RoAR-VAE: Improving Latent Diffusability through Robust Prediction

Anonymous authors
Paper under double-blind review

Abstract

An effective visual tokenizer must preserve image fidelity while producing latents that downstream diffusion models can learn efficiently. Prior work improves latent diffusability through semantic alignment or joint tokenizer–diffusion training. Revisiting REPA-E, a representative approach, we find that removing its diffusion objective and dedicated generative blocks while retaining semantic prediction from corrupted latents yields comparable downstream generation performance under a matched setting. This observation motivates a general principle of *robust prediction*: generation-relevant information should remain recoverable from corrupted latents. We instantiate this principle with two complementary objectives. *RobustAlign* predicts pretrained semantic features from corrupted latents to preserve high-level structure, while *RobustRecon* reconstructs the clean image to retain fine-grained spatial and appearance information. Together, they form **Ro**bust **A**lignment and **R**econstruction VAE (**RoAR-VAE**). We attribute the resulting improvement in diffusability to robust prediction making useful information more accessible in high-variance latent directions, which diffusion models tend to learn earlier. Empirically, RoAR-VAE improves average generation benchmark performance by 2.19 points over the strongest baseline tokenizer under the same training budget, while improving small-text reconstruction accuracy by 16.93 percentage points over FLUX.1-VAE. These results demonstrate that RoAR-VAE achieves a favorable balance between reconstruction fidelity and downstream generative learnability.

1 Introduction

Visual tokenizers define the representation on which latent diffusion models learn to generate images (Rombach et al., 2022; Esser et al., 2024). An effective tokenizer should preserve sufficient image information for faithful reconstruction, while organizing that information into a representation that a downstream diffusion model can learn efficiently. This latter property, often referred to as *diffusability*, can conflict with reconstruction fidelity: higher-capacity or weakly regularized latents may retain more visual detail yet be harder to model. Balancing reconstruction quality and diffusability is therefore a central challenge in visual tokenizer design.

Recent work improves latent diffusability through semantic supervision during VAE training (Yao et al., 2025; 2026) or joint optimization of the tokenizer and diffusion model (Heek et al., 2026; Duggal et al., 2026). A representative approach, REPA-E (Leng et al., 2025a), combines both ideas: as shown in Fig. 2(a), it jointly trains a VAE and a diffusion Transformer while aligning intermediate features with semantic targets from pretrained visual encoders (Yu et al., 2025; Oquab et al., 2024). However, only the semantic alignment objective directly updates the tokenizer encoder, while gradients from the diffusion objective are detached. This raises a natural question: *what actually drives the diffusability improvement in REPA-E, joint diffusion training or semantic supervision?* To isolate these factors, we simplify REPA-E by removing the diffusion objective and its dedicated generative blocks while retaining prediction of pretrained visual features from corrupted latents, yielding *RobustAlign* (Figure 2(b)). Under a matched evaluation, the resulting tokenizer achieves comparable downstream generation performance to the tokenizer trained with REPA-E (Table 1), indicating that noisy semantic prediction alone can provide an effective learning signal for shaping a diffusion-friendly latent space.

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.

RobustAlign suggests a broader principle of *robust prediction*: generation-relevant information should remain predictable from heavily corrupted latents. By backpropagating this prediction objective through the corrupted latent, the tokenizer encoder is encouraged to organize target-relevant information so that it remains accessible under corruption. RobustAlign primarily supervises object-level structure and high-level visual content captured by pretrained features, but may underrepresent fine-grained spatial and appearance variations. We therefore introduce **RobustRecon**, which applies the same principle to the clean image itself, encouraging detailed image content to remain recoverable from corrupted latents. Combining the two objectives yields **Robust Alignment and Reconstruction (RoAR)** (Figure 2(c)). Both auxiliary objectives shape the encoder through corrupted latents, while the standard VAE reconstruction objective preserves faithful reconstruction from clean latents. The latent PCA visualizations in Figure 3 further illustrate their complementary effects: RobustAlign emphasizes object-level organization and foreground–background separation, RobustRecon exposes richer local variation, and their combination exhibits both forms of structure.

We further ask: *why does robust prediction improve diffusability?* Under Gaussian corruption, higher-variance latent directions retain a higher signal-to-noise ratio and are therefore more robust to noise. Training the encoder to recover clean targets from corrupted latents consequently encourages generation-relevant information to remain accessible in these dominant directions, either by strengthening the principal modes or by enriching the image information they encode, as suggested by the PCA visualizations in Fig. 3. Prior analyses further show that higher-variance covariance modes tend to be learned earlier by diffusion models (Wang & Pehlevan, 2025). Together, these observations suggest that robust prediction organizes useful information into prominent latent structure that can be acquired more readily by a downstream diffusion model.

Empirically, we train RoAR-VAE on ImageNet and evaluate reconstruction and downstream text-to-image generation, using a common diffusion backbone and dataset across tokenizers. RoAR-VAE achieves a favorable balance between reconstruction fidelity and generative learnability (Fig. 1): it attains the best performance across three text-to-image benchmarks, with an average score 2.19 points above the strongest diffusability-focused baseline. At the same time, it maintains high-fidelity reconstruction, achieving 32.88 dB PSNR and 0.0349 LPIPS at 256^2 resolution, while improving small-text reconstruction accuracy over FLUX.1-VAE by 16.93 percentage points (Table 3). Ablations further show that RobustAlign and RobustRecon provide complementary gains, and that latent corruption is critical to downstream performance. These results support robust prediction as an effective principle for improving latent diffusability while retaining high reconstruction fidelity.

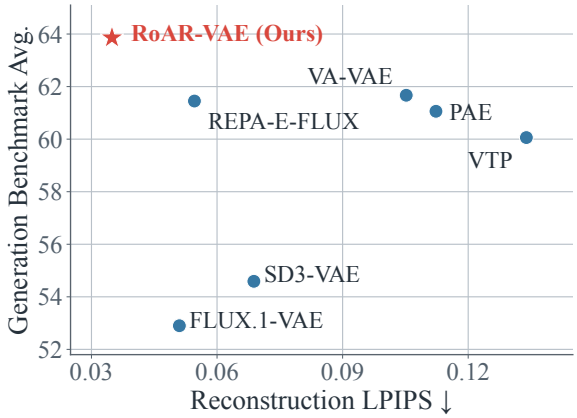

Figure 1: **Reconstruction versus generation.** RoAR-VAE achieves a favorable trade-off, combining low reconstruction LPIPS with strong diffusability.

2 Related Work

Improving diffusability of tokenizers. Latent diffusion separates perceptual compression from generative modeling (Rombach et al., 2022; Esser et al., 2024), motivating tokenizers that balance reconstruction fidelity with downstream learnability. Pre-trained visual representations (Oquab et al., 2024) have improved latent diffusion through representation alignment in the denoiser (Yu et al., 2025; Singh et al., 2026) as well as regularization of the tokenizer (Leng et al., 2025a; Yao et al., 2025). Recent approaches further build tokenizers directly upon frozen pre-trained encoders (Zheng et al., 2026), while subsequent methods address the loss of fine-grained information when repurposing such encoders for generation (Chen et al., 2026a; Zhang et al., 2026; Yue et al., 2026; Shi et al., 2026). A complementary line of work jointly optimizes the tokenizer and diffusion model (Vahdat et al., 2021; Heek et al., 2026; Duggal et al., 2026; Leng et al., 2025a) to better align the latent representation with downstream generative learning. Other approaches regularize tokenizer representations through

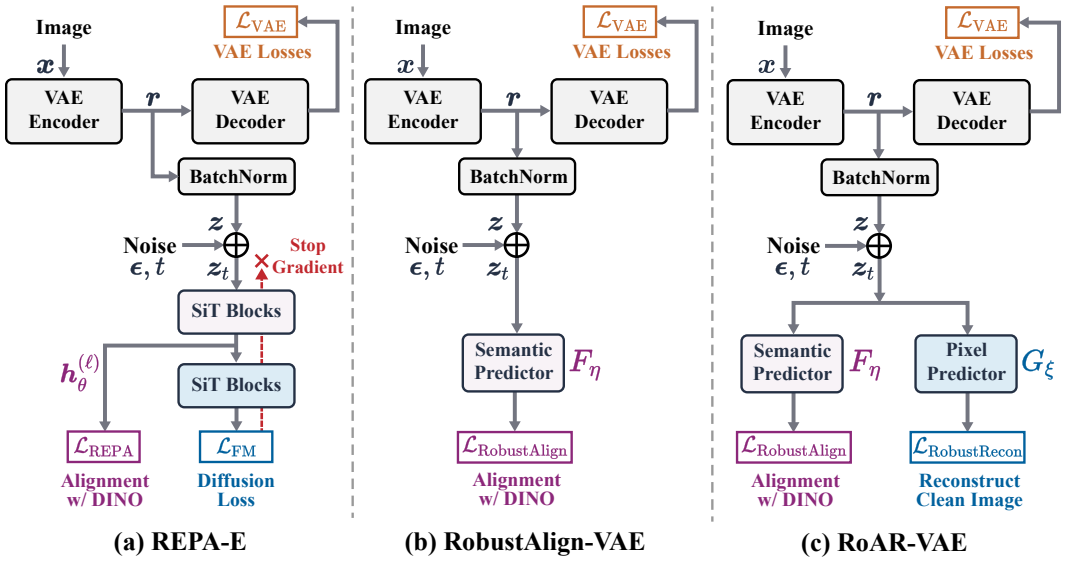

Figure 2: **From REPA-E to RoAR.** (a) REPA-E jointly trains the tokenizer and diffusion model, but only semantic alignment backpropagates through corrupted latents to the tokenizer encoder. (b) RobustAlign-VAE isolates this semantic-prediction pathway by removing the diffusion objective and its associated modules. (c) RoAR further introduces RobustRecon, combining semantic and image reconstruction targets under latent corruption alongside the standard VAE objective on clean latents.

structured perturbations in latent space, including equivariance-based objectives (Kouzelis et al., 2025; Skorokhodov et al., 2025) and reconstruction under masking or latent corruption (Chen et al., 2025; Yang et al., 2026; Liu et al., 2025).

Learning dynamics of diffusion models. Training dynamics influence both what diffusion models learn and when. Wang & Pehlevan (2025) derive an inverse-variance learning timescale for tractable denoisers, showing that higher-variance covariance modes are learned earlier.

Our work revisits REPA-E (Leng et al., 2025a), a notable approach that combines semantic enhancement with joint tokenizer–diffusion training, and identifies *robust prediction* as a key mechanism for improving latent learnability. The spectral bias of diffusion dynamics offers a complementary explanation: robust prediction makes generation-relevant information more accessible in dominant latent modes, which downstream diffusion models tend to learn earlier.

3 Method

Overview. Our goal is to learn a VAE latent space that supports both faithful reconstruction and efficient downstream diffusion learning. We begin by revisiting REPA-E (§ 3.1) and isolate the supervision that directly shapes the tokenizer encoder. This leads to *RobustAlign* (§ 3.2), which predicts pretrained semantic features from corrupted latents without jointly training a full diffusion model. We then introduce *RobustRecon* (§ 3.3), which instead reconstructs the clean image from corrupted latents to preserve fine-grained spatial and appearance details that semantic features may underrepresent. Combining the two objectives yields *RoAR-VAE*. Finally, we discuss how robust prediction improves diffusability by making semantic and spatial information more accessible in dominant latent modes (§ 3.4).

3.1 Preliminaries

Visual tokenizer. Given an image $\boldsymbol{x}$, the VAE encoder E_ϕ produces a latent representation $\boldsymbol{r} = E_\phi(\boldsymbol{x})$, from which the decoder D_ψ reconstructs $\boldsymbol{x}$. The tokenizer is trained with the standard VAE objective $\mathcal{L}_{\text{VAE}}$, which combines pixel reconstruction, perceptual, posterior regularization, and

Table 1: **Controlled comparison with REPA-E.** We train the same downstream text-to-image diffusion Transformer on the same data using latents from Vanilla-KL-VAE, REPA-E, and RobustAlign-VAE, and report the GenEval (Ghosh et al., 2023) results throughout training. Both REPA-E and RobustAlign substantially improve over the reconstruction-only Vanilla-KL baseline. Results are reported as mean ± standard deviation over four sampling seeds.

Method	4k	8k	16k	32k
T2I modeling w. Vanilla-KL-VAE	11.74±0.30	23.34±0.57	35.47±0.53	47.60±0.31
T2I modeling w. REPA-E-Tokenizer	26.33±0.25	36.42±0.56	49.46±0.31	60.60±0.90
T2I modeling w. RobustAlign-VAE	24.56±0.27	39.47±0.70	51.12±0.35	60.89±0.46

adversarial losses. Latent diffusion operates on the normalized representation $\boldsymbol{z} = \mathrm{Norm}(\boldsymbol{r})$, where $\mathrm{Norm}(\cdot)$ normalizes the latent statistics.

End-to-end representation alignment (REPA-E). REPA-E (Leng et al., 2025a) jointly trains the autoencoder and diffusion model, as illustrated in Figure 2(a). Let V_θ denote the diffusion Transformer (Ma et al., 2024; Peebles & Xie, 2023). Under the flow-matching formulation, the noisy latent and diffusion objective are

$$\begin{aligned} \boldsymbol{z}_t &= (1-t)\boldsymbol{z} + t\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \quad t \sim \mathcal{U}(0,1), \\ \mathcal{L}_{\mathrm{FM}}(\theta) &= \|V_\theta(\boldsymbol{z}_t, t) - (\boldsymbol{\epsilon} - \boldsymbol{z})\|_2^2 . \end{aligned} \tag{1}$$

REPA-E augments this generative objective with representation alignment (Yu et al., 2025), encouraging intermediate denoiser features to match pretrained visual representations, such as DINO features (Oquab et al., 2024):

$$\begin{aligned} \boldsymbol{z}_t(\phi) &= (1-t) \cdot \mathrm{Norm}(E_\phi(\boldsymbol{x})) + t\boldsymbol{\epsilon}, \\ \mathcal{L}_{\mathrm{REPA}}(\phi, \theta) &= \frac{1}{N} \sum_{n=1}^{N} \left(1 - \mathrm{Sim}\left(\boldsymbol{h}_\theta^{(\ell)}(\boldsymbol{z}_t(\phi), t)^{[n]}, T(\boldsymbol{x})^{[n]}\right)\right). \end{aligned} \tag{2}$$

Here, $\boldsymbol{z}_t(\phi)$ denotes the corrupted latent while retaining differentiability with respect to the encoder parameters ϕ. The feature map $\boldsymbol{h}_\theta^{(\ell)} \in \mathbb{R}^{N \times d}$ denotes the MLP-projected intermediate features from the ℓ-th layer of the diffusion Transformer, n indexes the N image patches, $T(\boldsymbol{x})$ denotes the pretrained visual target, and $\mathrm{Sim}(\cdot, \cdot)$ is cosine similarity. The alignment loss therefore supervises both the diffusion parameters θ and the tokenizer encoder parameters ϕ, while the standard VAE objective $\mathcal{L}_{\mathrm{VAE}}$ is retained throughout training.

To account for changing latent statistics during joint training, REPA-E applies channelwise batch normalization without affine parameters. It tracks the running latent mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$ and normalizes $\boldsymbol{z} = \mathrm{Norm}(\boldsymbol{r}) = (\boldsymbol{r} - \boldsymbol{\mu})/\boldsymbol{s}$, $\boldsymbol{s} = \sqrt{\boldsymbol{\sigma}^2 + \delta}$, where $\delta > 0$ ensures numerical stability.

A key detail of REPA-E is that the flow-matching loss does *not* backpropagate through the VAE encoder, as such gradients were found to cause latent variance collapse (Leng et al., 2025a). Specifically, the latent input to $\mathcal{L}_{\mathrm{FM}}$ is gradient-detached, whereas $\mathcal{L}_{\mathrm{REPA}}$ remains differentiable through $\boldsymbol{z}_t$ to the encoder. Consequently, among the two diffusion-side objectives, only $\mathcal{L}_{\mathrm{REPA}}$ provides additional supervision to the VAE encoder.

3.2 Robust Alignment

Simplifying REPA-E. REPA-E improves the VAE while jointly training a diffusion model, even though the flow-matching loss itself does not update the VAE encoder. This joint training requires a full generator and also prevents caching latents for large-scale training. We therefore ask: *is the improvement in diffusability driven by joint VAE–diffusion training, or by semantic prediction from corrupted VAE latents?* To isolate these factors, we introduce *RobustAlign*, shown in Figure 2(b). RobustAlign-VAE removes the flow-matching objective and all network components used only for diffusion prediction, retaining only the standard VAE objective together with semantic prediction from corrupted latents. Specifically, a shallow prediction network F_η recovers pretrained teacher

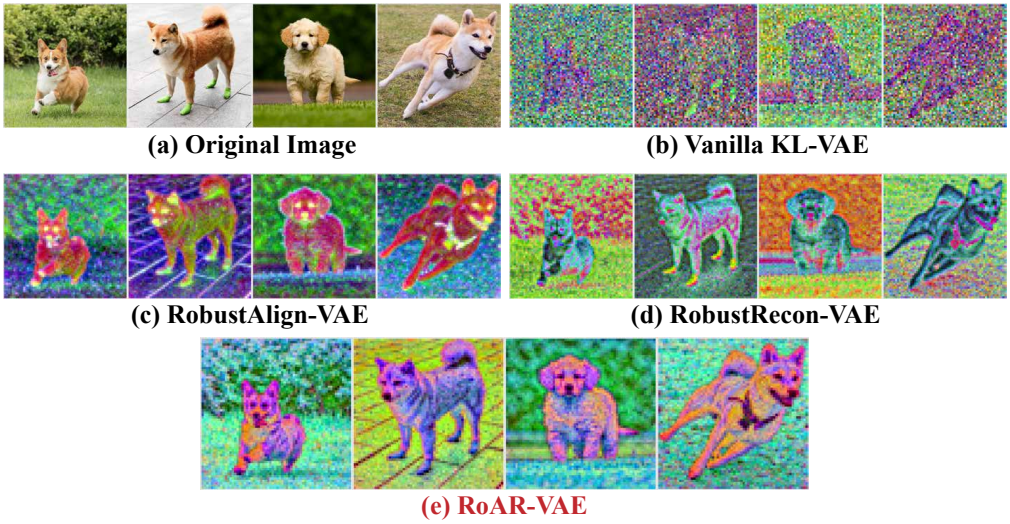

Figure 3: **PCA visualization of latent representations.** For the input images in (a), we visualize the top three principal components of normalized spatial latent features from (b) KL-VAE, (c) RobustAlign-VAE, (d) RobustRecon-VAE, and (e) RoAR-VAE. RobustAlign emphasizes object-level organization, RobustRecon preserves richer local variation, and RoAR combines both properties. PCA is fitted jointly across the four images for each tokenizer.

features $T(\boldsymbol{x})$ from $\boldsymbol{z}_t$:

$$\mathcal{L}_{\text{RobustAlign}}(\phi, \eta) = \frac{1}{N} \sum_{n=1}^{N} \left(1 - \text{Sim}\left(F_\eta(\boldsymbol{z}_t(\phi))^{[n]}, T(\boldsymbol{x})^{[n]}\right)\right). \tag{3}$$

The resulting RobustAlign-VAE is trained with

$$\mathcal{L}_{\text{RobustAlign-VAE}} = \mathcal{L}_{\text{VAE}} + \lambda_{\text{A}} \mathcal{L}_{\text{RobustAlign}}. \tag{4}$$

Table 1 compares the resulting tokenizer with REPA-E under matched downstream diffusion training[1]. RobustAlign-VAE performs comparably to REPA-E, indicating that semantic prediction from corrupted latents can recover the diffusability benefit without jointly training the full diffusion model.

Representation view. To examine how noisy semantic prediction changes the latent representation, Figure 3 visualizes the leading principal components of the latent space. Compared with the reconstruction-only baseline in Figure 3(b), RobustAlign in Figure 3(c) exhibits clearer object-level organization, with more coherent within-object regions and stronger foreground–background separation. This qualitative pattern is consistent with semantic prediction making target-relevant structure more accessible in dominant latent directions.

3.3 Robust Reconstruction and the RoAR Framework

RobustAlign encourages semantically organized latent representations, but semantic targets may underrepresent fine-grained spatial and appearance details that are also important for downstream generation. As illustrated in Fig. 3(c), RobustAlign produces coherent object-level structure, yet local variations within an object, such as texture and boundaries, are less prominent. This motivates a complementary objective that encourages detailed image information to remain recoverable from corrupted latents.

RobustRecon: Preserving fine-grained details. As shown in Figure 2(c), *RobustRecon* introduces an auxiliary decoder G_ξ that predicts the clean image $\boldsymbol{x}$ directly from the corrupted latent $\boldsymbol{z}_t$:

$$\mathcal{L}_{\text{RobustRecon}}(\phi, \xi) = \mathcal{L}_{\text{MAE}}\left(G_\xi(\boldsymbol{z}_t(\phi)), \boldsymbol{x}\right) + \mathcal{L}_{\text{LPIPS}}\left(G_\xi(\boldsymbol{z}_t(\phi)), \boldsymbol{x}\right), \tag{5}$$

[1]To isolate latent learnability, we train a separate text-to-image diffusion model using the tokenizer produced by REPA-E, without applying the REPA objective during downstream training.

where the two terms measure pixel and perceptual reconstruction error, respectively (Zhang et al., 2018). G_ξ shares the architecture of the VAE decoder but uses independent parameters and is discarded after tokenizer training. Beyond standard posterior sampling in VAEs (Kingma & Welling, 2013), RobustRecon applies corruption over a broad range of noise levels, including latents with very low signal-to-noise ratios. The resulting objective backpropagates through $\boldsymbol{z}_t$ to the encoder, encouraging fine-grained image information to remain accessible under corruption. As shown in Fig. 3(d), RobustRecon exposes richer local spatial and appearance variations than RobustAlign, although its object-level foreground–background organization is less pronounced.

Overall framework. RobustAlign encourages noisy latents to preserve high-level semantic structure, whereas RobustRecon encourages them to retain fine-grained image details. We therefore combine the two objectives into *RoAR-VAE*, short for **Ro**bust **A**lignment and **R**econstruction VAE, as illustrated in Fig. 2(c):

$$\mathcal{L}_{\mathrm{RoAR}} = \mathcal{L}_{\mathrm{VAE}} + \lambda_{\mathrm{A}}\mathcal{L}_{\mathrm{RobustAlign}} + \lambda_{\mathrm{R}}\mathcal{L}_{\mathrm{RobustRecon}}, \tag{6}$$

where λA and λ_{R} balance the two auxiliary objectives. Both objectives supervise the encoder through corrupted latents, but use complementary clean targets: pretrained semantic features for RobustAlign and the original image for RobustRecon. As shown in Figure 3(e), their combination yields latent representations that retain both coherent object-level organization and richer local structure, including clearer boundaries, textures, and intra-object variations.

3.4 Discussion

Robust prediction promotes key information in dominant modes. Prior work suggests that higher-variance latent modes are learned more rapidly by diffusion models (Wang & Pehlevan, 2025), indicating that diffusability depends not only on *what* information is encoded, but also on *how* prominently it is represented. Both RobustAlign and RobustRecon follow the same robust prediction principle: clean targets must remain predictable from corrupted latents. Under isotropic corruption, higher-variance latent directions retain a higher signal-to-noise ratio and are therefore more resistant to noise. Optimizing robust prediction thus encourages target-relevant semantics and spatial details to remain accessible in these dominant modes. This provides a possible explanation for the improved diffusability of the resulting latent representations.

4 Experiments

4.1 Experimental Setup

We briefly summarize the training and evaluation setup here; full implementation details are provided in Appendix A.

Tokenizer training. RoAR-VAE uses a convolutional architecture (Esser et al., 2021; Rombach et al., 2022) with $8\times$ spatial downsampling, 32 latent channels, and approximately 84M parameters. For controlled comparison with prior tokenizer studies, we train exclusively on ImageNet (Deng et al., 2009) at 256^2 resolution. Training proceeds in two stages: (1) a 200k-step VAE warm-up using $\mathcal{L}_{\mathrm{VAE}}$, followed by (2) 250k steps of RoAR refinement with $\lambda_A = 1.5$ and $\lambda_R = 0.5$. For RobustAlign, the semantic predictor F_η comprises an 8-layer Transformer (Ma et al., 2024) followed by an MLP projector, with frozen DINOv2-B (Oquab et al., 2024) features as targets. For RobustRecon, the pixel predictor G_ξ is initialized from the stage-one VAE decoder D_ψ. The corruption level is sampled uniformly as $t \sim \mathcal{U}(0, 1)$.

Downstream generative modeling. To evaluate latent diffusability, we freeze each tokenizer and train the same text-conditioned diffusion Transformer from scratch on a filtered and recaptioned LAION dataset (Schuhmann et al., 2022) at 256^2 resolution. We use an MMDiT backbone (Esser et al., 2024; Chen et al., 2026b) with 24 layers, a hidden dimension of 1024, and approximately 0.8B parameters, together with a frozen Qwen3-0.6B text encoder (Yang et al., 2025). For tokenizers with different spatial compression factors, we adjust the latent patch size to keep the diffusion sequence length fixed. RoAR-VAE uses shifted logit-normal timestep sampling with exponential shifting ($\mu = 1.5$), while each baseline uses its official timestep scheduler.

Table 2: **Reconstruction and downstream generation results.** We report reconstruction metrics evaluated on the ImageNet validation set at 256^2 resolution. Generation performance is evaluated at 32k downstream updates across three benchmarks: GenEval, GenEval2, and DPG-Bench. Avg. denotes the unweighted mean of the three generation scores.

Method	Reconstruction				Generation			
	rFID↓	PSNR↑	LPIPS↓	SSIM↑	GenEval	GenEval2	DPG-Bench	Avg.
SD3-VAE	0.2038	29.58	0.0688	0.8565	52.46	38.03	73.27	54.59
FLUX.1-VAE	0.1711	31.08	0.0510	0.8938	51.79	34.82	72.08	52.90
VA-VAE	0.2892	26.26	0.1052	0.7625	57.43	<u>51.28</u>	76.30	<u>61.67</u>
VTP	0.3457	25.75	0.1339	0.7383	53.79	50.31	76.08	60.06
PAE	0.2643	25.90	0.1123	0.7487	56.60	49.77	<u>76.82</u>	61.06
REPA-E-FLUX	0.1563	29.89	0.0546	0.8737	<u>58.92</u>	48.63	76.81	61.45
RoAR-VAE (Ours)	**0.1309**	**32.88**	**0.0349**	**0.9258**	**62.22**	**51.79**	**77.56**	**63.86**
FLUX.2-VAE	0.1085	31.52	0.0398	0.9052	60.04	51.28	76.88	62.73



Figure 4: **Qualitative comparison of text-to-image generation.** Rows compare FLUX.1-VAE, VA-VAE, and RoAR-VAE tokenizers; each subject is shown after 32k and 120k downstream updates.

Baselines. We compare against tokenizers from large-scale generative systems, including SD3-VAE (Esser et al., 2024) and FLUX.1-VAE (Labs, 2024), as well as recent methods targeting improved latent diffusability, including VA-VAE (Yao et al., 2025), VTP (Yao et al., 2026), PAE (Yue et al., 2026), and REPA-E-FLUX (Leng et al., 2025b). We additionally report FLUX.2-VAE (Labs, 2025) as a reference, but exclude it from baseline comparisons because its tokenizer training data and training recipe are undisclosed.

Evaluation protocol. For reconstruction, we report rFID, PSNR, LPIPS, and SSIM on the ImageNet validation set at 256^2 resolution, and evaluate text and facial detail preservation using TokBench (Wu et al., 2025). For generation, we train a downstream diffusion model for each tokenizer under the controlled setup above and evaluate on GenEval (Ghosh et al., 2023), GenEval2 (Kamath et al., 2025), and DPG-Bench (Hu et al., 2024). Each score is averaged over four independent sampling seeds, and we report the arithmetic mean across the three benchmarks as a summary metric.

4.2 Main Results

Reconstruction versus diffusion learnability. Table 2 compares RoAR-VAE with existing visual tokenizers in terms of reconstruction quality and downstream generation under a fixed 32k-step training budget. RoAR-VAE achieves a favorable balance between the two objectives, substantially improving reconstruction while maintaining strong downstream learnability. Compared with the reconstruction-oriented FLUX.1-VAE, RoAR-VAE improves PSNR by 1.80 dB and reduces LPIPS from 0.0510 to 0.0349, while outperforming the strongest competing learnability-enhanced tokenizer

Table 3: **Text and face reconstruction on TokBench,** using text accuracy (T-ACC), text normalized edit distance (T-NED), and face similarity (F-Sim), evaluated at an image resolution of 256^2. Subscripts s and m denote metrics averaged over small-scale and all instances, respectively. Higher is better for every column. Bold denotes the best displayed value.

Method	Text (%) ↑				Face ↑	
	T-ACC$_s$	T-ACC$_m$	T-NED$_s$	T-NED$_m$	F-Sim$_s$	F-Sim$_m$
SD3-VAE	37.29	68.32	59.86	81.36	0.43	0.71
FLUX.1-VAE	52.07	76.57	71.41	86.96	0.53	0.77
VA-VAE	7.05	37.15	25.12	56.30	0.22	0.49
VTP	0.14	6.88	5.87	20.03	0.10	0.24
PAE	3.31	28.97	16.73	46.97	0.19	0.45
RoAR-VAE (Ours)	**69.00**	**84.95**	**82.87**	**92.22**	**0.65**	**0.83**
FLUX.2-VAE	67.33	83.80	81.90	91.54	0.63	0.83



Figure 5: **Image reconstruction visualization.** Columns show the original image and reconstructions from baseline tokenizers and RoAR-VAE. Images are selected from TokBench (Wu et al., 2025).

by 2.19 points on the average downstream benchmark score. RoAR-VAE also approaches the reconstruction and generation performance of FLUX.2-VAE, a strong industrial tokenizer with a similarly balanced reconstruction–generation trade-off.

Generated image quality. Figure 4 qualitatively compares downstream generation across different tokenizers. RoAR-VAE produces clearer object shapes and finer visual structure, including more distinct animal contours, sharper suitcase and bag textures, and more regular soccer-ball patterns. Compared with the reconstruction-driven FLUX.1-VAE, these results indicate faster acquisition of object appearance and structure under the same downstream training budget. VA-VAE also exhibits strong learnability, but its generations are noticeably smoother and contain fewer details, consistent with its weaker reconstruction performance relative to RoAR-VAE.

Fine-grained reconstruction. Table 3 further evaluates preservation of challenging text and facial details. RoAR-VAE achieves the best performance across all reported text and face reconstruction metrics. Despite being trained from scratch solely on ImageNet, it improves small-text accuracy over FLUX.1-VAE by 16.93 percentage points and increases small-face similarity from 0.53 to 0.65. Although VA-VAE, VTP, and PAE achieve competitive generation scores in Table 2, their weaker text and face reconstruction results reveal a substantial loss of fine-grained information. Figure 5 provides qualitative examples of these differences, highlighting distortions in textual and facial patterns introduced by the compared tokenizers.

4.3 Analysis

Contribution of the objectives and latent corruption. Table 4 isolates the contributions of the two robust prediction objectives. Both RobustRecon (❷) and RobustAlign (❸) consistently improve over Vanilla-KL (❶) across all three benchmarks, increasing the average score from 50.50 to 59.38 and

Table 4: **Ablation of robust prediction objectives and latent corruption.** "Noisy Latent" indicates whether the auxiliary predictors receive noise-corrupted latents as input. Scores are measured after 32k downstream updates; Avg. denotes the mean over the three benchmarks.

	Noisy Latent	$\mathcal{L}_{\text{RobustAlign}}$	$\mathcal{L}_{\text{RobustRecon}}$	GenEval	GenEval2	DPG-Bench	Avg.
❶	✗	✗	✗	47.60	33.68	70.22	50.50
❷	✓	✗	✓	57.63	45.20	75.32	59.38 (↑ 8.88)
❸	✓	✓	✗	60.89	51.11	75.48	62.49 (↑ 11.99)
❹	✓	✓	✓	**62.22**	**51.79**	**77.56**	**63.86** (↑ 13.36)
❺	✗	✓	✓	57.54	44.01	73.86	58.47 (↓ 5.39)

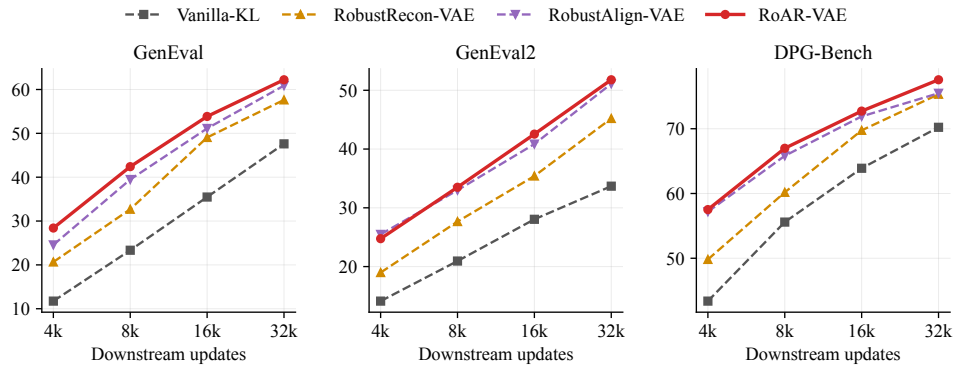

Figure 6: **Downstream learning curves.** RoAR-VAE consistently outperforms using RobustAlign or RobustRecon alone across different training budgets.

62.49, respectively. Combining the two objectives (❹) further improves performance, reaching an average score of 63.86 and demonstrating the complementary benefits of semantic- and pixel-level robust prediction. Removing latent corruption (❺) substantially degrades the average score from 63.86 to 58.47, highlighting prediction from corrupted latents as a key component of the framework. Figure 6 further compares downstream performance throughout training: RoAR-VAE consistently outperforms either robust prediction objective used alone across different training budgets.

Predictor design choices. Table 5 studies architectural choices for the auxiliary predictors. For RobustAlign, replacing the Transformer-based semantic predictor F_η with a simple MLP causes a substantial performance drop, suggesting that a more expressive prediction pathway is beneficial for shaping diffusion-friendly latents. For RobustRecon, sharing parameters between the pixel predictor G_ξ and the clean reconstruction decoder D_ψ also slightly degrades performance, indicating that the two decoders serve distinct roles in the framework.

Table 5: **Predictor architecture choices.**

Variant	Avg. Score
RobustAlign w. simple MLP F_η	52.70
RobustAlign-VAE	**62.49**
RoAR w. shared G_ξ and D_ψ	62.13
RoAR-VAE	**63.86**

5 Conclusion

In this paper, we present RoAR-VAE, a visual tokenizer built on the principle of *robust prediction*. By revisiting and simplifying REPA-E, we identify prediction of clean targets from corrupted latents as an effective source of supervision for shaping diffusion-friendly representations. We realize this principle with RobustAlign for semantic structure and RobustRecon for fine-grained spatial and appearance details. Our analysis suggests that robust prediction encourages generation-relevant information to remain accessible in prominent latent modes that are more resistant to corruption and tend to be learned earlier by downstream diffusion models. Empirically, the two objectives provide complementary gains, allowing RoAR-VAE to improve downstream generative learnability while maintaining high-fidelity reconstruction. These results highlight robust prediction as a simple and effective principle for designing visual tokenizers that better balance reconstruction and generation.

AI USE STATEMENT

We used generative AI tools for several research tasks that fall under the required disclosure categories, including assisting with implementation of the proposed method, cleaning and reformatting datasets and benchmark inputs, translation, providing feedback on research methodology and experimental design, and assisting with the interpretation of experimental results. We also used generative AI tools for recommended disclosure activities, including developing and editing research code, searching for and identifying relevant literature and research materials, summarizing related work, and drafting and editing portions of the manuscript for clarity and readability.

We did not use generative AI tools to generate synthetic datasets, develop the conceptual framework of the work, formulate or refine research hypotheses, formulate mathematical claims, provide ingredients for mathematical proofs, or write proofs. The proof-related categories are not applicable to this work.

All AI-assisted outputs were reviewed by the authors. AI-generated or AI-edited code was inspected, tested, and verified for correctness before use. The authors take full responsibility for the final content of the paper, including its methodology, experiments, analysis, claims, text, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work evaluates reconstruction performance on a published image reconstruction benchmark (Wu et al., 2025) that includes human faces. However, the proposed visual tokenizer is a general-purpose information compressor and is neither trained specifically on human-centric datasets nor optimized for human-related attributes. For downstream visual generation systems built upon our tokenizer, we recommend adopting appropriate safeguards, such as watermarking and monitoring mechanisms, to mitigate potential misuse.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training objectives, network architectures, experimental setup, and hyperparameter configurations in Section 3 and Appendix A. The implementation code and trained models will be released.

REFERENCES

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*, 2026a. URL https://openreview.net/forum?id=ajnBafpqmE.

Dong Chen, Fangyun Wei, Ziyu Wan, Dongdong Chen, Jiawei Zhang, Jinjing Zhao, Sirui Zhang, Yang Yue, Zhiyang Liang, Baining Guo, et al. Lens: Rethinking training efficiency for foundational text-to-image models. *arXiv preprint arXiv:2605.21573*, 2026b.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dzwUOiBlQW.

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. Ieee, 2009.

Shivam Duggal, Xingjian Bai, Zongze Wu, Richard Zhang, Eli Shechtman, Antonio Torralba, Phillip Isola, and William T Freeman. End-to-end training for unified tokenization and latent denoising. *arXiv preprint arXiv:2603.22283*, 2026.

Patrick Esser, Robin Rombach, and Bjorn 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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (ul): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

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

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UWhW5YYLo6.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

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

Xingjian Leng, Jaskirat Singh, Ryan Murdock, Ethan Smith, Rebecca Li, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Family of End-to-End Tuned VAEs for Supercharging T2I Diffusion Transformers. https://end2end-diffusion.github.io/repa-e-t2i/, 2025b.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

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.

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, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 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.

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*, 2026. URL https://openreview.net/forum?id=y0UxFtXqXf.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2hEDcA7xy4.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5dca4c6b9e244d24a30b4c45601d9720-Paper.pdf.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=SDhOClkyqC.

Junfeng Wu, Dongliang Luo, Weizhi Zhao, Zhihao Xie, Yuanhao Wang, Junyi Li, Xudong Xie, Yuliang Liu, and Xiang Bai. Tokbench: Evaluating your visual tokenizer before visual generation. *arXiv preprint arXiv:2505.18142*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1jBsi98fVe.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. In *European Conference on Computer Vision*, pp. 661–678. Springer Nature Switzerland Cham, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, et al. What matters for diffusion-friendly latent manifold? prior-aligned autoencoders for latent diffusion. *arXiv preprint arXiv:2605.07915*, 2026.

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

Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian GE, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, Daniil Pakhomov, Kai Zhang, Zhe Lin, and Ping Luo. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=UmlXZ4t44h.

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.

A Implementation Details

A.1 Tokenizer Training

RoAR-VAE uses the standard convolutional autoencoder architecture adopted by latent diffusion models (Esser et al., 2021; Rombach et al., 2022), with an $8\times$ spatial downsampling factor, 32 latent channels, and approximately 84M parameters. The encoder and decoder each contain four resolution stages with channel widths $(128, 256, 512, 512)$. All tokenizer training is performed on ImageNet (Deng et al., 2009) at 256×256 resolution.

Training proceeds in two stages. We first warm up the tokenizer for 200k optimization steps using the standard VAE objective $\mathcal{L}_{\mathrm{VAE}}$. The reconstruction objective uses unit weights for the L1 and LPIPS losses, together with a KL regularization weight of 10^{-6}. Adversarial supervision is provided by a DINOv2-S (Oquab et al., 2024)-based discriminator. We then refine the tokenizer for an additional 250k steps using the RoAR objectives with $\lambda_{\mathrm{A}} = 1.5$ and $\lambda_{\mathrm{R}} = 0.5$. The corruption level is sampled uniformly as $t \sim \mathcal{U}(0, 1)$.

For RobustAlign, the semantic predictor F_η consists of an 8-layer Transformer (Ma et al., 2024) followed by an MLP projector that matches the dimensionality of the target features. We use features from a frozen DINOv2-B (Oquab et al., 2024) encoder as the semantic prediction target. For RobustRecon, the pixel predictor G_ξ is initialized from the VAE decoder D_ψ obtained after the warm-up stage and predicts the clean image from corrupted latents using L1 and LPIPS losses.

The complete RoAR framework is optimized with AdamW (Loshchilov & Hutter, 2019), using a batch size of 256 and a learning rate of 1×10^{-5}.

A.2 Downstream Generative Modeling

For each tokenizer, we freeze its parameters and train a text-conditioned diffusion Transformer from scratch on a filtered and recaptioned version of LAION (Schuhmann et al., 2022). All downstream models are trained at 256×256 resolution with a global batch size of 1,024. The diffusion backbone is an MMDiT (Esser et al., 2024; Chen et al., 2026b) with 24 Transformer layers, a hidden dimension of 1,024, 16 attention heads, and approximately 0.8B parameters. Text conditioning is provided by a frozen Qwen3-0.6B encoder (Yang et al., 2025). Because the compared tokenizers use different spatial compression factors, we adjust the latent patch size to maintain the same sequence length for the diffusion Transformer. Specifically, we use a 1×1 latent patch size for tokenizers with a downsampling factor of 16 and a 2×2 patch size for tokenizers with a downsampling factor of 8. This keeps the number of latent tokens processed by the downstream model consistent across tokenizer variants.

For RoAR-VAE, diffusion timesteps are sampled using shifted logit-normal sampling with exponential shifting and $\mu = 1.5$. For each baseline tokenizer, we use its official or recommended timestep scheduler. All downstream models are optimized with AdamW (Loshchilov & Hutter, 2019), using $\beta_1 = 0.9$, $\beta_2 = 0.999$, a constant learning rate of 2×10^{-4}, and a caption dropout probability of 0.1. At inference time, we use an Euler sampler with 50 sampling steps and a classifier-free guidance weight of 5.0.

A.3 Evaluation Details

Baselines. SD3-VAE (Esser et al., 2024) and FLUX.1-VAE (Labs, 2024) serve as reconstruction-oriented baselines from large-scale text-to-image systems. VA-VAE (Yao et al., 2025) directly aligns latent representations with pretrained visual features using a margin-based similarity objective, while VTP (Yao et al., 2026) jointly optimizes reconstruction and visual representation learning. PAE (Yue et al., 2026) builds on pretrained visual encoders and regularizes the latent manifold for improved downstream generation. REPA-E-FLUX (Leng et al., 2025b) jointly refines the tokenizer and diffusion model through representation alignment; in our evaluation, we use only the resulting tokenizer and retrain the downstream diffusion model from scratch. We additionally include FLUX.2-VAE (Labs, 2025) as a strong reference, since its tokenizer training data and recipe are undisclosed.

Reconstruction. We evaluate reconstruction on the ImageNet (Deng et al., 2009) validation set at 256×256 resolution and report reconstruction FID (rFID), PSNR, LPIPS, and SSIM. To further assess preservation of fine-grained content, we evaluate on TokBench (Wu et al., 2025), which contains challenging text- and face-centric images. For text reconstruction, TokBench reports exact recognition accuracy (T-ACC) and normalized edit distance (T-NED), with separate metrics for small text instances and all text instances. For face reconstruction, it measures cosine similarity between face-recognition embeddings (F-Sim), likewise reporting both small-face and overall results. These metrics complement standard image-level reconstruction measures by directly evaluating whether tokenizers preserve small, semantically important details.

Text-to-image generation. For each tokenizer, we train a downstream diffusion model following Appendix A.2. Unless otherwise stated, benchmark results are measured after 32k downstream optimization steps, enabling comparison of latent diffusability under a fixed training budget. We evaluate text–image alignment on three complementary benchmarks. GenEval (Ghosh et al., 2023) evaluates compositional generation involving object co-occurrence, spatial relations, counting, and colors, while GenEval2 (Kamath et al., 2025) extends this setting with broader coverage of visual concepts and higher compositionality. DPG-Bench (Hu et al., 2024) focuses on dense prompts containing multiple objects, detailed attributes, and complex relationships. Each reported score is averaged over four independent sampling seeds. We additionally report the arithmetic mean of the three benchmark scores, denoted *Avg.*, as a compact summary.