GUAVAE: GUIDANCE-ALIGNED END-TO-END VAE TRAINING FOR LATENT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have demonstrated that pretrained visual encoders such as DINO can help to improve latent spaces for diffusion modeling. However, these approaches depend on an external representation model. We investigate whether diffusion transformers (DiTs) themselves can provide supervision for learning better latent spaces through VAE tuning. Prior work has shown that classifier-free guidance (CFG) can enhance internal DiT features, making them effective targets for representation alignment during training. We introduce GuaVAE, an end-to-end VAE tuning method that extends this objective to the VAE encoder. To encourage the VAE to learn a more generation-friendly latent space, we align a trainable DiT's features with CFG-enhanced features obtained at low noise levels from a frozen pretrained DiT. We backpropagate the alignment loss through the trainable DiT to the VAE encoder. Experiments show that the tuned VAE improves generation quality across DiT backbone sizes, achieving performance competitive with DINO-supervised VAEs. We also observe better generation quality when using VAEs tuned with larger DiTs, supporting diffusion representations as a scalable source of supervision. A DiT trained with the tuned VAE and CFG-enhanced feature alignment achieves a gFID of 4.97 on ImageNet 256×256 without CFG, outperforming self-contained representation alignment baselines as well as DINO-guided REPA. With CFG, it achieves a competitive gFID of 1.62.

1 INTRODUCTION

Diffusion models have become a widely used approach to high-quality image generation (Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021). To reduce the cost of modeling high-resolution images, latent diffusion models operate in a compressed latent space produced by an autoencoder, typically a variational autoencoder (VAE) (Rombach et al., 2022). The encoder maps images to latent representations, the diffusion model learns to generate samples in this latent space, and the decoder maps those samples back to images. The choice of latent space affects the quality of generated images: some latent distributions are easier for a diffusion model to learn than others. This ease of learning is referred to as *diffusability* (Skorokhodov et al., 2025).

Representation alignment offers one way to improve VAE latent spaces. VA-VAE (Figure 1b) aligns VAE latents with features from a pretrained visual encoder such as DINO (Yao et al., 2025). REPA-E (Figure 1a) instead aligns intermediate diffusion features with the visual encoder's features and backpropagates the alignment loss through the diffusion model to the VAE encoder (Leng et al., 2025). Both approaches improve generation with the resulting VAE, but require an external pretrained encoder to provide the alignment targets.

Representation alignment also improves diffusion training itself. REPA accelerates the training of diffusion transformers (DiTs) by aligning their intermediate features with DINO features (Yu et al., 2025). Self-contained alternatives such as SRA and LayerSync guide earlier diffusion layers with representations from deeper layers, avoiding this external dependency (Jiang et al., 2025; Haghighi et al., 2025). Self-Transcendence further improves the quality of internal targets (Sun et al., 2026). It first trains a diffusion teacher using an auxiliary VAE-based objective to improve its early-layer representations. It then applies classifier-free guidance (CFG) to the teacher's intermediate features to strengthen their class-specific information. Aligning a new diffusion model with these features achieves generation quality competitive with DINO-guided REPA. External features can improve

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.

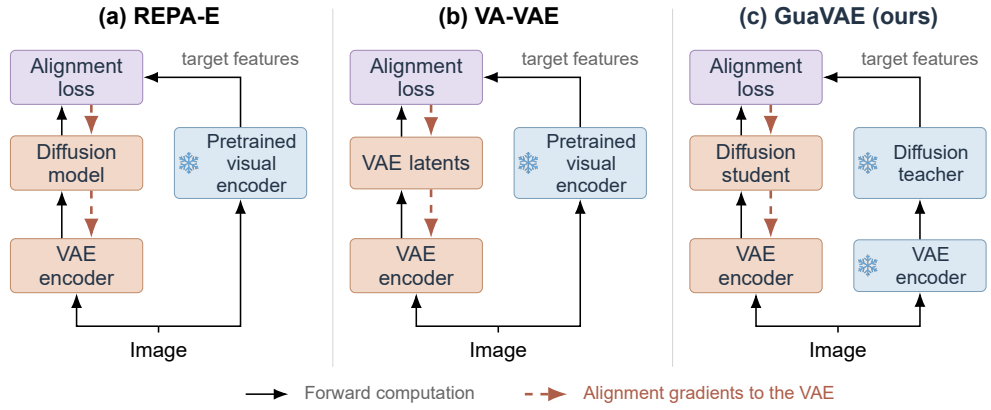


Figure 1: **Representation supervision for improving VAE latent spaces.** (a) REPA-E aligns diffusion features with a pretrained visual encoder and backpropagates the alignment loss through the diffusion model to the VAE. (b) VA-VAE aligns VAE latents with a pretrained visual encoder. (c) GuaVAE replaces the external visual encoder with a diffusion teacher. Snowflakes denote frozen modules; dashed arrows show alignment gradients to the VAE encoder. Projection heads, noising, feature guidance, and other objectives are omitted; Figure 2 details the diffusion alignment pathway.

VAE latent spaces, and internal diffusion features can replace external features in diffusion training. This motivates our central question: **Can internal diffusion representations be used to improve a VAE's latent space?**

We introduce GuaVAE (**Gu**idance-**A**ligned **VAE**), an end-to-end method for tuning a VAE using intermediate features from a frozen DiT teacher. To strengthen the semantic signal, we apply CFG to the teacher's features to obtain the alignment targets. We fine-tune the VAE and pass its latents to a trainable DiT student, which aligns its own intermediate features with these targets. Backpropagating the alignment loss through the student updates the VAE encoder, adapting its latent space for downstream generation (Figure 1c). The teacher's noise level is important: independently sampling its timestep from a low-noise interval yields substantially better VAEs than the naive approach of matching teacher and student timesteps.

An additional question is whether stronger diffusion models can produce better VAEs. The encoder-size studies in REPA and REPA-E report limited generation gains from larger DINO encoders in their tested settings (Yu et al., 2025; Leng et al., 2025). In our experiments, increasing the teacher and student backbones together improves downstream generation across the tested model sizes, with the evaluation model held fixed. The resulting VAEs narrow the gap in generation quality with DINO-supervised VAEs, suggesting that increasing diffusion-model capacity can improve the effectiveness of this supervision.

Our contributions are summarized as follows:

- We introduce GuaVAE, a method for tuning VAE latent spaces using internal diffusion representations without an external visual encoder. The resulting VAEs improve generation with DiTs and narrow the gap in generation quality with DINO-supervised VAEs as the DiT backbone used for VAE tuning increases in size.
- We combine GuaVAE with Self-Transcendence to obtain a self-contained generative method. On ImageNet 256×256, it achieves gFID 4.97 without sampling CFG, outperforming self-contained representation-alignment methods and DINO-guided REPA, and a competitive gFID of 1.62 with sampling CFG.

2 Related Work

Visual tokenizers and latent diffusion. Visual tokenizers map images to compact representations for generative modeling. VAEs learn continuous latent variables through variational inference (Kingma & Welling, 2014), while VQ-VAE learns a discrete codebook (van den Oord et al., 2017).

VQGAN combines quantization with perceptual and adversarial objectives to preserve visual detail under compression (Esser et al., 2021). Diffusion and score-based models generate samples by reversing a noise-corruption process (Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021). LSGM and latent diffusion move this modeling problem to an autoencoder's latent space, reducing the cost of high-resolution generation (Vahdat et al., 2021; Rombach et al., 2022). Transformer backbones such as U-ViT and DiT provide scalable architectures for diffusion modeling (Bao et al., 2023; Peebles & Xie, 2023). Flow matching learns transport vector fields (Lipman et al., 2023), and SiT studies flow and diffusion within a scalable interpolant framework (Ma et al., 2024).

Representation learning for diffusion. Auxiliary representation objectives can improve diffusion training. MaskDiT combines denoising with masked-patch reconstruction (Zheng et al., 2024), while SD-DiT adds self-supervised discriminative learning (Zhu et al., 2024). REPA directly aligns intermediate diffusion features with representations from pretrained visual encoders (Yu et al., 2025), including DINOv2 (Oquab et al., 2024). iREPA studies which properties make these targets effective and emphasizes spatial structure over global semantic information (Singh et al., 2025). Other methods obtain supervision from the generative system itself. SRA and LayerSync align internal diffusion representations (Jiang et al., 2025; Haghighi et al., 2025); SRA2 uses VAE representations (Wang et al., 2026). Self-Transcendence combines VAE-based teacher preparation with guided diffusion features to provide effective alignment targets without a separate visual encoder (Sun et al., 2026).

Autoencoders for downstream generation. Recent work studies latent-space properties that affect generation. EQ-VAE regularizes spatial equivariance (Kouzelis et al., 2025), while Skorokhodov et al. (2025) analyze latent frequency structure and introduce scale-equivariance regularization. A broader study of latent diffusability examines spatial structure and semantic organization as predictors of generation quality (Zhong et al., 2026). MAETok uses masked autoencoding to learn representations suitable for diffusion (Chen et al., 2025). Representation autoencoders pair pretrained representation encoders with learned decoders and adapt diffusion to their high-dimensional latents (Zheng et al., 2025). VA-VAE instead aligns VAE latents with pretrained visual features (Yao et al., 2025), and REPA-E propagates representation alignment through the diffusion model to the VAE encoder (Leng et al., 2025). Diffusion as Self-Distillation (DSD) addresses latent collapse through a self-distillation formulation, sharing a transformer backbone across encoding, reconstruction, and diffusion (Wang & Zhang, 2025). Concurrent work GenFirst stabilizes joint training through posterior-entropy regularization and a schedule that prioritizes generation before progressively strengthening reconstruction (Zheng et al., 2026).

3 METHOD

Our training system consists of a trainable VAE and diffusion student, and a diffusion teacher with its own VAE (Figure 2). We prepare the teacher following the first stage of Self-Transcendence (Sun et al., 2026): an auxiliary head predicts the diffusion velocity from an early layer, alongside the final diffusion prediction. Appendix A specifies this objective. This preparation improves downstream generation over a teacher trained only with the diffusion loss (Table 3(d)). During end-to-end tuning, the teacher and its VAE remain frozen. The alignment loss updates the student and reaches the trainable VAE encoder through it.

3.1 LATENT DIFFUSION SETUP

Let E and D denote the trainable VAE encoder and decoder, and f the diffusion student. For an image x with class label y, we sample $u \sim q_E(u \mid x)$ from the encoder posterior and normalize it to $z = B(u)$. Following REPA-E (Leng et al., 2025), B is a non-affine batch-normalization layer that tracks the changing latent statistics. Following SiT (Ma et al., 2024), we form the noisy latent

$$z_t = (1-t)z + t\epsilon, \qquad t \sim \mathcal{U}(0,1), \quad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

The student predicts the velocity $\epsilon - z$ using the diffusion loss

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\frac{1}{N}\left\|f(z_t, t, y) - (\epsilon - z)\right\|_2^2\right], \tag{2}$$

where N is the number of scalar latent entries per image and the norm is taken over the flattened tensor. Following REPA-E, we detach the VAE latents when computing this loss, so its gradients

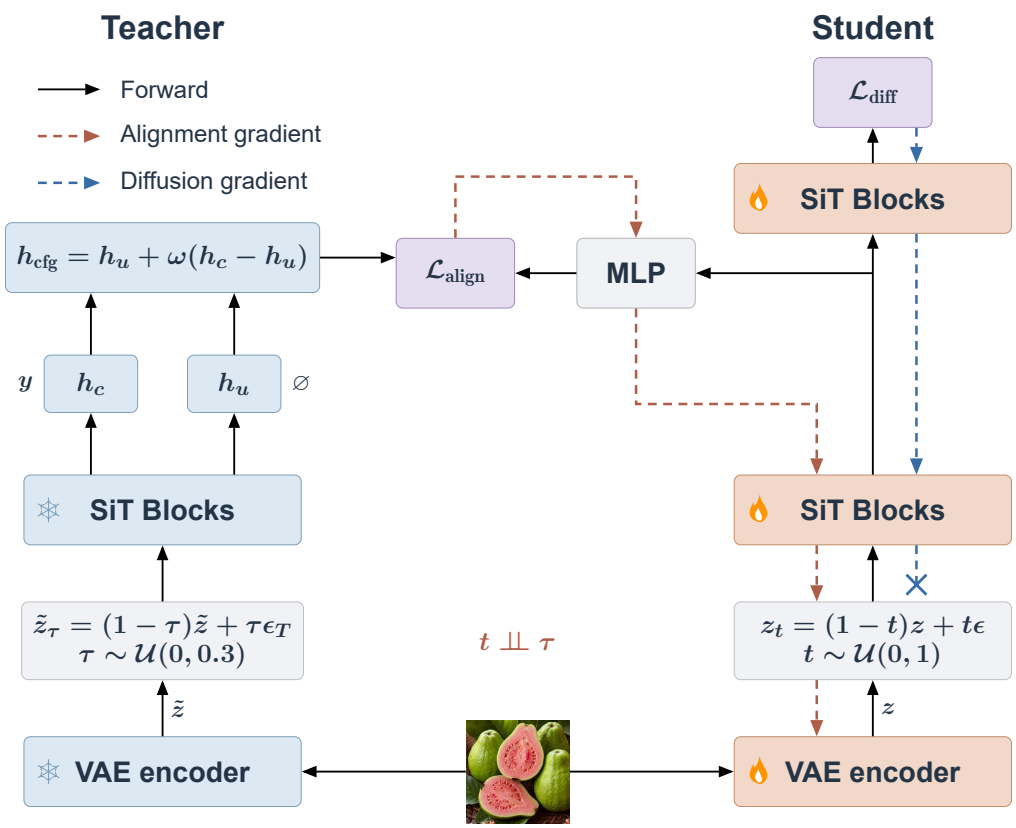


Figure 2: **GuaVAE tuning.** The prepared teacher supplies CFG-enhanced targets at independently sampled low-noise timesteps. Red dashed arrows trace alignment gradients; blue dashed arrows trace diffusion gradients. The cross marks the diffusion-gradient boundary; snowflakes and flames denote frozen and trainable modules. VAE and student updates alternate (Section 3.3); normalization and reconstruction objectives are omitted.

update only the diffusion student. The alignment loss defined below provides the gradient path through the student to the VAE encoder.

3.2 Constructing Diffusion Alignment Targets

Let $\tilde{z}$ denote the normalized latent of the same image x from the teacher VAE, and let f_T denote the diffusion teacher. We sample its timestep τ independently of the student timestep t:

$$\tilde{z}_\tau = (1-\tau)\tilde{z} + \tau\epsilon_T, \qquad \tau \sim \mathcal{U}(a, b). \tag{3}$$

For same-shape SD-VAE tuning, $\epsilon_T = \epsilon$ from Eq. (1). For cross-VAE tuning with different latent shapes, the two Gaussian draws are independent (Appendix A).

Feature guidance. Following Self-Transcendence (Sun et al., 2026), we enhance the teacher's intermediate features using CFG. Conditional and unconditional passes use the same teacher input:

$$\begin{aligned} h_c &= h_T(\tilde{z}_\tau, \tau, y), \qquad h_u = h_T(\tilde{z}_\tau, \tau, \varnothing), \\ h_{\text{cfg}} &= h_u + \omega(h_c - h_u), \end{aligned} \tag{4}$$

where h_T extracts features from a selected teacher layer and ω controls feature guidance. This CFG constructs training targets and is separate from CFG applied during image sampling. Our ablation shows that conditional teacher features alone ($\omega = 1$) do not improve downstream generation, whereas CFG-enhanced targets do (Table 3(c)).

Choosing the teacher timestep. Our aim is to obtain targets that improve the VAE as well as the student. In a preliminary shared-timestep run, PCA visualizations of the VAE latents show clearer

object structure at $1{,}650$ iterations, but this structure deteriorates by 100K iterations despite continued alignment, and the latents again resemble those of the teacher's VAE (Appendix D.1, Figure 4). We hypothesize that, when teacher and student noise levels match, alignment can be achieved mainly through updates to the student diffusion model. This provides little incentive for the VAE to learn a latent space better suited for generation. We therefore construct targets whose noise level does not track the student's timestep. Clean visual-encoder targets in REPA-E (Leng et al., 2025) also motivate using less noisy targets. We sample the teacher timestep independently from a low-noise interval, retaining the full range for the student. Unlike SRA (Jiang et al., 2025), this does not require lower teacher noise for every pair. Independently sampling the teacher timestep improves downstream generation even when both timesteps follow the same full-range distribution. Restricting the teacher to a low-noise interval provides a further improvement (Table 3(b)), supporting both components of our sampling strategy.

3.3 Updating the VAE through the Student

Let $h(z_t, t, y)$ denote the student features and P a trainable projection head. We align the projected features with the CFG-enhanced teacher target:

$$\mathcal{L}_{\text{align}} = \mathbb{E}\left[\frac{1}{N}\left\|P(h(z_t, t, y)) - h_{\text{cfg}}\right\|_2^2\right]. \tag{5}$$

Here N is the number of scalar feature entries per image, so the loss averages over both tokens and channels. The norm is taken over flattened features. The teacher target is fixed and receives no gradients. Since z_t depends on the trainable encoder, the alignment loss backpropagates through P and the student to the VAE encoder.

VAE regularization. To preserve reconstruction quality during tuning, we retain the VAE's standard regularization losses, denoted collectively by $\mathcal{L}_{\text{reg}}$: pixel reconstruction, LPIPS, adversarial, and KL terms, following REPA-E (Leng et al., 2025).

Training objectives and alternating updates. The two updates minimize separate sums of losses:

$$\begin{aligned}\mathcal{L}_{\text{VAE}} &= \mathcal{L}_{\text{reg}} + \lambda_V \mathcal{L}_{\text{align}},\\ \mathcal{L}_{\text{student}} &= \mathcal{L}_{\text{diff}} + \lambda_S \mathcal{L}_{\text{align}}.\end{aligned} \tag{6}$$

The VAE update optimizes E and D, holding f and P fixed while differentiating through their inputs. The student update optimizes f and P with detached VAE latents. The teacher remains frozen in both updates. The weights λ_V and λ_S control alignment in the respective objectives. Training details are in Appendix A.

3.4 Generation after VAE Tuning

After tuning, we freeze the VAE and its latent normalization. To assess the VAE independently of the jointly trained student, we train a new diffusion model from scratch using only $\mathcal{L}_{\text{diff}}$.

For our final generative model, we instead continue the existing student with $\mathcal{L}_{\text{diff}}$ alone. This retains the benefit of feature alignment during the initial stage, while sampling uses the trained diffusion model, inverse latent normalization, and the VAE decoder. Section 4.5 compares continuation with training a new feature-aligned student on the tuned VAE.

4 Experiments

We evaluate whether diffusion representations improve VAE latent spaces for generation, how the gains depend on model size, and how to build a generator from the tuned VAE and student.

Setup. We evaluate class-conditional generation on ImageNet at 256×256 resolution (Deng et al., 2009). We tune SD-VAE ($f8d4$) (Rombach et al., 2022) and IN-VAE ($f16d32$), the ImageNet-trained VAE provided by REPA-E (Leng et al., 2025). Both are baseline VAEs trained without DINO supervision. Here f denotes spatial downsampling and d the number of latent channels. In the frozen-VAE comparisons, we train a new SiT on each fixed VAE for 400K iterations using only the diffusion loss. Scaling increases teacher and student sizes together, from SiT-B to SiT-XL and SiT-2XL, while holding the downstream generator fixed.

Table 1: **VAE quality for downstream generation.** gFID of new generators trained on frozen VAEs. Both Ours VAEs use SiT-XL tuning with $\lambda_V = 0.05$ and an SD-VAE teacher. *: DINO supervision.

(a) SiT-B, $f8d4$

VAE	gFID↓
SD-VAE	37.02
EQ-VAE	31.20
Ours	22.53
REPA-E*	20.50

(b) SiT-B, $f16d32$

VAE	gFID↓
IN-VAE	39.31
Ours	19.30
REPA-E*	17.31
VA-VAE*	13.85

(c) SiT-XL, $f8d4$

VAE	gFID↓
SD-VAE	17.46
Ours	10.29

Our default schedule comprises 200K iterations of teacher preparation and 100K of VAE tuning, with teacher timesteps sampled from $(0, 0.3)$. Feature guidance $\omega = 30$ and student alignment $\lambda_S = 0.5$ follow Self-Transcendence (Sun et al., 2026). For the main runs, the VAE alignment weight is $\lambda_V = 0.25$ for SiT-B and 0.05 for SiT-XL. The smaller XL weight better preserves reconstruction (Appendix C). The scaling comparison instead holds $\lambda_V = 0.25$ fixed across model sizes. Full configurations are given in Appendix A. We evaluate 50K samples with the ADM evaluator (Dhariwal & Nichol, 2021) and 250-step Euler–Maruyama SDE sampling. Our sampling uses uniformly random class labels and no CFG except in the guided comparison. Differences in published sampling protocols are described in Appendix B. In the tables, * marks methods using DINO supervision. Baseline sources and evaluation details are given in Appendix B.

4.1 Improving VAE Quality for Downstream Generation

Table 1 evaluates the tuned VAEs with newly trained diffusion models. For SD-VAE, tuning lowers gFID from 37.02 to 22.53 with SiT-B and from 17.46 to 10.29 with SiT-XL. Because both generators are trained from scratch, these gains demonstrate the value of the resulting latent space independently of the jointly trained student. Our SD-VAE also improves on EQ-VAE, while DINO-supervised REPA-E remains stronger in the SiT-B comparison.

Table 2: **SD-VAE scaling.** SiT and BS describe tuning; evaluation uses a new SiT-B/2. Ours holds $\lambda_V = 0.25$ fixed.

Method	SiT	BS	gFID↓
Ours	B	256	23.82
Ours	XL	256	21.02
Ours	2XL	256	**20.70**
REPA-E*	B	32	20.91
REPA-E*	L	32	21.29
REPA-E*	XL	256	**20.50**

Generalization to IN-VAE. The method also improves IN-VAE, whose latents have greater spatial compression and more channels. To choose its teacher VAE, we compare SD-VAE and IN-VAE teachers under identical SiT-B tuning settings: SD-VAE gives gFID 21.06, versus 26.12 with IN-VAE. We therefore use SD-VAE as the teacher VAE for both latent formats (Appendix A).

With this teacher, our selected XL-tuned IN-VAE lowers downstream gFID from 39.31 to 19.30. A gap to REPA-E and VA-VAE remains, but the improvement shows that diffusion supervision is useful beyond SD-VAE. Both Ours checkpoints in Table 1 use $\lambda_V = 0.05$; reconstruction results for both latent formats are reported in Appendix C.

4.2 Scaling the Models Used for VAE Tuning

We next scale the teacher and student backbones used to tune SD-VAE, fixing the alignment weight, batch size, and downstream generator. gFID decreases consistently (Table 2), from 23.82 with SiT-B to 21.02 with SiT-XL and 20.70 with SiT-2XL, approaching REPA-E (20.50).

REPA-E does not exhibit the same monotonic trend in this frozen-VAE comparison. One possible explanation is the source of supervision: REPA-E uses the same DINO encoder across these runs, whereas our teacher grows together with the student. Increasing model size therefore also increases the capacity of the model supplying our alignment targets, which may make the supervision more effective. Separately, REPA and REPA-E report limited gains from increasing the DINO encoder size (Yu et al., 2025; Leng et al., 2025).

Table 3: **Effect of VAE tuning settings on downstream generation.** Color marks the varied setting. ω: feature-guidance weight for teacher targets; $\omega = 1$ uses conditional teacher features without guidance. VA: auxiliary velocity supervision; Std.: diffusion-only teacher. Iter. and BS refer to VAE tuning. PSNR/rFID: reconstruction by the tuned VAE.

λ_V	τ	ω	Teacher	Iter.	BS	PSNR↑	rFID↓	gFID↓	IS↑
Original SD-VAE (no tuning)						26.91	0.71	37.02	40.38
(a) Alignment weight									
0.1	$(0, 0.1)$	30	VA	100K	32	**25.42**	**0.85**	26.14	54.51
0.25	$(0, 0.1)$	30	VA	100K	32	25.13	1.00	25.15	55.83
0.5	$(0, 0.1)$	30	VA	100K	32	24.66	1.05	24.69	**56.79**
1	$(0, 0.1)$	30	VA	100K	32	24.37	1.11	**24.27**	56.53
2	$(0, 0.1)$	30	VA	100K	32	23.93	1.32	25.08	55.42
(b) Teacher timestep									
0.25	$\tau = t$	30	VA	100K	32	–	–	35.02	41.00
0.25	$(0, 1.0)$	30	VA	100K	32	–	–	27.68	51.25
0.25	$(0, 0.05)$	30	VA	100K	32	–	–	25.52	54.86
0.25	$(0, 0.1)$	30	VA	100K	32	–	–	25.15	55.83
0.25	$(0, 0.2)$	30	VA	100K	32	–	–	24.82	55.72
0.25	$(0, 0.3)$	30	VA	100K	32	–	–	**23.89**	**57.97**
0.25	$(0, 0.4)$	30	VA	100K	32	–	–	24.65	56.91
0.25	$(0.1, 0.3)$	30	VA	100K	32	–	–	24.86	56.65
(c) Feature guidance									
0.25	$(0, 0.3)$	1	VA	100K	32	–	–	37.27	40.39
0.25	$(0, 0.3)$	30	VA	100K	32	–	–	**23.89**	**57.97**
(d) Teacher preparation									
0.25	$(0, 0.3)$	30	Std.	100K	32	–	–	29.75	48.95
0.25	$(0, 0.3)$	30	VA	100K	32	–	–	**23.89**	**57.97**
(e) Tuning duration									
0.25	$(0, 0.3)$	30	VA	25K	32	–	–	29.11	50.97
0.25	$(0, 0.3)$	30	VA	100K	32	–	–	23.89	57.97
0.25	$(0, 0.3)$	30	VA	200K	32	–	–	**23.15**	**58.50**
(f) Batch size									
0.25	$(0, 0.3)$	30	VA	100K	32	25.07	0.97	23.89	**57.97**
0.25	$(0, 0.3)$	30	VA	100K	256	**25.20**	**0.81**	**23.82**	57.82

4.3 Component Analysis

Table 3 compares tuning choices using SiT-B and a common downstream evaluation.

Alignment weight. Increasing the alignment weight initially improves generation, but consistently degrades reconstruction, as shown by decreasing PSNR and increasing rFID (Table 3(a)). We choose $\lambda_V = 0.25$ for SiT-B to balance generation and reconstruction quality. For the main XL runs, we use 0.05: compared with 0.25, this weight better preserves reconstruction at the cost of a modest increase in downstream gFID (Appendix C).

Teacher timestep. Table 3(b) compares shared and independent timesteps. Shared-timestep alignment yields gFID 35.02. Sampling the teacher timestep independently from the same full range, $\tau \sim \mathcal{U}(0, 1)$, improves gFID to 27.68. Restricting the independent teacher timestep to $(0, 0.3)$ further improves it to 23.89, the best result among the tested intervals. These comparisons support both independent timestep sampling and the use of low-noise teacher targets.

Feature guidance. Using conditional teacher features without guidance ($\omega = 1$) yields gFID 37.27, providing no improvement over the original VAE (37.02). Increasing ω to 30 reduces gFID to 23.89 (Table 3(c)), demonstrating the importance of amplifying the class-conditional signal in the alignment targets.

Teacher preparation. Adding auxiliary velocity prediction during teacher preparation improves downstream gFID from 29.75 to 23.89 compared with diffusion-only preparation (Table 3(d)). We use this prepared teacher in the remaining experiments.

Tuning duration. Most gains occur between 25K and 100K iterations, with smaller gains at 200K (Table 3(e)); we therefore use 100K.

Batch size. Batch sizes 32 and 256 give similar gFID in the SiT-B configuration (Table 3(f)), but batch size 256 yields substantially better reconstruction. We use batch size 32 for the other ablations to reduce computational cost.

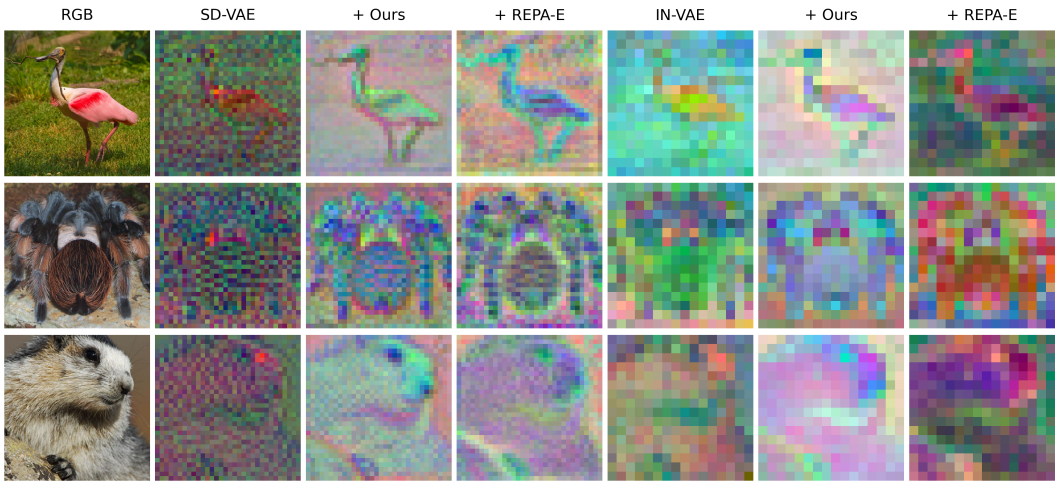


Figure 3: **PCA comparison of VAE latents.** Three examples compare the original SD-VAE and IN-VAE with GuaVAE and REPA-E. Both + *Ours* columns use SiT-XL tuning with $\lambda_V = 0.05$.

4.4 Latent Structure and Diffusion Features

We next examine how VAE tuning changes representations. Figure 3 shows more coherent object regions and less checkerboard-like variation after tuning, particularly for SD-VAE.

Semantic-region self-similarity (SRSS) measures similarity within object regions relative to their surroundings (Singh et al., 2025). Prior work links this spatial structure to downstream generation quality (Zhong et al., 2026). Table 4 shows that SRSS increases from 0.330 to 0.512 for SD-VAE and from 0.320 to 0.549 for IN-VAE, supporting the visual evidence of stronger spatial structure. Without DINO supervision, our tuning substantially narrows the SRSS gap to REPA-E on $f8d4$ and closes it on $f16d32$.

We also measure latent neighbor consistency (LNC) (Zhong et al., 2026) on SiT-B features rather than VAE latents. LNC@10 is the mean fraction of same-class neighbors among the ten nearest feature vectors. With a fixed layer and noise level, it rises from 8.23% to 13.42% for SD-VAE and from 10.35% to 14.04% for IN-VAE, reaching values close to REPA-E. Thus the spatial changes in VAE latents are accompanied by more class-consistent neighborhoods in the downstream generator. Appendix D specifies both protocols.

Table 4: **Representation analysis.** SRSS on VAE latents and LNC@10 (%) on SiT-B features.

VAE	SRSS↑	LNC@10↑
$f8d4$		
SD-VAE	0.330	8.23
Ours	0.512	13.42
REPA-E*	0.636	12.85
$f16d32$		
IN-VAE	0.320	10.35
Ours	0.549	14.04
REPA-E*	0.542	14.18

Table 5: **Student training.** Total counts updates across all stages.

Student	Teacher VAE	Total	gFID↓
New	Tuned	900K	17.17
New	SD-VAE	700K	17.83
Continued	–	600K	17.47

4.5 From VAE Tuning to a Generative Model

VAE tuning yields both a VAE and a diffusion student. Table 5 compares continuing this student with training a new Self-Transcendence student on the tuned VAE. Continuation gives nearly the best gFID with fewer training stages, so we freeze the VAE and continue the existing student using diffusion loss alone. Totals include teacher preparation and E2E; Appendix A gives the schedules.

Generation at different training durations. Table 6 compares generation after 80 epochs. Our procedure improves on SiT and Self-Transcendence with both B and XL backbones.

Table 6: **Short-run ImageNet-256 generation.** All generators are trained for 80 epochs (400K updates) and evaluated without guidance. –: not reported; sources in Appendix B.

(a) SiT-B, $f8d4$

Method	gFID↓	IS↑
SiT	37.02	40.38
Self-Transcendence	20.49	70.57
Ours	**17.47**	**75.48**

(b) SiT-XL, $f8d4$

Method	gFID↓	IS↑
SiT	17.46	75.53
Self-Transcendence	7.51	–
Ours	**6.79**	**131.14**

Table 7: **Class-conditional ImageNet-256 generation.** Epochs count generator training; 80 epochs corresponds to 400K updates. †: our evaluation of released weights; AG: AutoGuidance. Other guided entries use CFG. –: not reported. Sources and sampling protocols are in Appendix B.

			Without guidance		With guidance				
Method	Latents	Epochs	gFID↓	IS↑	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Without DINO supervision									
DiT-XL/2	$f8d4$	1400	9.62	121.50	2.27	4.60	278.24	0.83	0.57
SiT-XL/2	$f8d4$	1400	8.61	131.70	2.06	4.50	270.30	0.82	0.59
MaskDiT	$f8d4$	1600	5.69	177.99	2.28	5.67	276.56	0.80	0.61
U-ViT-H/2	$f8d4$	400	6.58	–	2.29	5.68	263.88	0.82	0.57
SRA	$f8d4$	400	–	–	1.85	4.50	297.20	0.82	0.61
SRA2	$f8d4$	400	7.00	–	1.84	4.50	306.50	0.82	0.61
LayerSync	$f8d4$	800	6.32	–	1.49	–	285.20	0.80	0.63
Self-Transcendence	$f8d4$	400	6.73†	149.27†	**1.44**	4.85	311.30	0.79	0.66
Ours	$f8d4$	400	4.97	160.41	1.62	4.96	293.73	0.795	0.638
Ours	$f16d32$	160	**4.36**	155.56	1.62	4.77	259.99	0.788	0.641
With DINO supervision									
REPA*	$f8d4$	800	5.90	157.80	1.42	4.70	305.70	0.80	0.65
REPA-E*	$f16d32$	800	1.83	217.30	**1.26**	–	314.90	–	–
VA-VAE*	$f16d32$	800	2.17	205.60	1.35	4.15	295.30	0.79	0.65
RAE*AG	$16^2 \times 768$	800	**1.60**	242.70	1.28	–	262.90	–	–

Table 7 compares complete generative systems at longer training durations. At 400 epochs, our SD-VAE system reaches gFID 4.97 without sampling guidance, compared with 6.73 for Self-Transcendence. It also improves on SRA2 at 400 epochs and LayerSync at 800 epochs, as well as DiT, SiT, and MaskDiT trained for substantially longer. Our IN-VAE system reaches unguided gFID 4.36 at 160 epochs, the lowest among the systems without DINO supervision in this table.

With CFG, our SD-VAE and IN-VAE systems achieve equal gFID 1.62. Both improve on the reported 400-epoch results of SRA and SRA2, while Self-Transcendence performs better. DINO-supervised systems retain an advantage: REPA improves diffusion training, while REPA-E, VA-VAE, and RAE also change the representation used for generation. We use random-label sampling results where available and explicitly mark RAE's AutoGuidance. These are system-level comparisons, complementing the controlled frozen-VAE experiments in Table 1.

5 Conclusion

We introduced GuaVAE, an end-to-end VAE tuning method that uses CFG-enhanced diffusion features as supervision without an external visual encoder. Independently sampling low-noise teacher timesteps improves the effectiveness of this supervision. Across VAEs with different latent shapes, the tuned latent spaces improve generation with diffusion models trained from scratch. Increasing the teacher and student sizes further improves the resulting VAEs, suggesting that stronger diffusion models can provide more effective supervision for tokenizer learning.

These results make diffusion representations a practical alternative to external visual features for improving VAE latent spaces. The tuning procedure produces both a reusable VAE and a diffusion model that can be continued with the diffusion objective alone. It therefore supports subsequent training of new generators while also providing a direct path to a competitive generative model through continuation of the existing model.

AI USE STATEMENT

We used generative AI tools to assist with polishing and revising the manuscript, including improving grammar, clarity, readability, and the organization of the text. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2209.12152.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2105.05233.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2012.09841.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. doi: 10.1109/TPAMI.2022.3218275.

Yasaman Haghighi, Bastien van Delft, Mariam Hassan, and Alexandre Alahi. LayerSync: Self-aligning intermediate layers. *arXiv preprint arXiv:2510.12581*, 2025. URL https://arxiv.org/abs/2510.12581.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.11239.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. URL https://arxiv.org/abs/2505.02831.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *International Conference on Machine Learning*, 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025. URL https://arxiv.org/abs/2504.10483.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

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*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé 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://arxiv.org/abs/2304.07193.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.09748.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/skorokhodov25a.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Ruibin Li, Yujing Sun, Shuaizheng Liu, and Lei Zhang. Beyond external guidance: Unleashing the semantic richness inside diffusion transformers for improved training. *arXiv preprint arXiv:2601.07773*, 2026. URL https://arxiv.org/abs/2601.07773.

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

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1711.00937.

Mengmeng Wang, Dengyang Jiang, Liuzhuozheng Li, Yucheng Lin, Guojiang Shen, Xiangjie Kong, Yong Liu, Guang Dai, and Jingdong Wang. SRA2: Variational autoencoder self-representation alignment for efficient diffusion training. *arXiv preprint arXiv:2601.17830*, 2026. URL https://arxiv.org/abs/2601.17830.

Xiyuan Wang and Muhan Zhang. Diffusion as self-distillation: End-to-end latent diffusion in one model. *arXiv preprint arXiv:2511.14716*, 2025. URL https://arxiv.org/abs/2511.14716.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. URL https://arxiv.org/abs/2501.01423.

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. URL https://arxiv.org/abs/2410.06940.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. URL https://arxiv.org/abs/2510.11690.

Guangting Zheng, Yiyuan Zhang, Tao Yang, Yunpeng Chen, Rui Zhu, Jiajun Deng, and Yanyong Zhang. GenFirst: Generation before reconstruction for stable end-to-end latent generative modeling. *arXiv preprint arXiv:2608.29335*, 2026. URL https://arxiv.org/abs/2608.29335.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2306.09305.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026. URL https://arxiv.org/abs/2606.03578.

Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. SD-DiT: Unleashing the power of self-supervised discrimination in diffusion transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2403.17004.