T2MD: Diffusion Transformer-to-Mamba Distillation for High-Resolution Image Generation

Yuan Yao[1,2] Yicong Hong[2] Difan Liu[2] Mai Long[2] Feng Liu[2] Jiebo Luo[1]

[1]University of Rochester [2]Adobe Research

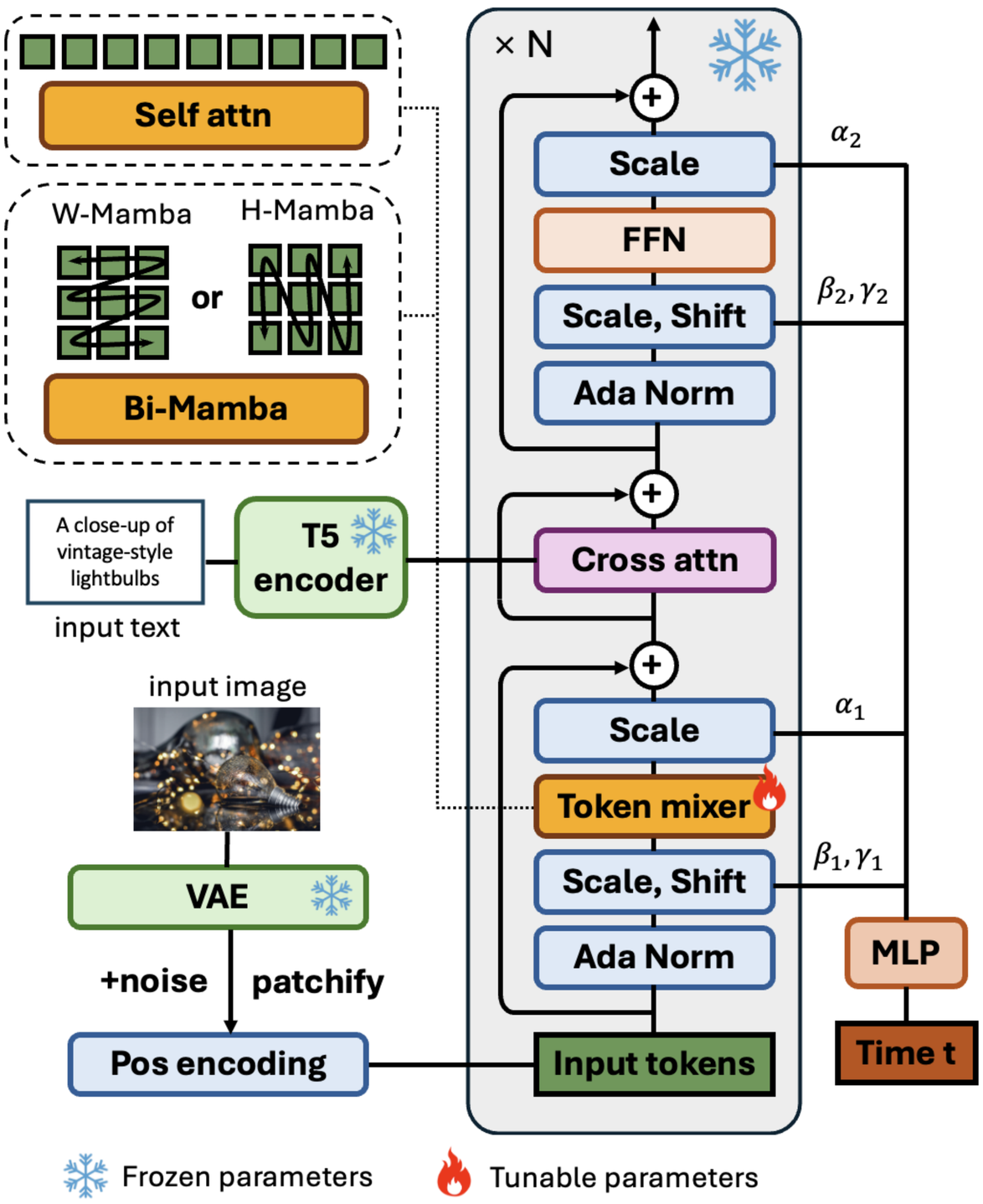
Quadratic complexity of DiT models are bad!

Let's solve it with linear-complexity model Mamba.

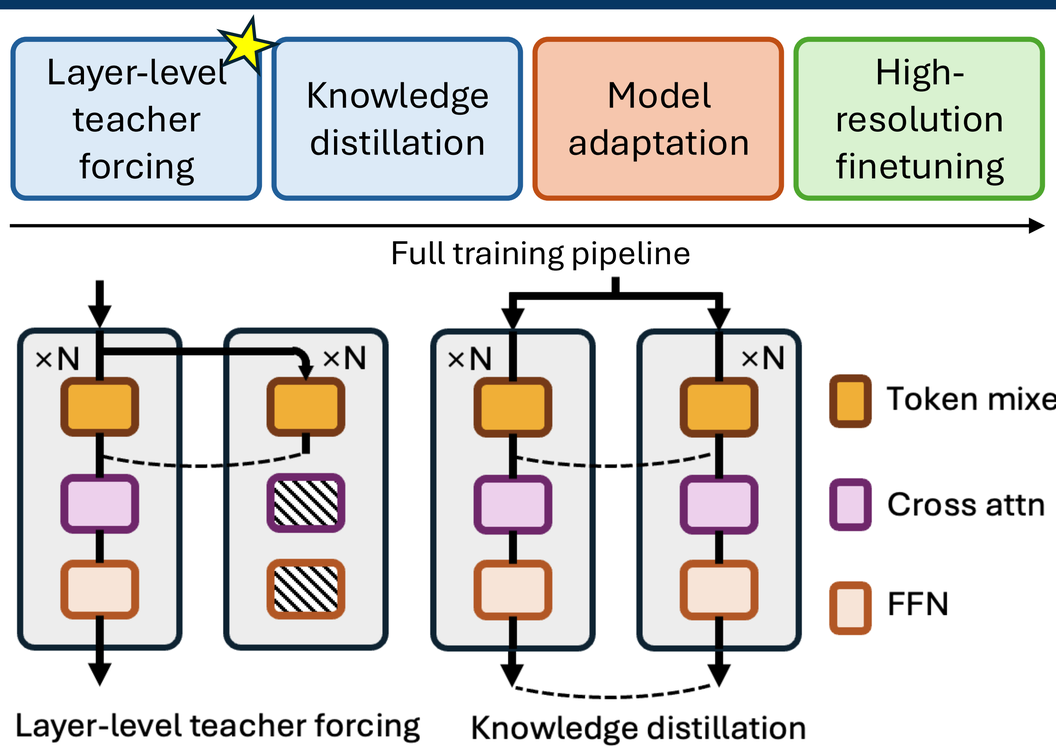
But Mamba models are difficult to train?

We build an **efficient** and **effective** training path for the Mamba-based model toward high-resolution text-to-image generation.

The model



How to effectively train?



Evaluation

Method	GenEval↑
SDv1.5	0.427
PixArt-α *(upper-bound)*	0.481
baseline hybrid Mamba	0.301
(+non-Mamba weight init)	0.397
(+$\mathcal{L}_{soft}$)	0.452
(+$\mathcal{L}_{feature}$)	0.462
(+teacher forcing) [**T2MD**]	**0.485**
Causal-to-causal Mamba init	0.433
Bi-dir → Uni-dir	0.448
No SA	0.420

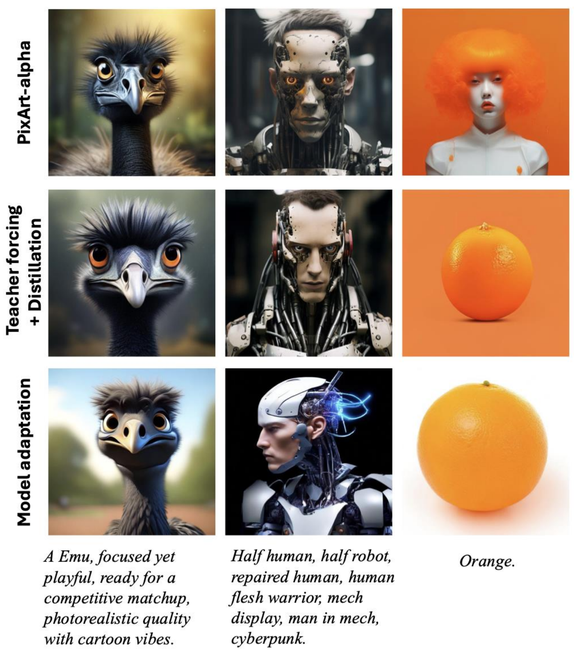
Ablation and Comparison. T2MD greatly improves the GenEval score by 0.184, and is comparable with the teacher model.

Method	#Params	FID-30K ↓
LDM	1.4B	12.64
DALLE 2	6.5B	10.39
SDv1.5	0.9B	9.62
Dimba	0.9B	8.93
LinFusion	0.9B	12.57
PixArt-α *(teacher)*	0.6B	7.32
T2MD (ours)	0.7B	8.63

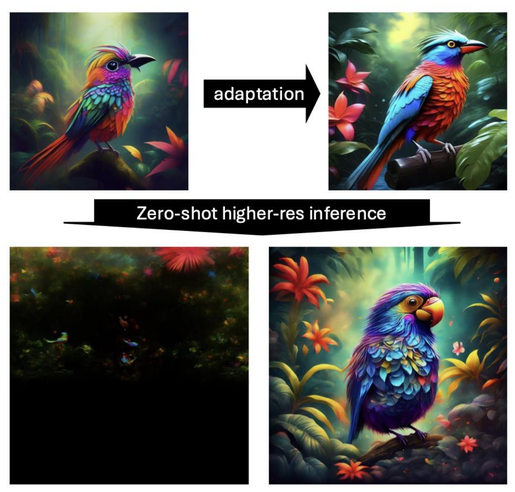
Comparison under MS-COCO 2014.

resolution	latency	SA latency	speedup
2048×2048	2.8s	4.2s	1.5×
3840×2160	6.5s	13.6s	2.1×

Speedup. T2MD achieves 1.5×-2.1× speedup in high-resolution generation.



Effects of different 512×512 stages



Our Mamba model has zero-shot capabilities after model adaptaion