Diffusion Transformers with Representation Autoencoders

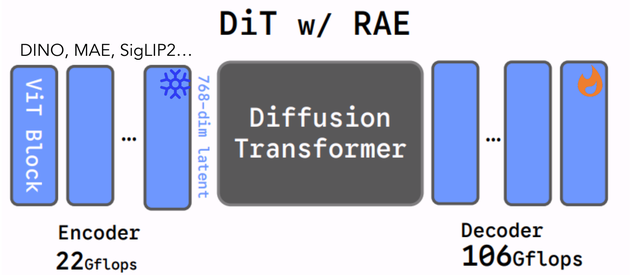
Boyang Zheng, Nanye Ma, Shengbang Tong, Saining Xie

NYU

ICLR
International Conference On Learning Representations

Semantic Representation Can be Reconstructed!

Representation Autoencoder:
Train decoder on top of frozen semantic encoders!



Standard reconstruction loss yields good reconstruction!

$$z = E(x), \hat{x} = D(z)$$

$$\mathcal{L}_{rec}(x) = \omega_L \,\mathrm{LPIPS}(\hat{x}, x) + \mathrm{L1}(\hat{x}, x) + \omega_G \lambda \,\mathrm{GAN}(\hat{x}, x),$$

Input DINOv2-B SigLIP2-B MAE-B SD-VAE

Model	rFID
DINOv2-B	0.49
SigLIP2-B	0.53
MAE-B	0.16
SD-VAE	0.62

Better Rep. Quality

Model	Top-1 Acc.
DINOv2-B	84.5
SigLIP2-B	79.1
MAE-B	68.0
SD-VAE	8.0

Larger Decoder, Better rFID

Decoder	rFID	GFLOPs
ViT-B	0.58	22.2
ViT-L	0.50	78.1
ViT-XL	0.49	106.7
SD-VAE	0.62	310.4

Works for different size encoders

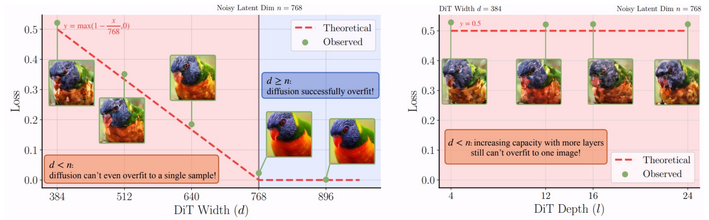
Encoder	rFID
DINOv2-S	0.52
DINOv2-B	0.49
DINOv2-L	0.52

Train Diffusion Transformers on High Dimensional Representation!

Naïve Flow Matching fails...

	RAE	SD-VAE
DiT-S	215.76	**51.74**
DiT-XL	23.08	**7.13**

Findings #1: DiT must be wider than RAE tokens

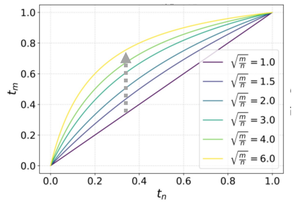


Findings #2: High-dimension diffusion requires noise schedule shift

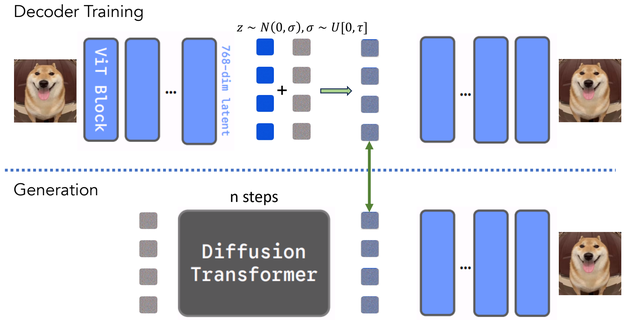
Same Noise std: 0.4

Becomes clearer and clearer when resolution(dimension) increases!

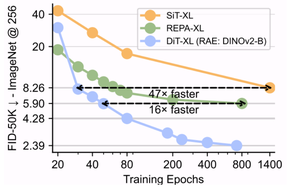
$$t_{RAE} = \frac{s \times t_{VAE}}{1 + (s-1)\, t_{VAE}}, s^2 = \frac{Dim_{RAE}}{Dim_{VAE}} = \frac{16\times16\times n}{32\times32\times4}$$



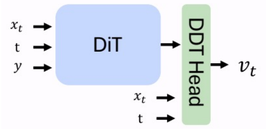
Findings #3: Add noise in RAE decoder training



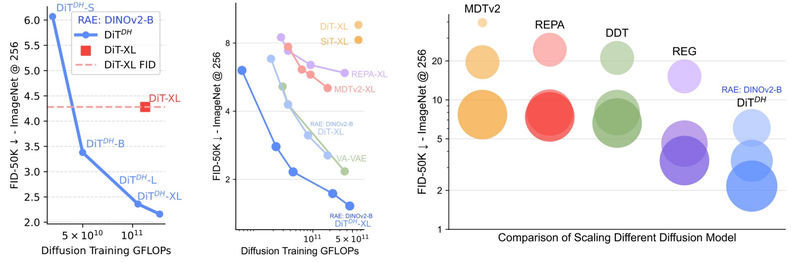
1+2+3 = Very fast convergence!



Even faster: DiT w/ wide DDT Head



Diffusion in Representation Space Converges Much Faster!



Method	Epochs	#Params	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Latent Diffusion with RAE (Ours)										
DiT-XL (DINOv2-S)	800	676M	1.87	209.7	0.80	0.63	1.41	309.4	0.80	0.63
DiT^{DH}-XL (DINOv2-B)	20	839M	3.71	198.7	**0.86**	0.50	–	–	–	–
	80		2.16	214.8	0.82	0.59	–	–	–	–
	800		**1.51**	**242.9**	0.79	0.63	**1.13**	262.6	0.78	**0.67**