CVPR
SEATTLE, WA JUNE 17-21, 2024

SD-DiT: Unleashing the Power of Self-supervised Discrimination in Diffusion Transformer

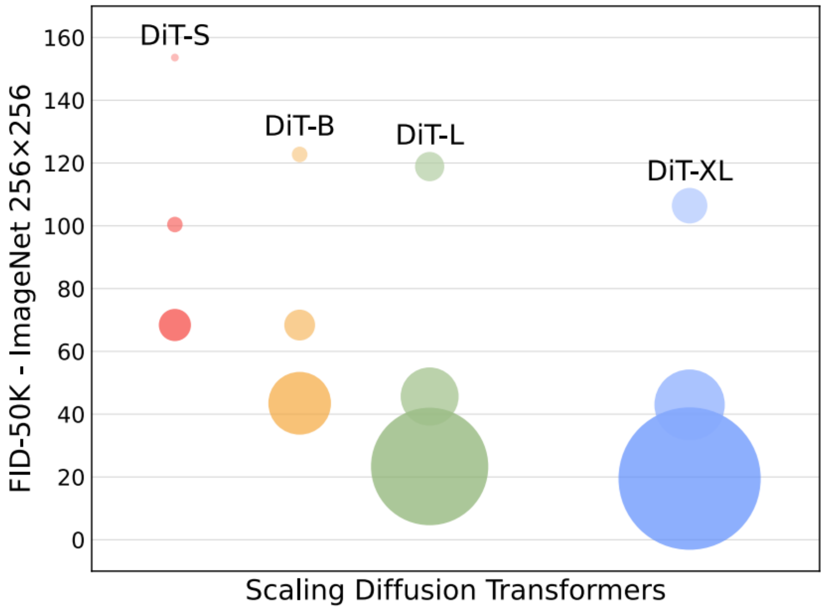
Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, Chang Wen Chen

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

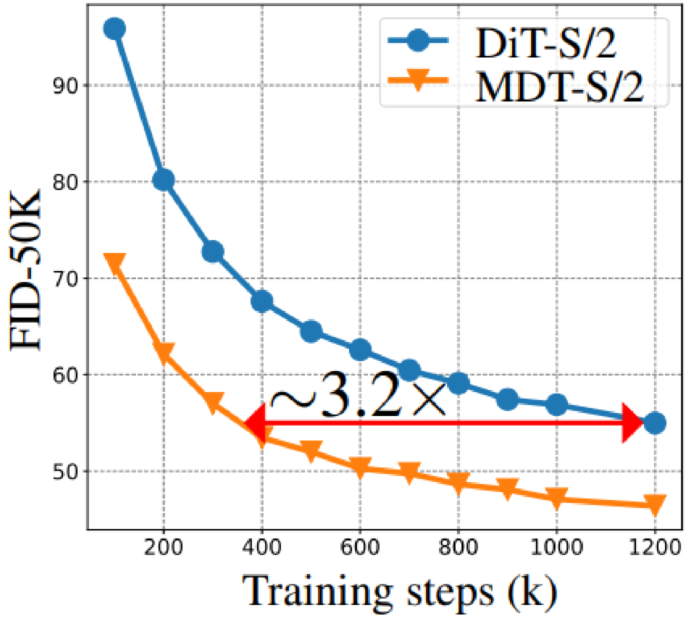
智象未来
HiDream.ai

The Hong Kong Polytechnic University
香港理工大學

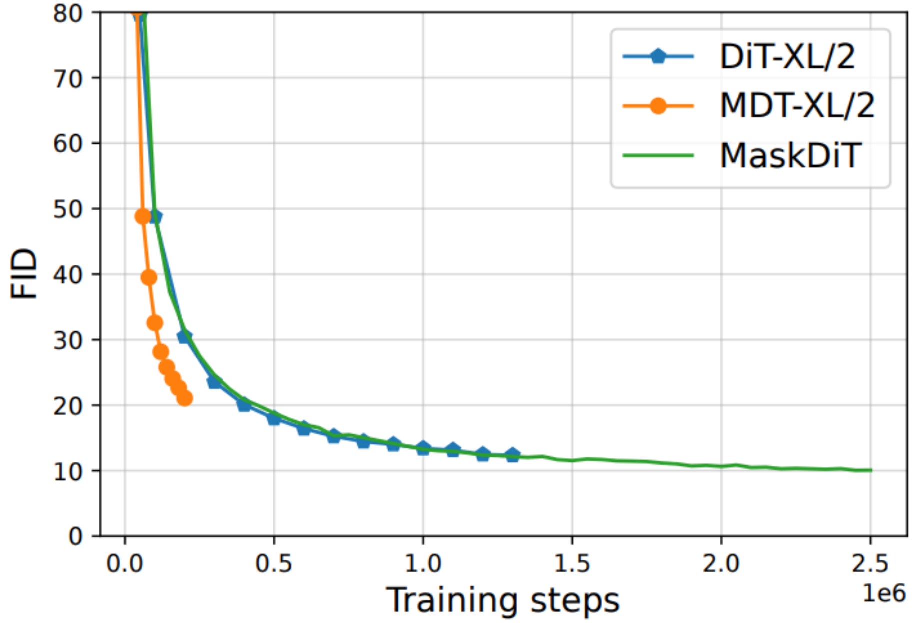
Diffusion Transformer - Scalable but Slow Convergence



DiT: Scalable
Arxiv 2022.12



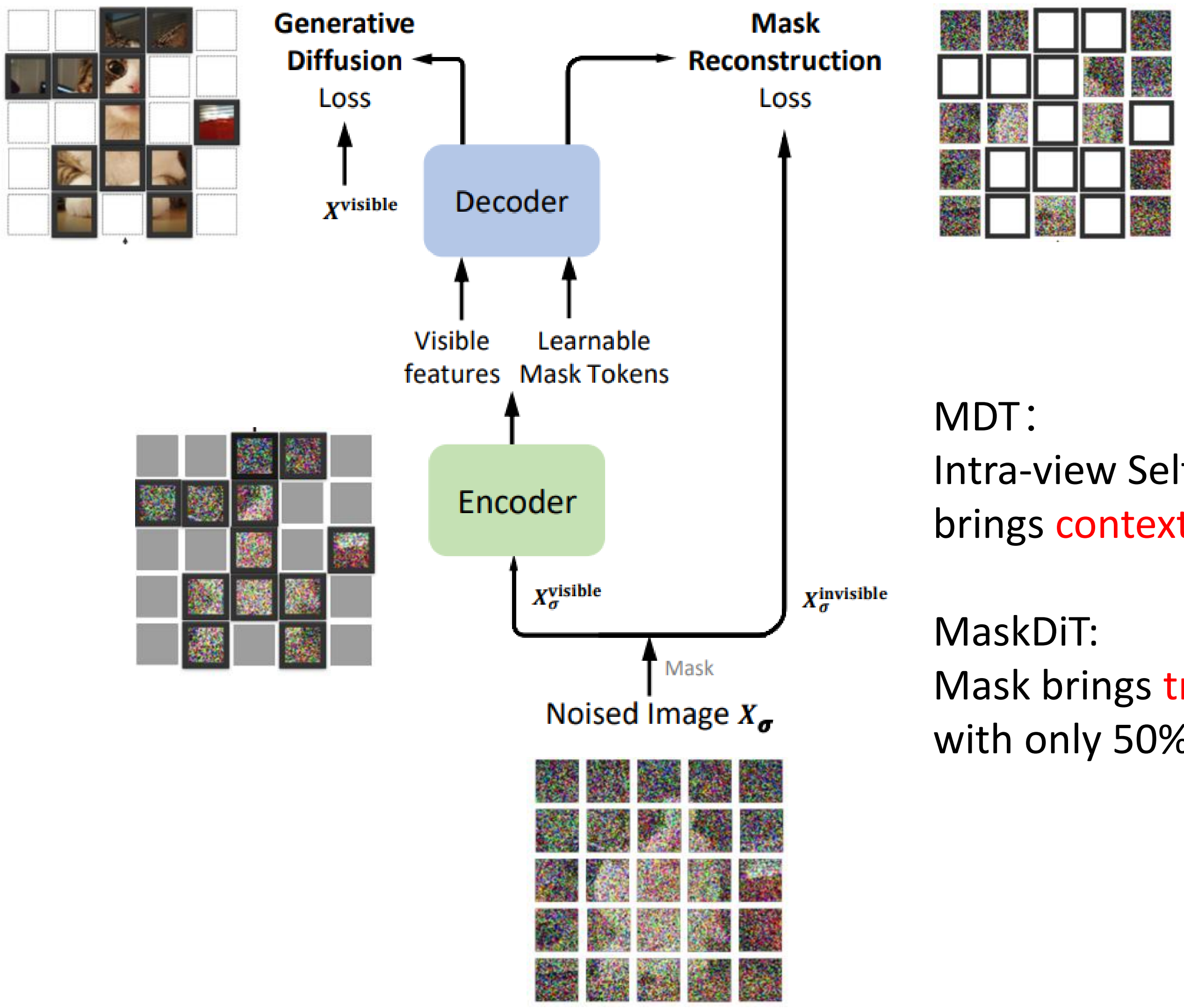
MDT: Fast Convergence
Arxiv 2023.03



MaskDiT: Efficiency
Arxiv 2023.06

[1] Peebles, William, and Saining Xie. "Scalable diffusion models with transformers." ICCV 2023.
[2] Gao, Shanghua, et al. "Masked diffusion transformer is a strong image synthesizer." ICCV 2023.
[3] Zheng, Hongkai, et al. "Fast training of diffusion models with masked transformers." TMLR 2024.

Why Bring Mask to DiT – Contextual Relation inside View



MDT：
Intra-view Self-Reconstruction via mask brings contextual relation learning

MaskDiT:
Mask brings training efficiency with only 50% input

[1] Gao, Shanghua, et al. "Masked diffusion transformer is a strong image synthesizer." ICCV 2023.
[2] Zheng, Hongkai, et al. "Fast training of diffusion models with masked transformers." TMLR 2024.

Can we impose Inter-view Discrimination to DiT?

HiDream.ai

Representation Learning

Generative Modeling

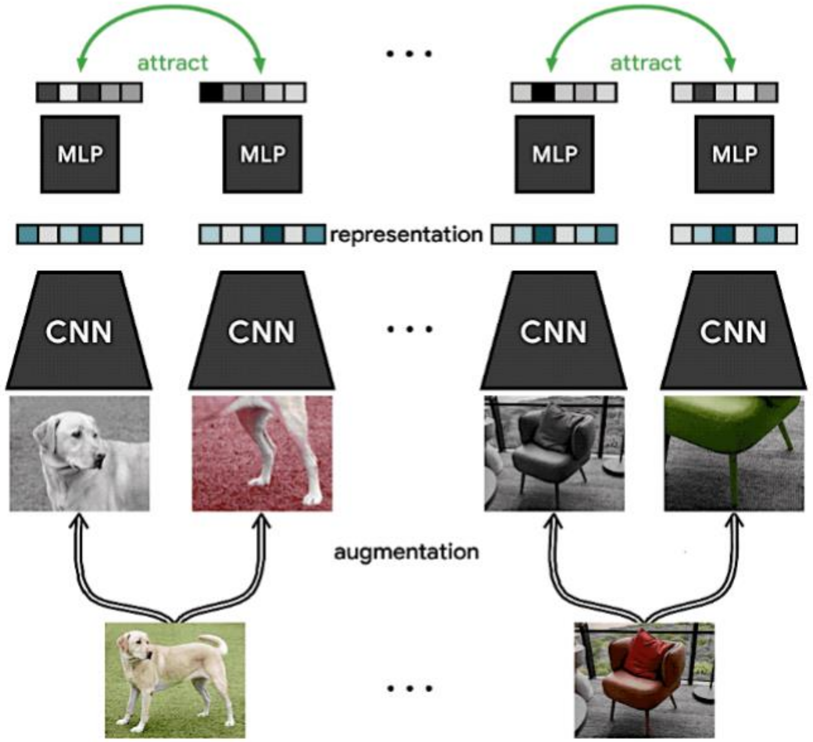
How to construct discriminative view pair for Generative Diffusion Transformer?

Contrastive Self-supervised Learning relies on Data Augmentation for positive pair

[1] https://github.com/google-research/simclr

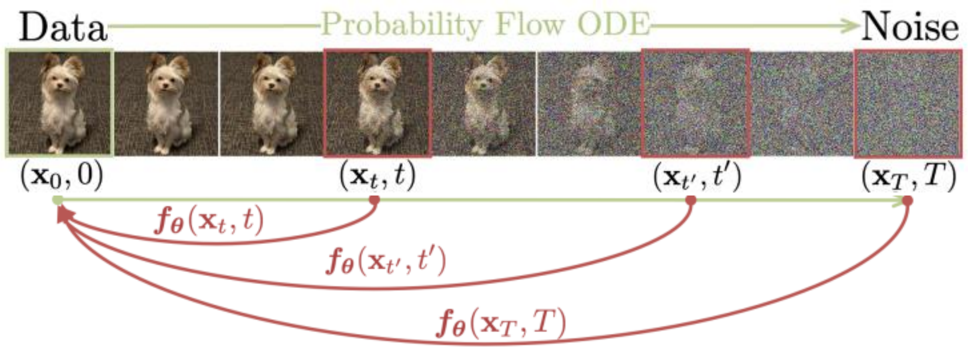
Can we impose Inter-view Discrimination to DiT?

Representation Learning



$$p_{data \odot Aug} \rightarrow p_{data}$$

Generative Modeling



Inspired by Consistency models, whose outputs of the points on the same PF-ODE trajectory are consistent

$$\boldsymbol{f}(\boldsymbol{x}_\sigma, \sigma) = \boldsymbol{f}(\boldsymbol{x}_{\sigma'}, \sigma'), \quad \sigma, \sigma' \in [\sigma_{\min}, \sigma_{\max}].$$

$$\boldsymbol{f} : (\boldsymbol{x}_\sigma, \sigma) \mapsto \boldsymbol{x}_{\sigma_{\min}}$$

We construct discriminative pair by adding noise $(x_{\sigma_S}, x_{\sigma_T})$ on the same PF-ODE

$$p_{\sigma_S} \rightarrow p_{\sigma_T}$$

[1] https://github.com/google-research/simclr

[2] Song, Yang, Dhariwal Prafulla, Chen Mark, Sutskever Ilya. "Consistency models." ICML 2023.

Preliminary

PF-ODE

$$\mathrm{d}\boldsymbol{x}_t = [\boldsymbol{\mu}(\boldsymbol{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)]\mathrm{d}t.$$

EDM utilizes $p_\sigma(\boldsymbol{x})$ instead of $p_t(\boldsymbol{x})$

$$\boldsymbol{\mu}(\boldsymbol{x}, t) := \mathbf{0} \text{ and } g(t) := \sqrt{2t}$$

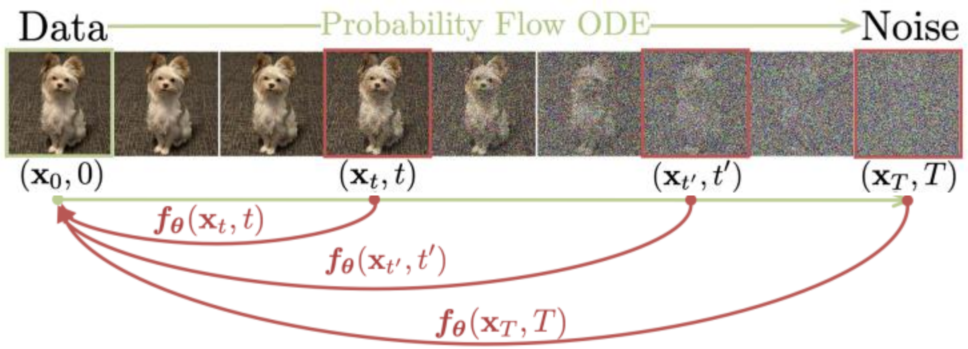
PF-ODE in EDM

$$d\boldsymbol{x} = -\sigma \nabla_{\boldsymbol{x}} \log p_\sigma(\boldsymbol{x}) d\sigma, \quad \sigma \in [\sigma_{\min}, \sigma_{\max}],$$

$$p_\sigma(\boldsymbol{x}) = p_{\text{data}}(\boldsymbol{x}) * \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\boldsymbol{x}_\sigma = \boldsymbol{x}_0 + \boldsymbol{n}, \ \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

Generative Modeling



Consistency models, whose outputs of the points on the same PF-ODE trajectory are consistent

$$\boldsymbol{f}(\boldsymbol{x}_\sigma, \sigma) = \boldsymbol{f}(\boldsymbol{x}_{\sigma'}, \sigma'), \quad \sigma, \sigma' \in [\sigma_{\min}, \sigma_{\max}].$$

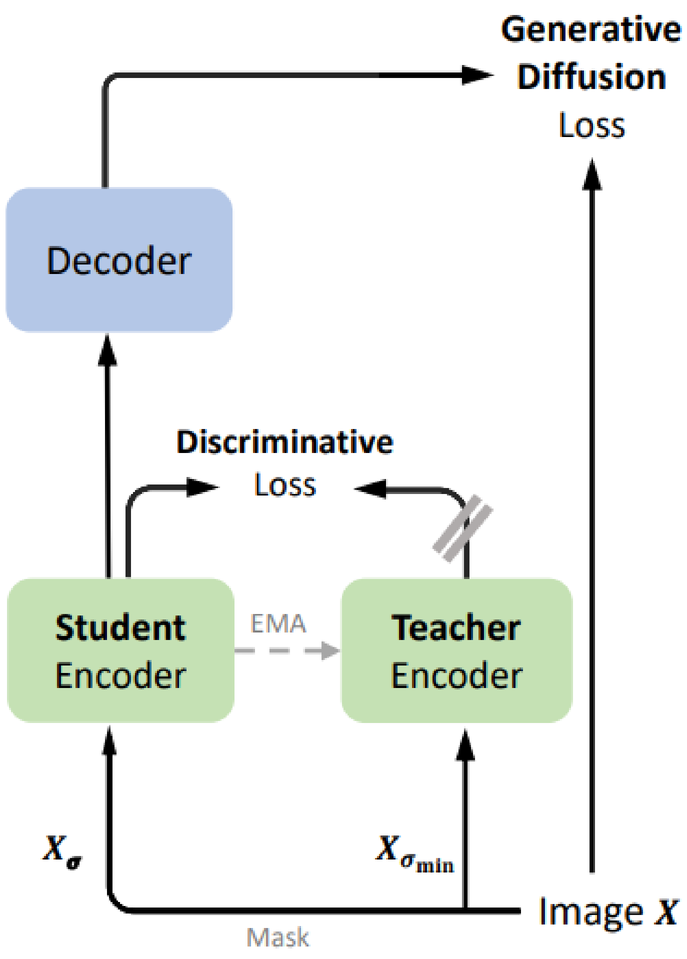
$$\boldsymbol{f} : (\boldsymbol{x}_\sigma, \sigma) \mapsto \boldsymbol{x}_{\sigma_{\min}}$$

We construct discriminative pair by adding noise $(x_{\sigma_\mathrm{S}}, x_{\sigma_\mathrm{T}})$ on the same PF-ODE

$$p_{\sigma_\mathrm{S}} \rightarrow p_{\sigma_\mathrm{T}}$$

[1] Song, Yang, Dhariwal Prafulla, Chen Mark, Sutskever Ilya. "Consistency models." ICML 2023.
[2] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." ICLR2021.
[3] Karras, Tero, et al. "Elucidating the design space of diffusion-based generative models." NeurIPS 2022.

SD-DiT: Discriminative Objective



(a) SD-DiT

$$\boldsymbol{x}_{\sigma_{\mathrm{S}}} = \boldsymbol{x}_0 + \boldsymbol{n},\ \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathrm{S}}^2 \mathbf{I}),\ \ \sigma_{\mathrm{S}} \in [\sigma_{\min}, \sigma_{\max}]$$

$$\boldsymbol{x}_{\sigma_{\mathrm{T}}} = \boldsymbol{x}_0 + \boldsymbol{n},\ \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\min}^2 \mathbf{I})$$

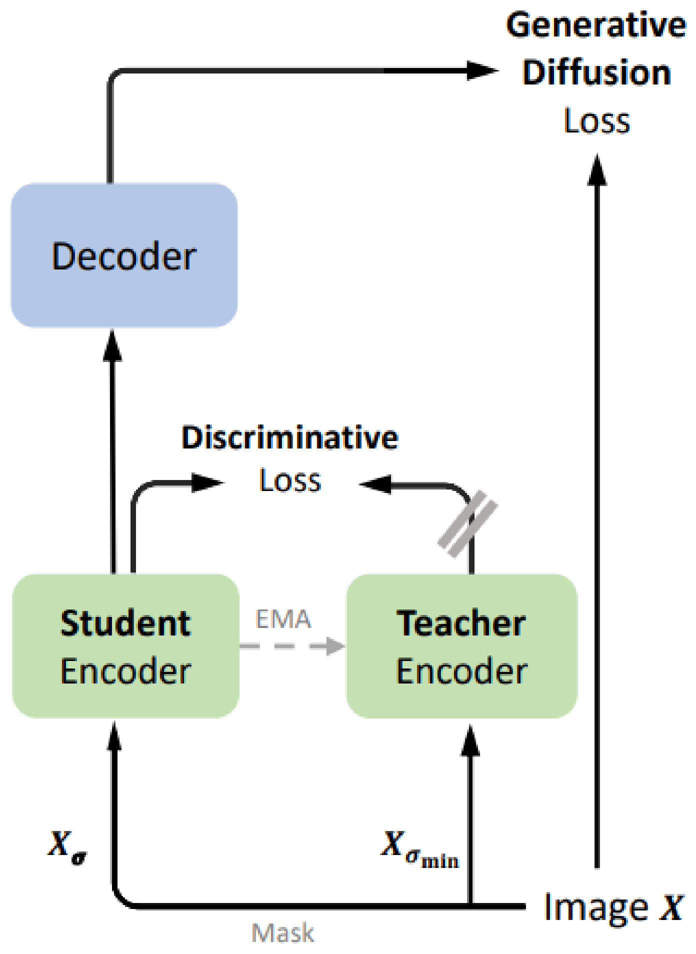
$$P_{\mathrm{S}_i} = \frac{\exp(j_\theta(\boldsymbol{e}_{\mathrm{S}_i})/\tau_{\mathrm{S}})[k]}{\sum_{k=1}^{K} \exp(j_\theta(\boldsymbol{e}_{\mathrm{S}_i})/\tau_{\mathrm{S}})[k]},$$

$$\mathcal{L}_{\mathrm{D}}(i) = -\sum_k P_{\mathrm{T}_i} \log(P_{\mathrm{S}_i}).$$

Loss on visible tokens and CLS token:

$$\mathcal{L}_{\mathrm{D}} = \frac{1}{(1-\mathcal{M})} \sum_{i \in (1-\mathcal{M})} \mathcal{L}_{\mathrm{D}}(i) + \mathcal{L}_{\mathrm{D}}(\texttt{[CLS]}).$$

Various Teacher Noise in Discriminative Pair



(a) SD-DiT

$$\boldsymbol{x}_{\sigma_S} = \boldsymbol{x}_0 + \boldsymbol{n},\ \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_S^2 \mathbf{I}),\ \ \sigma_S \in [\sigma_{\min}, \sigma_{\max}]$$

$$\boldsymbol{x}_{\sigma_T} = \boldsymbol{x}_0 + \boldsymbol{n},\ \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \sigma_{\min}^2 \mathbf{I})$$

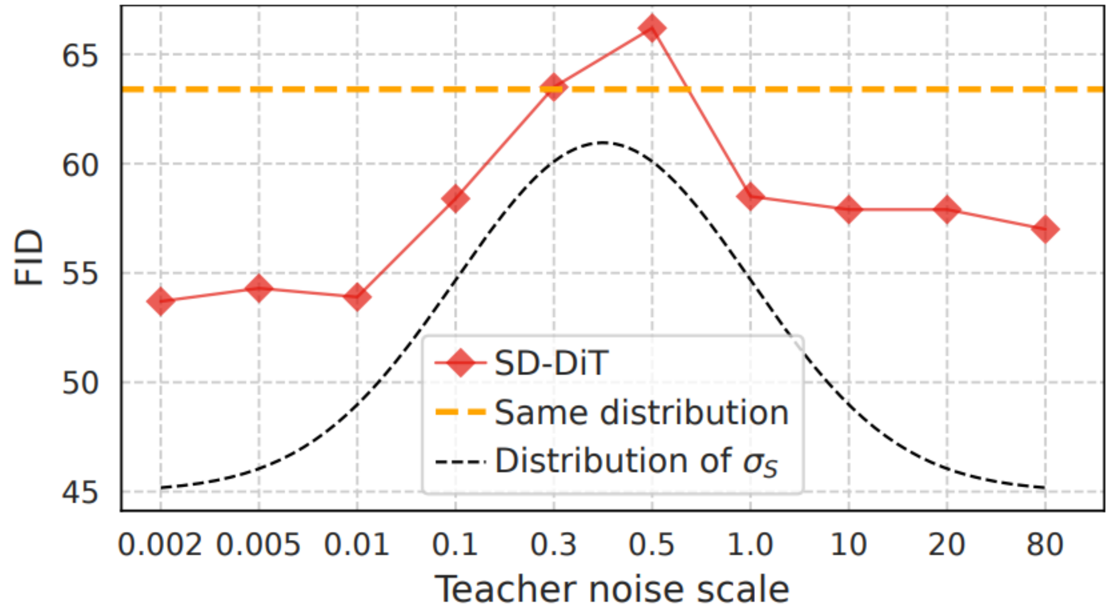
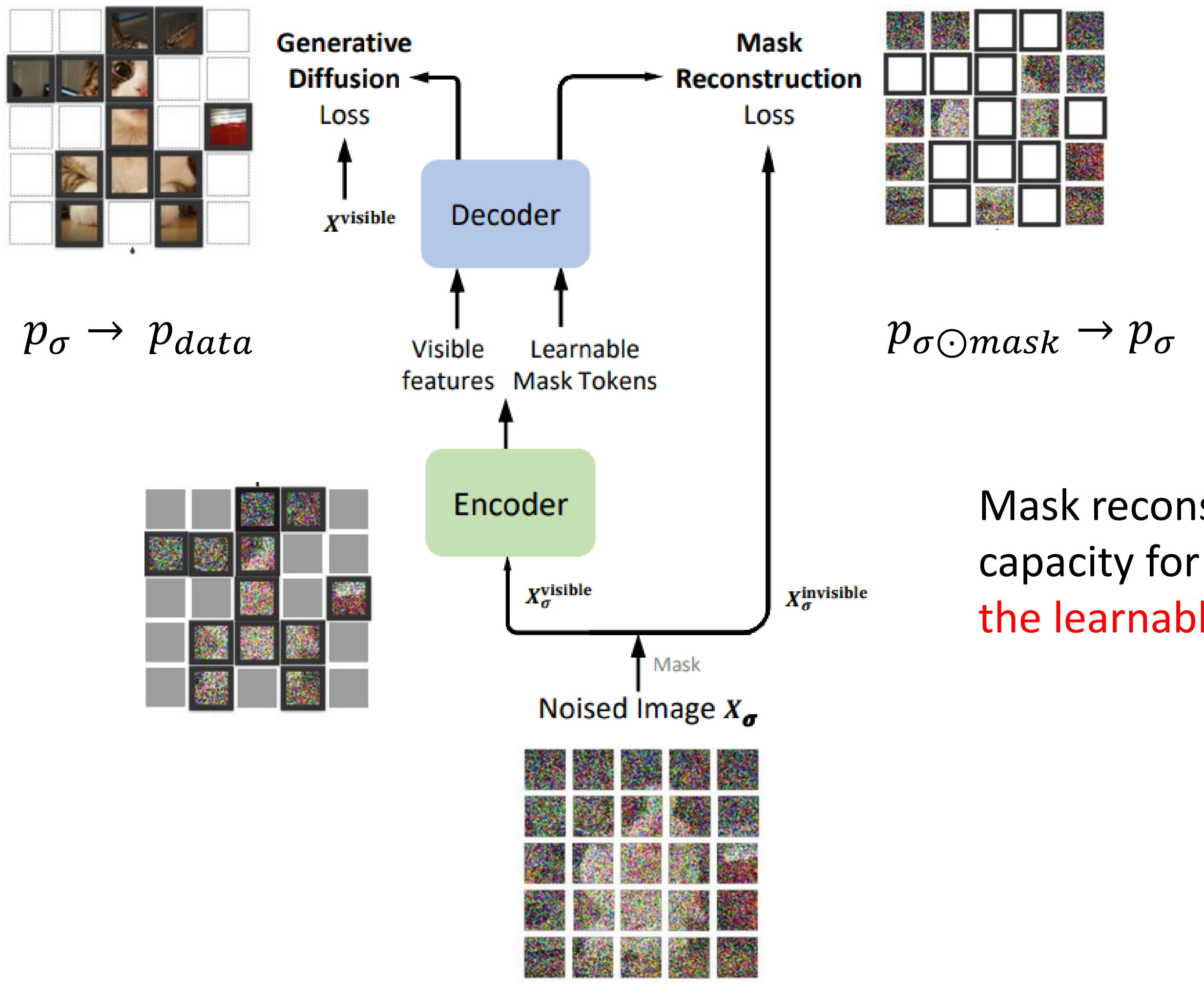


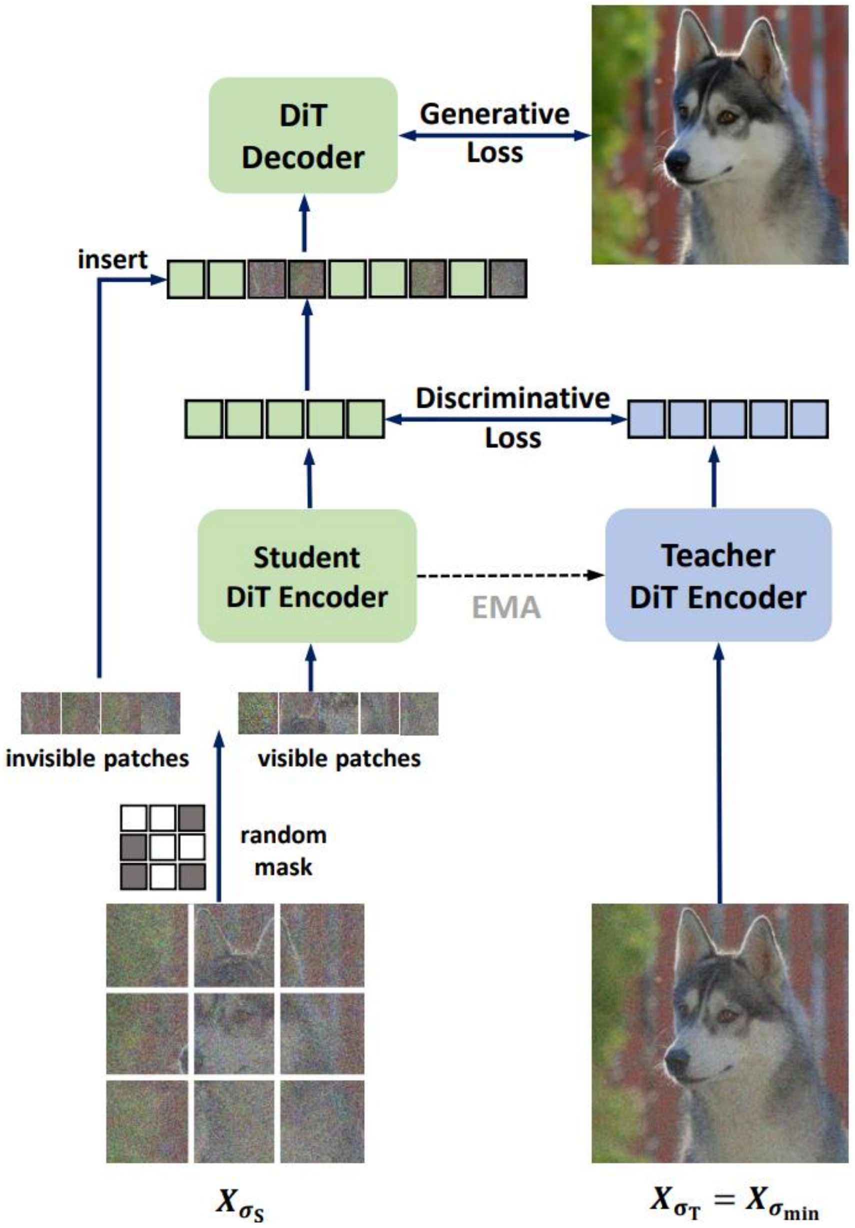
Figure 6. FID *vs.* teacher noise on SD-DiT-S/2 with 400k steps.

Fuzzy relations: Mask Reconstruction vs. Generative Diffusion



Mask reconstruction loss wastes model capacity for representation learning and the learnable mask tokens.

SD-DiT: Decoupled Encoder-decoder w/o mask tokens



- Decoder for generative loss: $p_\sigma \to p_{data}$
- Encoder for discriminative loss: $p_\sigma \to p_{\min}$
- Keep masks for training efficiency and location contextual awareness.
- Remove the mask reconstruction loss $p_{\sigma \odot mask} \to p_\sigma$

(which wastes model capacity for representation learning)

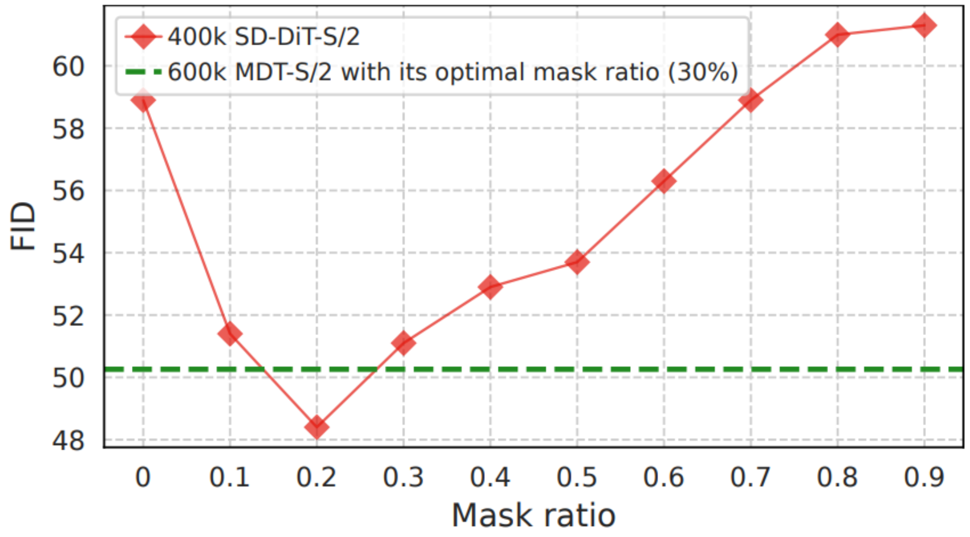


Figure 5. FID *vs.* mask ratio on SD-DiT-S/2 with 400k steps.

Experiments on ImageNet: Fast Convergence

HiDream.ai

Method	Training Steps(k)	FID-50K↓
DiT-S/2 [45]	400	68.40
MDT-S/2 [19]	400	53.46
SD-DiT-S/2	400	**48.39**
DiT-B/2 [45]	400	43.47
MDT-B/2 [19]	400	34.33
SD-DiT-B/2	400	**28.62**
DiT-XL/2 [45]	7000	9.62
MaskDiT-XL/2 [73]	1300	12.15
MDT-XL/2 [19]	1300	9.60
SD-DiT-XL/2	1100	9.66
SD-DiT-XL/2	1300	**9.01**

Table 1. Performance comparison with state-of-the-art DiT-based approaches under various model sizes on ImageNet 256×256 for class-conditional image generation (batch size: 256).

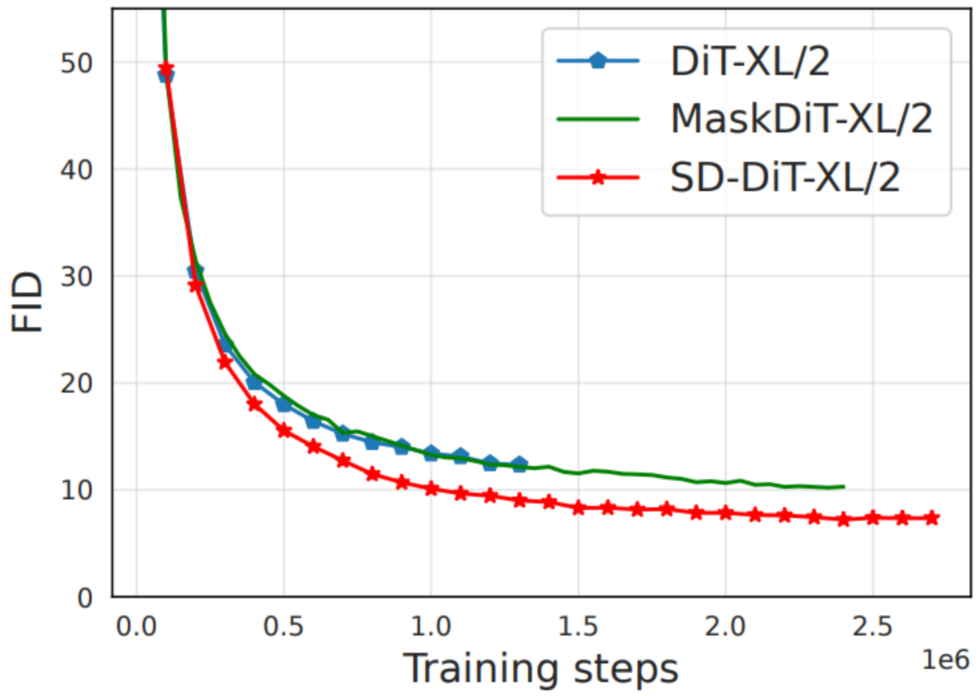


Figure 4. Comparison of convergence speed with SOTA DiT-based approaches in DiT-XL backbone (batch size: 256). The results of DiT and MaskDiT are directly cited from MaskDiT [81]. Our SD-DiT-XL/2 consistently outperforms DiT-XL/2 and MaskDiT-XL/2 across training steps, leading to better training convergence.

Experiments on ImageNet: Compare with SOTAs

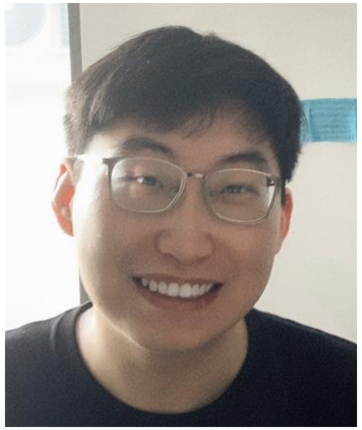
HiDream.ai

Method	Cost(Iter×BS)	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
VQGAN [16]	-	15.78	78.3	-	-	-
BigGAN-deep [5]	-	6.95	7.36	171.4	0.87	0.28
StyleGAN [57]	-	2.30	4.02	265.12	0.78	0.53
I-DDPM [43]	-	12.26	-	-	0.70	0.62
MaskGIT [9]	1387k×256	6.18	-	182.1	0.80	0.51
CDM [29]	-	4.88	-	158.71	-	-
ADM [14]	1980k×256	10.94	6.02	100.98	0.69	0.63
ADM-U [14]	—	7.49	5.13	127.49	0.72	0.63
LDM-8 [50]	4800k×64	15.51	-	79.03	0.65	0.63
LDM-4 [50]	178k×1200	10.56	-	103.49	0.71	0.62
MaskDiT-XL/2 [73]	2000k×**1024**	5.69	10.34	177.99	0.74	0.60
DiT-XL/2 [45]	7000k×256	9.62	6.85	121.50	0.67	**0.67**
MDT-XL/2 [19]	2500k×256	7.41	**4.95**	121.22	0.72	0.64
SD-DiT-XL/2	2400k×256	**7.21**	5.17	**144.68**	**0.72**	0.61

Table 2. Performance comparison with state-of-the-art methods on ImageNet 256×256 for class-conditional image generation. Similar to most DiT-based approaches, here we report the results of our SD-DiT in DiT-XL backbone with 256 batch size, while MaskDiT reports results with the largest batch size (1024).

Thanks for Listening!

HiDream.ai



Rui Zhu
The Chinese University of Hong Kong, Shenzhen

Yingwei Pan
HiDream.ai

Yehao Li
HiDream.ai

Ting Yao
HiDream.ai

Zhenglong Sun
The Chinese University of Hong Kong, Shenzhen

Tao Mei
HiDream.ai

Chang Wen Chen
The Hong Kong Polytechnic University

Any question: ruizhu@link.cuhk.edu.cn