Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think

Submission of ICLR25' – 10,10,10,8,8,8

Sihyun Yu[1], Sangkyung Kwak[1], Huiwon Jang[1], Jongheon Jeong[2], Jonathan Huang[3], Jinwoo Shin[1*] , Saining Xie[4*]

[1]KAIST, [2]Korea University, [3]Scaled Foundations, [4]NYU
* Equal Supervision

马逸扬
2024.12.01

Content

- Authors
- Background
- Method

Content

- Authors
- Background
- Method

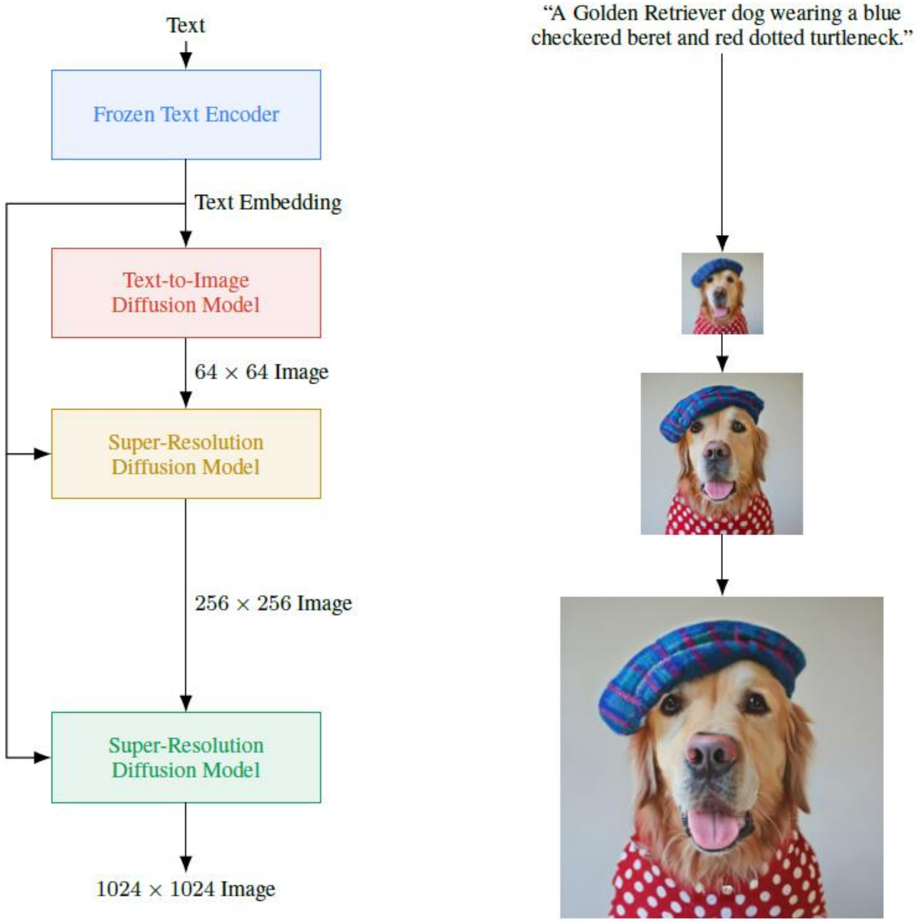
Content

- Authors
- Background
- Method

Background

Speeding up building diffusion models / frameworks

Cascaded models: DALLE2, Imagen.

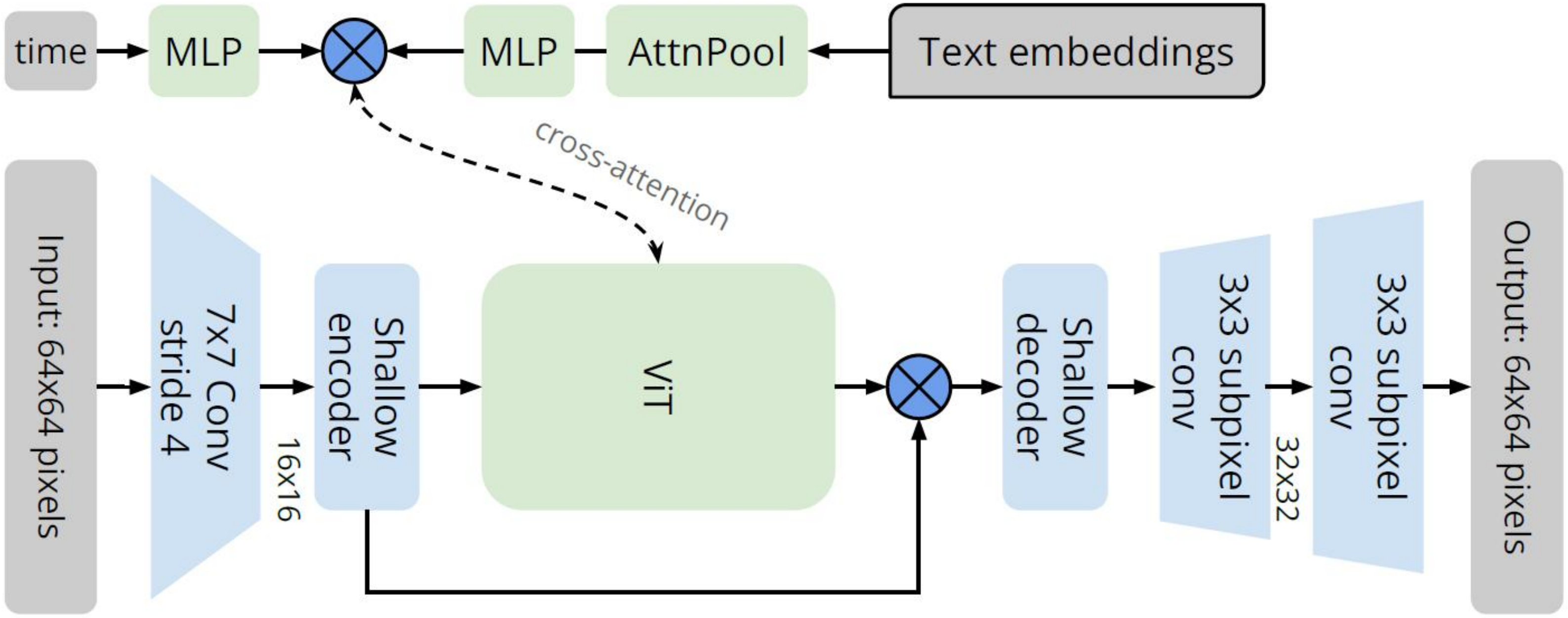


[1] Aditya Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv 2204.
[2] Chitwan Saharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, NIPS22'.

Background

Speeding up building diffusion models / frameworks

Greedy growing: Vermeer.

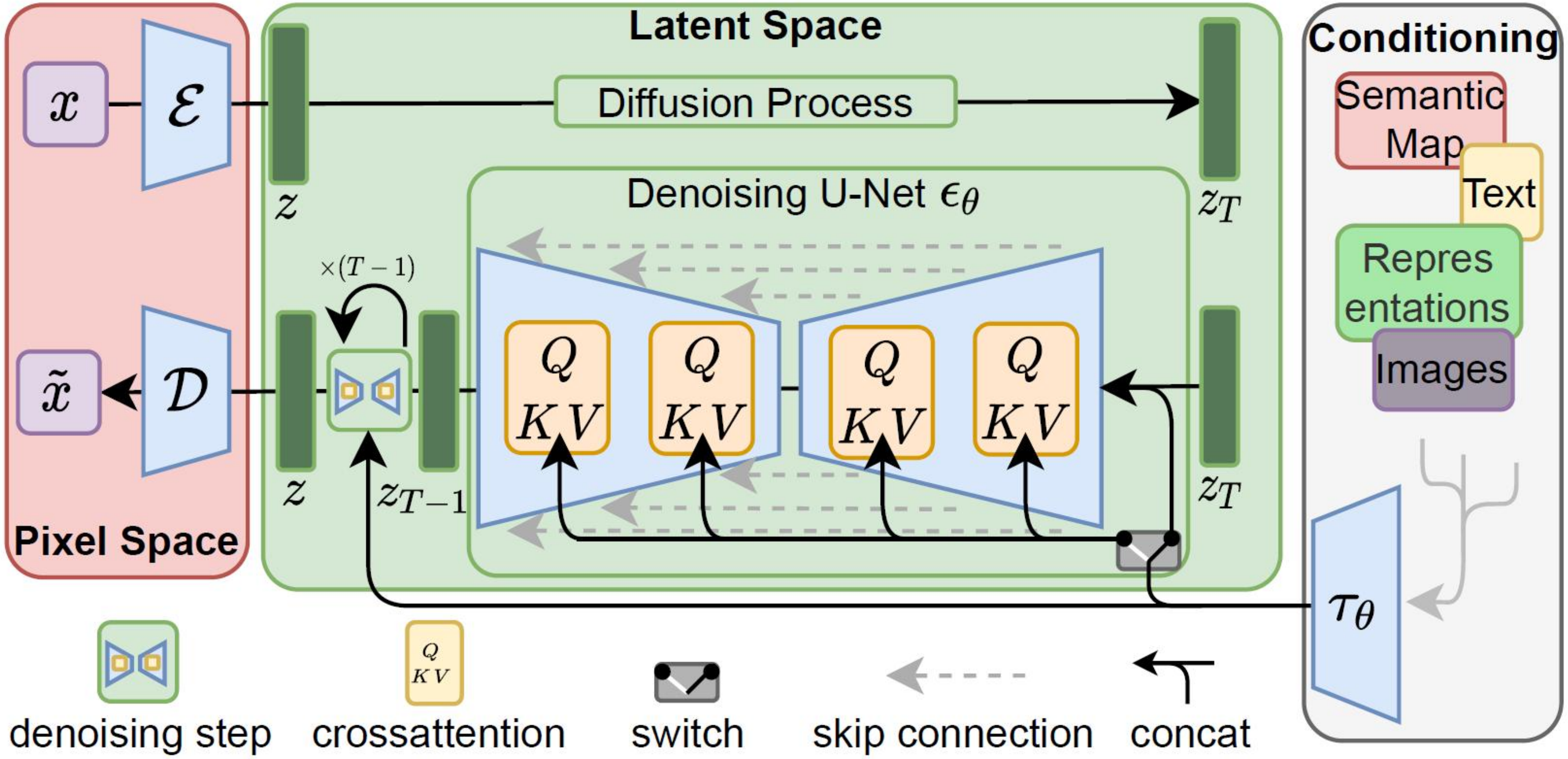


[3] Cristina N. Vasconcelos et al. Greedy Growing Enables High-Resolution Pixel-Based Diffusion Models, TMLR24'.

Background

Speeding up building diffusion models / frameworks

Employing latent spaces: StableDiffusion, FLUX.

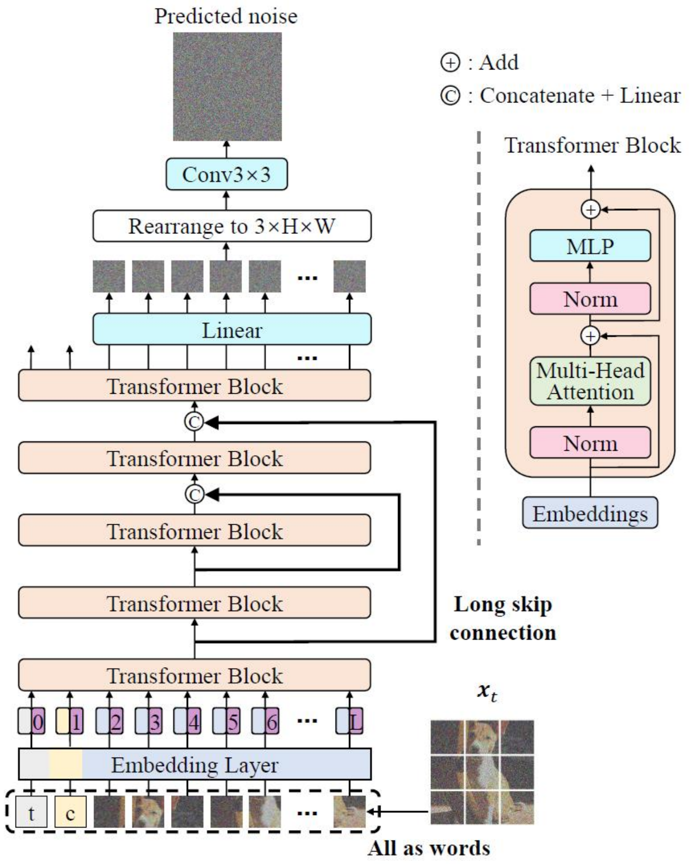


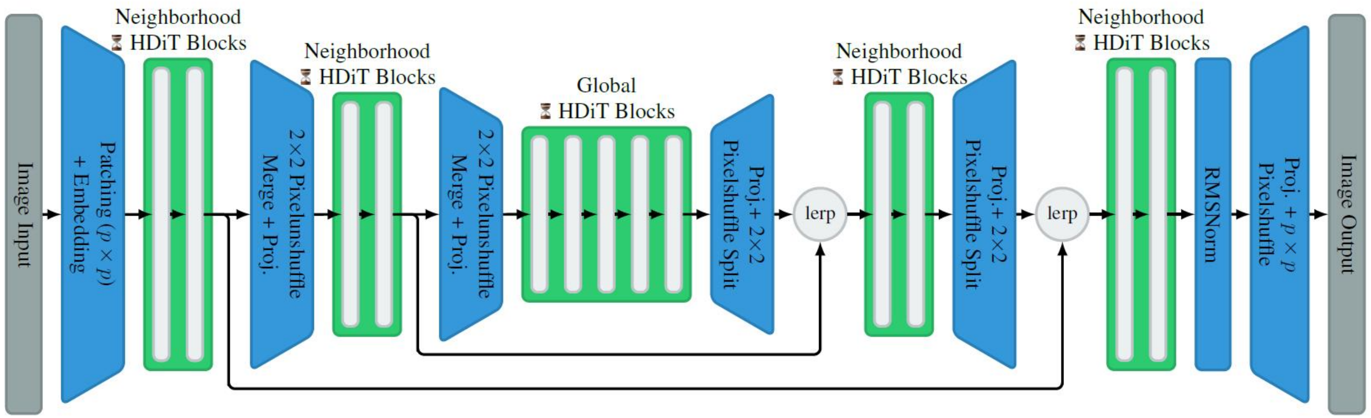
[4] Robin Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models, CVPR22'.
[5] Black Forest Labs. FLUX, 2024.

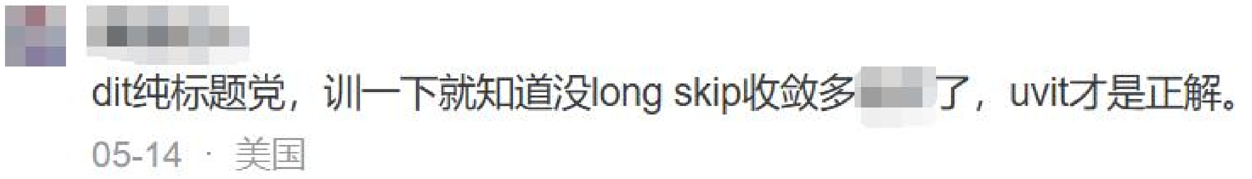
Background

Speeding up building diffusion models / frameworks

Adding long skip connections: from DiT to U-ViT, HDiT.







[6] Fan Bao et al. All are Worth Words: A ViT Backbone for Diffusion Models, CVPR23'.
[7] Katherine Crowson et al. Scalable High-Resolution Pixel-Space Image Synthesis with Hourglass Diffusion Transformers, ICML24'.

Background

Speeding up building diffusion models / frameworks

Reweighting the loss items: min-SNR-γ.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$$

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$$

$$\mathrm{SNR}(t) = \frac{\alpha_t^2}{\sigma_t^2}$$

$$w_t = \min\{\mathrm{SNR}(t), \gamma\}$$

where w_t is the weight of

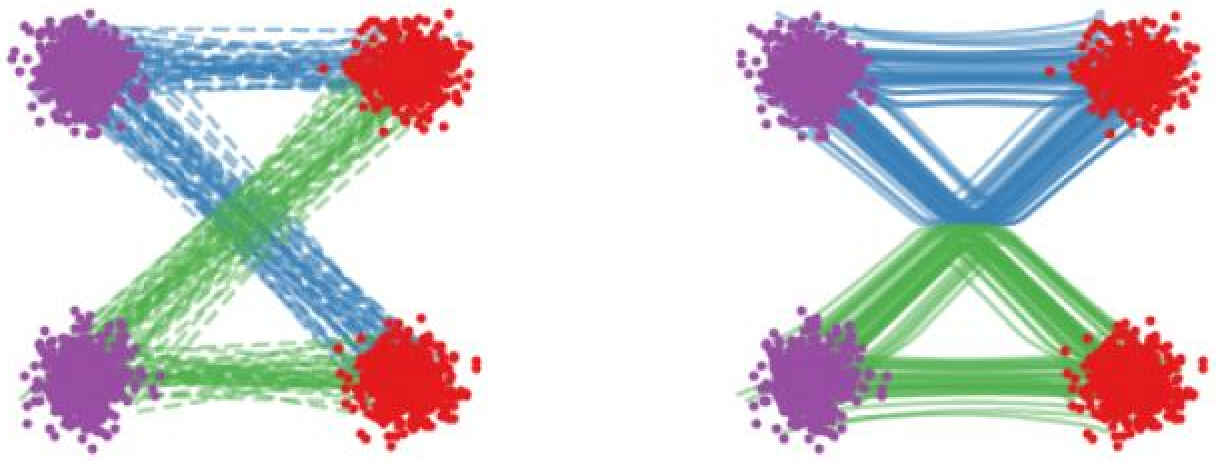
$$\mathcal{L}_{\text{simple}}^t(\theta) = \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\|\mathbf{x}_0 - \hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon)\|_2^2\right]$$

[8] Tiankai Hang et al. Efficient Diffusion Training via Min-SNR Weighting Strategy, ICCV23'.

Background

Speeding up building diffusion models / frameworks

Improved modeling: from DDPM to Rectified Flow.



[9] Xingchao Liu et al. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, ICLR23'.

Background

Speeding up building diffusion models / frameworks

Reschedule the sampling of t: StableDiffusion3.

The probability density function of sampling t is:

$$\pi_{\ln}(t; m, s) = \frac{1}{s\sqrt{2\pi}} \frac{1}{t(1-t)} \exp\Big(-\frac{(\text{logit}(t) - m)^2}{2s^2}\Big)$$

In practice, the sampling of t is achieved by:

- sample $u \sim \mathcal{N}(u; m, s)$;
- map it through standard logistic function.

[10] Patrick Esser et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, ICML23'.

Background

Speeding up building diffusion models / frameworks

Combining different strategies to achieve optimal performances.

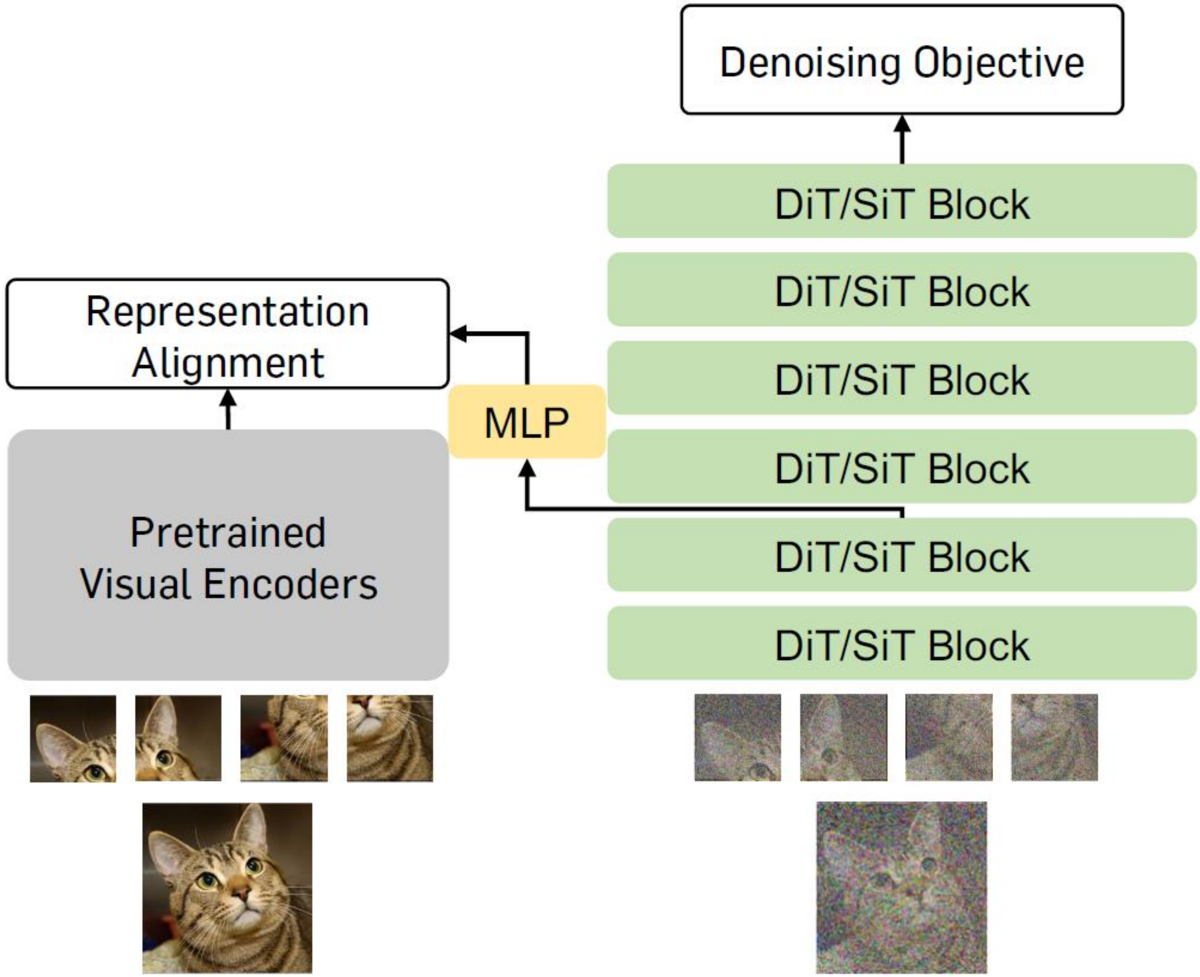
e.g., latent generation + skip connection + RF + rescheduled timestep.

Content

- Authors
- Background
- Method

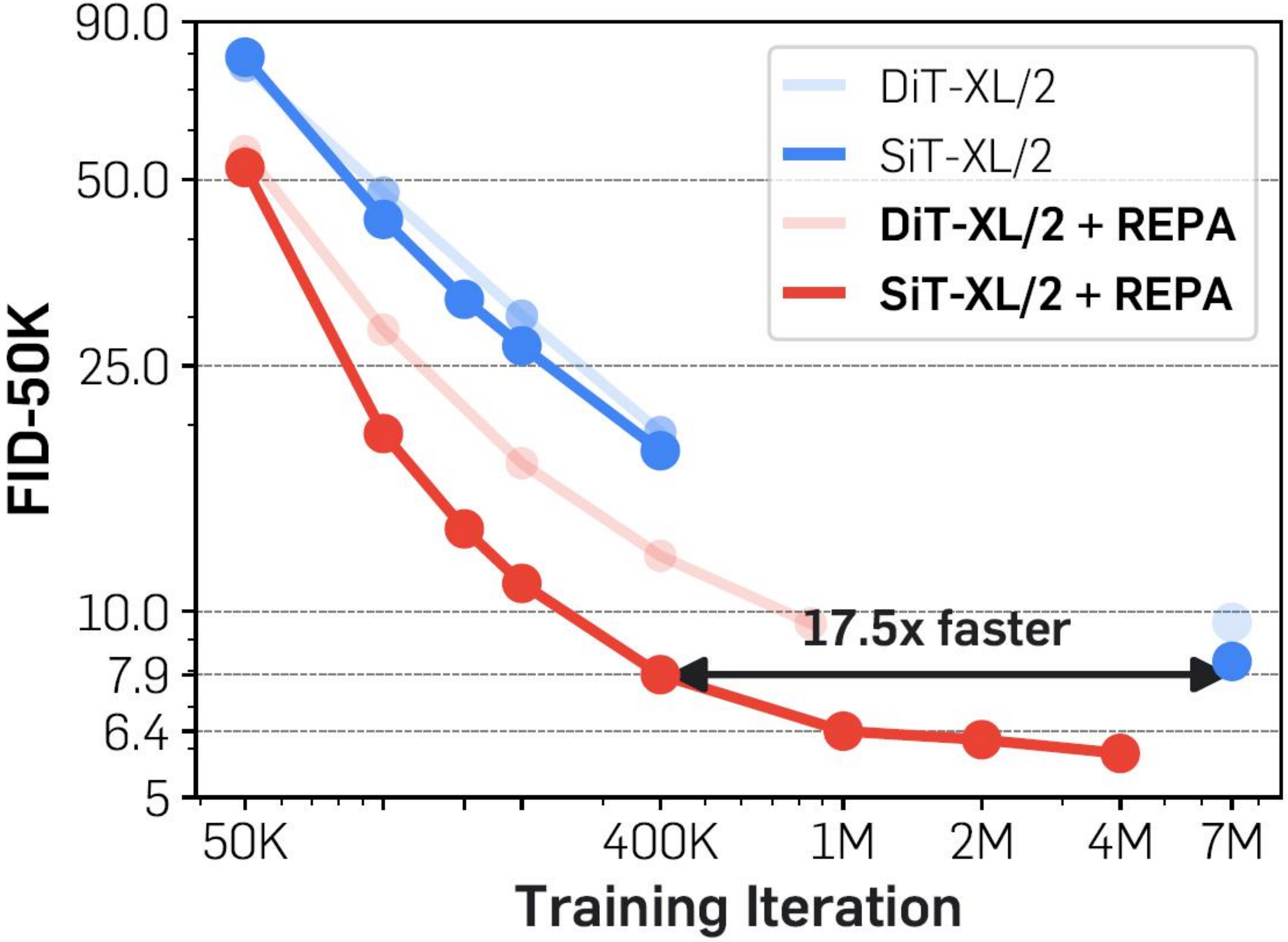
Method

REPresentation Alignment (REPA) for accelerating diffusion training.



Method

REPA speeds up the training and improves the quality significantly.



Method

REPA speeds up the training and improves the quality significantly.

Table 3: **FID comparisons with vanilla DiTs and SiTs** on ImageNet 256×256. We do not use classifier-free guidance (CFG). ↓ denotes lower values are better. Iter. indicates the training iteration.

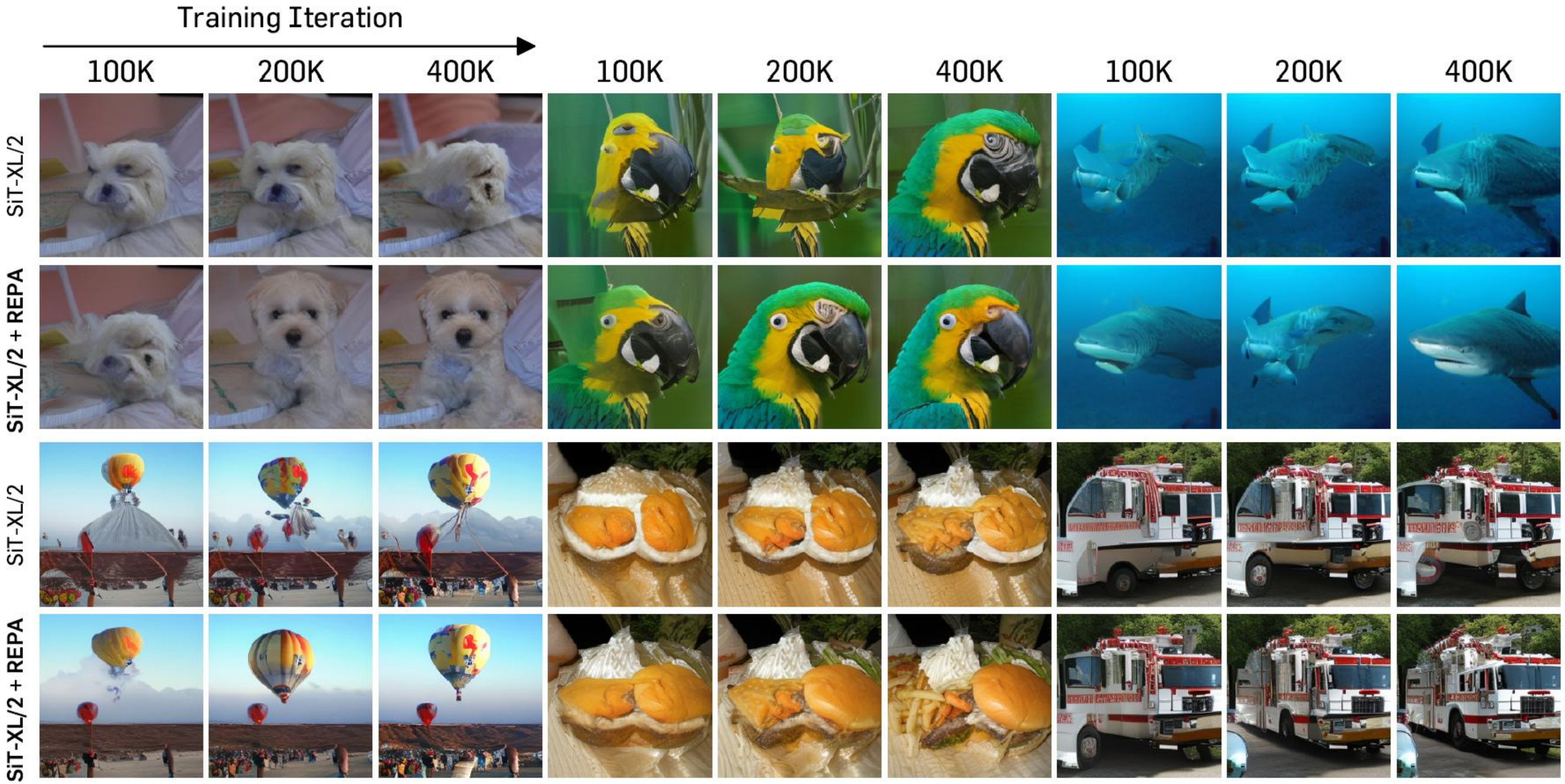
Model	#Params	Iter.	FID↓
DiT-L/2	458M	400K	23.3
+ REPA (ours)	458M	**400K**	**15.6**
DiT-XL/2	675M	400K	19.5
+ REPA (ours)	675M	**400K**	**12.3**
DiT-XL/2	675M	**7M**	9.6
+ REPA (ours)	675M	**850K**	**9.6**
SiT-B/2	130M	400K	33.0
+ REPA (ours)	130M	**400K**	**24.4**
SiT-L/2	458M	400K	18.8
+ REPA (ours)	458M	**400K**	**9.7**
+ REPA (ours)	458M	**700K**	**8.4**
SiT-XL/2	675M	400K	17.2
+ REPA (ours)	675M	**150K**	**13.6**
SiT-XL/2	675M	7M	8.3
+ REPA (ours)	675M	**400K**	**7.9**
+ REPA (ours)	675M	**1M**	**6.4**
+ REPA (ours)	675M	**4M**	**5.9**

Table 4: **System-level comparison** on ImageNet 256×256 with CFG. ↓ and ↑ indicate whether lower or higher values are better, respectively. Results that include additional CFG scheduling are marked with an asterisk (*), where the guidance interval from (Kynkäänniemi et al., 2024) is applied for REPA.

Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Pixel diffusion						
ADM-U	400	3.94	6.14	186.7	0.82	0.52
VDM++	560	2.40	-	225.3	-	-
Simple diffusion	800	2.77	-	211.8	-	-
CDM	2160	4.88	-	158.7	-	-
Latent diffusion, U-Net						
LDM-4	200	3.60	-	247.7	0.87	0.48
Latent diffusion, Transformer + U-Net hybrid						
U-ViT-H/2	240	2.29	5.68	263.9	0.82	0.57
DiffiT*	-	1.73	-	276.5	0.80	0.62
MDTv2-XL/2*	1080	1.58	4.52	314.7	0.79	0.65
Latent diffusion, Transformer						
MaskDiT	1600	2.28	5.67	276.6	0.80	0.61
SD-DiT	480	3.23	-	-	-	-
DiT-XL/2	1400	2.27	4.60	278.2	**0.83**	0.57
SiT-XL/2	1400	2.06	4.50	270.3	0.82	0.59
+ REPA (ours)	200	1.96	**4.49**	264.0	0.82	0.60
+ REPA (ours)	800	1.80	4.50	284.0	0.81	0.61
+ REPA (ours)*	**800**	**1.42**	4.70	**305.7**	0.80	**0.65**

Method

REPA speeds up the training and improves the quality significantly.



Method

The ablation studies in VE, depth and objective.

Iter.	Target Repr.	Depth	Objective	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑	Acc.↑
400K	Vanilla SiT-L/2 (Ma et al., 2024a)			18.8	5.29	72.0	0.64	0.64	N/A
400K	MAE-L	8	NT-Xent	12.5	4.89	90.7	0.68	0.63	57.3
400K	DINO-B	8	NT-Xent	11.9	5.00	92.9	0.68	0.63	59.3
400K	MoCov3-L	8	NT-Xent	11.9	5.06	92.2	0.68	0.64	63.0
400K	I-JEPA-H	8	NT-Xent	11.6	5.21	98.0	0.68	0.64	62.1
400K	CLIP-L	8	NT-Xent	11.0	5.25	100.4	0.67	0.66	67.2
400K	SigLIP-L	8	NT-Xent	10.2	5.15	107.0	0.69	0.64	68.8
400K	DINOv2-L	8	NT-Xent	10.0	5.09	106.6	0.68	0.65	68.1
400K	DINOv2-B	8	NT-Xent	9.7	5.13	107.5	0.69	0.64	65.7
400K	DINOv2-L	8	NT-Xent	10.0	5.09	106.6	0.68	0.65	68.1
400K	DINOv2-g	8	NT-Xent	9.8	5.22	108.9	0.69	0.64	65.7
400K	DINOv2-L	6	NT-Xent	10.3	5.23	106.5	0.69	0.65	66.2
400K	DINOv2-L	8	NT-Xent	10.0	5.09	106.6	0.68	0.65	68.1
400K	DINOv2-L	10	NT-Xent	10.5	5.50	105.0	0.68	0.65	68.6
400K	DINOv2-L	12	NT-Xent	11.2	5.14	100.2	0.68	0.64	69.4
400K	DINOv2-L	14	NT-Xent	11.6	5.61	99.5	0.67	0.65	70.0
400K	DINOv2-L	16	NT-Xent	12.1	5.34	96.1	0.67	0.64	71.1
400K	DINOv2-L	8	NT-Xent	10.0	5.09	106.6	0.68	0.65	68.1
400K	DINOv2-L	8	Cos. sim.	9.9	5.34	111.9	0.68	0.65	68.2

Method

The authors addressed almost all the concerns of the reviewers.

Common Response

Official Comment by Authors 22 Nov 2024, 11:53 (modified: 22 Nov 2024, 12:12) Everyone Revisions

Comment:
Dear reviewers and AC,

We sincerely appreciate your valuable time and effort spent reviewing our manuscript.

We propose REPA, a regularization for improving diffusion transformers by aligning their representations with powerful pretrained target representations. As reviewers highlighted, our method is well-motivated (K7ig, vfsY, i9uK, MZyg), simple yet effective (vfsY, i9uk), providing a new meaningful insight (XnZq, K7ig), demonstrated by strong empirical results and comprehensive experiments and analysis (all). We appreciate your constructive feedback on our manuscript. In response to the comments, we have carefully revised and enhanced the manuscript as follows:

- ImageNet 512x512 experiment (Appendix J)
- Text-to-image experiment (Appendix K)
- Feature map visualization (Appendix L)
- Dataset analysis used for training target visual encoders (Appendix C.3)
- State-of-the-art FID on ImageNet 256x256 with guidance interval (Table 2)
- Limitation and Future Work (Appendix M)
- Fixing typos (L290)

In the revised manuscript, these updates are temporarily highlighted in blue for your convenience to check.

We sincerely believe that these updates may help us better deliver the benefits of the proposed REPA to the ICLR community.

Thank you very much,
Authors.

Add: Public Comment

Discussion

Can REPA be integrated with other speeding up techniques?

JanusFlow has conducted experiments of combining:

- REPA
- Skip connection
- Reschedule the sampling of t

and demonstrated the effectiveness of employing REPA.

[10] Yiyang Ma et al. JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation, arXiv 2411.