\text{VFM}; \text{Detail}]$ target using our diffusion training adaptation. Unlike REPA, this setting makes the complete VFM feature part of the diffusion target, so the model learns its denoising process in context with the detail feature.

Directly training diffusion on the detail feature gives poor generation (gFID 11.51), and REPA alignment improves gFID to 9.25, whereas the joint target effectively reduces gFID to 2.54. This shows that CoAE's gain largely comes from using the VFM feature as a denoising target. Finally, the complete CoAE uses the joint $[\text{VFM}; \text{Detail}]$ latent in both tokenizer and diffusion training, yield-

Table 3: Latent organization analysis under different alignment strengths. gFID is evaluated at 80 epochs without CFG.

Organization	Latent Property			Reconstruction			Generation	
	CKA_v	CKA_d	L.P. Acc.↑	PSNR↑	SSIM↑	rFID↓	gFID↓	IS↑
DINOv2-B (VFM)	1.00	–	81.01	–	–	–	–	–
Concat	0.99	0.10	81.01	26.79	0.78	**0.21**	**2.16**	207.08
+ $\lambda_a = 0.05$	0.99	0.11	81.01	25.14	0.71	0.30	2.19	**210.27**
+ $\lambda_a = 1.0$	0.99	0.11	**81.05**	24.87	0.70	0.31	2.29	209.29
Additive	0.31	0.94	80.79	27.16	0.79	**0.21**	5.15	158.18
+ $\lambda_a = 0.05$	0.38	0.90	81.01	26.41	0.76	0.23	4.43	169.30
+ $\lambda_a = 1.0$	0.74	0.68	81.03	25.23	0.72	0.29	3.44	183.67
Projection	0.14	0.95	75.52	**27.35**	0.79	**0.21**	5.78	145.50
+ $\lambda_a = 0.05$	0.18	0.91	79.12	26.24	0.75	0.26	4.42	166.01
+ $\lambda_a = 1.0$	0.26	0.85	79.42	24.80	0.70	0.31	3.44	179.97
Modulation	0.16	0.97	71.54	27.27	0.79	**0.21**	5.42	148.49
+ $\lambda_a = 0.05$	0.26	0.94	79.00	26.12	0.75	0.25	4.12	168.34
+ $\lambda_a = 1.0$	0.41	0.81	80.24	24.74	0.70	0.31	3.45	178.87

Table 4: Analysis with different tokenizer latents and diffusion targets. Directly denoising VFM features is more effective than using VFM features for alignment.

Tokenizer Latent	PSNR↑	rFID↓	Diffusion Target	gFID↓
VFM (RAE)	18.93	0.57	VFM	**2.16**
Detail	26.72	0.25	Detail	11.51
Detail	26.72	0.25	Detail w/ REPA	9.25
Detail	26.72	0.25	[VFM;Detail]	2.54
[VFM;Detail] (CoAE)	**26.79**	**0.21**	[VFM;Detail]	**2.16**

Table 5: Ablation of the diffusion training strategy on ImageNet 256×256 at 80 epochs. Group-wise MSE largely improves generation quality, while separate prediction heads provide a further gain.

Configuration	gFID↓	IS↑
Standard training	16.75	112.25
+ Group-wise MSE	2.51	203.32
+ Separate prediction heads	**2.16**	**207.08**

ing the best performance. This indicates that introducing the VFM feature in tokenizer training forms complementary subspaces with the detail feature, further improving both reconstruction and generation.

4.3 Ablation and Generalization

Diffusion Training Strategy Ablation. We ablate the diffusion training strategy proposed for heterogeneous CoAE latents. As shown in Table 5, directly applying the standard diffusion objective performs poorly, with a gFID of 16.75. The group-wise MSE loss reduces gFID to 2.51 without adding any parameters, showing that balancing the two subspaces is critical. Using separate prediction heads further improves gFID to 2.16, while adding only 2M parameters.

Backbone and Detail Dimension. Table 6 compares different tokenizer configurations. We vary the detail dimension C_d and test different VFM backbones, including DINOv2-B (Oquab et al., 2024), SigLIP2 (Tschannen et al., 2025), and DINOv3-L (Siméoni et al., 2025). Increasing C_d improves reconstruction, while $C_d = 32$ already provides strong generation quality. CoAE also remains effective across different VFM backbones.

Detail Source. We further compare different detail sources. Besides the default detail branch, we use shallow VFM features as the detail source. Specifically, we

Table 6: Generalization to different tokenizer configurations on ImageNet 256×256 at 80 epochs.

Config	PSNR↑	rFID↓	gFID↓	IS↑
Detail dimension C_d				
$C_d = 16$	25.39	0.28	**2.16**	**211.05**
$C_d = 32$	26.79	0.21	**2.16**	207.08
$C_d = 64$	**27.34**	**0.19**	2.34	204.02
VFM backbone ($C_d = 32$)				
DINOv2-B	26.79	0.21	**2.16**	207.08
SigLIP2	26.18	0.27	3.04	189.93
DINOv3-L	**26.93**	**0.20**	2.21	**219.15**

Table 7: Effect of detail source and latent organization on ImageNet 256×256 at 80 epochs. Keeping semantic and detail information as distinct subspaces consistently improves generation across different detail sources.

Detail Source	Organization	PSNR↑	rFID↓	gFID↓	IS↑
Independent branch	Additive	**27.16**	**0.21**	5.15	158.18
Independent branch	Concat	26.79	**0.21**	2.16	207.08
Shallow VFM	Additive	25.13	0.23	4.06	161.94
Shallow VFM	Concat	25.04	0.23	**1.97**	**210.51**



Figure 6: Qualitative text-to-image generated at 256×256 resolution using the finetuned models.

take the layer-4 VFM feature, reduce it to $C_d = 32$, and concat it with the semantic VFM feature. More detail source experiments are provided in Appendix B.3.

Table 7 shows that CoAE is not tied to a specific detail-source architecture. With both the independent branch and shallow VFM features, *Concat* gives much better generation than additive fusion. This supports our main claim that the key factor is how semantic and detail information are organized, rather than the particular architecture used to obtain detail features.

4.4 Text-to-Image Generation

We further evaluate CoAE on text-to-image generation at 256×256 resolution, comparing FLUX VAE (Labs, 2024), RAE, Additive Variant, and CoAE under the same protocol. Following the setup of RAEv2 (Tong et al., 2026), we use Qwen3-0.6B (Yang et al., 2025) to encode prompts into 256 conditioning tokens for DiT^{DH}-XL. We train on JourneyDB (Sun et al., 2023) and BLIP3-o (Chen et al., 2025) for 200K iterations, then finetune on BLIP3-o-60K for 50 epochs. We evaluate on GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024).

Table 8 reports the quantitative comparison after finetuning under the same protocol. CoAE performs better than FLUX VAE, RAE, and additive fusion, indicating that the separated semantic-detail organization also benefits text-conditioned generation. Figure 6 provides qualitative samples generated by CoAE. Additional text-to-image training details are provided in Appendix A.3, while qualitative comparisons across different latent representations are provided in Appendix D.3.

Table 8: Text-to-image comparison at 256×256 resolution after finetuning under the same training protocol.

Tokenizer	GenEval	DPG
FLUX VAE	82.82	83.43
RAE	83.00	85.07
Additive Variant	84.92	84.27
CoAE	**86.62**	**85.29**

5 Conclusion

We propose CoAE, which preserves the VFM feature and detail information as complementary subspaces. We further introduce a diffusion training strategy tailored to this latent organization. Ablation experiments show that fused latents require alignment and follow a clear reconstruction–generation tradeoff, whereas CoAE performs best without alignment. On ImageNet 256×256, CoAE achieves an rFID of 0.21 and a gFID of 2.16 after only 80 epochs, and transfers effectively to text-to-image generation. These results suggest that latent organization is an important factor in visual tokenizers. By preserving both semantic structure and visual details, CoAE also shows potential as a tokenizer for future unified multimodal models.

AI USE STATEMENT

We used generative AI tools to polish writing, find related works, and assist with code checking and refinement of experiments. All AI-assisted suggestions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vfm-vae: Vision foundation models can be good tokenizers for latent diffusion models, 2026. URL https://arxiv.org/abs/2510.18457.

Hun Chang, Byunghee Cha, and Jong Chul Ye. Hyperspherical autoencoder for high-fidelity image reconstruction and generation, 2026. URL https://arxiv.org/abs/2601.22904.

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

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. URL https://arxiv.org/abs/2505.09568.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. URL https://arxiv.org/abs/2012.09841.

Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding, 2025. URL https://arxiv.org/abs/2512.19693.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Wbr51vK331.

Yue Gong, Hongyu Li, Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Manyuan Zhang, Dawei Leng, Yuhui Yin, and Lijun Zhang. Rpiae: A representation-pivoted autoencoder enhancing both image generation and editing, 2026. URL https://arxiv.org/abs/2603.19206.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

Mingkai Jia, Mingxiao Li, Zhijian Shu, Anlin Zheng, Liaoyuan Fan, Jiaxin Guo, Tianxing Shi, Dongyue Lu, Zeming Li, Xiaoyang Guo, Xiaojuan Qi, Xiao-Xiao Long, Qian Zhang, Ping Tan, and Wei Yin. Dino-tok: Adapting dino for visual tokenizers, 2026. URL https://arxiv.org/abs/2511.20565.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 52996–53021. Curran Associates, Inc., 2024. doi: 10.52202/079017-1679. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5ee7ed60a7e8169012224dec5fe0d27f-Paper-Conference.pdf.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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

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

Hangyu Liu, Jianyong Wang, and Yutao Sun. Geometric autoencoder for diffusion models, 2026a. URL https://arxiv.org/abs/2603.10365.

Siyu Liu, Chujie Qin, Hubery Yin, Qixin Yan, Zheng-Peng Duan, Chen Li, Jing Lyu, Chun-Le Guo, and Chongyi Li. Improving reconstruction of representation autoencoder, 2026b. URL https://arxiv.org/abs/2602.08620.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024. URL https://arxiv.org/abs/2401.08740.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

John Page, Xuesong Niu, Kai Wu, and Kun Gai. Boosting latent diffusion models via disentangled representation alignment, 2026. URL https://arxiv.org/abs/2601.05823.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kdpeJNbFyf.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=y0UxFtXqXf.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders, 2026b. URL https://arxiv.org/abs/2605.18324.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. JourneyDB: A benchmark for generative image understanding. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=vfzXDRTcF4.

Hao Tang, Chen-Wei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting CLIP for unified multimodal understanding, generation and editing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6tx4BGjwJP.

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, 2026. URL https://arxiv.org/abs/2601.16208.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40633–40642, June 2026a.

Yuqing Wang, Zhijie Lin, Ceyuan Yang, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Zihan Ding, Fuyun Wang, Shuai Wang, Youliang Zhang, Haoqi Fan, and Xihui Liu. Representation forcing for bottleneck-free unified multimodal models, 2026b. URL https://arxiv.org/abs/2605.31604.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think, 2025. URL https://arxiv.org/abs/2507.01467.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation, 2026. URL https://arxiv.org/abs/2604.28190.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation, 2026. URL https://arxiv.org/abs/2512.13687.

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

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

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive generation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PESrAH82Zh.

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

Xuanyu Zhu, Yan Bai, Yang Shi, Yihang Lou, Yuanxing Zhang, Jing Jin, and Yuan Zhou. Beyond the last layer: Multi-layer representation fusion for visual tokenization, 2026. URL https://arxiv.org/abs/2605.10780.

A Implementation Details

A.1 CoAE Tokenizer

CoAE uses a frozen DINOv2-B/14 (Oquab et al., 2024) as the VFM backbone, which produces a $16 \times 16 \times 768$ semantic feature. The detail branch is a lightweight 6-layer ViT with hidden dimension 768 and 12 heads, followed by a compressor block composed of MLP, attention, and RMSNorm. It produces a $16 \times 16 \times 32$ detail feature. We concatenate the semantic and detail features along the channel dimension to obtain a $16 \times 16 \times 800$ latent. The decoder uses a ViT-XL backbone with 28 layers, hidden dimension 1152, and 16 heads, and the discriminator uses a DINO-S/8 backbone. Following RAE (Zheng et al., 2026), we train the tokenizer with L1, LPIPS, and GAN losses; the discriminator and GAN schedule also follow the same recipe. The training configuration is summarized in Table 9.

GAN discriminator. Following StyleGAN-style adversarial training, we use a DINO ViT-S/8 backbone discriminator (kernel size 9, BatchNorm, spectral normalization, recipe `S_8`), optimized with the same AdamW + cosine schedule as the tokenizer ($\mathrm{lr}{=}2{\times}10^{-4} \rightarrow 2{\times}10^{-5}$). Discriminator augmentation uses probability 1.0 with no cutout. The adversarial loss is hinge loss for the discriminator and vanilla (non-saturating) loss for the generator, with discriminator weight 0.75. LPIPS perceptual loss (VGG16 backbone) is active from epoch 0 with weight 1.0. The discriminator itself is only updated starting at epoch 6, and its loss only contributes to the tokenizer's objective starting at epoch 8 ($\mathrm{disc_start}{=}8$), with a maximum discriminator-loss weight cap of 10^4 and one discriminator update per generator step.

A.2 Diffusion Training

Table 10 reports the Stage-2 DiT setup. Before diffusion training, we compute per-channel statistics of the CoAE latents on the ImageNet validation set and use them to normalize latents during diffusion training. The table shows the 80-epoch training configuration used in the main experiments. For the long-training result reported in the main paper, we train for 640 epochs with a cosine learning-rate schedule, 40 warmup epochs, and decay the learning rate from 2×10^{-4} to 2×10^{-5}. We train the model with the group-wise objective described in the main paper. The split prediction heads are used only at the output layer.

A.3 Text-to-Image Training

For text-to-image experiments, we keep the DiT backbone and training protocol fixed and only change the latent space. Captions are encoded by Qwen3-0.6B (Yang et al., 2025) into 256 text tokens, which are used as conditioning tokens for $\mathrm{DiT}^{\mathrm{DH}}$-XL. We first train on the mixed large-scale captioned image data and then finetune on the high-quality BLIP3-o-60K subset. The same pretraining, finetuning, and sampling settings are used for FLUX VAE (Labs, 2024), RAE, Additive, and CoAE. During evaluation, samples are generated from the EMA checkpoint with classifier-free guidance and measured on GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024). Table 11 summarizes the full text-to-image training and evaluation configuration.

B Implementation Details for Ablation Experiments

B.1 Details of Latent Organization Analysis

All variants use the same frozen VFM encoder, detail branch, decoder, data, and optimizer. The only changed component is the organization operator applied after the semantic and detail features are extracted. Table 12 lists the concrete operators used for the three fusion variants and the concatenation baseline.

For the alignment sweep, we add the alignment loss only during tokenizer training and train a new diffusion model for each tokenizer. The loss is computed between the pre-compression detail-branch feature and the frozen VFM feature, both with shape $16 \times 16 \times 768$. If alignment were applied to the final latent, an additional linear projection would be needed to map each latent to the VFM

Table 9: Stage-1 tokenizer training hyperparameters. These settings are shared by the default CoAE tokenizer and the tokenizer variants unless otherwise specified.

Hyperparameter	**Value**
Epochs	16
Global batch size	256
Optimizer	AdamW (β_1=0.9, β_2=0.95)
LR schedule	Cosine, warmup 1 epoch (from 0)
Base LR / Final LR	2×10^{-4} / 2×10^{-5}
EMA decay	0.9978
L1 / LPIPS weights	1.0 / 1.0
Discriminator update	From epoch 6
GAN loss	From epoch 8
GAN weighting	Adaptive
Discriminator weight	0.75

Table 10: Stage-2 DiT training configuration.

Field	**Setting**
Backbone	DiT^{DH}-XL
Input / output dim.	$16 \times 16 \times 800$
Encoder / decoder depth	28 / 2 layers
Encoder / decoder width	1152 / 2048
Attention heads	16 / 16
MLP ratio	4.0
Patch size	1 (operates on the 16×16 latent grid)
Total params	841M
Epochs	80
Global batch size	1024
Optimizer	AdamW (β_1=0.9, β_2=0.95, wd=0)
Learning rate	2×10^{-4}
EMA decay	0.9995
Precision	BF16
Gradient clipping	1.0
Objective	Group-wise velocity prediction
Prediction heads	Semantic head + detail head
Transport	Linear path
Timestep sampling	Logit-normal $(0, 1)$
Class dropout	0.1
Sampler	Euler ODE, 50 steps
AutoGuidance (Karras et al., 2024)	DiT^{DH}-S, 20 epochs, scale 1.37

dimension before computing the ℓ_2 loss. We therefore align the same-dimensional features before compression, which keeps the implementation simple and comparable across variants. We sweep $\lambda_a \in \{0, 0.05, 1.0\}$ and keep all other Stage-1 and Stage-2 hyperparameters unchanged.

We compute CKA (Kornblith et al., 2019) by uniformly sampling 10,000 ImageNet (Deng et al., 2009) validation images with batch size 64. For each representation, we mean-pool the spatial tokens and compute centered linear CKA between the resulting feature matrices. For linear probing, the tokenizer is frozen and its spatial tokens are mean pooled. A single linear classifier is trained on ImageNet-1K for 30 epochs with AdamW, batch size 256, learning rate 0.001, and zero weight decay; top-1 accuracy is reported on the validation set.

B.2 Stage-Wise VFM Supervision

For the detail-only tokenizer, we remove the VFM branch from Stage 1 and keep the remaining tokenizer architecture unchanged. In the joint-target setting, the frozen VFM feature is added only to the Stage-2 diffusion target and is not used by the decoder. For the REPA (Yu et al., 2025)

Table 11: Text-to-image training and evaluation configuration.

Hyperparameter	Value
Resolution	256×256
Backbone	DiT^{DH}-XL
Text encoder	Qwen3-0.6B
Text tokens	256
Condition dropout	0.1
Pretraining data	JourneyDB (Sun et al., 2023) + BLIP3-o (Chen et al., 2025) long/short subsets
Pretraining iterations	200K
Finetuning data	BLIP3-o-60K (Chen et al., 2025)
Finetuning epochs	50
Global batch size	1024
LR schedule	Linear, warmup 5 epochs
Base LR / Final LR	2×10^{-4} / 2×10^{-5}
EMA decay	0.9995
Gradient clipping	1.0
Evaluation	GenEval, DPG-Bench
Sampler	Euler ODE, 50 steps
Guidance	CFG scale 6

Table 12: Fusion operators used in the latent organization analysis.

Variant	Fusion operator	Extra params.
Additive	$\mathbf{z}^{\text{sem}} + W_a\mathbf{z}^{\text{det}}$	Linear$(32, 768)$
Projection	$W_m[\mathbf{z}^{\text{sem}}; \mathbf{z}^{\text{det}}]$	Linear$(800, 768)$
Modulation	$\mathbf{z}^{\text{sem}} \odot (1+\gamma) + \beta$	Linear$(32, 1536)$
Concat	$[\mathbf{z}^{\text{sem}}; \mathbf{z}^{\text{det}}]$	None

baseline, we add an auxiliary alignment loss to the hidden state after the 8th DiT block and set its loss weight to 0.5.

B.3 Shallow VFM Detail Source

The shallow-feature variant uses the layer-4 DINOv2-B (Oquab et al., 2024) feature as the detail source and applies the same compressor as CoAE to reduce it to 32 channels. For the shallow + additive variant, the 32-channel shallow feature is projected back to 768 channels and then added to the final-layer VFM feature, instead of directly concatenating shallow and final-layer features. This keeps the reconstruction level comparable across variants, making the generation comparison cleaner. All other tokenizer and diffusion training settings are kept unchanged.

C Additional Experiments

C.1 Group-Wise Loss Weights

The main paper uses a group-wise objective that computes the MSE separately for the semantic and detail subspaces. We further test whether the result depends on the exact relative weight between the two group losses:

$$\mathcal{L}_{\text{FM}} = \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{det}}\mathcal{L}_{\text{det}}.$$

The default setting uses equal group weights $(\lambda_{\text{sem}}, \lambda_{\text{det}}) = (1, 1)$. Table 13 shows that the results are stable across these weight choices. Using equal group weights already achieves a strong gFID of 2.16, and increasing the detail weight to 1.5 only brings a marginal change. This suggests that the important factor is to compute the MSE separately for the two subspaces, rather than relying on a carefully tuned loss-weight ratio.



Figure 7: Reconstruction comparisons across RAE and CoAE. CoAE better preserves fine details, text-like patterns, and complex local structures.

C.2 Sampling Steps

We evaluate the effect of Euler ODE sampling steps using the same 80-epoch EMA checkpoint. Since CoAE preserves the VFM feature as part of the latent, the diffusion model inherits a generation-friendly representation structure. As a result, CoAE already achieves strong generation quality with relatively few sampling steps, unlike conventional VAE latents that often require more sampling steps for high-quality results. Table 14 reports the quantitative results across different numbers of Euler steps.

D Additional Qualitative Results

D.1 Reconstruction Comparisons

Figure 7 provides additional reconstruction comparisons between RAE and CoAE. RAE uses only frozen VFM latents, so its reconstructions often miss low-level information, leading to blurred textures, distorted text, and inaccurate local structures, while CoAE recovers these details faithfully.

D.2 ImageNet Generation

Figures 8 and 9 show more class-conditional ImageNet generation results. Each block corresponds to one ImageNet class and contains multiple generated samples from the same class condition.

D.3 Text-to-Image Results

Figure 10 shows outputs generated from the same prompts under the common text-to-image protocol, including FLUX VAE, RAE, the additive variant, and CoAE. The comparison shows that CoAE more faithfully follows the input prompts, especially in preserving the specified objects, attributes, and compositions. Figure 11 shows more samples generated by CoAE.

Table 13: Effect of semantic and detail loss weights in the group-wise diffusion objective.

λ_{sem}	λ_{det}	gFID↓	IS↑
0.5	0.5	2.23	209.31
1.0	1.0	2.16	207.08
1.0	1.5	**2.15**	**210.57**

Table 14: Effect of Euler ODE sampling steps using the 80-epoch EMA model.

Euler steps	gFID↓	IS↑	Pre.↑	Rec.↑
10	3.97	190.73	0.77	0.58
25	2.32	206.01	0.79	0.60
50	**2.16**	207.08	0.79	0.61
100	2.23	**207.18**	0.79	0.61
250	2.25	206.42	0.79	0.61

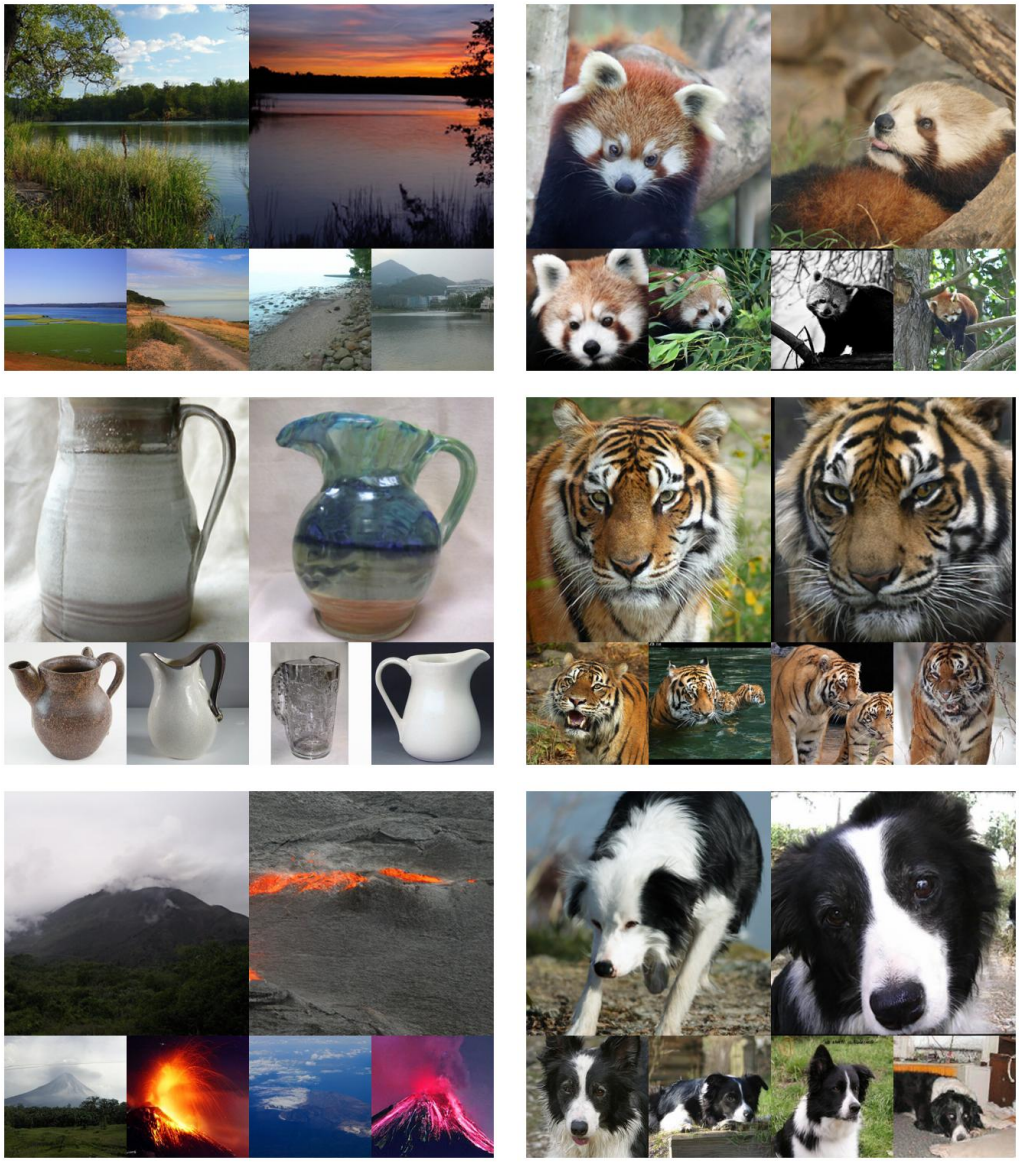

Figure 8: Unselected class-conditional ImageNet 256 × 256 samples from the 640-epoch CoAE model with AutoGuidance scale 1.37. Classes are shown from left to right and top to bottom: lakeside/lakeshore (975), lesser panda/red panda (387), pitcher/ewer (725), tiger (292), volcano (980), and Border collie (232).



Figure 9: Unselected class-conditional ImageNet 256 $\times$ 256 samples from the 640-epoch CoAE model with AutoGuidance scale 1.37. Classes are shown from left to right and top to bottom: tabby cat (281), daisy (985), geyser (974), giant panda (388), golden retriever (207), and koala (105).

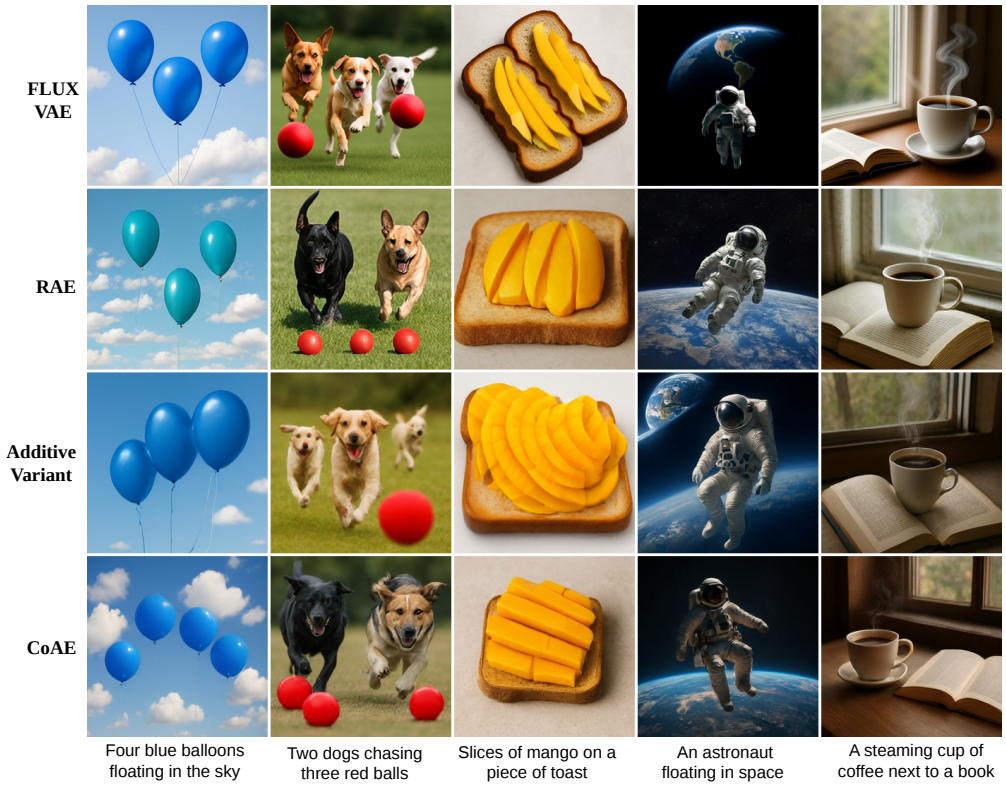

Figure 10: Qualitative text-to-image comparison at 256×256 resolution under the common fine-tuning protocol. Each row uses the same prompt across FLUX VAE, RAE, the additive variant, and CoAE.

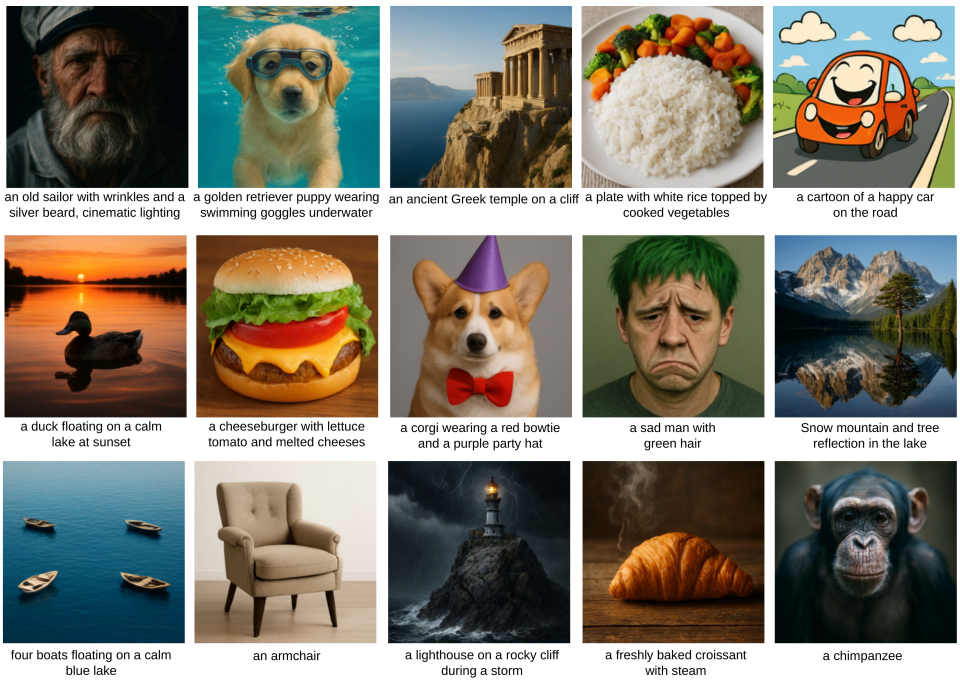

Figure 11: Text-to-image samples generated by the finetuned CoAE model with CFG scale 6.