BEYOND THE LAST LAYER: MULTI-LAYER REPRESENTATION FUSION FOR VISUAL TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation autoencoders that reuse frozen pretrained vision encoders as visual tokenizers have achieved strong reconstruction and generation quality. However, using only the last encoder layer leaves intermediate features unavailable to the decoder, limiting the reconstruction of fine visual details. We improve detail reconstruction by supplying the decoder with fused intermediate-layer features. We propose **DRoRAE** (Depth-Routed Representation AutoEncoder), which uses energy-constrained routing for adaptive multi-layer aggregation and incremental correction to retain a contribution from the original representation. Starting from a pretrained base model, we first train fusion with the decoder frozen, then fine-tune the decoder with fusion fixed. On ImageNet-256, DRoRAE reduces rFID from 0.570 to 0.290 and improves class-conditional generation FID from 1.740 to 1.650 (with AutoGuidance). For text-to-image generation, DRoRAE achieves a overall GenEval score of 0.600, compared with 0.560 for the base model. With encoder features and all other settings fixed, increasing fusion expert capacity improves reconstruction. Over the measured range, rFID decreases approximately log-linearly with fusion expert capacity (R^2=0.86).

1 INTRODUCTION

The image tokenizer maps pixels into a compact latent space and defines the quality ceiling of modern visual generation systems Rombach et al. (2022); Peebles & Xie (2023). Recent work Yu et al. (2025); Yao et al. (2025); Zheng et al. (2026); Gong et al. (2026) incorporates features from pretrained vision foundation models (VFMs), such as DINOv2 Oquab et al. (2024), to improve reconstruction fidelity and downstream generation quality over tokenizers trained from scratch.

RAE and several related VFM-based tokenizers extract features from the *last layer* of the encoder. Different encoder depths provide representations with different spatial and semantic properties Raghu et al. (2021); Amir et al. (2021). A recent analysis by Meituan LongCat Team (2026) discusses how residual connections can retain low-level information while successive layers emphasize semantic abstractions. Figure 1 illustrates how learned multi-layer fusion gives the decoder direct access to intermediate features while preserving the latent token count and channel dimension.

Realizing multi-layer fusion in practice, however, requires addressing two challenges. *(1) Content-adaptive fusion.* The fusion module must combine features from different encoder depths while allowing their contributions to vary with spatial position and image content. We learn these contributions from the reconstruction objective, allowing each token to use a signed combination of layer-specific expert outputs. *(2) Decoder and generator compatibility.* The pretrained decoder reconstructs images from the original RAE representation. Fusion must supply latents that this decoder can use while retaining the token count and channel dimension of the existing interface. Downstream generation further requires a diffusion model that can learn the fused representation.

We propose **DRoRAE** (Depth-Routed Representation AutoEncoder), the fusion module in Figure 2, to address both challenges. For content-adaptive fusion, we design an energy-constrained routing mechanism. Per-layer expert MLPs transform features from different depths, and a learned router assigns signed aggregation weights to each token, permitting both additive and subtractive combinations. Incremental correction retains a contribution from the original RAE representation by

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.

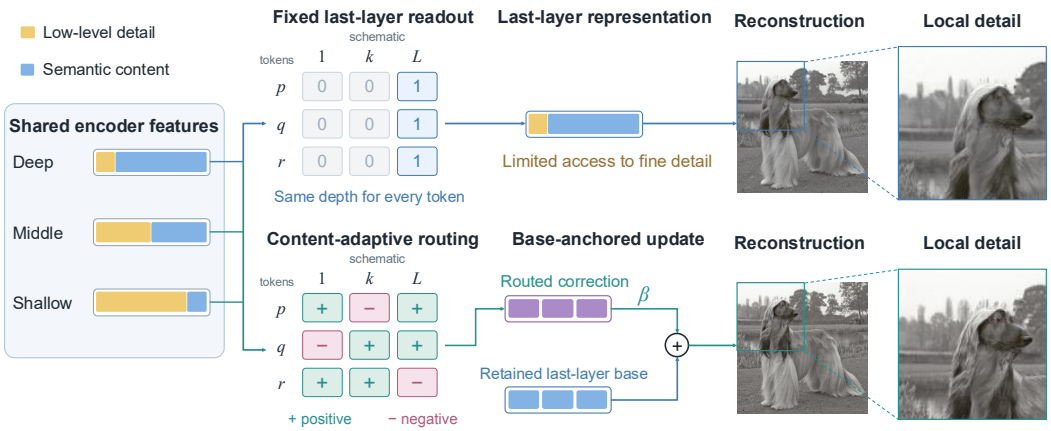


Figure 1: **Multi-layer fusion motivation.** Schematic comparison of last-layer tokenization and learned combinations of encoder layers. Signs denote positive and negative routing coefficients.

interpolating between the base and fused latents. Starting from a pretrained RAE, we learn fusion through the frozen decoder and then adapt the decoder with fusion fixed. On ImageNet-256, DRoRAE reduces rFID from 0.570 to 0.290 and improves class-conditional generation (gFID with AutoGuidance: 1.740→1.650), with gains also transferring to text-to-image synthesis. We also find that increasing fusion expert capacity yields an approximately log-linear reduction in reconstruction FID over the measured range (R^2=0.86).

Our contributions are as follows:

- We extend last-layer representation autoencoders with DRoRAE, a depth-routed fusion module that combines intermediate features through energy-constrained routing, incremental correction, and decoder-constrained training.
- DRoRAE improves ImageNet-256 reconstruction (rFID: 0.570→0.290) and class-conditional generation with AutoGuidance (gFID: 1.740→1.650), and raises GenEval from 0.560 to 0.600 over RAE in text-to-image models trained on CC12M-LLaVA-Next.
- We identify an approximately log-linear reconstruction scaling trend as fusion expert capacity increases, with the encoder, decoder, and training configuration fixed. The fit holds over the measured range (R^2=0.86).

2 Related Work

2.1 Image Tokenizers for Latent Generation

Image tokenizers map images into compact latent representations on which generative models operate. Early approaches, such as VQGAN Esser et al. (2021) and SD-VAE Rombach et al. (2022), jointly learn encoders and decoders with discrete codebooks or KL-regularized continuous latents, forming the foundation of latent generative models Peebles & Xie (2023); Ma et al. (2024). Although effective for reconstruction, these objectives do not explicitly encourage the semantic structure useful for downstream generation. Recent work therefore incorporates pretrained visual representations: REPA Yu et al. (2025) aligns intermediate diffusion features, while VA-VAE Yao et al. (2025) introduces semantic supervision into tokenizer training. Representation autoencoders take a more direct approach by reusing frozen pretrained vision encoders and learning decoders to reconstruct their inputs Zheng et al. (2026), with subsequent extensions exploring alternative organizations of the representation space Gong et al. (2026); Singh et al. (2026). Together, these approaches demonstrate the value of pretrained visual features for reconstruction and generative modeling, motivating a closer examination of which encoder representations a tokenizer should expose.

The original RAE extracts features from the final layer of the pretrained encoder Zheng et al. (2026). Different layers of a Vision Transformer encode different information, ranging from fine-grained textures and edges in shallow layers to high-level semantics in deep layers Raghu et al. (2021);

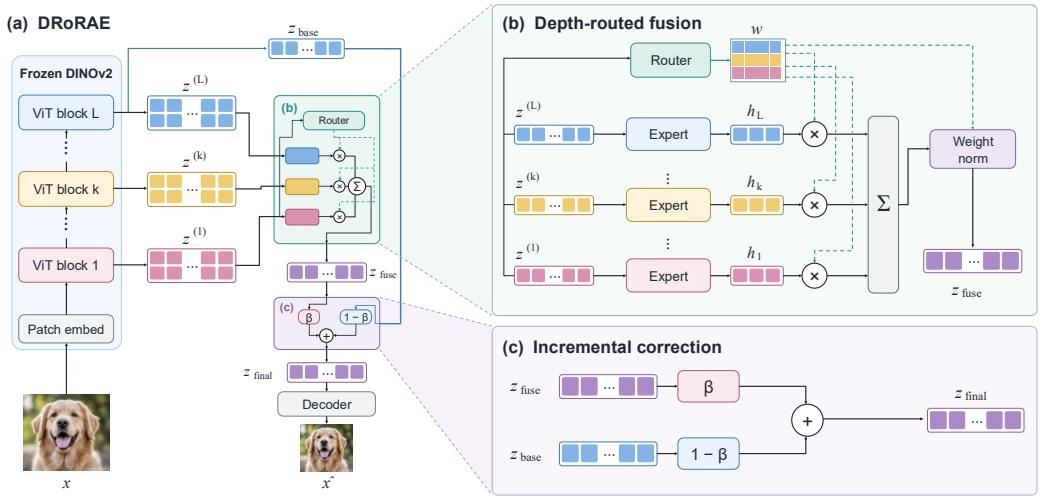


Figure 2: **DRoRAE overview.** Layer-wise experts and signed routing fuse frozen DINOv2 features; incremental correction retains the base representation.

Amir et al. (2021). This design leaves intermediate representations unavailable as direct inputs to the decoder, motivating an evaluation of explicit multi-layer fusion.

2.2 Multi-Layer Feature Utilization in Vision Models

The complementary nature of features at different depths is well established in visual understanding. Feature Pyramid Networks Lin et al. (2017), Dense Prediction Transformers Ranftl et al. (2021), and hypercolumns Hariharan et al. (2015) all aggregate multi-layer features for dense prediction tasks. Studies of ViT representations describe differences in spatial and semantic properties across depth Raghu et al. (2021); Amir et al. (2021), while residual connections can retain low-level information in the final output Meituan LongCat Team (2026). In multimodal large language models (MLLMs), Dense Connector Yao et al. (2024), MMFuser Cao et al. (2024), Chen et al. (2025), and Instruction-Guided Fusion Li et al. (2026) have further demonstrated that fusing multi-layer ViT features improves fine-grained visual understanding. For image tokenization, multi-layer fusion must also supply a representation that supports reconstruction and downstream generation.

Several learned tokenizers Esser et al. (2021); Rombach et al. (2022) and representation-based tokenizers Zheng et al. (2026); Gong et al. (2026) use a single encoder output without explicitly aggregating intermediate-layer features. We study how to integrate multi-layer fusion into the existing RAE latent interface, evaluating its benefits for reconstruction, class-conditional generation with AutoGuidance, and text-to-image synthesis.

3 Method

We present DRoRAE, an extension to the Representation Autoencoder framework that fuses multi-layer features from a frozen pretrained encoder into an enriched latent representation.

3.1 Preliminaries

We build upon the Representation Autoencoder (RAE) framework Zheng et al. (2026), which uses a frozen pretrained vision encoder $\mathcal{E}$ and trains a decoder $\mathcal{D}$. For a 256×256 target image $\mathbf{x}$, let $P(\mathbf{x})$ denote the 224×224 encoder input after resizing and preprocessing. Let $\mathbf{z}^{(k)} \in \mathbb{R}^{N \times C}$ denote the per-layer patch readouts supplied to fusion. Here, N is the number of spatial tokens, C is their channel dimension, and $k \in \{1, \ldots, L\}$ indexes the L encoder layers. We omit the batch dimension and exclude auxiliary tokens from these feature tensors. The original RAE latent is

$$\mathbf{z}_{\text{base}} = \mathcal{E}(P(\mathbf{x})) \in \mathbb{R}^{N \times C}. \tag{1}$$

Here, $\mathcal{E}(P(\mathbf{x}))$ denotes the final-layer readout that the original RAE supplies to its decoder. The baseline decoder produces the reconstruction $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_{\text{base}})$. DRoRAE supplies intermediate-layer information through fusion while keeping $\mathbf{z}_{\text{base}}$ as the reference latent.

3.2 Depth-Routed Fusion Module

The fusion module $\mathcal{F}$ combines features from all L encoder layers and uses $\mathbf{z}_{\text{base}}$ as the reference for an incremental correction. Its output $\mathbf{z}_{\text{final}} \in \mathbb{R}^{N\times C}$ retains the baseline latent shape, and the decoder reconstructs the image as $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_{\text{final}})$.

Layer-wise experts. Each encoder layer k has a separate expert e_k that transforms each token independently:

$$\mathbf{h}_k = e_k(\mathbf{z}^{(k)}) \in \mathbb{R}^{N\times C}, \quad k = 1, \ldots, L. \tag{2}$$

Each expert transforms 768-channel tokens through a hidden width of 3072 and returns the same channel dimension. Appendix A gives the expert dimensions and training settings.

Energy-constrained routing. A linear router $R : \mathbb{R}^{LC} \to \mathbb{R}^{L}$ assigns a raw score to each layer at each token position. The router takes the encoder features $\mathbf{z}^{(k)}$ and concatenates them along the channel dimension. The concatenated tensor has shape $N \times LC$, and R applies the same learned affine map to each row, producing $\mathbf{w} \in \mathbb{R}^{N\times L}$. Let $i \in \{1, \ldots, N\}$ index tokens, w_{ik} denote the raw score for token i and layer k, and $\mathbf{h}_{k,i} \in \mathbb{R}^{C}$ denote token i of $\mathbf{h}_k$. We write the core routing and aggregation operations as

$$\mathbf{w} = R\Big(\Big[\mathbf{z}^{(1)}; \ldots; \mathbf{z}^{(L)}\Big]\Big), \tag{3}$$

$$\mathbf{z}_{\text{fuse},i} = \frac{\sum_{k=1}^{L} w_{ik}\mathbf{h}_{k,i}}{\sqrt{\sum_{k=1}^{L} w_{ik}^2 + \epsilon}}, \tag{4}$$

where $\epsilon > 0$ prevents division by zero. Stacking the fused tokens gives $\mathbf{z}_{\text{fuse}} \in \mathbb{R}^{N\times C}$. The effective coefficient is $a_{ik} = w_{ik}/\sqrt{\sum_j w_{ij}^2 + \epsilon}$. Here, energy refers to the squared ℓ_2 norm of the coefficient vector, with $\sum_k a_{ik}^2 \le 1$ for each token. Negative coefficients subtract expert outputs.

Incremental correction. The core incremental update interpolates between the base and fused representations:

$$\mathbf{z}_{\text{final}} = \mathbf{z}_{\text{base}} + \beta \cdot (\mathbf{z}_{\text{fuse}} - \mathbf{z}_{\text{base}}) \tag{5}$$

Here, $\beta \in [0, 1]$ controls the interpolation and applies across tokens and channels; the default is $\beta = 0.2$. The update combines $0.8\mathbf{z}_{\text{base}}$ with $0.2\mathbf{z}_{\text{fuse}}$. The interpolation preserves token count and channel dimension. Appendix A.1 describes preprocessing and Appendix B.1 describes the diffusion interface.

3.3 Training Strategy

Figure 3 illustrates our training procedure, which separates fusion learning from decoder adaptation. Starting from the RAE-provided decoder checkpoint, we train fusion with the decoder frozen, then adapt the decoder to the learned latent. The encoder backbone remains frozen throughout all phases. Appendix A provides the hyperparameters.

Phase 1: RAE decoder initialization. We use the pretrained decoder checkpoint provided by RAE Zheng et al. (2026) as the Phase 1 model. Standard RAE trains this decoder $\mathcal{D}$ with the backbone $\mathcal{E}$ frozen and no fusion module, reconstructing images from $\mathbf{z}_{\text{base}}$ with the objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_p \mathcal{L}_{\text{LPIPS}} + \lambda_g \alpha_{\text{adapt}} \mathcal{L}_{\text{GAN}} \tag{6}$$

Here, $\mathcal{L}_{\text{total}}$ is the tokenizer training loss. The terms $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{LPIPS}}$ measure ℓ_1 reconstruction error and perceptual distance Zhang et al. (2018) between $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively. The term $\mathcal{L}_{\text{GAN}}$ is the tokenizer's adversarial loss from a DINO-based discriminator Zheng et al. (2026). The nonnegative scalar coefficients λ_p and λ_g weight the perceptual and adversarial terms; α_{adapt} adjusts the adversarial weight using gradient norms Esser et al. (2021). We reuse this objective for adaptation after loading the pretrained decoder.

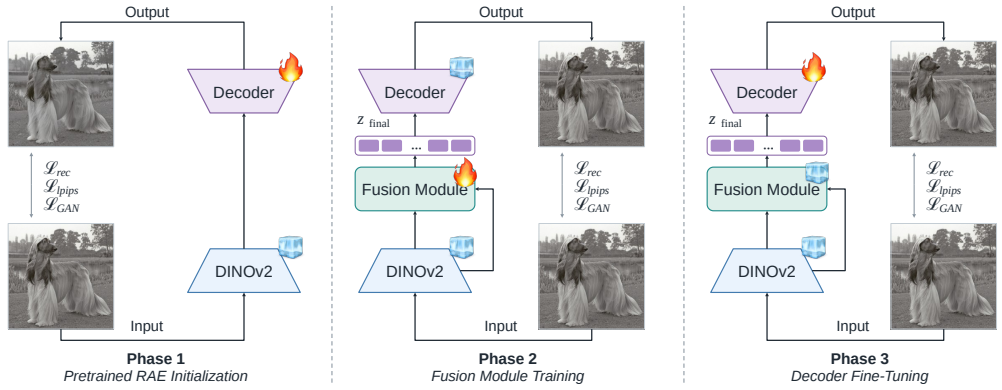


Figure 3: **Initialization and staged adaptation.** (1) Load the pretrained RAE decoder; (2) train fusion through the frozen decoder; (3) adapt the decoder with fusion frozen. The encoder stays frozen.

Phase 2: Fusion module training. With the backbone and Phase 1 decoder frozen, fusion contains the trainable autoencoder parameters. The discriminator updates separately under the adversarial objective. We fix the correction strength β at 0.2 and use the same reconstruction objective (Eq. 6). The layer-wise experts and router jointly learn to supply the fixed decoder with a latent from which it can reconstruct the input image.

Phase 3: Decoder fine-tuning. We freeze the fusion module after Phase 2 and train the decoder with Eq. 6. The decoder adapts to the fixed latent $\mathbf{z}_{\text{final}}$. Section 4.4 compares fusion-only training, the complete staged configuration, and joint continuation of fusion and decoder.

The two adaptation stages act on opposite sides of the latent interface. Phase 2 adjusts fusion for a fixed decoder, while Phase 3 adjusts the decoder for fixed fusion. For a given input, the encoder and fusion module produce the same latent throughout Phase 3. Decoder fine-tuning then optimizes reconstruction from this fixed representation.

4 Experiments

4.1 Experimental Setup

Datasets. We train and evaluate across three settings. **(1) Image reconstruction:** The tokenizer is trained on the ImageNet-1K Deng et al. (2009) training set (1.28M images, 1000 classes) at 256×256 resolution. Evaluation is performed on the 50K validation set. **(2) Class-conditional generation:** A DiT Peebles & Xie (2023) diffusion model is trained on the same ImageNet-1K training set, operating in each tokenizer's latent space. We follow the ADM Dhariwal & Nichol (2021) evaluation protocol and generate 50K images for FID computation. **(3) Text-to-image generation:** A unified multimodal model is trained on CC12M-LLaVA-Next Changpinyo et al. (2021).

Evaluation metrics. For reconstruction, we report rFID, LPIPS Zhang et al. (2018), PSNR, and SSIM, which together capture distributional fidelity, learned perceptual similarity, pixel-level distortion, and structural preservation respectively. For class-conditional generation, we report generation FID (gFID), Inception Score (IS), Precision, and Recall Dhariwal & Nichol (2021), which together reflect overall distributional similarity, sample quality and diversity, and the trade-off between fidelity and coverage. For text-to-image generation, we report GenEval Ghosh et al. (2023), which evaluates compositional generation ability across six dimensions.

Implementation details. Our encoder backbone is DINOv2-B Oquab et al. (2024), and the decoder is ViT-XL. For class-conditional generation, we use DiT$^{\text{DH}}$-XL Zheng et al. (2026). For text-to-image, we use the Bagel Deng et al. (2025) Mixture-of-Transformers (MoT) framework with a Qwen2.5-0.5B Yang et al. (2024) backbone. Full training hyperparameters are in Appendix A.

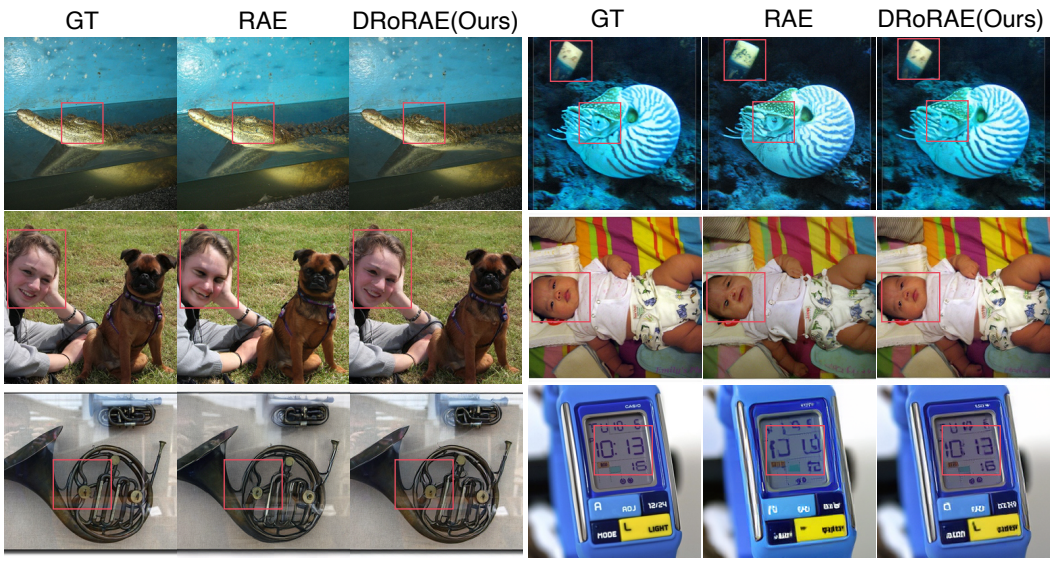


Figure 4: **ImageNet-256 reconstruction examples.** Selected spatial comparisons complement the aggregate metrics in Table 2.

Table 1: **Guided ImageNet-256 generation.** Epochs count generator training. DRoRAE uses AutoGuidance with scale 1.5. DRoRAE denotes the final Phase 3 model.

Tokenizer	Generator	Params	Epochs	gFID ↓	IS ↑	Prec.	Rec.
VQGAN (Esser et al., 2021)	MaskGiT	227M	555	–	–	–	–
SD-VAE (Rombach et al., 2022)	DiT-XL	675M	1400	2.270	278.200	0.830	0.570
REPA (Yu et al., 2025)	SiT-XL	675M	800	1.420	305.700	0.800	0.650
VA-VAE (Yao et al., 2025)	SiT-XL	675M	80	3.630	290.600	–	–
FAE-d32 (Gao et al., 2025)	LightningDiT	675M	80	1.700	243.800	0.820	0.610
SVG (Shi et al., 2026)	SVG-XL	675M	80	3.540	207.600	–	–
RAE (Zheng et al., 2026)	DiT^{DH}-XL	839M	80	1.740	235.000	0.810	0.600
RPiAE (Gong et al., 2026)	LightningDiT	675M	80	1.510	225.900	0.790	0.650
DRoRAE	DiT^{DH}-XL	839M	80	1.650	230.600	0.810	0.600

Comparison conditions. Our main RAE comparison uses the same encoder backbone and decoder architecture for reconstruction, and the same DiT^{DH}-XL generator architecture and 80 training epochs for class-conditional generation. We report the final DRoRAE tokenizer after decoder adaptation; the frozen-decoder experiments separately evaluate the learned fusion. DRoRAE uses AutoGuidance at scale 1.5 with DiT^{DH}-S as the guidance model. Table 1 also lists published results with their generator architectures and training durations; these entries use the respective methods' training and guidance protocols.

4.2 Reconstruction and Class-Conditional Generation

We compare DRoRAE with existing methods on ImageNet-256. Table 2 reports image reconstruction quality. Table 1 reports class-conditional generation quality with guidance, along with the generator architecture and number of training epochs used by each method.

Reconstruction. Table 2 shows that the complete DRoRAE tokenizer improves reconstruction across all four metrics relative to RAE. With the same DINOv2-B backbone and ViT-XL decoder architecture, it nearly halves rFID (0.570→0.290) and increases PSNR from 18.800 to 24.320 dB. LPIPS decreases from 0.256 to 0.134, and SSIM increases from 0.483 to 0.701. These results evaluate the combined effect of multi-layer fusion and decoder adaptation.

Table 2: **ImageNet-256 reconstruction.** We compare reconstruction quality using rFID, PSNR, LPIPS, and SSIM. DRoRAE denotes the final Phase 3 model.

Tokenizer	rFID ↓	PSNR ↑	LPIPS ↓	SSIM ↑
VQGAN (Esser et al., 2021)	2.230	17.900	0.202	0.422
SD-VAE (Rombach et al., 2022)	0.610	26.900	0.130	0.736
REPA (Yu et al., 2025)	0.610	26.900	0.130	0.736
VA-VAE (Yao et al., 2025)	0.270	27.700	0.097	0.779
FAE-d32 (Gao et al., 2025)	0.680	–	–	–
SVG (Shi et al., 2026)	0.650	23.890	0.190	0.650
RAE (Zheng et al., 2026)	0.570	18.800	0.256	0.483
RPiAE (Gong et al., 2026)	0.500	21.300	0.216	0.525
DRoRAE	**0.290**	24.320	0.134	0.701

Fusion improves reconstruction before decoder adaptation. With the pretrained decoder frozen, DRoRAE achieves an rFID of 0.470, compared with 0.570 for the last-layer RAE baseline (Section 4.4). These results show that learned multi-layer fusion improves reconstruction through the frozen decoder. Section 4.5 examines expert width and layer count; increasing layer count adds intermediate representations together with their experts.

Figure 4 illustrates the complete method's improvements in local image content. DRoRAE more faithfully reconstructs the displayed digits, facial details, and the curved tubing of the horn. These examples connect the aggregate gains to the recovery of fine structures and local content.

Generation. In Table 1, DRoRAE improves guided generation alongside reconstruction: its gFID reaches 1.650, compared with 1.740 for RAE, using the same generator architecture and 80 training epochs. AutoGuidance reduces DRoRAE's gFID from 2.680 to 1.650; the corresponding RAE results are 2.160 and 1.740. Appendix B.3 reports the unguided evaluations. This larger response to guidance motivates examining which statistical component contributes most to the improvement.

The main-model ADM decomposition in Appendix C.2 attributes most of the AutoGuidance improvement to the covariance term of FID in Inception feature space. The mean term decreases from 0.399 to 0.181, while the covariance term decreases from 2.279 to 1.470. The covariance term accounts for approximately 78.8% of the FID reduction in this decomposition. Thus, AutoGuidance brings the generated-image distribution closer to the real-image reference primarily through better covariance agreement in the evaluation feature space.

4.3 Text-to-Image Generation

We evaluate DRoRAE on text-to-image generation as a more complex downstream task that requires images to follow compositional text prompts. Following RPiAE Gong et al. (2026), we use the Bagel Deng et al. (2025) MoT architecture with a Qwen2.5-0.5B backbone, training on CC12M-LLaVA-Next with identical configurations except for tokenizer-specific adaptations. Appendix A.3 details these adaptations and the shared training settings. Table 3 reports an overall GenEval score of 0.600 for DRoRAE, compared with 0.560 for RAE. Counting increases from 0.460 to 0.560 and color accuracy from 0.700 to 0.790, extending the downstream gains to generation from compositional text prompts.

The RAE and DRoRAE comparison keeps the latent interface fixed: both tokenizers produce 16×16 tokens with 768 channels per token. Their MoT vision experts consequently use the same interface dimensions under the shared training configuration. The GenEval improvement thus accompanies a change in the latent representation at the same token count and channel width.

4.4 Ablation Studies

Ablation of energy-constrained routing. Replacing softmax routing with energy-constrained routing lowers rFID from 0.475 to 0.447 without incremental correction, and from 0.512 to 0.470 with incremental correction (Table 5). The consistent improvement across both settings supports the routing design, which combines signed coefficients with per-token ℓ_2 normalization.

Table 3: **GenEval** (Ghosh et al., 2023). SO/TO: single/two objects; CT: counting; CL: colors; POS: position; ATTR: color attribution.

Tokenizer	SO↑	TO↑	CT↑	CL↑	POS↑	ATTR↑	Overall↑
FLUX-VAE (Black Forest Labs, 2024)	0.830	0.300	0.250	0.650	0.080	0.180	0.380
VA-VAE (Yao et al., 2025)	0.960	0.720	0.460	0.790	0.250	0.490	0.610
RAE (Zheng et al., 2026)	0.960	0.690	0.460	0.700	0.230	0.330	0.560
RPiAE-VB (Gong et al., 2026)	0.970	0.720	0.560	0.790	0.260	0.380	0.610
DRoRAE (ours)	0.980	0.690	0.560	0.790	0.220	0.350	0.600

Table 4: **Training strategy ablation.** Phase 1 denotes initialization from pretrained RAE (Zheng et al., 2026); Phase 2 trains fusion with the encoder and decoder frozen; Phase 3 fine-tunes the decoder with fusion frozen. Generation uses DiT^{DH}-XL and 80 training epochs; DRoRAE uses AutoGuidance at scale 1.5.

	Reconstruction				Guided generation			
Configuration	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓	IS↑	Prec.	Rec.
Phase 1 only	0.570	18.800	0.256	0.483	1.740	235.000	0.810	0.600
Phase 1 + Phase 2	0.470	21.790	0.195	0.583	1.700	222.600	0.810	0.610
Phase 1 + Phase 2 + Phase 3	0.290	24.320	0.134	0.701	1.650	230.600	0.810	0.600

Ablation of incremental correction. We introduce incremental correction to retain a contribution from the last-layer representation while incorporating multi-layer information. With energy-constrained routing, adding correction reduces the epoch-12 diffusion training loss from 0.810 to 0.470. The same comparison under softmax routing reduces the loss from 0.790 to 0.480. All four variants use the same diffusion training settings and standardize their latents using separately estimated means and variances. Appendix B.1 describes this standardization. These reductions measure improved early optimization of the generative model. With energy-constrained routing, reconstruction rFID changes from 0.447 to 0.470, showing the trade-off between reconstruction fidelity and diffusion training.

Table 5: **Fusion design ablation.** Energy/Inc.: energy routing/incremental correction. Loss: epoch-12 DiT. Cross-Attn: last-layer query.

Method	Energy	Inc.	rFID↓	Loss↓
RAE	–	–	0.570	0.430
Cross-Attn	✗	✗	0.498	–
Softmax	✗	✗	0.475	0.790
Energy	✓	✗	0.447	0.810
Softmax	✗	✓	0.512	0.480
Ours	✓	✓	0.470	0.470

Ablation of correction strength β. The correction-strength sweep in Figure 5 also demonstrates a benefit in generated-image quality. Setting $\beta = 1$ removes the direct contribution of the last-layer baseline in Eq. 5, yielding $\mathbf{z}_{\text{final}} = \mathbf{z}_{\text{fuse}}$. Reducing β to 0.3 retains this baseline contribution and lowers gFID from 10.618 to 5.257. Both results use generators trained for 20 epochs and evaluate 50K samples with CFG at scale 1.5. Reconstruction rFID changes from 0.448 to 0.494. This comparison shows that retaining the last-layer baseline can improve generation despite a reconstruction trade-off. Together with the diffusion-loss results, it supports incremental correction as a design for downstream generative modeling. Our main experiments use $\beta = 0.2$.

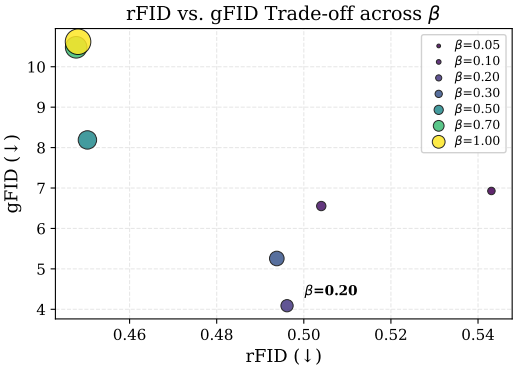


Figure 5: **Effect of β.** Point size scales with β.

Ablation of training strategy. Fusion-only and staged adaptation use a common pretrained RAE initialization. Fusion-only training (Phase 1 + Phase 2) improves rFID from the RAE baseline's 0.570 to 0.470 while keeping the encoder and decoder frozen (Table 4). Compared with fusion-only training, the complete staged configuration (Phase 1 + Phase 2 + Phase 3) achieves lower rFID (0.290

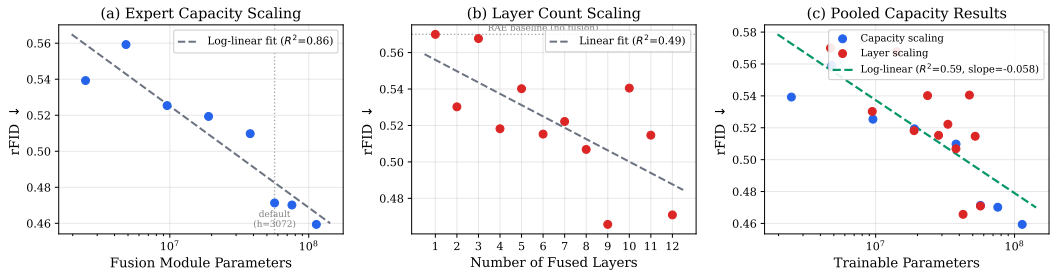


Figure 6: **Reconstruction scaling.** (a) Expert capacity scaling, (b) layer count scaling, and (c) pooled capacity results; respective fits have $R^2 = 0.86$, 0.49, and 0.59.

vs. 0.470) and guided gFID (1.650 vs. 1.700). PSNR, LPIPS, and SSIM also improve. These results support staged fusion learning and decoder adaptation for reconstruction and guided generation.

We also compare staged adaptation with jointly updating fusion and decoder while keeping the encoder backbone frozen. Appendix B.2 details the adaptation schedules and their results. At the same 20-epoch generator budget without guidance, staged adaptation achieves gFID 4.385, compared with 18.071 for joint updates. Joint updates yield a lower reconstruction rFID of 0.163, compared with 0.290 for staged adaptation. These results support the staged configuration for its balance of reconstruction and generation quality under the evaluated training schedules.

We next compare frozen and fine-tuned encoders in Table 6 to examine this reconstruction–generation balance. Unfreezing the encoder reduces rFID from 0.290 to 0.130 but increases unguided gFID from 2.680 to 18.360. This comparison supports keeping the encoder frozen for downstream generation. Our training strategy therefore learns fusion and adapts the decoder while retaining the frozen encoder, balancing reconstruction fidelity with generative modeling.

Table 6: **Effect of encoder fine-tuning.** Both configurations use learned fusion and decoder adaptation. +AG: gFID with AutoGuidance.

Encoder	rFID↓	gFID↓	+AG↓
Frozen (DRoRAE)	0.290	2.680	1.650
Fine-tuned	0.130	18.360	–

4.5 Reconstruction Scaling with Fusion Capacity

Figure 6a shows an approximately log-linear reduction in reconstruction FID as fusion expert capacity increases over the measured range. We fix the encoder, decoder, Phase 2 training configuration, and all 12 input layers, then vary expert hidden width from 128 to 6144. Fusion capacity grows from approximately 3M to 113M parameters, with rFID decreasing from 0.540 to 0.460 and a log-linear fit of $R^2 = 0.86$. This trend shows how additional capacity for transforming the same encoder features can improve reconstruction. We use width 3072 as a capacity–reconstruction compromise: doubling it to 6144 reduces rFID from 0.471 to 0.459.

The layer-count sweep in Figure 6b complements this finding by expanding access to encoder depth at a fixed expert width of 3072. Increasing the count from 1 to 12 reduces rFID from 0.570 to 0.470, with greater variation across intermediate configurations. Each added layer supplies both an intermediate representation and its expert. The pooled view in Figure 6c compares both sweeps against total parameter count; the expert-width sweep exhibits the clearest log-linear trend.

5 Conclusion

DRoRAE extends representation autoencoders with learned fusion across the frozen encoder hierarchy. Signed routing, incremental correction, and staged decoder adaptation reduce ImageNet-256 rFID from 0.570 to 0.290 and guided gFID from 1.740 to 1.650 with AutoGuidance. The downstream gains extend to text-to-image generation, where GenEval increases from 0.560 to 0.600 over RAE. With all other settings fixed, increasing fusion expert capacity yields an approximately log-linear reduction in reconstruction FID over the measured range. We focus on DINOv2-B; extending DRoRAE to other encoder architectures and scales remains future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language editing, as well as figure preparation based on recorded experimental results and model outputs. These tools supported the preparation of the research materials, while the authors retained responsibility for all research decisions. The authors are responsible for the design of the method and experiments, the verification and interpretation of the results, and the final content of the manuscript.

REPRODUCIBILITY STATEMENT

Appendix A specifies the fusion configuration and training and sampling settings. Appendix B reports diffusion normalization, training strategy comparisons, and additional generation results. Appendix C describes routing visualizations and the main-model AutoGuidance FID decomposition. Appendix D provides selected qualitative examples.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

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

Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, Wenhai Wang, Danhuai Zhao, and Tong Lu. MMFuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding. *arXiv preprint arXiv:2410.11829*, 2024. doi: 10.48550/arXiv.2410.11829.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6677–6695, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.339. URL https://aclanthology.org/2025.emnlp-main.339/.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. doi: 10.48550/arXiv.2505.14683.

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. URL https://api.semanticscholar.org/CorpusID:57246310.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021. doi: 10.1109/CVPR46437.2021.01268.

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *arXiv preprint arXiv:2512.07829v2*, 2025. URL https://arxiv.org/abs/2512.07829v2.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023.

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. *arXiv preprint arXiv:2603.19206*, 2026. doi: 10.48550/arXiv.2603.19206.

Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 447–456, 2015. doi: 10.1109/CVPR.2015.7298642.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1679. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5ee7ed60a7e8169012224dec5fe0d27f-Abstract-Conference.html.

Xu Li, Yi Zheng, Haotian Chen, Xiaolei Chen, Yuxuan Liang, Chenghang Lai, Bin Li, and Xiangyang Xue. Instruction-guided fusion of multi-layer visual features in large vision-language models. *Pattern Recognition*, 170:111932, 2026. ISSN 0031-3203. doi: 10.1016/j.patcog.2025.111932. URL https://www.sciencedirect.com/science/article/pii/S0031320325005928.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 936–944, 2017.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024. doi: 10.1007/978-3-031-72980-5_2.

Meituan LongCat Team. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://dblp.org/rec/journals/tmlr/OquabDMVSKFHMEA24.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12159–12168, 2021. doi: 10.1109/ICCV48922.2021.01196.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

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 *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa5ddd6bac0d665c72969d79221b680a-Abstract-Conference.html.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026. doi: 10.48550/arXiv.2605.18324. URL https://arxiv.org/abs/2605.18324.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, Yuxin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for MLLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33108–33140, 2024.

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*, pp. 15703–15712, 2025. doi: 10.1109/CVPR52734.2025.01464.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

APPENDIX OVERVIEW

Appendix A collects implementation and training settings. Appendix B reports diffusion normalization, training strategy comparisons, and additional generation results. Appendix C examines learned routing and the response to AutoGuidance. Appendix D presents selected frequency-domain reconstruction examples.

A IMPLEMENTATION AND TRAINING DETAILS

A.1 TOKENIZER ARCHITECTURE AND TRAINING

DRoRAE trains fusion and decoder adaptation in separate phases while keeping the encoder frozen. Table 7 lists the component architectures, and Table 8 lists the optimization settings.

Table 7: **Tokenizer component architectures.** The fusion input contains all 12 encoder layers; its output retains the RAE latent shape.

Field	Encoder	Decoder	Fusion
Input dim.	$224 \times 224 \times 3$	$16 \times 16 \times 768$	$16 \times 16 \times 768 \times 12$
Output dim.	$16 \times 16 \times 768$	$256 \times 256 \times 3$	$16 \times 16 \times 768$
Hidden dim.	768	1152	3072
Num. layers	12	28	12 (experts)
FFN hidden dim.	3072	4096	3072
Dim. per head	64	72	–
Num. heads	12	16	–
Total params	86M	415.35M	56.85M

Table 8: **Tokenizer adaptation schedules.** Epochs count tokenizer training. Joint continuation uses 6 fusion-only epochs followed by 10 joint epochs; its reconstruction evaluation uses joint-training epoch 6.

Setting	Fusion-only	Staged	Joint continuation
Fusion epochs	16	6	6 + 10 (joint)
Decoder epochs	0	10	10 (joint)
Batch size	384	192	192
Fusion LR	2×10^{-4}	2×10^{-4}	2×10^{-4}
Decoder LR	–	5×10^{-5}	5×10^{-5}
LR schedule	Cosine	Cosine	Cosine
Warmup epochs	1	1	0
EMA decay	0.9978	0.9978	0.9978

Image preprocessing and token extraction. During tokenizer training, we resize each image's shorter side to 384 pixels with bicubic interpolation and take a random 256×256 crop. Reconstruction evaluation uses a center crop at 256×256. The encoder pipeline resizes this target to 224×224 with bicubic interpolation and `align_corners=False`, then applies the encoder's image mean and standard deviation. DINOv2-B-reg uses 14×14 patches; we retain 256 patch tokens and exclude class and register tokens. The decoder reconstructs the 256×256 target. Training uses tokenizer noise scale $\tau = 0.8$; inference disables this noise.

Phase 1: RAE decoder initialization. We load the ViT-XL decoder checkpoint provided by RAE (Zheng et al., 2026) and retain the frozen DINOv2-B-reg encoder. The decoder reconstructs 256×256 images from $\mathbf{z}_{\text{base}} \in \mathbb{R}^{16 \times 16 \times 768}$.

Phase 2: Fusion module training. With both encoder and decoder frozen, we train the fusion module. Each of the 12 experts e_k in Eq. 2 maps 768-channel tokens through a hidden width of 3072 and back to 768 channels. The experts use GELU between their two affine projections, with dropout probability zero. Both expert projections and the linear router include biases. We initialize their weights from $\mathcal{N}(0, 0.02^2)$ and biases to zero. The router maps the concatenated layer readouts

through Linear(12×768, 12) independently at each token, with routing stabilizer $\epsilon = 10^{-6}$. We set $\beta = 0.2$ and use the Phase 1 loss.

Phase 3: Decoder fine-tuning. In the staged run, we freeze fusion after six training epochs and fine-tune the decoder for ten epochs. Table 8 specifies the adaptation schedules.

Adversarial optimization. The discriminator uses frozen DINO-S/8 features and trainable classification heads. We update these heads separately from the autoencoder in both adaptation stages, using AdamW with learning rate 2×10^{-4}, betas $(0.9, 0.95)$, zero weight decay, and cosine scheduling. We use a hinge discriminator loss, a vanilla generator loss, perceptual weight 1.0, and adversarial weight 0.75. Fusion-only and staged training activate adversarial training at epoch indices 6 and 3, respectively, with one warmup epoch. Joint continuation uses zero warmup, starts discriminator updates at index 0, and activates the generator adversarial loss at index 1. The discriminator retains its state when staged training switches to decoder adaptation.

A.2 Class-Conditional Generator

We train DiT$^{\text{DH}}$-XL (Zheng et al., 2026) on the DRoRAE latent space for class-conditional generation. Table 9 lists its dual-head architecture, optimization, and sampling settings. AutoGuidance uses a DiT$^{\text{DH}}$-S model at scale 1.5.

Table 9: **Class-conditional generator configuration.** Paired architecture entries give the main branch and prediction head, respectively.

Architecture	Value	Training / sampling	Value
Input dim.	$16 \times 16 \times 768$	Training data	ImageNet-1K (1.28M)
Output dim.	$16 \times 16 \times 768$	Training epochs	80
Hidden dim.	[1152, 2048]	Global batch per update	1024
Num. layers	[28, 2]	Optimizer	AdamW
MLP ratio	4	Peak LR	2×10^{-4}
Dim. per head	[72, 128]	(β_1, β_2)	(0.900, 0.950)
Num. heads	[16, 16]	Weight decay	0
Total params	839M	LR schedule	Linear
		Training objective	v-prediction
		Gradient accumulation	2
		Warmup epochs	40
		Time distribution	Logit-normal(0,1)
		Training seed	42
		Sampling steps	50
		Unguided scale	1.000
		AutoGuidance	DiT$^{\text{DH}}$-S, scale=1.500

The global batch of 1024 includes two gradient-accumulation steps. With W workers, each worker processes $1024/(2W)$ images per microbatch. We sample with the 50-step Euler solver. Appendix B.1 specifies representation-specific standardization. Before decoding generated latents, the pipeline reverses that standardization and reshapes the tensor into 256 tokens with 768 channels each.

Image evaluation protocols. Generation FID uses 50,000 samples, Inception-2048 features, and the ImageNet-256 training-reference moments stored in `VIRTUAL_imagenet256_labeled.npz`. We evaluate reconstructions of the 50,000 ImageNet validation images. PSNR averages per-image values computed from RGB mean squared error in $[0, 1]$. LPIPS uses the VGG16 implementation with inputs in $[-1, 1]$ and pretrained linear weights. SSIM uses the TorchMetrics implementation with data range 1, computes a value per image, and then averages. Appendix C.2 decomposes the main DRoRAE FID under this ADM generation protocol.

A.3 TEXT-TO-IMAGE GENERATOR

For text-to-image evaluation (Section 4.3), we use the Bagel (Deng et al., 2025) Mixture-of-Transformers (MoT) architecture. Table 10 summarizes the training configuration.

Table 10: Text-to-image training configuration.

Hyperparameter	**Value**
LLM backbone	Qwen2.5-0.5B
Architecture	Mixture-of-Transformers (MoT)
Training data	CC12M-LLaVA-Next
Image resolution	256×256
Batch size	128
Optimizer	AdamW (β_1=0.900, β_2=0.950)
Learning rate	5×10^{-5}
LR schedule	Cosine decay
Training steps	100k
Noise schedule	Flow matching (logit-normal)
Denoising head	DDT
Sampler	Euler ODE
Sampling steps	50
Evaluation	GenEval (553 prompts)

All tokenizers share the training data, optimization schedule, and sampling settings in Table 10. Latent token counts and vision-expert channel dimensions follow each tokenizer (Table 11). RAE and DRoRAE share the same latent shape and vision-expert interface dimensions.

Table 11: Tokenizer-specific latent configurations for T2I experiments.

Tokenizer	**Latent Shape**	**Downsample**	**Latent Dim**
FLUX-VAE	$32 \times 32 \times 16$	$8\times$	16
VA-VAE	$32 \times 32 \times 32$	$8\times$	32
RAE / DRoRAE	$16 \times 16 \times 768$	$16\times$	768
RPiAE-VB	$16 \times 16 \times 64$	$16\times$	64

B Additional Quantitative Results

B.1 Latent Standardization

For each routing/correction variant in Table 5, let $\mathbf{u}$ denote the latent before dataset-level standardization. Each variant uses its own estimated mean $\mathbf{m}$ and variance $\mathbf{v}$ to form the diffusion input

$$\mathbf{u}_{\text{std}} = \frac{\mathbf{u} - \mathbf{m}}{\sqrt{\mathbf{v} + \epsilon_{\text{std}}}}, \tag{7}$$

where $\epsilon_{\text{std}} > 0$ stabilizes the division. Both statistics have shape $768 \times 16 \times 16$, with one value per channel and spatial position.

The four variants share the generator training settings in Appendix A.2.

B.2 Staged Adaptation and Joint Continuation

Fusion-only and staged adaptation share the pretrained RAE initialization and follow the schedules in Table 8. Table 12 compares their reconstruction and generation results with joint continuation.

We first train fusion for six epochs, then jointly train fusion and the decoder for ten epochs while keeping the encoder frozen. The reported reconstruction result uses joint-training epoch 6; Table 8 gives its optimization settings.

Table 12: **Reconstruction and generation under different adaptation schedules.** Generator evaluations use 20 training epochs, 50,000 samples, 50 sampling steps, and no guidance (scale 1).

Configuration	Tokenizer training	rFID↓	gFID↓
Fusion-only	16 fusion epochs	0.470	4.089
Staged	6 fusion + 10 decoder epochs	0.290	4.385
Joint continuation	6 fusion + 10 joint epochs	0.163	18.071

At the same 20-epoch generator budget, staged adaptation combines rFID 0.290 with gFID 4.385, while joint continuation reaches rFID 0.163 and gFID 18.071. A separate evaluation of the joint-continuation tokenizer with a 40-epoch generator yields gFID 7.635 using AutoGuidance at scale 1.5.

B.3 Class-Conditional Generation without Guidance

Table 13 complements the guided results in Table 1 by retaining class conditioning and disabling guidance. Appendix C.2 examines the response to guidance using the main DRoRAE checkpoint and the main generation evaluation protocol.

Table 13: **Unguided ImageNet-256 generation.** Class conditioning retained; DRoRAE rows specify the tokenizer training configuration.

Tokenizer	Epochs	gFID ↓	IS ↑	Prec.	Rec.
VQGAN	555	6.180	182.100	0.800	0.510
SD-VAE	1400	9.620	121.500	0.670	0.670
REPA	80	7.900	–	–	–
VA-VAE	80	5.960	128.000	–	–
FAE-d32	80	2.080	207.600	0.820	0.590
SVG	80	6.570	137.900	–	–
RAE	80	2.160	214.800	0.820	0.590
RPiAE	80	2.250	208.700	0.810	0.600
DRoRAE (fusion-only)	80	2.450	197.800	0.800	0.600
DRoRAE (staged)	80	2.678	197.800	0.800	0.590

B.4 Generation with an Extended Training Budget

Extending the staged-tokenizer generator to 800 epochs yields unguided gFID 2.071 and gFID 1.127 with AutoGuidance at scale 1.5. Evaluation uses 50,000 samples, 50 Euler steps, and the small guidance model associated with this run. These longer-training results supplement the 80-epoch comparison in the main text.

C Routing and AutoGuidance Analysis

C.1 Signed Routing across Depth and Space

The learned router assigns signed weights that vary across encoder depth and spatial position. Figure 7 illustrates this behavior for four selected inputs.

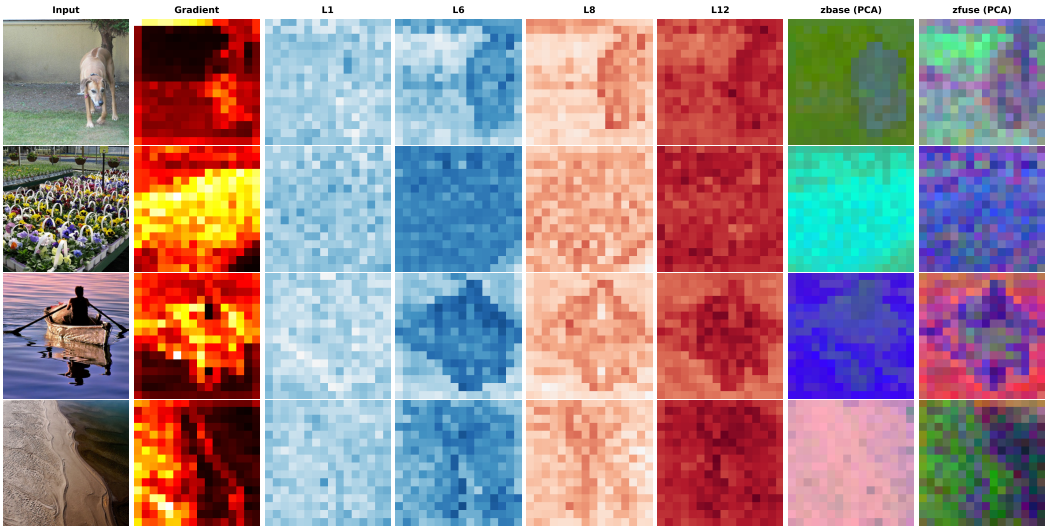


Figure 7: **Routing and latent visualizations for four selected inputs.** Columns show the input, image gradient, raw router scores for L1, L6, L8, and L12, and principal component analysis (PCA) views of $\mathbf{z}_{\text{base}}$ and $\mathbf{z}_{\text{fuse}}$. Red and blue denote positive and negative scores on the 16×16 token grid. The fused branch precedes incremental correction. We fit PCA separately to the two representations.

Spatially varying, signed routing. Across the four examples, L1 is predominantly weakly negative, L6 is more strongly negative, L8 is positive, and L12 is strongly positive. The L6 and L8 maps show related object- or region-dependent patterns with opposite signs, particularly around the dog and rower.

Interpreting the weights. The plotted values are raw router scores. Equation 4 converts them to effective coefficients through per-token ℓ_2 normalization. Each coefficient acts on a separately learned expert output $\mathbf{h}_k$, so a negative coefficient denotes a subtractive contribution to the fused latent.

C.2 AutoGuidance Primarily Reduces the FID Covariance Term

AutoGuidance (AG) (Karras et al., 2024) reduces the main DRoRAE model's gFID from 2.678 to 1.651. We decompose this change into mean and covariance terms in Inception feature space to identify the dominant component of the improvement.

Evaluation setup. We evaluate the staged DRoRAE tokenizer with its 80-epoch DiT^{DH}-XL generator and AG scale 1.5. Both conditions use 50,000 generated images and the ImageNet-256 training-reference statistics from the main ADM evaluation. The decomposition uses the same checkpoint and protocol as Table 1 and Appendix B.3.

FID decomposition. Let (μ_r, Σ_r) and (μ_g, Σ_g) denote the empirical mean and covariance of real and generated Inception features. FID separates into

$$\begin{aligned} \text{FID} &= D_\mu + D_\Sigma, \\ D_\mu &= \|\mu_g - \mu_r\|_2^2, \\ D_\Sigma &= \operatorname{tr}\left[\Sigma_g + \Sigma_r - 2\left(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\right)^{1/2}\right]. \end{aligned} \tag{8}$$

Table 14: **Main-model FID decomposition under ADM.** All entries use the staged DRoRAE model at generator epoch 80. Values show three decimal places; decreases and percentages use unrounded measurements.

Sampling	FID ↓	Mean term	Covariance term
Unguided	2.678	0.399	2.279
AG, scale 1.5	1.651	0.181	1.470
Decrease	1.027	0.218	0.809

AG reduces both terms (Table 14). The covariance term contributes 0.809 of the 1.027 FID decrease, or approximately 78.8%; the mean term contributes 0.218. Thus, most of the measured gain comes from closer covariance agreement between generated and real Inception features.

D Qualitative Reconstruction

D.1 Frequency-Domain Reconstruction

Figure 8 complements the spatial comparisons in Figure 4 with frequency-domain views of three selected images.

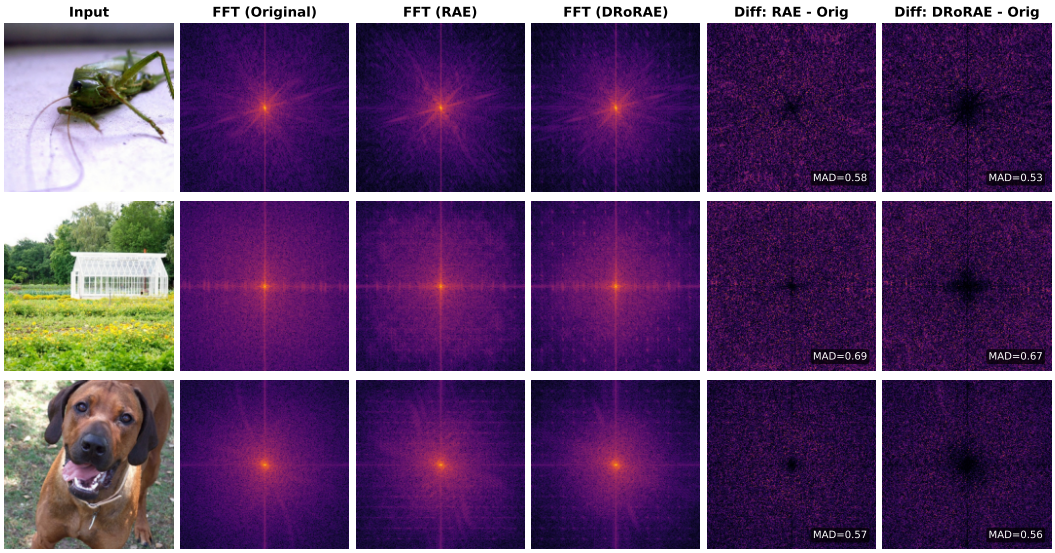


Figure 8: **Selected spectral reconstruction examples.** Rows show an insect, a greenhouse, and a dog. Columns show the input; two-dimensional fast Fourier transform (FFT) log-magnitude spectra of the input and both reconstructions; and each reconstruction's spectral difference map relative to the input. Darker difference-map regions indicate smaller deviations.