

# REPA-E: Unlocking VAE for End-to-End Tuning with Latent Diffusion Transformers

Xingjian Leng<sup>α\*</sup> Jaskirat Singh<sup>α\*</sup> Yunzhong Hou<sup>α</sup> Zhenchang Xing<sup>β</sup>  
Saining Xie<sup>γ</sup> Liang Zheng<sup>α</sup>

<sup>α</sup>Australian National University <sup>β</sup>Data61 CSIRO <sup>γ</sup>New York University

{xingjian.leng\*, jaskirat.singh\*, yunzhong.hou, liang.zheng}@anu.edu.au

zhenchang.xing@data61.csiro.au saining.xie@nyu.edu

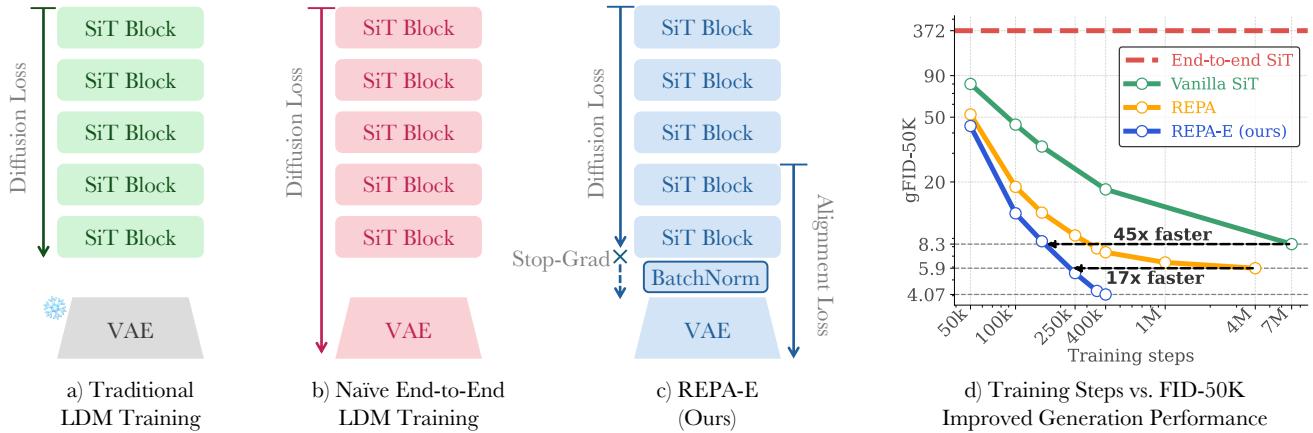


Figure 1. **Can we unlock VAE for end-to-end tuning with latent-diffusion models?** – Traditional deep learning wisdom dictates that end-to-end training is often preferable when possible. However, latent diffusion models usually only update the generator network while keeping the variational auto-encoder (VAE) fixed (a). This is because directly using the diffusion loss to update the VAE (b) causes the latent space to collapse. We show that while direct diffusion-loss is ineffective, end-to-end training can be unlocked through the representation-alignment (REPA) loss – allowing both encoder and diffusion model to be jointly tuned during the training process (c). Notably, this allows for significantly accelerated training; speeding up training by over 17 $\times$  and 45 $\times$  over REPA and vanilla training recipes, respectively (d).

## Abstract

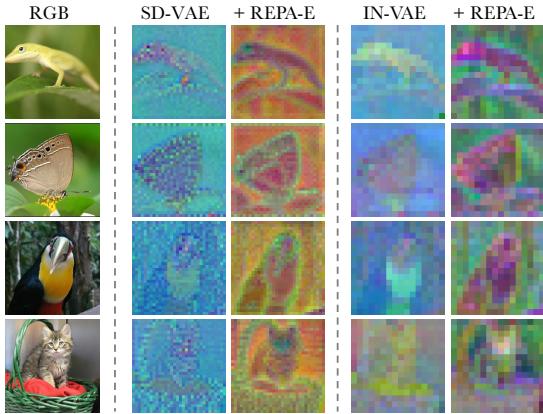
In this paper we tackle a fundamental question: “Can we train latent diffusion models together with the variational auto-encoder (VAE) tokenizer in an end-to-end manner?” Traditional deep-learning wisdom dictates that end-to-end training is often preferable when possible. However, for latent diffusion transformers, it is observed that end-to-end training both VAE and diffusion-model using standard diffusion-loss is ineffective, even causing a degradation in final performance. We show that while diffusion loss is ineffective, end-to-end training can be unlocked through the representation-alignment (REPA) loss – allowing both VAE and diffusion model to be jointly tuned during the training process. Despite its simplicity, the proposed training recipe (REPA-E) shows remarkable performance; speeding up diffusion model training by over 17 $\times$  and 45 $\times$  over REPA and

vanilla training recipes, respectively. Interestingly, we observe that end-to-end tuning with REPA-E also improves the VAE itself; leading to improved latent space structure and downstream generation performance. In terms of final performance, our approach sets a new state-of-the-art; achieving FID of 1.12 and 1.69 with and without classifier-free guidance on ImageNet 256  $\times$  256. Code is available at <https://end2end-diffusion.github.io>.

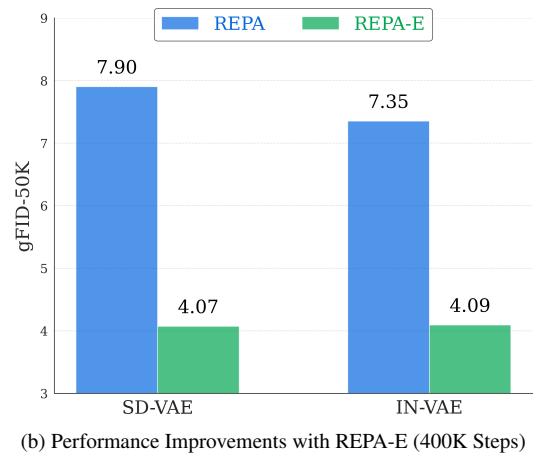
## 1. Introduction

End-to-end training has propelled the field forward for the past decade. It is understood that incorporating more components into end-to-end training can lead to increased performance, as evidenced by the evolution of the RCNN family [14, 15, 38]. With that said, training schemes of latent diffusion models (LDMs) [40] remain two-stage: first, the variational auto-encoder (VAE) [22] is trained with the re-

\* Equal Contribution.



(a) PCA Analysis on VAE Latent-Space Structure



**Figure 2. End-to-End Training Automatically Improves VAE Latent-Space Structure.** (a) Following [24], we visualize latent space structure from different VAEs before and after end-to-end training using principal component analysis (PCA) that projects them to three channels colored by RGB. We consider SD-VAE [40], and IN-VAE<sup>1</sup>, a 16× downsampling, 32-channel VAE trained on ImageNet [6]. For SD-VAE we find that latent representations have high-frequency noise. Applying end-to-end tuning helps learning a more smooth and less noisy latent representation. Interestingly to the contrast, the latent space for IN-VAE is over-smoothed (*e.g.*, row-2). Applying end-to-end tuning automatically helps learn a more detailed latent space structure to best support final generation performance. (b) Jointly tuning both VAE and latent diffusion model (LDM) significantly improves final generation performance (gFID) across different VAE architectures.

construction loss; then, the diffusion model is trained with the diffusion loss while keeping the VAE fixed (see Fig. 1a).

The above two-stage division of the LDM training process, though popular, leads to a challenging optimization task: “How to best optimize the representation from first stage (VAE) for optimal performance while training the second stage (diffusion model)?” While recent works study the interplay between the performance of the two stages [24, 44], they are often limited to empirical analysis, which may vary depending on the architecture and training setting for both the VAE and the diffusion model. For instance, in a concurrent work [44] show that the latent space of popular autoencoders *e.g.*, SD-VAE [40] suffer from high-frequency noise / components. However, as seen in Fig. 2 & 6, while the same holds for some VAEs (*e.g.* SD-VAE), it might not be true for other VAE architectures — which instead might suffer from an over-smoothed latent space (Fig. 2, 6).

In this paper, we therefore ask a fundamental question: “*Can we jointly tune both VAE and LDM in an end-to-end manner to best optimize final generation performance?*” Technically, it is straightforward to do end-to-end LDM training by simply back-propagating the diffusion loss to the VAE tokenizer. However, experiments (§3) reveal that this naive approach for end-to-end training is ineffective. The diffusion loss encourages learning a simpler latent space structure which is easier for denoising objective (refer §3.1), but leads to reduced generation performance (Fig. 1d).

To address this, we propose REPA-E; an end-to-end training recipe using representation alignment loss [54]. We show that while the diffusion loss is ineffective, end-to-end tuning can be unlocked through the recently proposed representation-alignment (REPA) loss - allowing both VAE

and diffusion model to be jointly tuned during training process. Through extensive evaluations, we demonstrate that end-to-end tuning with REPA-E offers several advantages;

**End-to-End Training Leads to Accelerated Generation Performance:** speeding up diffusion training by over 17× and 45× over REPA and vanilla training recipes (Fig. 1d). Furthermore, it also helps significantly improve the final generation performance. For instance as seen in Fig. 1d, we find that when using the popular SiT-XL [30] architecture, REPA-E reaches an FID of 4.07 within 400K steps, significantly boosting final performance over even REPA which only reaches a final FID for 5.9 after 4M steps [54].

**End-to-End Training improves VAE latent-space structure.** As seen in Fig. 2 and §4.4, we find that jointly tuning the VAE and latent diffusion model during training , automatically improves the latent space structure across different VAE architectures. For instance, for SD-VAE [40], it is observed that the original latent space suffers from high-frequency noise (Fig. 2). Applying end-to-end tuning helps learn a more smooth latent space representation. In contrast, the latent space for IN-VAE<sup>1</sup> is over-smoothed. Applying REPA-E automatically helps learn more detailed latent space structure to best support generation performance.

**End-to-End Tuning Improves VAE Performance.** Finally, we find that once tuned using REPA-E, the end-to-end tuned VAE can be used as a drop-in replacement for their original counterparts (*e.g.* SD-VAE) showing improved generation performance across diverse training settings and model architectures (refer §4.4).

To summarize, key contributions of this paper are: 1) We propose REPA-E; an end-to-end training recipe for jointly

<sup>1</sup>trained on imagenet at f16d32 using official training code from [40].

tuning both VAE and LDM using representation alignment loss (§3). 2) We find that despite its simplicity, REPA-E leads to accelerated generation performance; speeding up diffusion training by over  $17\times$  and  $45\times$  over REPA and vanilla training recipes, respectively (§4.2). 3) We show that end-to-end training is able to adaptively improve the latent space structure across diverse VAE architectures. 4) We demonstrate that once tuned using REPA-E, the end-to-end tuned VAE can be used as a drop-in replacement for their original counterparts (*e.g.*, SD-VAE), exhibiting significantly better downstream generation performance (§4.4).

## 2. Related Work

**Tokenizers** or autoencoders (AE) [3] use either the variational objective [22] for continuous tokenization or a vector quantization objective [9, 48] for discrete tokenization [8–10, 16, 21, 22, 36, 40, 48, 53, 55]. However, current tokenizers are primarily trained for minimizing the reconstruction error, which maybe not provide the optimal latent space for generation [24]. We show that improved latent space structure is achieved by end-to-end training of LDMs.

**Latent diffusion models** leverage pre-trained image tokenizers to compress images into a lower-dimensional latent space to simplify the generative task [5, 10, 10, 11, 26, 32, 36, 40, 43, 47]. Despite their effectiveness, existing tokenizers and diffusion models are trained separately [10, 36, 40]. In this paper, we explore jointly optimizing tokenizers and diffusion models to achieve faster convergence and improved generation performance (Sec. 4).

**Representation alignment for generative learning** has recently shown huge promise for improving the training speed and performance of diffusion models [35, 50, 54]. We find that instead of applying the REPA loss separately over LDM [54] or VAE [50], significantly better performance and training speed can be achieved through E2E training.

**End-to-End Diffusion.** LSGM [47] explores joint training with score-based generative models, which uses a variational lower bound objective with an entropy term for preventing latent space collapse while backpropagating the diffusion loss. We emperically find that while this helps prevent latent space collapse, REPA-E shows significantly faster convergence during E2E training (refer App. B).

## 3. REPA-E: Unlocking VAE for Joint Training

**Overview.** Given a variational autoencoder (VAE) and latent diffusion transformer (*e.g.*, SiT [30]), we wish to jointly tune the VAE latent representation and diffusion model features in an end-to-end manner to best optimize the final generation performance. To this end, we first make three key insights in §3.1: 1) Naive end-to-end tuning - directly back-propagating the diffusion loss to the VAE is ineffective. The diffusion loss encourages learning a more simpler latent space structure (Fig. 3a) which is easier for min-

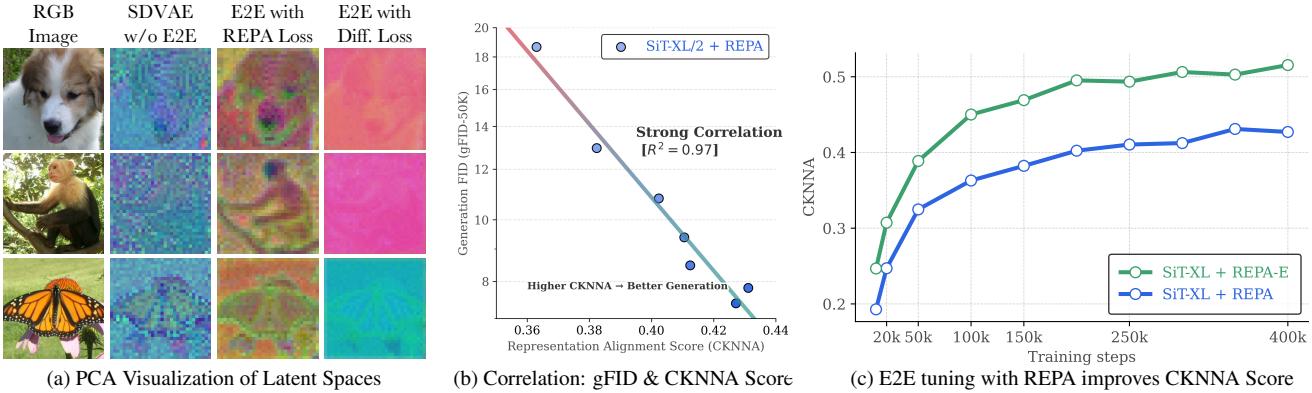
imizing the denoising objective [40], but degrades the final generation performance. We next analyze the recently proposed representation-alignment loss [54] showing that; 2) Higher representation-alignment score [54] correlates with improved generation performance (Fig. 3b). This offers an alternate path for improving final generation performance using representation-alignment score as a proxy. 3) The maximum achievable alignment score with vanilla-REPA is bottlenecked by the VAE latent space features. We further show that backpropagating the REPA loss to the VAE during training can help address this limitation, significantly improving final representation-alignment score (Fig. 3c).

Given the above insights, we finally propose REPA-E (§3.2); an end-to-end tuning recipe for both VAE and LDM features. Our key idea is simple: instead of directly using diffusion loss for end-to-end tuning, we can use the representation alignment score as a proxy for the final generation performance. This motivates our final approach, where instead of the diffusion loss, we propose to perform end-to-end training using the representation-alignment loss. The end-to-end training with REPA loss helps better improve the final representation-alignment score (Fig. 3b), which in turn leads to improved final generation performance (§3.1).

### 3.1. Motivating End-to-End Training with REPA

**Naive End-to-End Tuning is Ineffective.** We first analyze the naive approach for end-to-end tuning; directly backpropagating the diffusion loss to the VAE tokenizer. As shown in Fig. 3a, we observe that directly backpropagating the diffusion loss encourages learning a more simpler latent space structure with lower variance along the spatial dimensions (Tab. 10). The simpler latent-space structure poses an easier problem for the denoising objective [40], but leads to reduced generation performance (Fig. 1). Consider an intermediate latent  $z_t = \alpha_t z_{\text{VAE}} + \sigma_t \epsilon_{\text{orig}}$  for any timestep  $t$ . The denoising objective [34] mainly aims to predict  $\epsilon_{\text{pred}}$ ; estimating the originally added noise  $\epsilon_{\text{orig}}$  from VAE features  $z_{\text{VAE}}$  and timestep  $t$ . As the variance along the spatial dimensions for VAE latent  $z_{\text{VAE}}$  goes down, the denoising objective effectively reduces to predicting a bias term for recovering back the originally added noise  $\epsilon_{\text{orig}}$ . Thus, backpropagation the diffusion loss effectively hacks the latent space structure to create an easier denoising problem, but leads to a reduced generation performance (Fig. 1).

**Higher Representation Alignment Correlates with Better Generation Performance.** Similar to the findings of [54], we also measure representation alignment using CKNN scores [19] across different model sizes and training iterations. As seen in Fig. 3b, we observe that higher representation alignment during the training process leads to improved generation performance. This suggests an alternate path for improving generation performance by using the representation alignment objective instead of the diffusion loss for end-to-end training (refer §3.2).



**Figure 3. Motivating End-to-End Tuning using Representation Alignment (REPA) Loss.** We make three key insights: 1) Naive end-to-end (E2E) tuning using diffusion loss is ineffective. The diffusion encourages learning a more simpler latent space structure (a) which is easier for denoising objective (refer §3.1) but degrades final generation performance (Fig. 1). We next analyze the recently proposed representation alignment (REPA) loss [54] showing: 2) Higher representation alignment (CKNNA) leads to better generation performance. This suggests an alternate path for improving performance by using representation-alignment (CKNNA) as proxy for generation performance. 3) The maximum achievable CKNNA score with vanilla-REPA is bottlenecked by the VAE features (c) saturating around  $\sim 0.42$ . Back-propagating the REPA-loss to the VAE helps address this limitation and improve the final CKNNA score. Given the above insights: we propose REPA-E (§3.2) for end-to-end LDM training. The key idea is simple: instead of using the diffusion loss, we perform end-to-end training using the REPA loss. The end-to-end training with REPA loss helps improve the final representation-alignment (CKNNA), which in turn leads to improved generation performance (§4).

### Representation Alignment is Bottlenecked by the VAE Features.

Fig. 3c shows that while the naive application of REPA loss [54] leads to improved representation-alignment (CKNNA) score, the maximum achievable alignment score is still bottlenecked the VAE features saturating around a value of 0.4 (maximum value of 1). Furthermore, we find that backpropagating the representation-alignment loss to the VAE helps address this limitation; allowing end-to-end optimization of the VAE features to best support representation-alignment objective [54].

### 3.2. End-to-End Training with REPA

Given the above insights, we next propose REPA-E (§3.2); an end-to-end tuning recipe for jointly training both VAE and LDM features. Instead of directly using diffusion loss, we propose to perform end-to-end training using the representation-alignment loss. The end-to-end training with REPA loss helps better improve the final representation-alignment score (Fig. 3c), which in turn leads to improved final generation performance (refer §4.2). We next discuss key details for implementation of REPA-E for training.

#### Batch-Norm Layer for VAE Latent Normalization.

To enable end-to-end training, we first introduce a batch-norm layer between the VAE and latent diffusion model (Fig. 1). Typical LDM training involves normalizing the VAE features using precomputed latent statistics (*e.g.*,  $\text{std} = 1 / 0.1825$  for SD-VAE [40]). This helps normalize the VAE latent outputs to zero mean and unit variance for more efficient training for the diffusion model. However, with end-to-end training the statistics need to be recomputed whenever the VAE model is updated - which is expensive. To address this, we propose the use of a batch-

norm layer [20] which uses the exponential moving average (EMA) mean and variance as a surrogate for dataset-level statistics. The batch-norm layer thus acts as a differentiable normalization operator without the need for recomputing dataset level statistics after each optimization step.

**End-to-End Representation-Alignment Loss.** We next enable end-to-end training, by using the REPA loss [54] for updating the parameters for both VAE and LDM during training. Formally, let  $\mathcal{V}_\phi$  represent the VAE,  $\mathcal{D}_\theta$  be the diffusion model,  $f$  be the fixed pretrained perceptual model (*e.g.*, DINO-v2 [33]) for REPA [54] and  $x$  be a clean image. Also similar to REPA, consider  $h_\omega(\mathbf{h}_t)$  be the projection of diffusion transformer output  $\mathbf{h}_t$  through a trainable projection layer  $h_\omega$ . We then perform end-to-end training by applying the REPA loss over both LDM and VAE as,

$$\mathcal{L}_{\text{REPA}}(\theta, \phi, \omega) = -\mathbb{E}_{\mathbf{x}, \epsilon, t} \left[ \frac{1}{N} \sum_{n=1}^N \text{sim}(\mathbf{y}^{[n]}, h_\omega(\mathbf{h}_t^{[n]})) \right],$$

where  $\mathbf{y} = f(\mathbf{x})$  is the output of the pretrained perceptual model (*e.g.*, DINO-v2 [33]),  $N$  is number of patches,  $\text{sim}(\cdot, \cdot)$  computes the patch-wise cosine similarities between pretrained representation  $\mathbf{y}$  from perceptual model (*e.g.*, DINO-v2) and diffusion transformer hidden state  $\mathbf{h}_t$ .

**Diffusion Loss with Stop-Gradient.** As discussed in Fig. 3a and §3.1, backpropagating the diffusion loss to the VAE causes a degradation of latent-space structure. To avoid this, we introduce a simple stopgrad operation which limits the application of diffusion loss  $\mathcal{L}_{\text{DIFF}}$  to only the parameters  $\theta$  of the latent diffusion model  $\mathcal{D}_\theta$ .

**VAE Regularization Losses.** Finally, we introduce regularization losses  $\mathcal{L}_{\text{REG}}$  for VAE  $\mathcal{V}_\phi$ , to ensure that the end-to-end training process does not impact the reconstruction

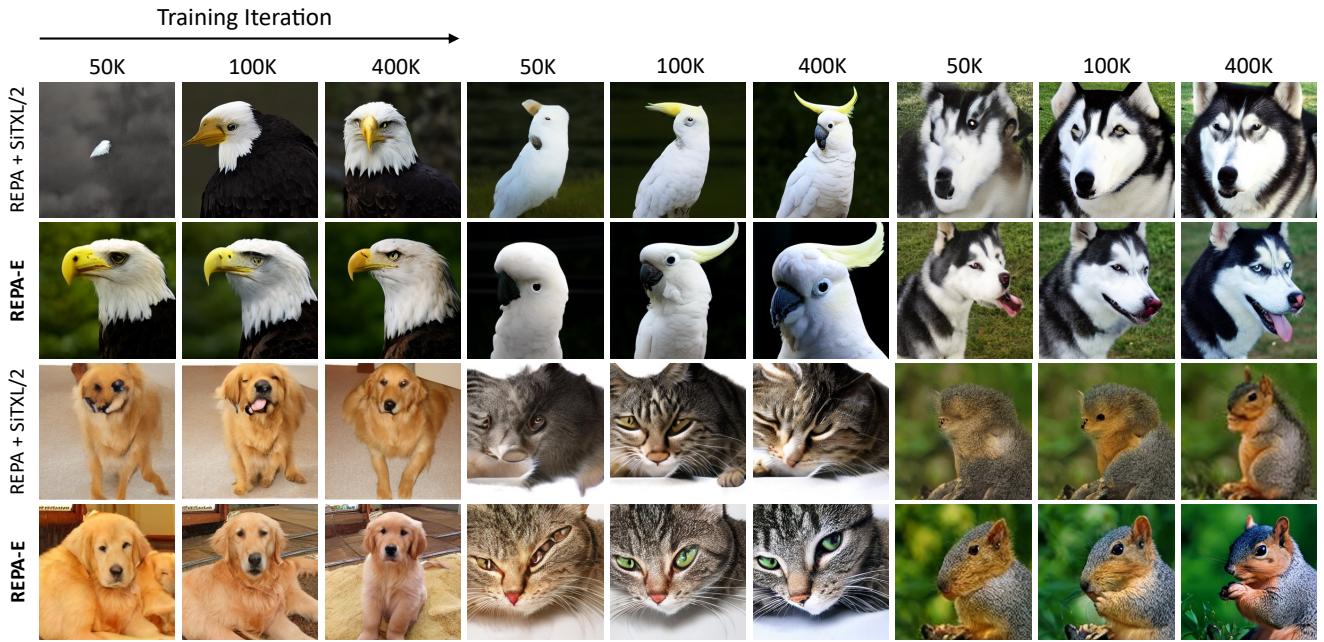


Figure 4. **End-to-End Tuning (REPA-E) Improves Visual Scaling.** We observe that REPA-E produces higher-quality images at 400K steps compared with the vanilla-REPA and generates more structurally meaningful images even in the early stages of training. Results for both methods are sampled using the same seed, noise and class label. We use a classifier-free guidance scale of 4.0 during sampling.

performance (rFID) of the original VAE. In particular, following [1], we use three losses, 1) Reconstruction Losses ( $\mathcal{L}_{\text{MSE}}, \mathcal{L}_{\text{LPIPS}}$ ), 2) GAN Loss ( $\mathcal{L}_{\text{GAN}}$ ), 3) KL divergence loss ( $\mathcal{L}_{\text{KL}}$ ) as regularization loss  $\mathcal{L}_{\text{REG}}$  for the VAE  $\mathcal{V}_\phi$ .

**Overall Training.** The overall training is then performed in an end-to-end manner using the following loss,

$$\mathcal{L}(\theta, \phi, \omega) = \mathcal{L}_{\text{DIFF}}(\theta) + \lambda \mathcal{L}_{\text{REPA}}(\theta, \phi, \omega) + \eta \mathcal{L}_{\text{REG}}(\phi),$$

where  $\theta, \phi, \omega$  refer to the parameters for the LDM, VAE and trainable REPA projection layer [54], respectively. Further implementation details are provided in §4.1 and Appendix.

## 4. Experiments

We next validate the performance of REPA-E and the effect of proposed components through extensive evaluation. In particular, we investigate three key research questions:

1. Can REPA-E significantly improve generation performance and training speed? (Sec. 4.2, Tab. 1, Fig. 1, 4)
2. Does REPA-E generalize across variations in training settings including model-scale, architecture, encoder model for REPA *etc.*? (Sec. 4.3, Tab. 2, 3, 4, 5, 6, 7)
3. Analyze the impact of end-to-end tuning (REPA-E) on VAE latent-space structure and downstream generation performance. (please refer Sec. 4.4, Fig. 6, Tab. 8, 9)

### 4.1. Setup

**Implementation Details.** We follow the same setup as in SiT [30] and REPA [54] unless otherwise specified. All training is conducted on the ImageNet [6] training split. We adopt the same data preprocessing protocol as

in ADM [7], where original images are center-cropped and resized to  $256 \times 256$  resolution. We experiment with publicly available VAEs, including SD-VAE (*f8d4*) [40], VA-VAE (*f16d32*) [40], and our own *f16d32* VAE trained on ImageNet, referred to as IN-VAE. Depending on the VAE downsampling rate, we adopt SiT-XL/1 and SiT-XL/2 for  $4\times$  and  $16\times$  downsampling rates, respectively, where 1 and 2 denote the patch sizes in the transformer embedding layer. We disable affine transformations in the BN [20] layer between the VAE and SiT, relying solely on the running mean and standard deviation. The VAE regularization loss combines multiple objectives and is defined as:  $\mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{GAN}}$ . For alignment loss, we use DINOv2 [33] as external visual features and apply alignment to the eighth layer of the SiT model. Empirically, we set the alignment loss coefficient to  $\lambda_{\text{REPA}_g} = 0.5$  for updating SiT and  $\lambda_{\text{REPA}_v} = 1.5$  for VAE. For optimization, we use AdamW [23, 29] with a constant learning rate of  $1 \times 10^{-4}$ , and a global batch size of 256. During training, we apply gradient clipping and exponential moving average (EMA) to the generative model for stable optimization. All experiments are conducted on 8 NVIDIA H100 GPUs.

**Evaluation.** For image generation evaluation, we strictly follow the ADM setup [7]. We report generation quality using Fréchet inception distance (gFID) [17], structural FID (sFID) [31], inception score (IS) [42], precision (Prec.) and recall (Rec.) [25], measured on 50K generated images. For sampling, we follow the approach in SiT [30] and REPA [54], using the SDE Euler-Maruyama sampler with 250 steps. In terms of VAE benchmark, we measure the reconstruction FID (rFID) on 50K images from the Im-



Figure 5. Qualitative Results on Imagenet  $256 \times 256$  using E2E-VAE and SiT-XL. We use a classifier-free guidance scale  $\alpha_{\text{cfg}} = 4.0$ .

Method	Tokenizer	Epochs	gFID $\downarrow$	sFID $\downarrow$	IS $\uparrow$
<b>Without End-to-End Tuning</b>					
MaskDiT [56]	SD-VAE	1600	5.69	10.34	<b>177.9</b>
DiT [34]		1400	9.62	6.85	121.5
SiT [30]		1400	8.61	6.32	131.7
FasterDiT [51]		400	7.91	5.45	131.3
REPA [54]	SD-VAE	20	19.40	6.06	67.4
		40	11.10	6.06	67.4
		80	7.90	5.06	122.6
		800	5.90	5.73	157.8
<b>With End-to-End Tuning (Ours)</b>					
REPA-E	SD-VAE*	20	12.83	5.04	88.8
		40	7.17	<b>4.39</b>	123.7
		80	<b>4.07</b>	<b>4.60</b>	<b>161.8</b>

Table 1. **REPA-E for Accelerated Generation Performance.** End-to-End training with REPA-E achieves significantly better performance (lower gFID) while using fewer epochs. Notably, REPA-E with only 80 epochs surpasses vanilla REPA using 10× epochs. \* indicates that VAE is updated during end-to-end training. All results are w/o classifier-free guidance on ImageNet  $256 \times 256$ . Additional system-level comparisons with classifier-free guidance and **state-of-the-art** results are provided in Tab. 9.

ageNet [6] validation set at a resolution of  $256 \times 256$ .

## 4.2. Impact on Training Performance and Speed

We first analyze the impact of end-to-end tuning using REPA-E (Sec. 3.2) for improving generation performance and speed when training latent-diffusion transformers.

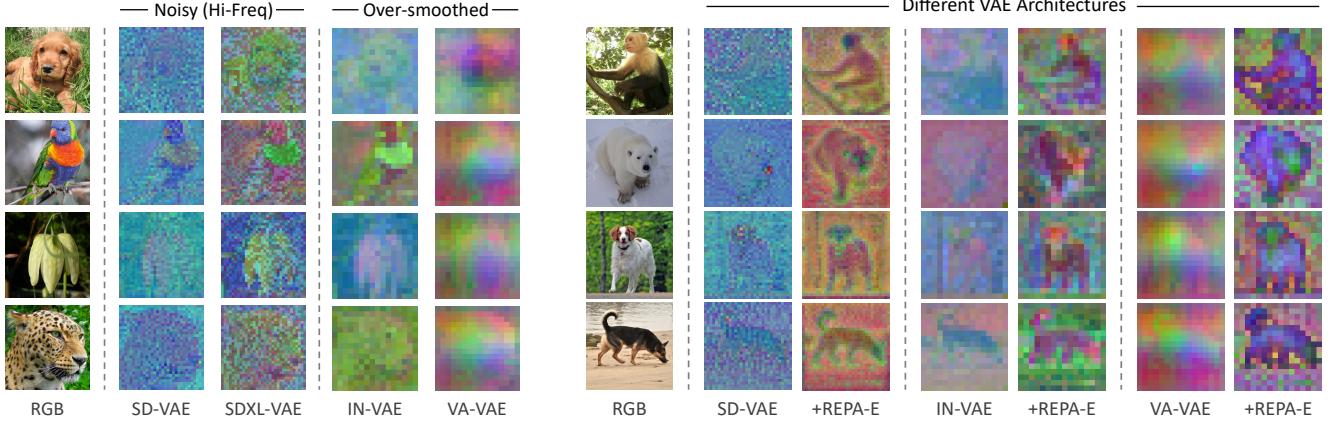
**Quantitative Evaluation.** We compare REPA-E against various latent diffusion model (LDM) baselines in Tab. 1. We evaluate models of similar sizes ( $\sim 675$ M parameters)

Diff. Model	gFID $\downarrow$	sFID $\downarrow$	IS $\uparrow$	Prec. $\uparrow$	Rec. $\uparrow$
SiT-B (130M)	49.5	7.00	27.5	0.46	<b>0.59</b>
<b>+REPA-E (Ours)</b>	<b>34.8</b>	<b>6.31</b>	<b>39.1</b>	<b>0.57</b>	<b>0.59</b>
SiT-L (458M)	24.1	6.25	55.7	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.3</b>	<b>5.69</b>	<b>75.0</b>	<b>0.68</b>	<b>0.60</b>
SiT-XL (675M)	19.4	6.06	67.4	0.64	<b>0.61</b>
<b>+REPA-E (Ours)</b>	<b>12.8</b>	<b>5.04</b>	<b>88.8</b>	<b>0.71</b>	0.58

Table 2. **Variation in Model-Scale.** We find that REPA-E brings substantial performance improvements across all model-scales. All baselines are reported using vanilla-REPA [54] for training.

on ImageNet  $256 \times 256$  generation task. All results are reported without classifier-free guidance [18] using popular SiT-XL [30] model for training. We make two observations; **1)** End-to-End tuning leads to faster training: consistently improving generation FID (gFID) from  $19.40 \rightarrow 12.83$  (20 epochs),  $11.10 \rightarrow 7.17$  (40 epochs), and  $7.90 \rightarrow 4.07$  (80 epochs), even when comparing with REPA [54]. **2)** End-to-End training leads to better final performance: REPA-E at 80 epochs surpasses FasterDiT [51] (gFID=7.91) trained for 400 epochs and even MaskDiT [56], DiT [34], and SiT [30] which are trained over 1400 epochs. For instance, REPA-E reaches an FID of 4.07 within 400K steps, significantly boosting final performance over even REPA which only reaches a final FID for 5.9 after 4M steps [54].

**Qualitative Evaluation.** We provide qualitative comparisons between REPA [54] and REPA-E in Fig. 4. We generate images from the same noise and label using checkpoints at 50K, 100K, and 400K training iterations, respectively. As seen in Fig. 4, we observe that REPA-E demonstrates superior image generation quality compared to the



(a) PCA Visualization of Latent Space Structure [24]

(b) Impact of End-to-End Tuning for Automatically Improving Latent Space Structure

**Figure 6. End-to-End Training Improves Latent Space Structure.** (a) We observe that the latent space of pretrained VAEs can suffer either high noise components (e.g., SDXL-VAE, SD-VAE [40]), or, be over-smoothed and lack details (e.g., VA-VAE [50]). (b) The use of end-to-end tuning (§3.2) automatically helps improve the latent space structure in a model-agnostic manner across different VAE architectures. For instance, similar to findings of concurrent work [44], we observe that SD-VAE suffers from high noise components in the latent space. Applying end-to-end training automatically helps adjust the latent space to reduce noise. In contrast, other VAEs such as recently proposed VA-VAE [50] suffer from an over-smoothed latent space. The use of end-to-end tuning with REPA-E automatically helps learn a more detailed latent-space structure to best support generation performance.

Target Repr.	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
I-JEPA-H [2]	23.0	5.81	60.3	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.5</b>	<b>5.18</b>	<b>73.6</b>	<b>0.68</b>	<b>0.60</b>
CLIP-L [37]	29.2	<b>5.98</b>	46.4	0.59	<b>0.61</b>
<b>+REPA-E (Ours)</b>	<b>23.4</b>	6.44	<b>57.1</b>	<b>0.62</b>	0.60
DINOv2-B [33]	24.1	6.25	55.7	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.3</b>	<b>5.69</b>	<b>75.0</b>	<b>0.68</b>	<b>0.60</b>
DINOv2-L [33]	23.3	5.89	59.9	0.61	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.0</b>	<b>5.59</b>	<b>77.7</b>	<b>0.68</b>	0.58

**Table 3. Variation in Representation Encoder.** REPA-E yields consistent performance improvements across different choices for the representation-encoder used for representation-alignment [54]. All baselines are reported using vanilla-REPA [54] for training.

Autoencoder	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
SD-VAE [40]	24.1	6.25	55.7	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.3</b>	<b>5.69</b>	<b>75.0</b>	<b>0.68</b>	<b>0.60</b>
IN-VAE ( <i>f16d32</i> )	22.7	<b>5.47</b>	56.0	0.62	<b>0.62</b>
<b>+REPA-E (Ours)</b>	<b>12.7</b>	5.57	<b>84.0</b>	<b>0.69</b>	<b>0.62</b>
VA-VAE [50]	12.8	6.47	83.8	0.71	0.58
<b>+REPA-E (Ours)</b>	<b>11.1</b>	<b>5.31</b>	<b>88.8</b>	<b>0.72</b>	<b>0.61</b>

**Table 4. Variation in VAE Architecture.** All baselines are reported using vanilla-REPA [54] for training.

REPA baseline, while also generating more structurally meaningful images during early stages of training process.

#### 4.3. Generalization and Scalability of REPA-E

We next analyze the generalization of the proposed approach to variation in training settings including model-size, tokenizer architecture, representation encoder, alignment depth [54] etc. Unless otherwise specified, all analysis and ablations use SiT-L [30] as the generative model,

Aln. Depth	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
6th layer	23.0	<b>5.72</b>	59.2	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.4</b>	6.64	<b>74.3</b>	<b>0.67</b>	0.59
8th layer	24.1	6.25	55.7	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.3</b>	<b>5.69</b>	<b>75.0</b>	<b>0.68</b>	<b>0.60</b>
10th layer	23.7	5.91	56.9	0.62	<b>0.60</b>
<b>+REPA-E (Ours)</b>	<b>16.2</b>	<b>5.22</b>	<b>74.7</b>	<b>0.68</b>	0.58

**Table 5. Variation in Alignment Depth.** End-to-End tuning (REPA-E) gives consistent performance improvements over original REPA [54] across varying alignment-depths.

Component	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
w/o stopgrad	444.1	460.3	1.49	0.00	0.00
w/o batch-norm	18.1	<b>5.32</b>	72.4	0.67	0.59
w/o $\mathcal{L}_{GAN}$	19.2	6.47	68.2	0.64	0.58
<b>REPA-E (Ours)</b>	<b>16.3</b>	<b>5.69</b>	<b>75.0</b>	<b>0.68</b>	<b>0.60</b>

**Table 6. Ablation Study on Role of Different Components.**

SD-VAE as the VAE, and DINOv2-B [33] as the pretrained vision model for REPA loss [54]. Default REPA alignment-depth of 8 is used. We train each variant for 100K iterations and report results without classifier-free guidance [18]. All baseline numbers are reported using vanilla REPA and compared with end-to-end training using REPA-E.

**Impact of Model Size.** Tab. 2 compares SiT-B, SiT-L, and SiT-XL to evaluate the effect of model size. We make two key observations. First, across all configurations, REPA-E consistently improves performance over the REPA baseline. Specifically, it reduces gFID from 49.5 → 34.8 for SiT-B, 24.1 → 16.3 for SiT-L, and 19.4 → 12.8 for SiT-XL, demonstrating the effectiveness. Second, surprisingly the percentage gains in gFID achieved with REPA-E (over REPA) improve with increasing model size. For in-

Method	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
<b>100K Iterations (20 Epochs)</b>					
REPA [54]	19.40	6.06	67.4	0.64	<b>0.61</b>
<b>REPA-E (scratch)</b>	14.12	7.87	83.5	0.70	0.59
<b>REPA-E (VAE init.)</b>	<b>12.83</b>	<b>5.04</b>	<b>88.8</b>	<b>0.71</b>	0.58
<b>200K Iterations (40 Epochs)</b>					
REPA [54]	11.10	5.05	100.4	0.69	<b>0.64</b>
<b>REPA-E (scratch)</b>	7.54	6.17	120.4	<b>0.74</b>	0.61
<b>REPA-E (VAE init.)</b>	<b>7.17</b>	<b>4.39</b>	<b>123.7</b>	<b>0.74</b>	0.62
<b>400K Iterations (80 Epochs)</b>					
REPA [54]	7.90	5.06	122.6	0.70	<b>0.65</b>
<b>REPA-E (scratch)</b>	4.34	<b>4.44</b>	154.3	0.75	0.63
<b>REPA-E (VAE init.)</b>	<b>4.07</b>	4.60	<b>161.8</b>	<b>0.76</b>	0.62

**Table 7. End-to-End Training from Scratch.** We find that while initializing the VAE with pretrained weights (SD-VAE [40]) helps slightly improve performance, REPA-E can be used to train both VAE and LDM from scratch in an end-to-end manner; still achieving significantly superior performance over REPA which requires a separate stage for training VAE in addition to LDM training.

stance, for SiT-B model REPA-E leads to a 29.6% improvement in gFID over REPA. Surprisingly even more gains are achieved for bigger models improving gFID by 32.3% and 34.0% for SiT-L and SiT-XL models respectively. This trend highlights the scalability of REPA-E; larger models achieve better percentage gains over vanilla-REPA.

**Variation in Representation Encoder.** We report results across different perception model encoders (CLIP-L, I-JEPA-H, DINOv2-B, and DINOv2-L) Tab. 3. We observe that REPA-E gives consistent performance improvements over REPA, across different choices of the perceptual encoder model. In particular, with DINOv2-B and DINOv2-L, REPA-E significantly reduces gFID from  $24.1 \rightarrow 16.3$  and from  $23.3 \rightarrow 16.0$ , respectively.

**Variation in VAE.** Tab. 4 evaluates the impact of different VAEs on REPA-E performance. In particular, we report results using three different VAEs 1) SD-VAE [1], 2) VA-VAE [50] and 3) IN-VAE (a  $16 \times$  downsampling, 32-channel VAE trained on ImageNet [6] using official training code from [40]). Across all variations, REPA-E consistently improves performance over the REPA baseline. REPA-E reduces gFID from  $24.1 \rightarrow 16.3$ , from  $22.7 \rightarrow 12.7$ , and  $12.8 \rightarrow 11.1$ , for SD-VAE, IN-VAE and VA-VAE, respectively. The results demonstrate that REPA-E robustly improves generative quality across diverse variations in architecture, pretraining dataset and training setting of the VAE.

**Variation in Alignment Depth.** Tab. 5 investigates the effect of applying the alignment loss at different layers the diffusion model. We observe that REPA-E consistently enhances generation quality over the REPA baseline across variation in choice of alignment depth; with gFID improving from  $23.0 \rightarrow 16.4$  (6th layer),  $24.1 \rightarrow 16.3$  (8th layer), and  $23.7 \rightarrow 16.2$  (10th layer).

VAE	Diffusion model	REPA	gFID-50K
SD-VAE [40]	DiT-XL [34]	✗	19.82
VA-VAE [50]	DiT-XL [34]	✗	<b>6.74</b>
<b>E2E-VAE (Ours)</b>	DiT-XL [34]	✗	6.75
SD-VAE [40]	SiT-XL [30]	✗	17.20
VA-VAE [50]	SiT-XL [30]	✗	5.93
<b>E2E-VAE (Ours)</b>	SiT-XL [30]	✗	<b>5.26</b>
SD-VAE [40]	DiT-XL [34]	✓	12.29
VA-VAE [50]	DiT-XL [34]	✓	4.71
<b>E2E-VAE (Ours)</b>	DiT-XL [34]	✓	<b>4.20</b>
SD-VAE [40]	SiT-XL [30]	✓	7.90
VA-VAE [50]	SiT-XL [30]	✓	4.88
<b>E2E-VAE (Ours)</b>	SiT-XL [30]	✓	<b>3.46</b>

**Table 8. Impact of End-to-End Tuning on VAE Performance.** We find that once tuned using REPA-E, the finetuned VAEs can be used as a drop-in replacement for their original counterparts offering significantly accelerated generation performance. We fix all the VAEs and only train the diffusion models (with and w/o REPA). E2E-VAE is obtained from REPA-E fine-tuning (VA-VAE + SiT-XL). All results are reported at 80 epochs (400K iterations).

**Ablation on Design Components.** We also perform ablation studies analyzing the importance of each component discussed in Sec. 3.2. Results are shown in Tab. 6. We observe that each component plays a key role in the final performance for REPA-E. In particular, we observe that the `stop-grad` operation on the diffusion loss helps prevent degradation of the latent-space structure. Similarly, the use of batch norm is useful adaptively normalizing the latent-statistics and helps improve the gFID from  $18.09 \rightarrow 16.3$ . Similarly, the regularization losses play a key role in maintaining the reconstruction performance of the finetuned VAE, thereby improving the gFID from  $19.07 \rightarrow 16.3$ .

**End-to-End Training from Scratch.** We next analyze the impact of VAE initialization on end-to-end training. As shown in Tab. 7, we find that while initializing the VAE from pretrained weights helps slightly improve performance, REPA-E can be used to train both VAE and LDM from scratch still achieving superior performance over REPA, which technically requires a separate stage for VAE training in addition to LDM training. For instance, while REPA achieves a FID of 5.90 after 4M iterations, REPA-E while training entirely from scratch (for both VAE and LDM) achieves much faster and better generation FID of 4.34 within just 400K iterations.

#### 4.4. Impact of End-to-End Tuning on VAE

We next analyze the impact of end-to-end tuning on the VAE. In particular, we first show that end-to-end tuning improves the latent-space structure (Fig. 6). We next show that once tuned using REPA-E, the finetuned VAEs can be used as a drop-in replacement for their original counterparts offering significantly improved generation performance.

**End-to-End Training improves Latent Space Structure.** Results are shown in Fig. 6. Following [24], we visu-

Tokenizer	Method	Training Epochs	#params	rFID $\downarrow$	Generation w/o CFG						Generation w/ CFG						
					gFID $\downarrow$	sFID $\downarrow$	IS $\uparrow$	Prec. $\uparrow$	Rec. $\uparrow$	gFID $\downarrow$	sFID $\downarrow$	IS $\uparrow$	Prec. $\uparrow$	Rec. $\uparrow$			
<b>AutoRegressive (AR)</b>																	
MaskGiT	MaskGIT [4]	555	227M	2.28	6.18	-	182.1	0.80	0.51	-	-	-	-	-	-	-	-
VQGAN	LlamaGen [45]	300	3.1B	0.59	9.38	8.24	112.9	0.69	0.67	2.18	5.97	263.3	0.81	0.58			
VQVAE	VAR [46]	350	2.0B	-	-	-	-	-	-	1.80	-	365.4	0.83	0.57			
LFQ tokenizers	MagViT-v2 [52]	1080	307M	1.50	3.65	-	200.5	-	-	1.78	-	319.4	-	-			
LDM	MAR [27]	800	945M	0.53	2.35	-	227.8	0.79	0.62	1.55	-	303.7	0.81	0.62			
<b>Latent Diffusion Models (LDM)</b>																	
SD-VAE [40]	MaskDiT [56]	1600	675M	0.61	5.69	10.34	177.9	0.74	0.60	2.28	5.67	276.6	0.80	0.61			
	DiT [34]	1400	675M		9.62	6.85	121.5	0.67	0.67	2.27	4.60	278.2	<b>0.83</b>	0.57			
	SiT [30]	1400	675M		8.61	6.32	131.7	0.68	0.67	2.06	4.50	270.3	0.82	0.59			
	FasterDiT [51]	400	675M		7.91	5.45	131.3	0.67	<b>0.69</b>	2.03	4.63	264.0	0.81	0.60			
	MDT [12]	1300	675M		6.23	5.23	143.0	0.71	0.65	1.79	4.57	283.0	0.81	0.61			
	MDTv2 [13]	1080	675M		-	-	-	-	-	1.58	4.52	314.7	0.79	0.65			
<b>Representation Alignment Methods</b>																	
VA-VAE [50]	LightningDiT [50]	80	675M	<b>0.28</b>	4.29	-	-	-	-	1.25	4.15	295.3	0.80	0.65			
		800	675M		2.05	4.37	207.7	0.77	0.66								
SD-VAE	REPA [54]	80	675M	0.61	7.90	5.06	122.6	0.70	0.65	-	-	-	-	-			
		800	675M		5.84	5.79	158.7	0.70	<b>0.68</b>	1.28	4.68	<b>305.7</b>	0.79	0.64			
E2E-VAE (Ours)	REPA	80	675M	<b>0.28</b>	3.46	<b>4.17</b>	159.8	0.77	0.63	1.67	4.12	266.3	0.80	0.63			
		800	675M		<b>1.69</b>	<b>4.17</b>	<b>219.3</b>	<b>0.77</b>	0.67	<b>1.12</b>	<b>4.09</b>	302.9	0.79	<b>0.66</b>			

Table 9. **System-Level Performance on ImageNet  $256 \times 256$**  comparing our end-to-end tuned VAE (E2E-VAE) with other VAEs for traditional LDM training. Note that all representation alignment methods at 800 epochs are evaluated using a class-balanced sampling protocol, as detailed in App. C. We observe that in addition to improving VAE latent space structure (Fig. 6), end-to-end tuning significantly improves VAE downstream generation performance. Once tuned using REPA-E, the improved VAE can be used as drop-in replacement for their original counterparts for accelerated generation performance. Overall, our approach helps improve both LDM and VAE performance — achieving a new *state-of-the-art* FID of 1.12 and 0.28, respectively for LDM generation and VAE reconstruction performance.

alize latent space structure using principal component analysis (PCA) that projects them to three channels colored by RGB. We consider three different VAEs: 1) SD-VAE [40], 2) IN-VAE (a  $16 \times$  downampling, 32-channel VAE trained on ImageNet [6]). 3) VA-VAE from recent work from [50]. We observe that end-to-end tuning using REPA-E automatically improves the latent space structure of the original VAE. For instance, similar to findings of concurrent work [44], we observe that SD-VAE suffers from high noise components in the latent space. Applying end-to-end training automatically helps adjust the latent space to learn reduce noise. In contrast, other VAEs such as recently proposed VA-VAE [50] suffer from over-smoother latent space. Application of E2E tuning automatically helps learn a more detailed latent-space to best support generation performance.

#### End-to-End Training Improves VAE Performance.

We next evaluate the impact of end-to-end tuning on downstream generation performance of the VAE. To this end, we first use end-to-end tuning for finetuning the recently proposed VA-VAE [50]. We then use the resulting end-to-end finetuned-VAE (named E2E-VAE), and compare its downstream generation performance with current state-of-the-art VAEs; including SDVAE [40] and VA-VAE [50]. To do this, we conduct traditional latent diffusion model training (w/o REPA-E), where only the generator network is updated while keeping the VAE frozen. Tab. 8 shows the comparison of VAE downstream generation across diverse train-

ing settings. We observe that end-to-end tuned VAEs consistently outperform their original counterparts for downstream generation tasks across variations in LDM architecture and training settings. Interestingly, we observe that a VAE tuned using SiT-XL yields performance improvements even when using a different LDM architecture such as DiT-XL; thereby demonstrating the robustness of our approach.

## 5. Conclusion

*“Can we unlock VAE’s for performing end-to-end training with latent diffusion transformers?”* Directly backpropagating diffusion loss to the VAE is ineffective and even degrades final performance. We show that while diffusion loss is ineffective, end-to-end training can be unlocked using REPA loss. Our end-to-end training recipe (REPA-E), significantly improves latent-space structure, shows remarkable performance; speeding up diffusion model training by over  $17 \times$  and  $45 \times$  over REPA and vanilla training recipes. Overall, our approach achieves a new state-of-the-art results with generation FID of 1.12 and 1.69 with and without use of classifier-free guidance. We hope that our work can help foster further research for enabling end-to-end training with latent diffusion transformers.

## Acknowledgments

We would like to extend our deepest appreciation to Zeyu Zhang, Qinyu Zhao, and Zhanhao Liang for insightful discussions. We would also like to thank all reviewers for their constructive feedback. This work was supported in part by the Australian Research Council under Discovery Project DP210102801 and Future Fellowship FT240100820. SX acknowledges support from the OpenPath AI Foundation, IITP grant funded by the Korean Government (MSIT) (No. RS-2024-00457882) and NSF Award IIS-2443404.

## References

- [1] Stability AI. Improved autoencoders ... <https://huggingface.co/stabilityai/sd-vae-ft-mse>, n.d. Accessed: April 11, 2025. 5, 8
- [2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 7
- [3] Dana H Ballard. Modular learning in neural networks. In *Proceedings of the sixth National conference on Artificial intelligence-Volume 1*, pages 279–284, 1987. 3
- [4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 9
- [5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3
- [6] 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*, pages 248–255. Ieee, 2009. 2, 5, 6, 8, 9, 13
- [7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 5
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [9] 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*, pages 12873–12883, 2021. 3
- [10] 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. 3, 13
- [11] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 3
- [12] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23164–23173, 2023. 9
- [13] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 9
- [14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1
- [15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5
- [18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6, 7
- [19] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024. 3
- [20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 4, 5
- [21] Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. *arXiv preprint arXiv:2501.07730*, 2025. 3
- [22] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [24] Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509*, 2025. 2, 3, 7, 8

- [25] Tuomas Kynkänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 5
- [26] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 3
- [27] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2025. 9, 14
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 13
- [29] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [30] 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*, pages 23–40. Springer, 2024. 2, 3, 5, 6, 7, 8, 9, 13
- [31] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *International Conference on Machine Learning*, pages 7958–7968. PMLR, 2021. 5
- [32] OpenAI. Sora. <https://openai.com/sora>, 2024. 3
- [33] Maxime Oquab, Timothée Darcret, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 4, 5, 7
- [34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 6, 8, 9
- [35] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 3
- [36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7
- [38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1
- [39] Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. *arXiv preprint arXiv:2502.20388*, 2025. 14
- [40] 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*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 7, 8, 9, 13, 14
- [41] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 13
- [42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 5
- [43] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. *arXiv preprint arXiv:2211.17084*, 2022. 3
- [44] Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025. 2, 7, 9
- [45] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 9
- [46] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 9
- [47] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, pages 11287–11302. Curran Associates, Inc., 2021. 3, 13
- [48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3
- [49] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 14
- [50] Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. 3, 7, 8, 9, 14
- [51] Jingfeng Yao, Wang Cheng, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *arXiv preprint arXiv:2410.10356*, 2024. 6, 9
- [52] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Vervari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 9
- [53] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2025. 3

- [54] 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. *arXiv preprint arXiv:2410.06940*, 2024. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [13](#)
- [55] Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. *arXiv preprint arXiv:2412.05796*, 2024. [3](#)
- [56] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. [6](#), [9](#)

# REPA-E: Unlocking VAE for End-to-End Tuning with Latent Diffusion Transformers

## Supplementary Material

Training Strategy	Spatial Variance	Total Variation
w/o E2E Tuning	17.06	6627.35
E2E w/ REPA Loss	18.02	5516.14
E2E w/ Diff. Loss	0.02	89.80

Table 10. **Impact of Naive End-to-End Training with Diffusion Loss.** We report total variation [41] and mean variance along each VAE latent channel for three training settings: 1) Standard LDM training (w/o end-to-end (E2E) tuning), 2) Naive E2E tuning with Diffusion loss, 3) E2E tuning with REPA loss [54]. All experiments use SDVAE for VAE initialization. We observe that using diffusion loss for end-to-end tuning encourages learning a simpler latent space with lower variance along the spatial dimensions (Fig. 3a). The simpler latent space is easier for denoising objective (§3.1), but degrades final generation performance (Fig. 1). All results are reported at 400K iterations with SiT-XL/2 [30] as LDM.

### A. Impact of Diffusion Loss on Latent Space

We analyze the effect of naively using diffusion loss for end-to-end tuning, focusing on how it alters the latent space structure. All experiments here use SD-VAE for tokenizer initialization and SiT-XL/2 [30] as the latent diffusion model, trained for 400K iterations without classifier-free guidance. We report two metrics to quantify latent structure, 1) **Spatial Variance**, computed as the mean per-channel variance across spatial dimensions, and 2) **Total Variation** [41], which captures local spatial differences in the latent map.

As shown in Tab. 10 and Fig. 3, directly backpropagating the diffusion loss leads to reduced spatial variance, which creates an easier denoising problem by hacking the latent space but leads to reduced image generation performance. In contrast, end-to-end training with REPA-E not only leads to improved generation performance but also improves the latent space structure for the underlying VAE (Fig. 3, 6).

### B. Additional Analysis

Method	gFID ↓	sFID ↓	IS ↑	Prec. ↑	Rec. ↑
REPA + E2E-Diffusion	444.1	460.3	1.49	0.00	0.00
REPA + E2E-LSGM	9.89	5.07	107.5	0.72	0.61
<b>REPA-E (Ours)</b>	<b>4.07</b>	<b>4.60</b>	<b>161.8</b>	<b>0.76</b>	<b>0.62</b>

Table 11. **Comparison with LSGM Objective.** REPA-E shows better generation performance and convergence speed.

**Comparison of End-to-End Training Objectives.** We provide additional results comparing different objectives for

Method	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
REPA + SiT-L	22.2	5.68	58.3	0.74	0.60
<b>REPA-E + SiT-L</b>	<b>12.8</b>	<b>4.60</b>	<b>90.6</b>	<b>0.79</b>	<b>0.61</b>

Table 12. **Scaling REPA-E to Higher Resolution.** System-level results on ImageNet-512 with  $64 \times 64$  latents using SiT-L at 100K steps without classifier-free guidance. We observe that REPA-E leads to significant performance improvements over vanilla-REPA [54] even at high resolutions.

Sampler	ODE, NFE=50		SDE, NFE=250	
	VA-VAE	E2E-VAE	VA-VAE	E2E-VAE
<b>gFID</b>	5.43	<b>5.02</b>	5.57	<b>4.97</b>

Table 13. **Generalization to T2I Tasks.** FID results on MSCOCO text-to-image generation using MMDiT + REPA. We find that end-to-end tuned VAEs (E2E-VAE) also generalizes to T2I tasks showing improved generation performance.

end-to-end training of VAE and LDM. Specifically, we evaluate: 1) naive E2E training by backpropagating diffusion loss to VAE encoder, 2) the LSGM entropy-regularized objective [47], 3) our proposed REPA-E. All methods are trained with SiT-XL for 400K steps under consistent settings.

The LSGM objective prevents feature collapse by maximizing entropy of the latent space. However, as shown in Tab. 11, our REPA-E formulation yields better performance across all metrics at just 400K steps, with significantly faster convergence and stronger generation quality.

**Scaling REPA-E to Higher Latent Resolution.** We conduct experiments on ImageNet-512 [6] to evaluate the performance of REPA-E under higher-resolution latent settings ( $64 \times 64$ ). We use SD-VAE [40] as the tokenizer and SiT-L as the diffusion model, trained for 100K steps and we report the performance without classifier-free guidance. As shown in Tab. 12, our approach yields significant improvements in generation quality compared to REPA.

#### MSCOCO Text-to-Image Generation with E2E-VAE.

To further evaluate the utility of the tuned VAE beyond ImageNet, we assess its performance in a text-to-image generation (T2I) setting on MSCOCO [28]. Following REPA [54], we adopt MMDiT [10] as the diffusion backbone and apply REPA loss across all variants. All models are trained for 100K steps and evaluated using classifier-free guidance with  $\alpha_{\text{cfg}} = 2.0$  and EMA weights during inference. We report generation FID, and observe that replacing VA-VAE with our E2E-VAE consistently improves downstream text-to-image generation quality (Tab. 13).

<b>Autoencoder</b>	<b>PSNR<math>\uparrow</math></b>	<b>SSIM<math>\uparrow</math></b>	<b>LPIPS<math>\downarrow</math></b>	<b>rFID<math>\downarrow</math></b>
SD-VAE [40]	25.67	0.72	0.13	0.74
<b>+REPA-E (Ours)</b>	<b>24.84</b>	<b>0.71</b>	<b>0.15</b>	<b>0.53</b>
IN-VAE ( <i>f16d32</i> )	27.40	0.80	0.09	0.26
<b>+REPA-E (Ours)</b>	<b>26.87</b>	<b>0.78</b>	<b>0.11</b>	<b>0.27</b>
VA-VAE [50]	26.32	0.76	0.11	0.28
<b>+REPA-E (Ours)</b>	<b>26.25</b>	<b>0.75</b>	<b>0.11</b>	<b>0.28</b>

Table 14. **VAE Reconstruction Evaluation on ImageNet-256.**

While REPA-E primarily improves the generative capability of the VAE (see Tab. 9), it also maintains competitive reconstruction quality across all metrics.

### C. Remarks on FID Evaluation

Throughout the paper, we follow the standard ImageNet conditional evaluation protocol, where 50,000 images are generated by randomly sampling class labels. Recent papers [27, 39, 49] have adopted class-balanced generation for evaluation, where 50 images per class are generated across the 1,000 categories. To our surprise, we found that using class-balanced sampling yields slightly better FID performance. Therefore, for the results in Tab. 9, we adopt this class-balanced sampling strategy. Accordingly, all representation alignment methods at the 800-epoch checkpoint in this table are computed under the balanced sampling protocol to ensure a fair and consistent comparison.