Representation Alignment for Diffusion Transformers without External Components

Dengyang Jiang[2,1‡] Mengmeng Wang[3,1*] Liuzhuozheng Li[1] Lei Zhang[2] Haoyu Wang[2] Wei Wei[2] Guang Dai[1] Yanning Zhang[2] Jingdong Wang[4†]

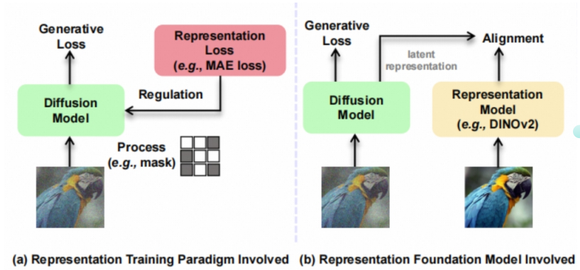
1. SGIT AI Lab, State Grid Corporation of China 2. Northwestern Polytechnical University 3. Zhejiang University of Technology 4. Baidu Inc.

Project: https://wwwjdy.github.io/sra *Code: https://github.com/vvvvjdy/SRA* *Arxiv: https://arxiv.org/abs/2505.02831*

ICLR International Conference On Learning Representations

Contact us

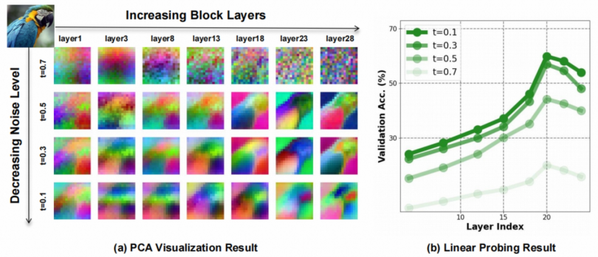
Background: External Representation Components Involved for Better Training Diffusion Transfoemer



(a) Representation Training Paradigm Involved (b) Representation Foundation Model Involved

While recent studies show that learning a meaningful internal representation can accelerate generative training, existing approaches either introducing an auxiliary external representation task or relying on a large-scale pre-trained external encoder to provide representation guidance.

Inspration and Observation: Representations in Pre-trained Diffusion Transformer

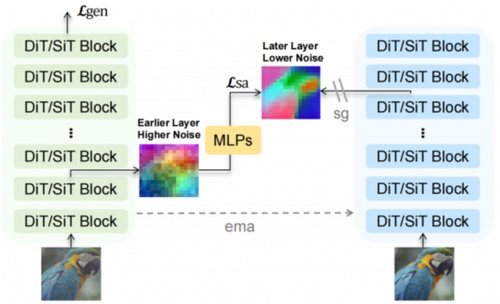


(a) PCA Visualization Result (b) Linear Probing Result

Inspration: The generative mechanism by which the diffusion model operates can be generally considered as a coarse to fine process, and how about the representations?

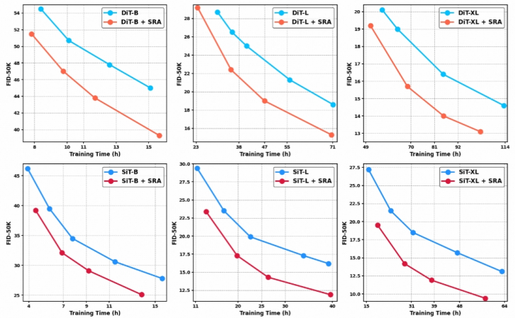
Observation: Diffusion transformer already got a roughly from bad to good discriminative process when only generative training is performed.

Method: Self-Representation Alignment (SRA)



SRA aligns the output latent representation in earlier layer with higher noise to that in later layer with lower noise to progressively enhance the overall representation learning during only generative process.

Results: Fast Convergence and Better Scalability



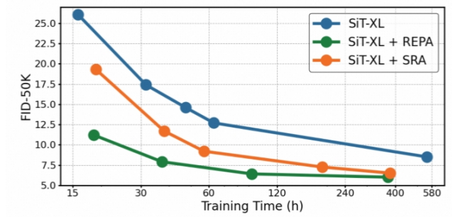
SRA shows benifit on different baselines across different model size. Meanwhile, As the size of the model increases, the gain also increases.

Future Work: Scale-up and Efficiency

Scaling up: Applying SRA to large scale T2I and T2V domain where the data distribution is more complex and do not exists an dominant representation encoder.

Efficiency optimizing: SRA requires an additional forward pass. As model size increase, this incurs increasingly computational and memory cost. Developing more efficient variants of self-representation alignment constitutes a vita direction for future research.

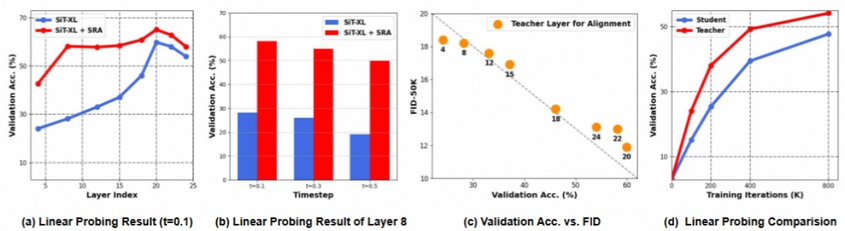
Results: Improved Generation and Representation Capacity



Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Pixel diffusion						
VDM++	-	2.65	-	278.1	-	-
Simple diffusion	800	4.28	-	171.0	-	-
Latent diffusion, Transformer						
DiT-XL/2	600	3.04	5.02	240.8	0.84	0.54
SiT-XL/2	600	2.62	4.18	252.2	**0.84**	0.57
MaskDiT	600	2.50	5.10	256.3	0.84	0.56
SiT + REPA	200	**2.08**	4.19	274.6	0.83	0.58
SiT + SRA (ours)	**200**	2.17	**4.15**	**279.3**	0.83	**0.59**

Method	FID↓	PickScore↑
ODE, NFE=50, Trained for 150K iter		
MMDiT	5.86	20.05
MMDiT + REPA	4.60	20.88
MMDiT + SRA	4.85	21.14

Model	Epochs	Tokenizer	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	LDM-VAE	3.60	-	247.7	**0.87**	0.48
Latent diffusion, Transformer							
DiT-XL/2	1400	SD-VAE	2.27	4.60	278.2	0.83	0.57
SiT-XL/2	1400	SD-VAE	2.06	**4.50**	270.3	0.82	0.59
SD-DiT	480	SD-VAE	3.23	-	-	-	-
MaskDiT	1600	SD-VAE	2.28	5.67	276.6	0.80	0.61
DiT + TREAD	740	SD-VAE	1.69	4.73	292.7	0.81	0.63
SiT + REPA	800	SD-VAE	**1.42**	4.70	305.7	0.80	**0.65**
SiT + MAETok	800	MAE-Tok	1.67	-	311.2	-	-
SiT + SRA (ours)	**400**	SD-VAE	1.85	**4.50**	297.2	0.82	0.61
SiT + SRA (ours)	**800**	SD-VAE	1.58	4.65	**311.4**	0.80	0.63



(a) Linear Probing Result (t=0.1) (b) Linear Probing Result of Layer 8 (c) Validation Acc. vs. FID (d) Linear Probing Comparision

SRA shows comparable or superior performance against other methods that leverage either auxiliary representation training task or external representation encoder on ImageNet 256/512 and COCO T2I. The representation capacity of the model is also by SRA during training.