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 9. **Impact of Naive End-to-End Training with Diffusion Loss.** We report total variation [40] 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 [52]. 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 degrages 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** [40], which captures local spatial differences in the latent map.

As shown in Tab. 9 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, 5).

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
REPA-E (Ours)	**4.07**	**4.60**	**161.8**	**0.76**	**0.62**

Table 10. **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 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 [46], 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. 10, our REPA-E formulation yields better performance across all metrics at just 400K steps, with significantly faster convergence and stronger generation quality.

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

Table 11. **Scaling REPA-E to Higher Resolution**. System-level results on ImageNet-512 with 64×64 latents using SiT-L at 100K steps without classifier-free guidance. We observe that REPA-E leads to signficant performance improvements over vanilla-REPA [52] even at high resolutions.

Sampler	ODE, NFE=50		SDE, NFE=250	
gFID	VA-VAE 5.43	E2E-VAE **5.02**	VA-VAE 5.57	E2E-VAE **4.97**

Table 12. **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.

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 (64×64). We use SD-VAE [39] 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. 11, 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 [52], 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. 12).

Figure 6. **Qualitative Results on Imagenet 256 × 256** using E2E-VAE and SiT-XL. We use a classifier-free guidance scale $\alpha_{\text{cfg}} = 4.0$.

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 [44]	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 [45]	350	2.0B	-	-	-	-	-	-	1.80	-	365.4	0.83	0.57
LFQ tokenizers	MagViT-v2 [50]	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
Latent Diffusion Models (LDM)														
SD-VAE [39]	MaskDiT [54]	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 [49]	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 [48]	LightningDiT [48]	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 [52]	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**

Table 13. **System-Level Performance on ImageNet 256 × 256** comparing our end-to-end tuned VAE (E2E-VAE) with other VAEs for traditional LDM training. We observe that in addition to improving VAE latent space structure (Fig. 5), 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 1.26 and 0.28, respectively for LDM generation and VAE reconstruction performance.

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, 6

[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. 13

[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, 12

[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, 12

[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. 13

[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. 13

[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. 5

[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 Kynkäänniemi, 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. 13

[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. 12

[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, 8, 12, 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 Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin 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, 5, 6, 8, 13

[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] 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, 6, 7, 8, 12, 13

[40] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. 12

[41] 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

[42] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. *arXiv preprint arXiv:2211.17084*, 2022. 3

[43] 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, 8

[44] 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. 13

[45] 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. 13

[46] 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, 12

[47] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[48] Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. 3, 6, 7, 8, 13

[49] Jingfeng Yao, Wang Cheng, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *arXiv preprint arXiv:2410.10356*, 2024. 5, 6, 13

[50] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, 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. 13

[51] 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

[52] 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, 12, 13

[53] 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

[54] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 5, 6, 13