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

Xingjian Leng Jaskirat Singh Yunzhong Hou
Australian National University

Zhenchang Xing Saining Xie Liang Zheng
Data61 CSIRO New York University Australian National University

2025.5.25

Presenter:唐果

目录

北京大学 PEKING UNIVERSITY

- Author
- Background
- Method
- Experiments

02 Background

Background

北京大学 PEKING UNIVERSITY

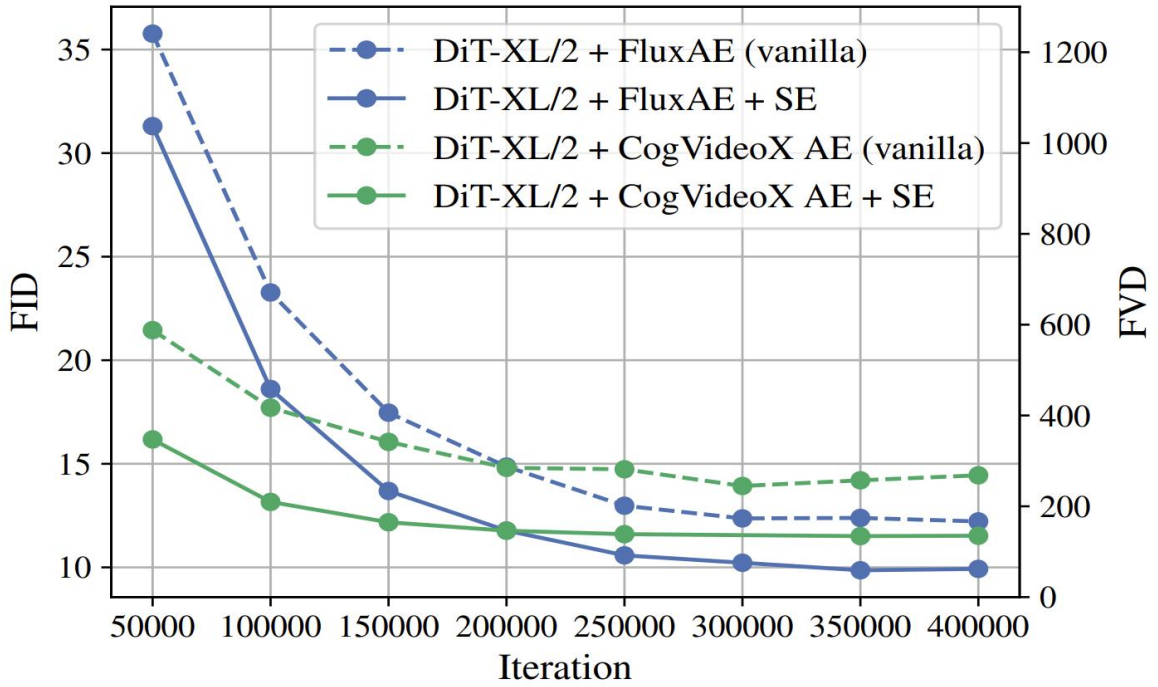
- Confines diffusion model training to a fixed latent space
- Interact between AE & LDM
- REPA[1] align early-layer features of DiT with clean image features from pre-trained vision encoders

[1] Yu S, Kwak S, Jang H, et al. Representation alignment for generation: Training diffusion transformers is easier than you think. (ICLR 25)

Improving the Diffusability of Autoencoders

Insights

- High-frequency noise in AE
- Scale equivariance: a simple regularization

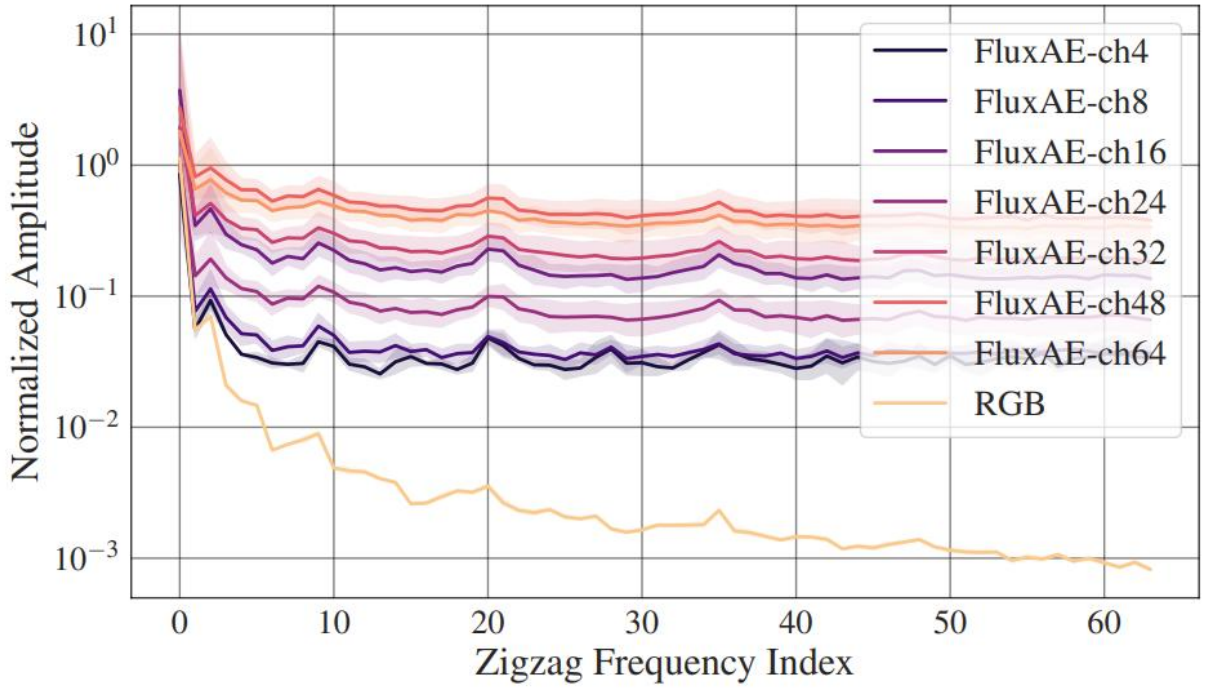


[1] Skorokhodov I, Girish S, Hu B, et al. Improving the diffusability of autoencoders[J]. (arXiv)

Improving the Diffusability of Autoencoders

Blockwise 2D DCT

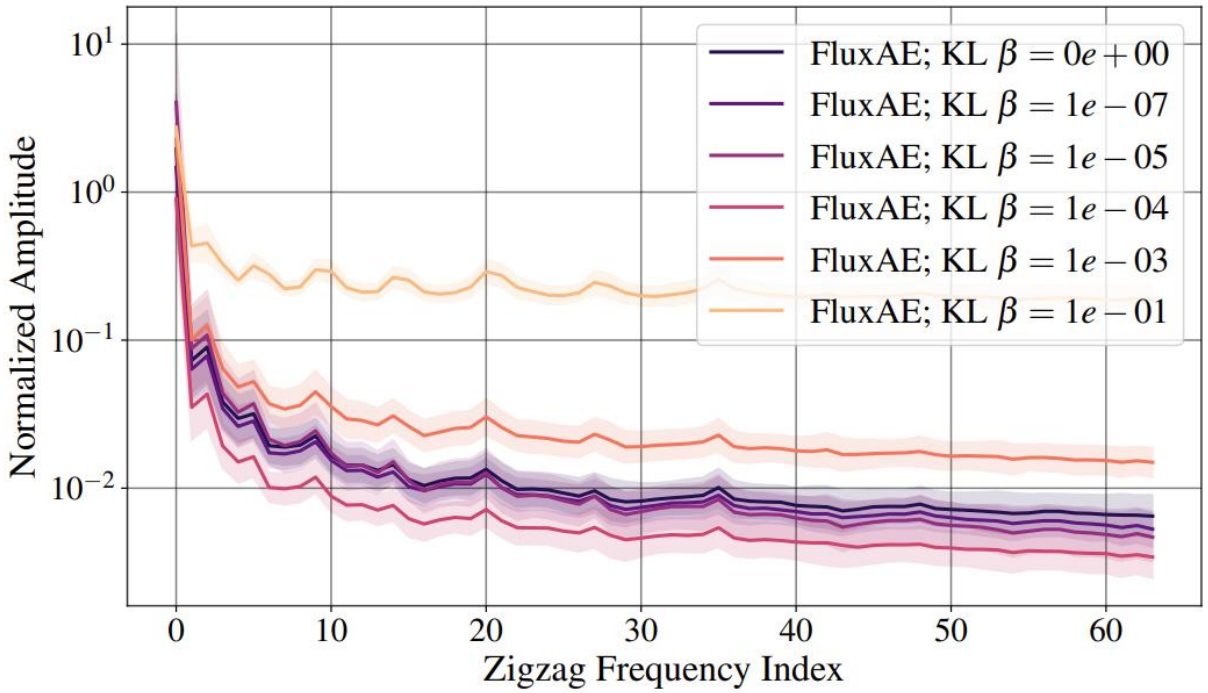
- Autoencoder is different from RGB
- Bigger channel has higher frequencies



[1] Skorokhodov I, Girish S, Hu B, et al. Improving the diffusability of autoencoders[J]. (arXiv)

Improving the Diffusability of Autoencoders

Spectral Analysis



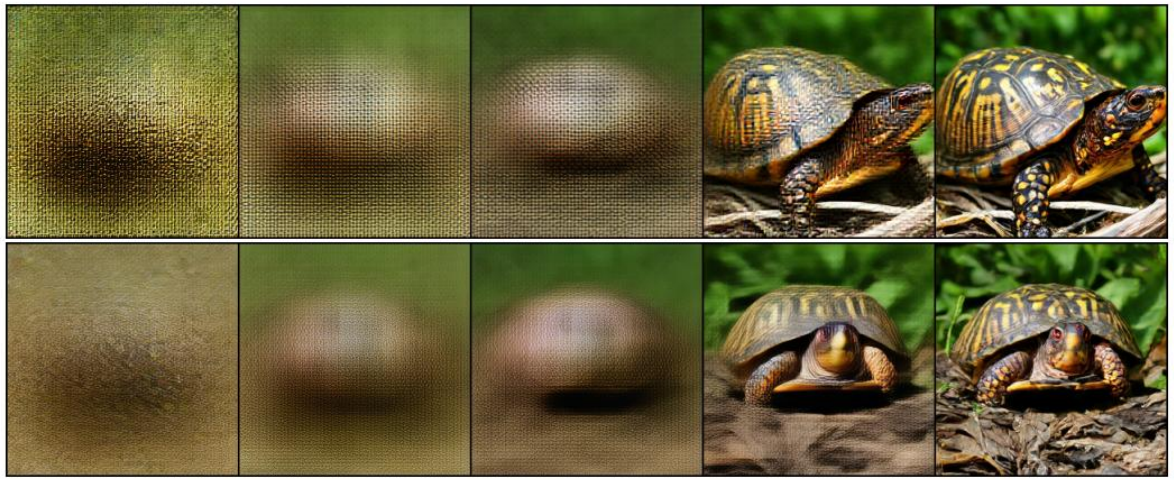


Figure 6. Denoising trajectories (steps 1, 16, 32, 128 and 256 out of 256) for DiT-XL trained with FluxAE (top) and FluxAE+SE. DiT-XL with vanilla FluxAE exhibits prominent high-frequency artifacts early on in the trajectory.

[1] Skorokhodov I, Girish S, Hu B, et al. Improving the diffusability of autoencoders[J]. (arXiv)

Improving the Diffusability of Autoencoders

北京大学 PEKING UNIVERSITY

Scale Equivariance Regularization

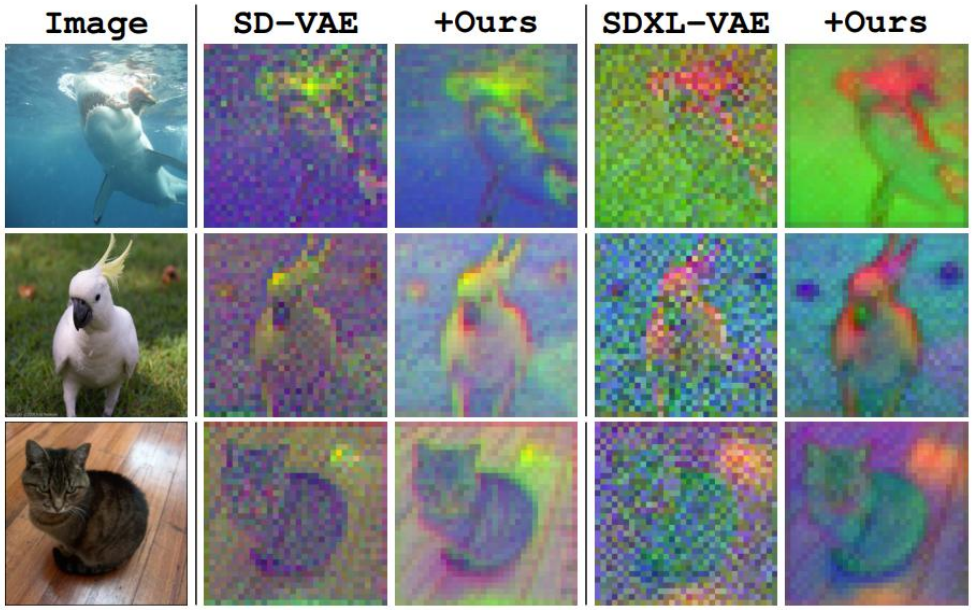
- ×2 – 4 bilinear downsampling $\tilde{x}$
- Removes high-frequency & remain low-frequency

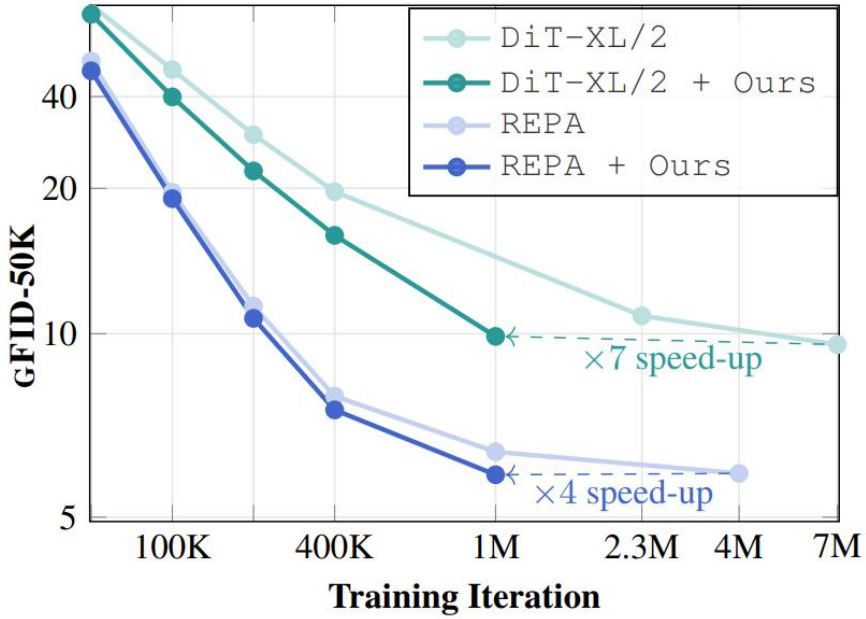
$$\mathcal{L}(x) = d(x, \mathrm{Dec}(z)) + \alpha d(\tilde{x}, \mathrm{Dec}(\tilde{z})) + \beta \mathcal{L}_{\mathrm{KL}}$$

[1] Skorokhodov I, Girish S, Hu B, et al. Improving the diffusability of autoencoders[J]. (arXiv)

Insights

- Lack semantic-preserving in AE
- Equivariance Regularization: a simple regularization





[1] Kouzelis T, Kakogeorgiou I, Gidaris S, et al. EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling. (arXiv)

北京大学 PEKING UNIVERSITY

Semantic-preserving

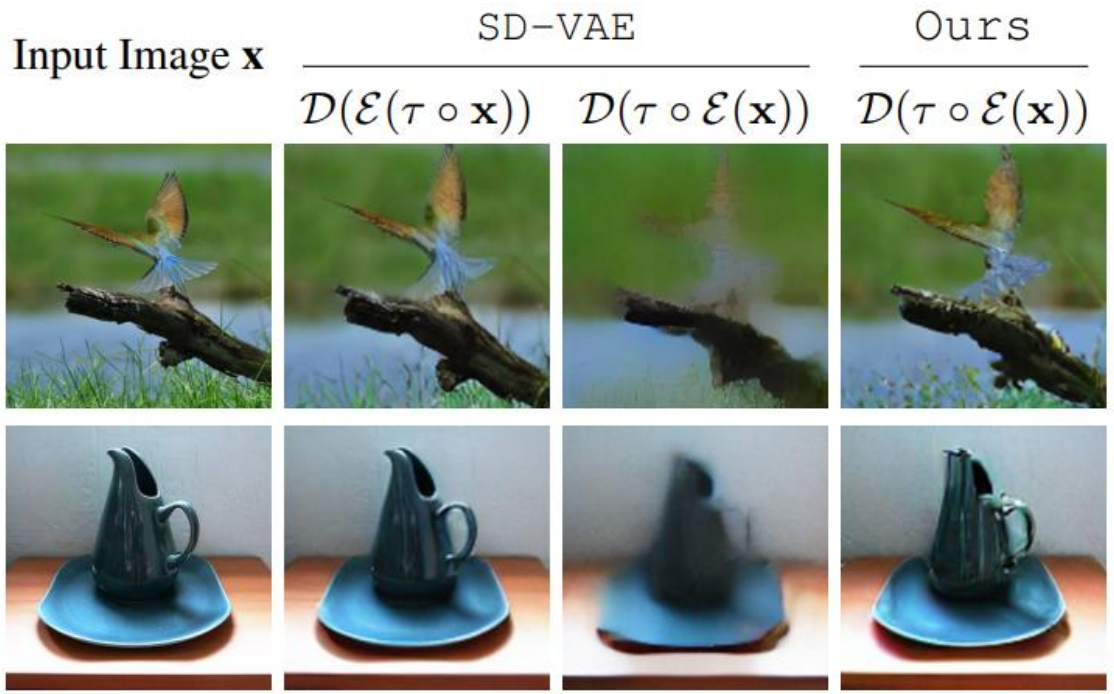
- Spatial Transformation $\tau \circ \mathbf{x}$
- Equivariance $\mathcal{E}(\tau \circ \mathbf{x}) = \tau \circ \mathcal{E}(\mathbf{x})$

$$\mathcal{L}_{\text{explicit}}(\mathbf{x}) = \|\tau \circ \mathcal{E}(\mathbf{x}) - \mathcal{E}(\tau \circ \mathbf{x})\|_2^2$$

[1] Kouzelis T, Kakogeorgiou I, Gidaris S, et al. EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling. (arXiv)

Equivariance Regularization

$$\mathcal{L}_{\text{EQ-VAE}}(\mathbf{x}, \tau) = \mathcal{L}_{rec}\Big(\tau \circ \mathbf{x}, \mathcal{D}\big(\tau \circ \mathcal{E}(\mathbf{x})\big)\Big) + \lambda_{gan}\mathcal{L}_{gan}\Big(\mathcal{D}\big(\tau \circ \mathcal{E}(\mathbf{x})\big)\Big) + \lambda_{reg}\mathcal{L}_{reg}$$

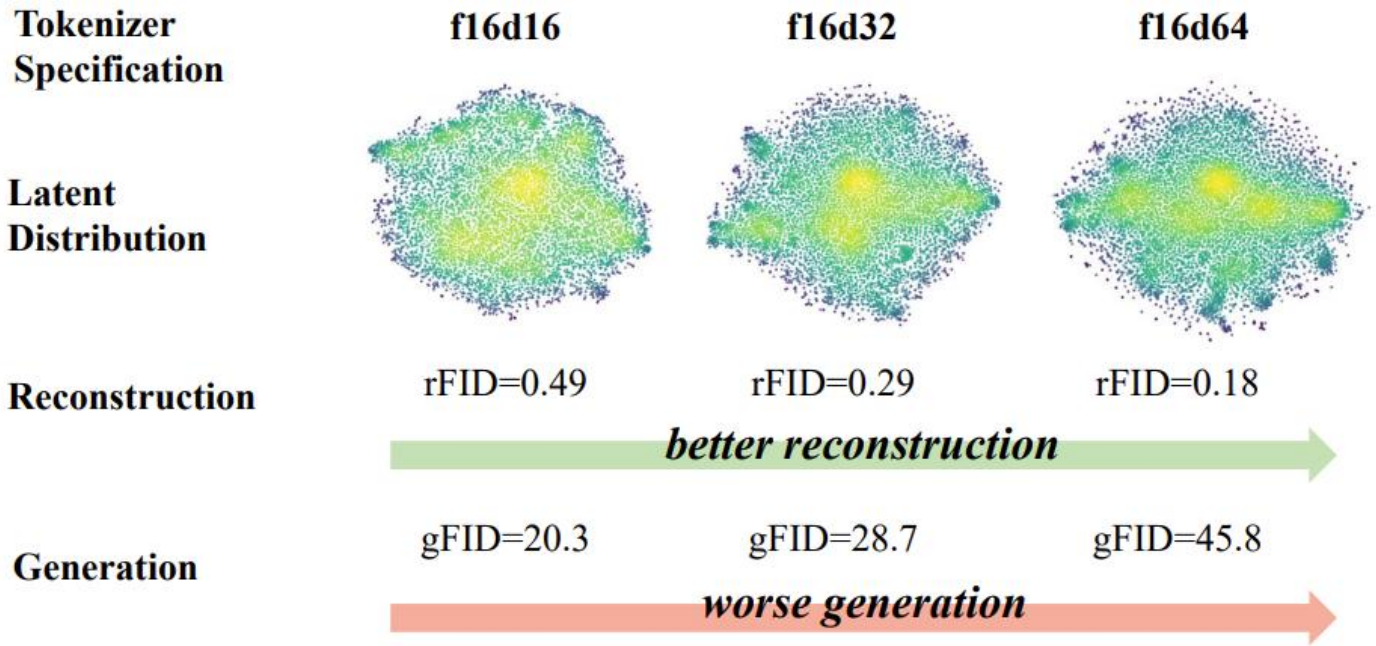


[1] Kouzelis T, Kakogeorgiou I, Gidaris S, et al. EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling. (arXiv)

北京大学 PEKING UNIVERSITY

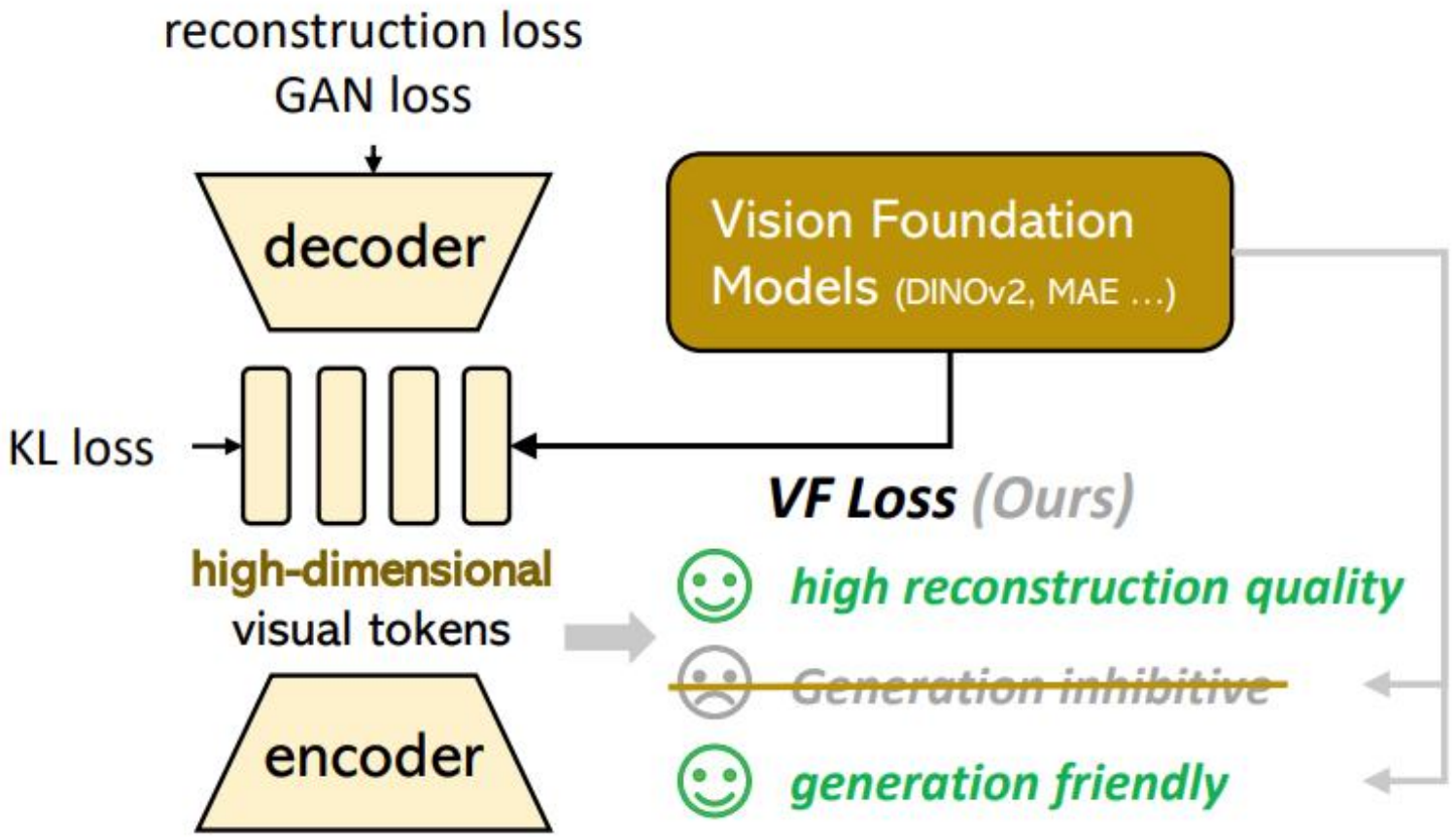
Motivation

- Uneven distribution of potential features in AE
- Contradiction between reconstruction and generation



[1] Yao J, Yang B, Wang X. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models.(CVPR 25)

Vision Foundation model alignment loss



[1] Yao J, Yang B, Wang X. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models.(CVPR 25)

Marginal Cosine Similarity Loss

- Image latents Z & foundational visual representations F

$$Z' = WZ$$

$$\mathcal{L}_{\text{mcos}} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \text{ReLU}\left(1 - m_1 - \frac{z'_{ij} \cdot f_{ij}}{\|z'_{ij}\| \|f_{ij}\|}\right)$$

[1] Yao J, Yang B, Wang X. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models.(CVPR 25)

Marginal Distance Matrix Similarity Loss

- Reduce relative distribution distance

$$\mathcal{L}_{\text{mdms}} = \frac{1}{N^2} \sum_{i,j} \text{ReLU}\left(\left| \frac{z_i \cdot z_j}{\|z_i\| \|z_j\|} - \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|} \right| - m_2 \right)$$

[1] Yao J, Yang B, Wang X. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models.(CVPR 25)

Adaptive Weighting

- Ensuring similar impacts on model optimization

$$w_{\text{adaptive}} = \frac{\|\nabla L_{\text{rec}}\|}{\|\nabla L_{\text{vf}}\|}$$

$$\mathcal{L}_{\text{vf}} = w_{\text{hyper}} * w_{\text{adaptive}}(\mathcal{L}_{\text{mcos}} + \mathcal{L}_{\text{mdms}})$$

[1] Yao J, Yang B, Wang X. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models.(CVPR 25)

03 Method

Motivation

北京大学 PEKING UNIVERSITY

- Recent works fintune the VAE to improve the equivariance
- Jointly tune both VAE and LDM in an end-to-end manner
- End-to-end training can lead to increased performance

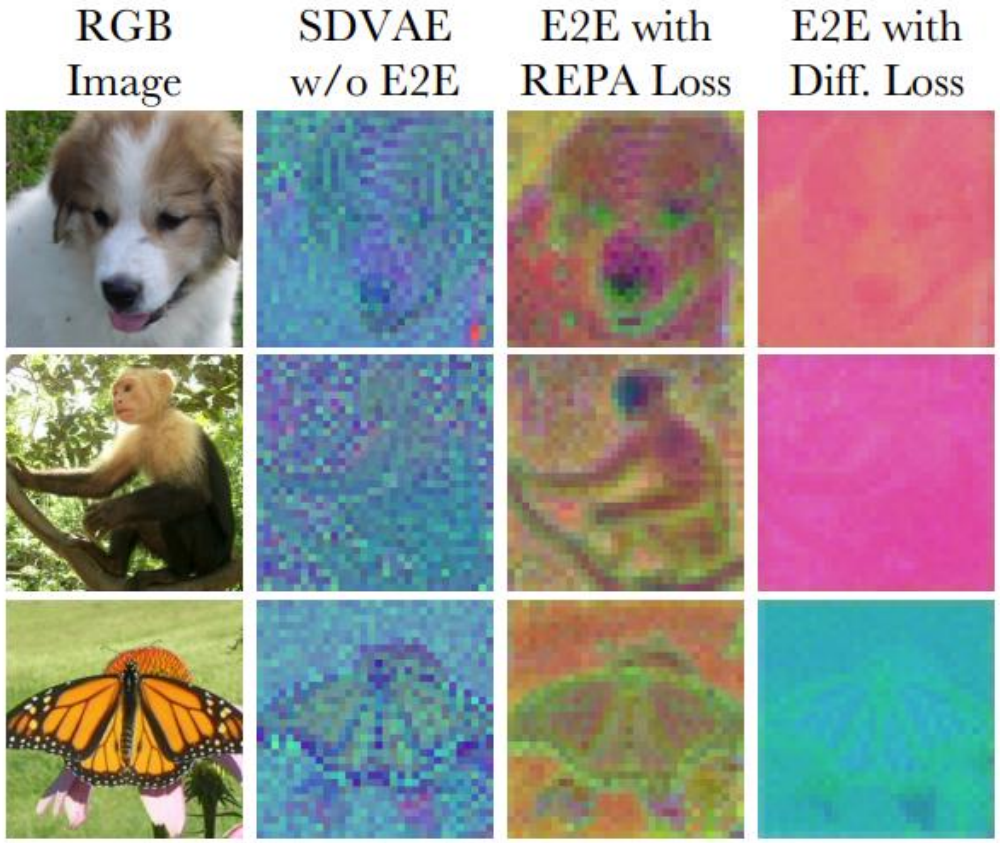
北京大学 PEKING UNIVERSITY

Naive End-to-End Tuning is Ineffective

- Simpler latent-space

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

- Reduced generation performance



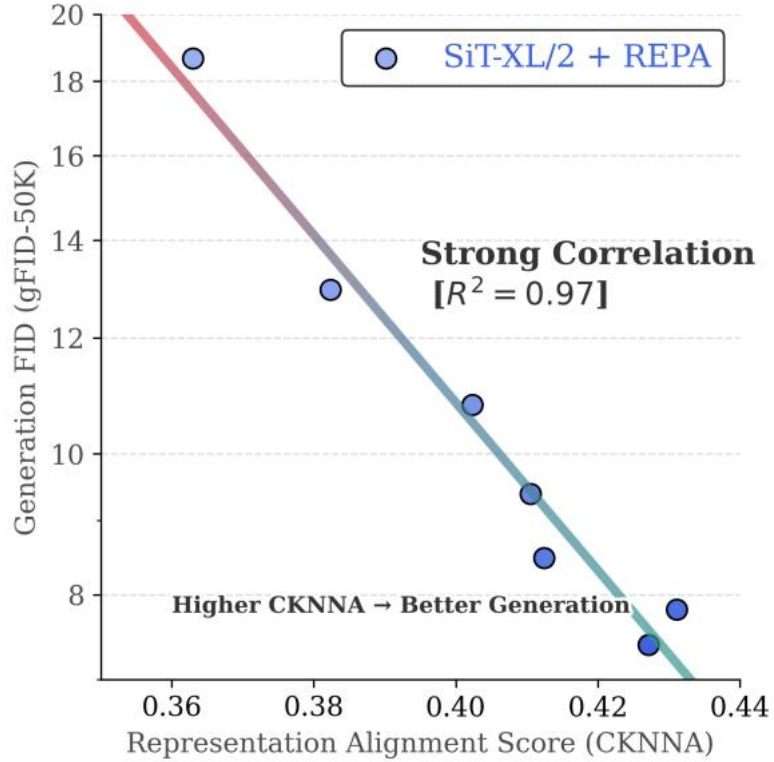
(a) PCA Visualization of Latent Spaces

北京大学 PEKING UNIVERSITY

CKNNA scores

- High CKNNA, better generation

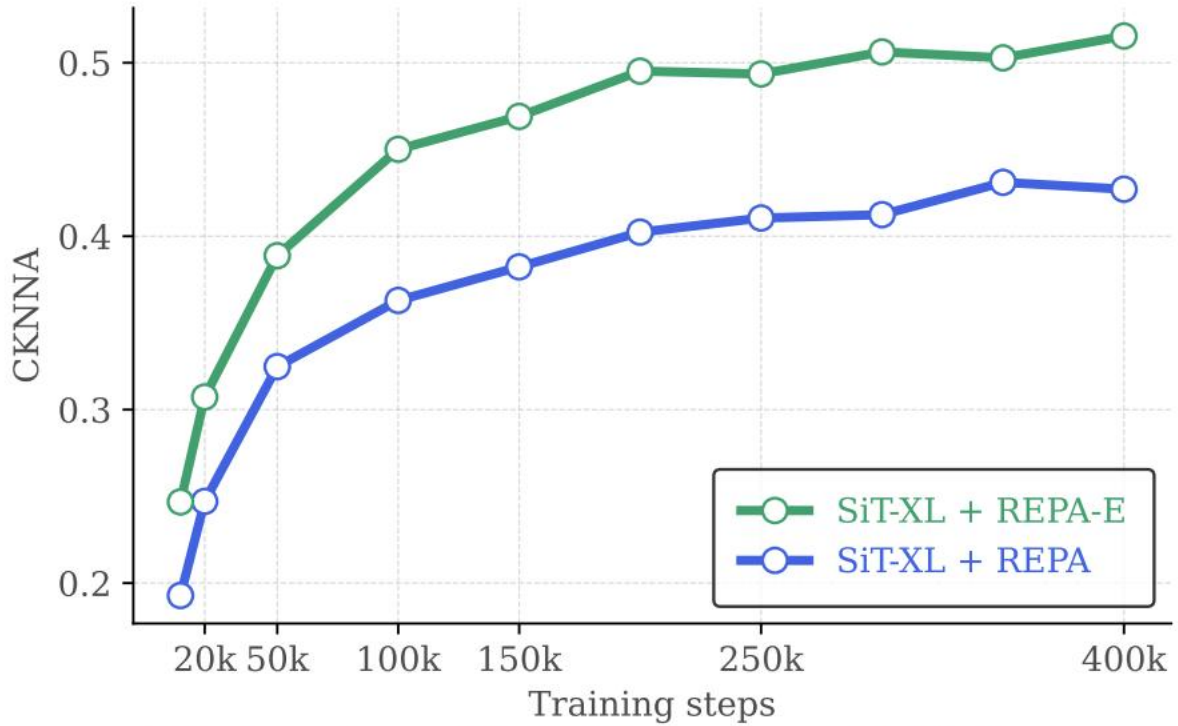
- Optimize CKNNA score(REPA loss)



(b) Correlation: gFID & CKNNA Score

Representation Alignment is Bottlenecked by VAE

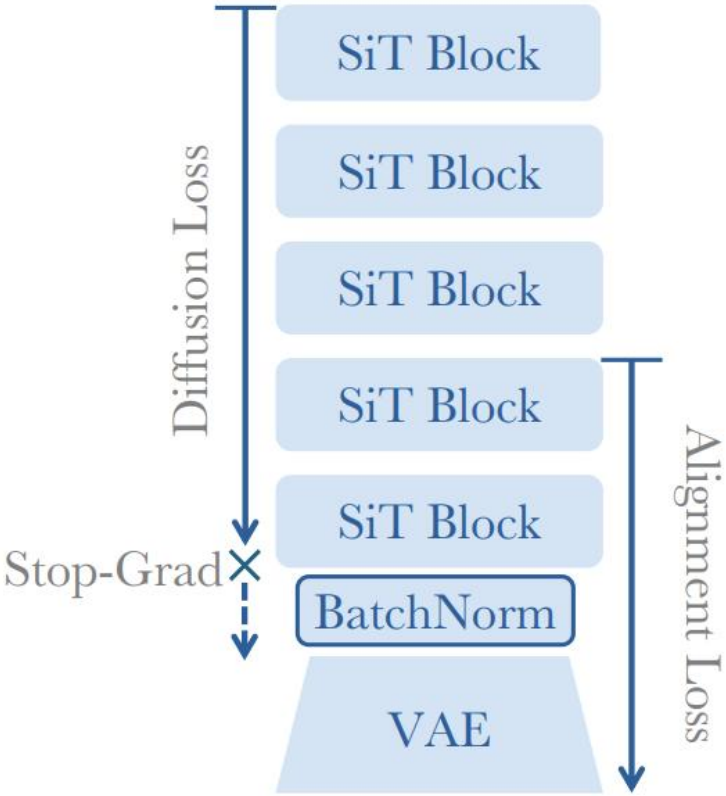
- Frozen VAE has limitation



(c) E2E tuning with REPA improves CKNNA Score

Batch-Norm Layer for VAE Latent Normalization

- From precomputed latent statistics(1/ 0.1825 for SD-VAE) to
- A new batch-norm layer



北京大学 PEKING UNIVERSITY

End-to-End Representation-Alignment Loss

- DiT hidden state gets projection

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

- Overall Training

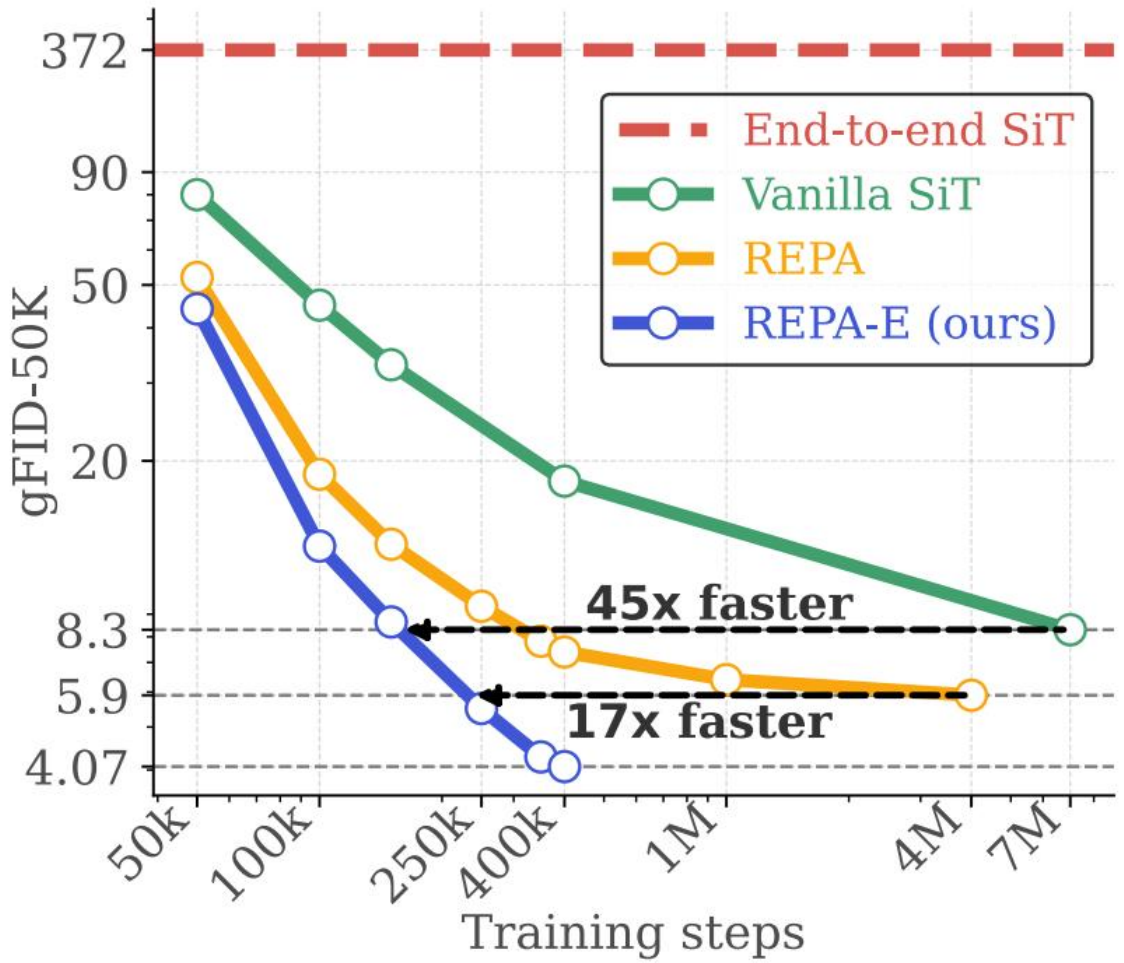
$$\mathcal{L}(\theta, \phi, \omega) = \mathcal{L}_{\text{DIFF}}(\theta) + \lambda\mathcal{L}_{\text{REPA}}(\theta, \phi, \omega) + \eta\mathcal{L}_{\text{REG}}(\phi)$$

$$\lambda_{\text{REPA}_g} = 0.5 \qquad \lambda_{\text{REPA}_v} = 1.5$$

04 Experiments

北京大学 PEKING UNIVERSITY

Significantly improve generation performance and training speed



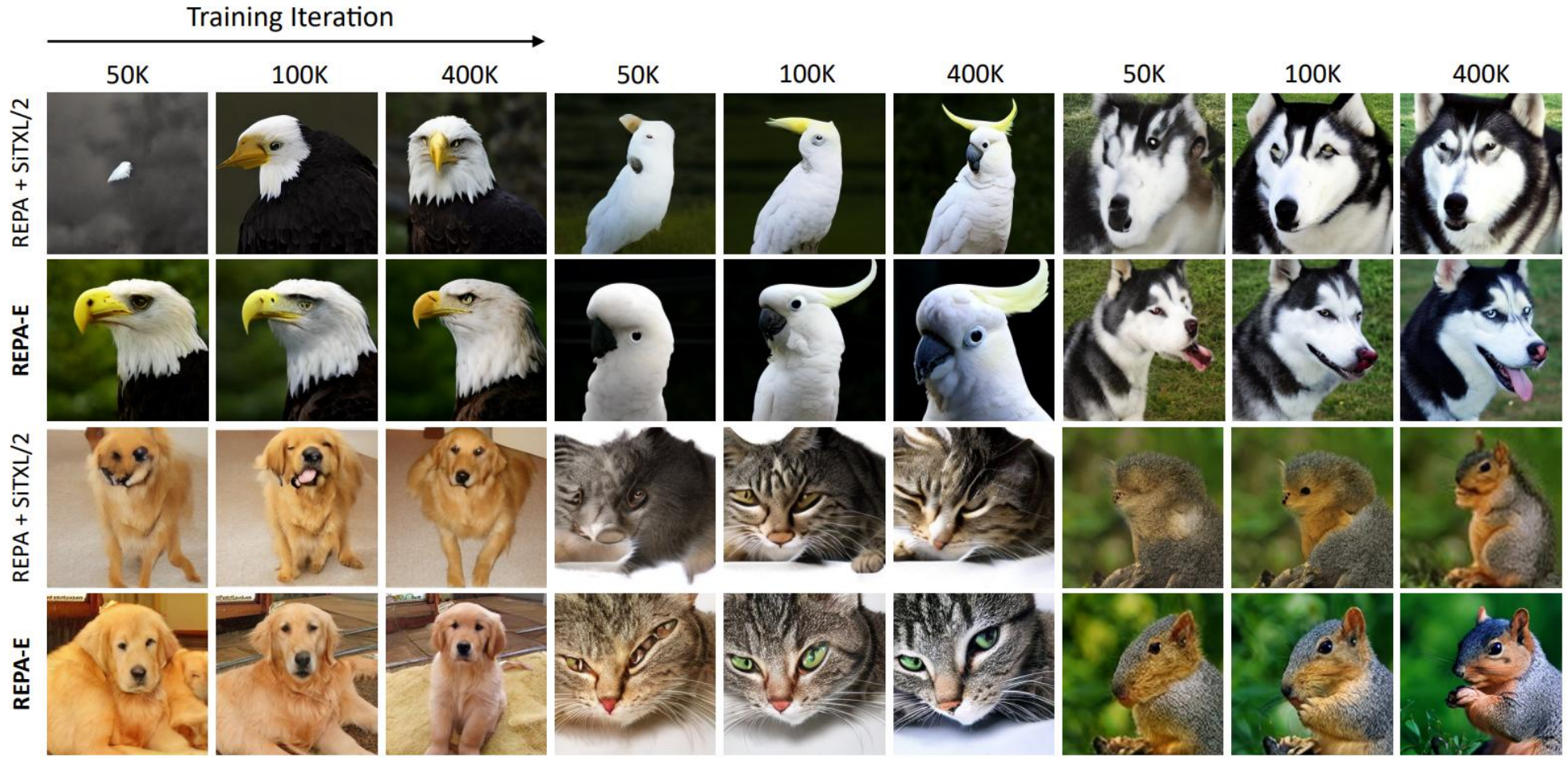
北京大学 PEKING UNIVERSITY

Significantly improve generation performance and training speed

Method	Tokenizer	Epochs	gFID↓	sFID↓	IS↑
Without End-to-End Tuning					
MaskDiT [55]	SD-VAE	1600	5.69	10.34	**177.9**
DiT [34]		1400	9.62	6.85	121.5
SiT [30]		1400	8.61	6.32	131.7
FasterDiT [50]		400	7.91	5.45	131.3
REPA [53]	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
With End-to-End Tuning (Ours)					
REPA-E	SD-VAE*	20	12.83	5.04	88.8
		40	7.17	**4.39**	123.7
		80	**4.07**	<u>4.60</u>	<u>161.8</u>

北京大学 PEKING UNIVERSITY

Significantly improve generation performance and training speed



北京大学 PEKING UNIVERSITY

REPA-E generalize across variations

Method	Tokenizer	Epochs	gFID↓	sFID↓	IS↑
Without End-to-End Tuning					
MaskDiT [55]	SD-VAE	1600	5.69	10.34	**177.9**
DiT [34]		1400	9.62	6.85	121.5
SiT [30]		1400	8.61	6.32	131.7
FasterDiT [50]		400	7.91	5.45	131.3
REPA [53]	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
With End-to-End Tuning (Ours)					
REPA-E	SD-VAE*	20	12.83	5.04	88.8
		40	7.17	**4.39**	123.7
		80	**4.07**	<u>4.60</u>	<u>161.8</u>

北京大学 PEKING UNIVERSITY

REPA-E generalize across variations

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

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

北京大学 PEKING UNIVERSITY

REPA-E generalize across variations

Autoencoder	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
SD-VAE [39]	24.1	6.25	55.7	0.62	**0.60**
+REPA-E (Ours)	**16.3**	**5.69**	**75.0**	**0.68**	**0.60**
IN-VAE *(f16d32)*	22.7	**5.47**	56.0	0.62	**0.62**
+REPA-E (Ours)	**12.7**	5.57	**84.0**	**0.69**	**0.62**
VA-VAE [49]	12.8	6.47	83.8	0.71	0.58
+REPA-E (Ours)	**11.1**	**5.31**	**88.8**	**0.72**	**0.61**

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

北京大学 PEKING UNIVERSITY

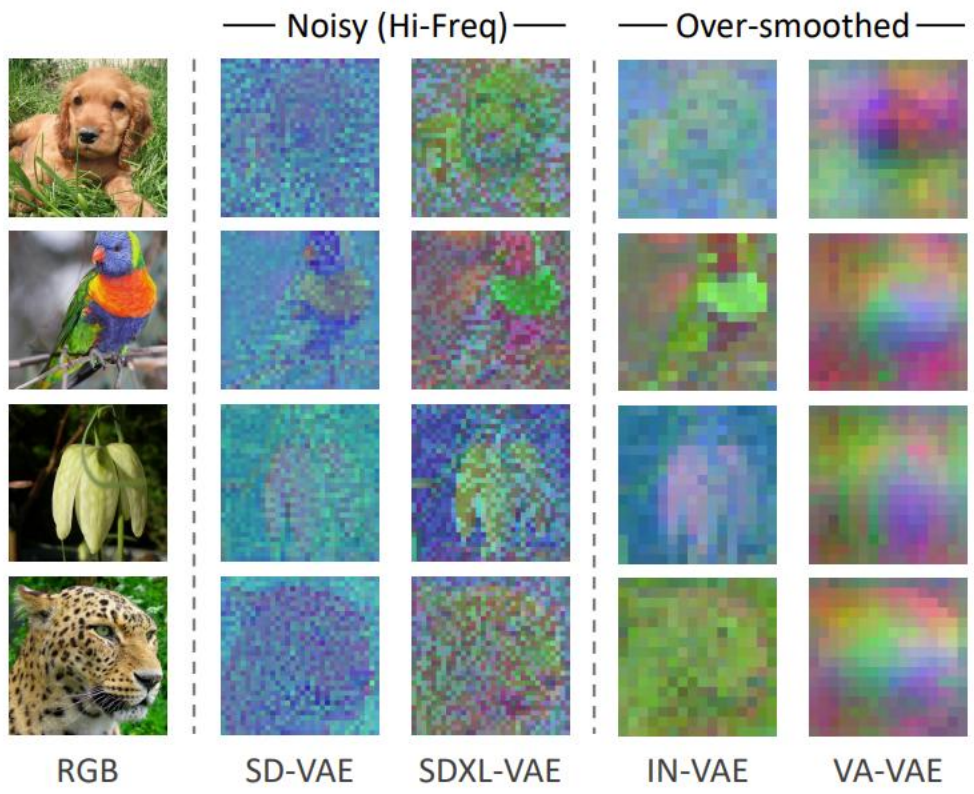
REPA-E generalize across variations

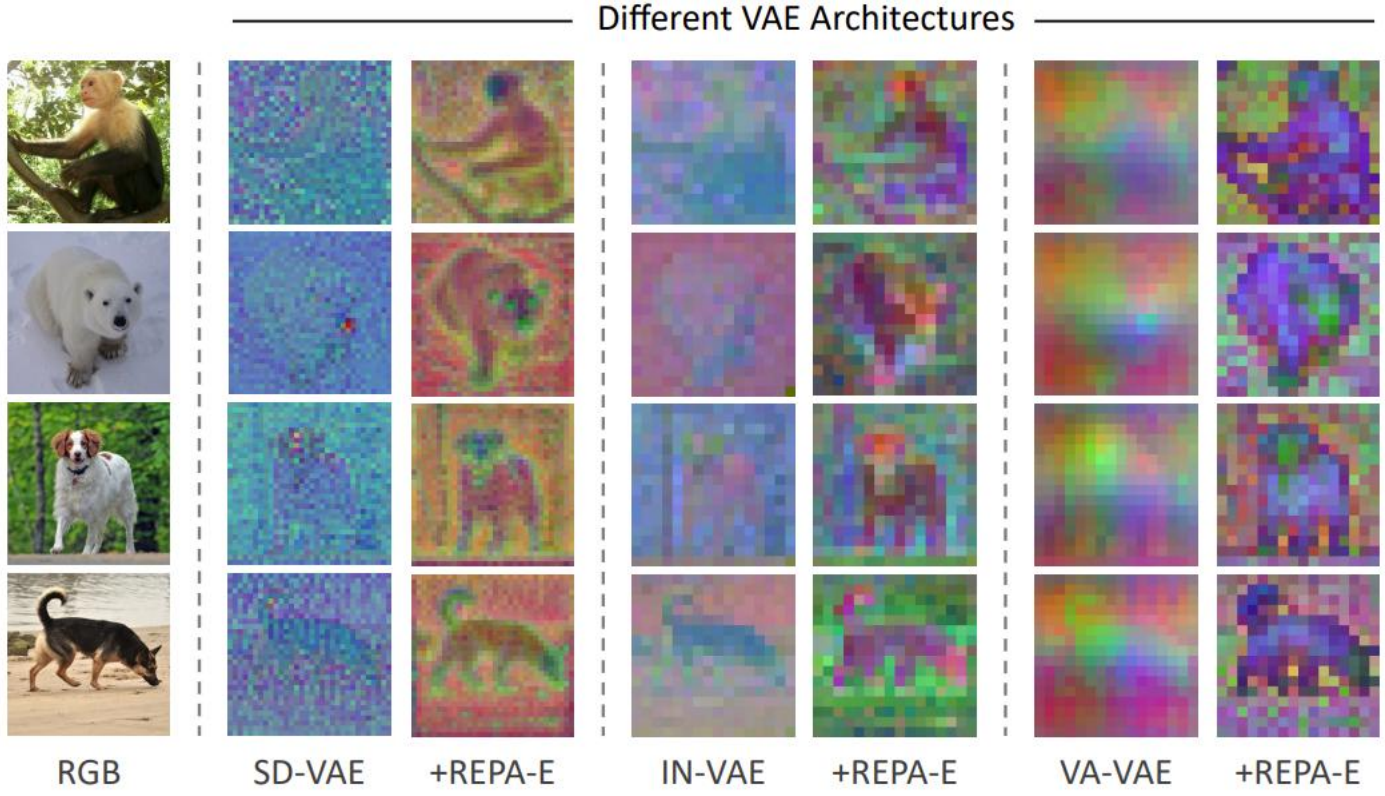
Component	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
w/o `stopgrad`	444.1	460.3	1.49	0.00	0.00
w/o `batch-norm`	18.1	**5.32**	72.4	0.67	0.59
w/o $\mathcal{L}_{\text{REG}}$	19.2	6.47	68.2	0.64	0.58
REPA-E (Ours)	**16.3**	5.69	**75.0**	**0.68**	**0.60**

Method	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
100K Iterations (20 Epochs)					
REPA [53]	19.40	6.06	67.4	0.64	**0.61**
REPA-E (scratch)	14.12	7.87	83.5	0.70	0.59
REPA-E (VAE init.)	**12.83**	**5.04**	**88.8**	**0.71**	0.58
200K Iterations (40 Epochs)					
REPA [53]	11.10	5.05	100.4	0.69	**0.64**
REPA-E (scratch)	7.54	6.17	120.4	**0.74**	0.61
REPA-E (VAE init.)	**7.17**	**4.39**	**123.7**	**0.74**	0.62
400K Iterations (80 Epochs)					
REPA [53]	7.90	5.06	122.6	0.70	**0.65**
REPA-E (scratch)	4.34	**4.44**	154.3	0.75	0.63
REPA-E (VAE init.)	**4.07**	4.60	**161.8**	**0.76**	0.62

北京大学 PEKING UNIVERSITY

The impact of end-to-end tuning





The impact of end-to-end tuning

VAE	Diffusion model	REPA	gFID-50K
SD-VAE [39]	DiT-XL [34]	✗	19.82
VA-VAE [49]	DiT-XL [34]	✗	**6.74**
E2E-VAE (Ours)	DiT-XL [34]	✗	6.75
SD-VAE [39]	SiT-XL [30]	✗	17.20
VA-VAE [49]	SiT-XL [30]	✗	5.93
E2E-VAE (Ours)	SiT-XL [30]	✗	**5.26**
SD-VAE [39]	DiT-XL [34]	✓	12.29
VA-VAE [49]	DiT-XL [34]	✓	4.71
E2E-VAE (Ours)	DiT-XL [34]	✓	**4.20**
SD-VAE [39]	SiT-XL [30]	✓	7.90
VA-VAE [49]	SiT-XL [30]	✓	4.88
E2E-VAE (Ours)	SiT-XL [30]	✓	**3.46**

北京大学 PEKING UNIVERSITY

The impact of end-to-end tuning

Tokenizer	Method	Training Epoches	#params	rFID↓	Generation w/o CFG					Generation w/ CFG				
					gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
AutoRegressive (AR)														
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 [51]	1080	307M	1.50	3.65	-	200.5	-	-	1.78	-	319.4	-	-
LDM	MAR [28]	800	945M	0.53	2.35	-	227.8	0.79	0.62	1.55	-	303.7	0.81	0.62
Latent Diffusion Models (LDM)														
SD-VAE [39]	MaskDiT [55]	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	**0.83**	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 [50]	400	675M		7.91	5.45	131.3	0.67	**0.69**	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
Representation Alignment Methods														
VA-VAE [49]	LightningDiT [49]	80	675M	**0.28**	4.29	-	-	-	-	-	-	-	-	-
		800	675M		2.17	4.36	205.6	0.77	0.65	1.35	4.15	295.3	0.79	0.65
SD-VAE	REPA [53]	80	675M	0.61	7.90	5.06	122.6	0.70	0.65	-	-	-	-	-
		800	675M		5.90	5.73	157.8	0.70	**0.69**	1.42	4.70	305.7	0.80	0.65
E2E-VAE (Ours)	REPA	80	675M	**0.28**	3.46	**4.17**	159.8	0.77	0.63	1.67	4.12	266.3	0.80	0.63
		800	675M		**1.83**	4.22	**217.3**	**0.77**	0.66	**1.26**	**4.11**	**314.9**	0.79	**0.66**

Thanks for listening

北京大學
PEKING UNIVERSITY