Reconstruction *vs.* Generation: Taming Optimization Dilemma in Latent Diffusion Models

Supplementary Material

A. More Implementation Details

Our models and codes are released in https://github.com/hustvl/LightningDiT. Please refer to our repo for repository implementation details.

B. More Ablations

B.1. Additional ablation on Loss Design

We provide a more detailed study of the parameter design of the loss function. Specifically, we employ DINOv2 [1] as the alignment target for the base model and train it with different settings of the VF loss for 50 epochs. The resulting VAE [2] is then used to train LightningDiT-B for 80 epochs to evaluate its generative performance. The results are shown in Table A1, where the combination of both L_{mcos} and L_{mdms} losses achieves the best generative performance. Furthermore, applying the margin operation leads to a slight additional improvement in performance.

Tokenizer	L_{mcos}	L_{mdms}	m_1	m_2	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓
f16d32	-	-	-	-	0.40	26.34	0.10	0.74	29.35
f16d32 + VF loss	✓	✓	0.5	0.25	0.42	25.88	0.10	0.72	**22.11** *(-7.24)*
	✓	✓	0.4	0.25	0.41	25.91	0.10	0.73	22.70
	✓	✓	0	0	0.44	25.91	0.10	0.73	23.43
	✓	-	0.5	-	0.37	25.38	0.12	0.71	25.70
	-	✓	-	0.25	0.39	25.55	0.12	0.71	26.99

Table A1. Detail ablation of loss design.

B.2. Detailed Comparison of VAE Dimension

Another potential issue should be clarified is whether finer-grained tuning of parameter s, given the trade-off it presents between reconstruction and generation performance, could yield superior results compared to VF loss. Our answer is in the negative. We conduct a more comprehensive analysis of the impact of dimensionality variations on the reconstruction and generative performance of VAE [2]. We train three different tokenizers with specifications of {f16d16, f16d24, f16d28}, and the results are presented in Table A2. Increasing the dimensionality of the tokenizer consistently improves reconstruction performance but gradually degrades generative performance. The VF loss effectively addresses this dilemma. The VA-VAE with the specification of f16d32 significantly enhances generative performance while maintaining reconstruction performance.

B.3. VF loss on Larger Dimensions

Furthermore, we explore whether the VF loss remains effective under higher spatial compression rates and deeper dimensions. To this end, we conduct two more aggressive sets of experiments with specifications of {f16d128 and f32d128}. As shown in Table A3, the VF loss continues to effectively enhance generative performance even under higher spatial compression rates and deeper dimensions.

Tokenizer	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓
f16d32 + VF loss	**0.33**	25.81	0.110	0.72	**19.93**
f16d16	0.49	24.45	0.142	0.66	21.20
f16d24	0.34	25.40	0.118	0.71	24.59
f16d28	**0.33**	**26.13**	**0.108**	**0.73**	26.87

Table A2. Comparing to f16d24/f16d28.

Tokenizer	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓
f16d128	0.13	30.00	0.047	0.86	63.92
f16d128 + VF loss	0.14	29.55	0.050	0.85	**46.13** *(-17.79)*
f32d128	0.38	25.21	0.119	0.69	54.52
f32d128 + VF loss	0.37	24.98	0.125	0.68	**38.20** *(-16.32)*

Table A3. Scaling VA-VAE to dim-128.

B.4. Visualization of VA-VAE

In addition to qualitative results, we conduct a visual analysis of the reconstruction outcomes from VA-VAE. As illustrated in Figure A1, the visual quality of VA-VAE with the specification of f16d32 surpasses that of f16d16 and approaches that of the f16d32 VAE without VF loss. This observation aligns consistently with the analysis of quantitative results.

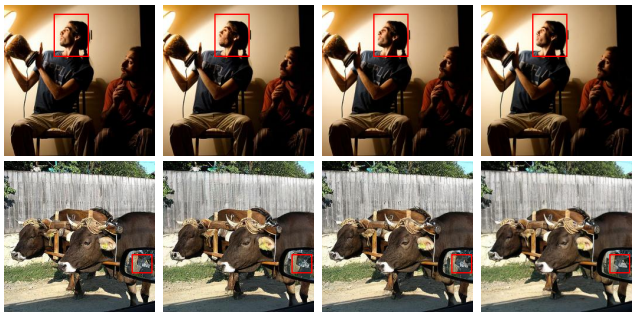

Figure A1. Reconstruction Performance of VA-VAE.

Tokenizer	Training GFLOPs	Inference GFLOPs
LDM [2]	1170	390
VA-VAE	1325 *(+13%)*	390

Table A4. FLOPs of VA-VAE.

B.5. Computation of VF loss

We conduct a computational analysis of VA-VAE in Table A4. During the training process, the VF loss introduces a visual foundation model, which remains unaltered by gradient updates. Compared to previous training procedures, the VF loss incurs an additional 13% training cost. In the inference phase, since the model architecture remains entirely unchanged, the computational consumption of VA-VAE is identical to that of a standard VAE.

References

[1] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 0

[2] 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. 0, 1