DIC: RETHINKING CONV3X3 DESIGNS IN DIFFUSION MODELS

TL;DR: Fully 3x3 Convolutional Diffusion Models WORK!

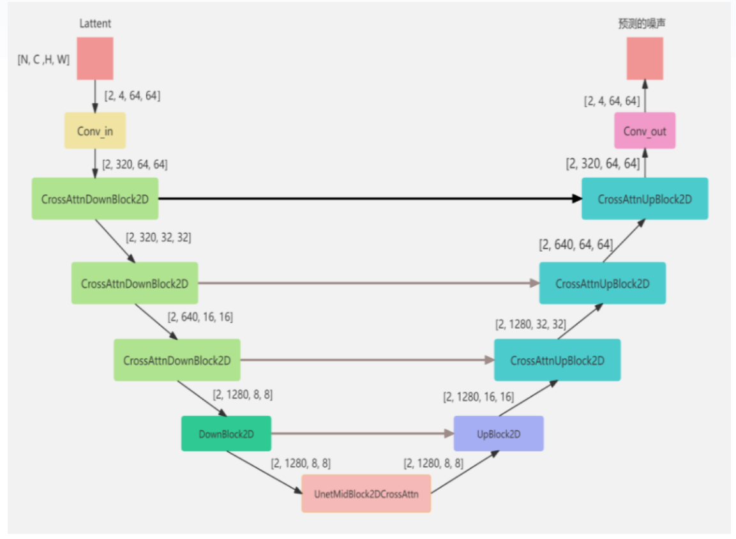
Yuchuan Tian, Jing Han*, Chengcheng Wang, Yuchen Liang, Chao Xu, Hanting Chen*

SIST Peking University, BUPT, Noah's Ark Lab

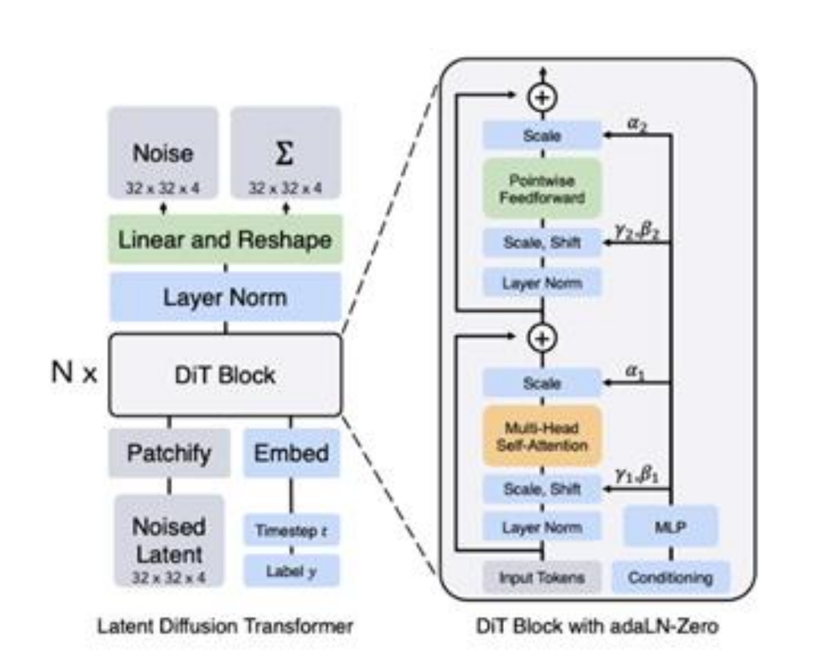
北京大学 PEKING UNIVERSITY

CVPR Nashville JUNE 11-15, 2025

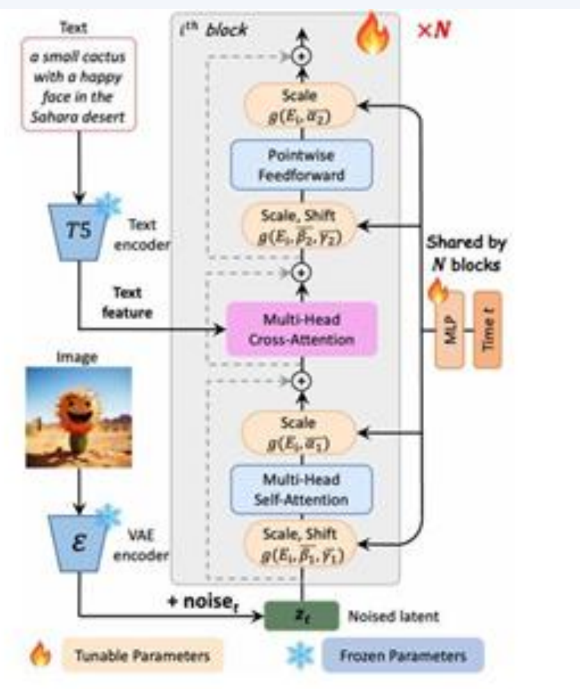
Current Trend: Diffusion Transformers



SongUNet



DiT



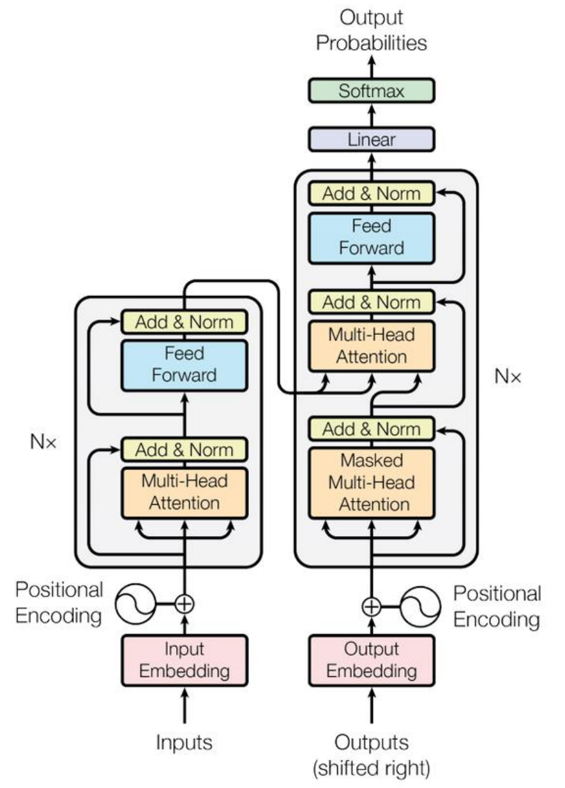
PixArt

All of the models above have self-attention…

- Low latency
- $O(N^2)$ Complexity

=> SLOW 🐌

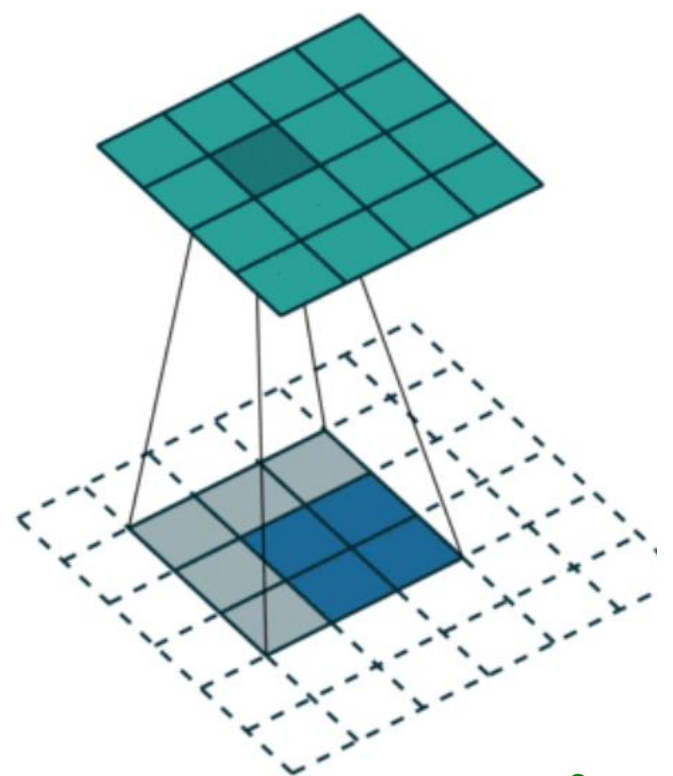
Self-Attention ❌



MAKE CONVS GREAT AGAIN

In Diffusion!

👊 🌀 🔥



Conv 3x3 ✓

Our Aim: A Conv3x3 Diffusion Model

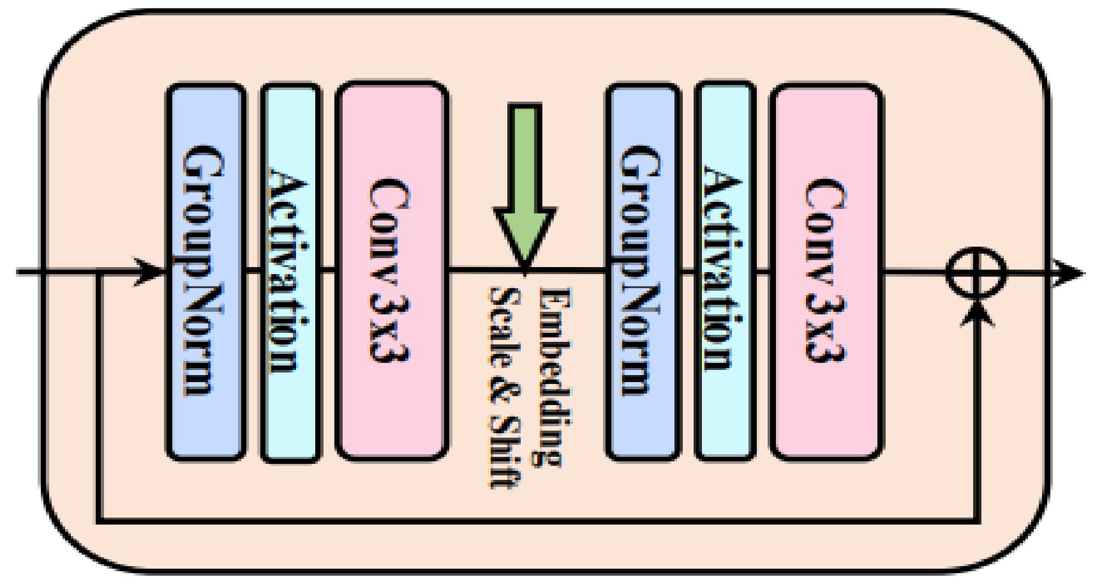
Conv3x3 Denoisers could match the performance of DiTs while maintaining a speed advantage.

➢ Macro & Micro-level design improvements

Improvements on Conv3x3 "Basic Blocks"

Starting from a "Basic Block" from U-Net:

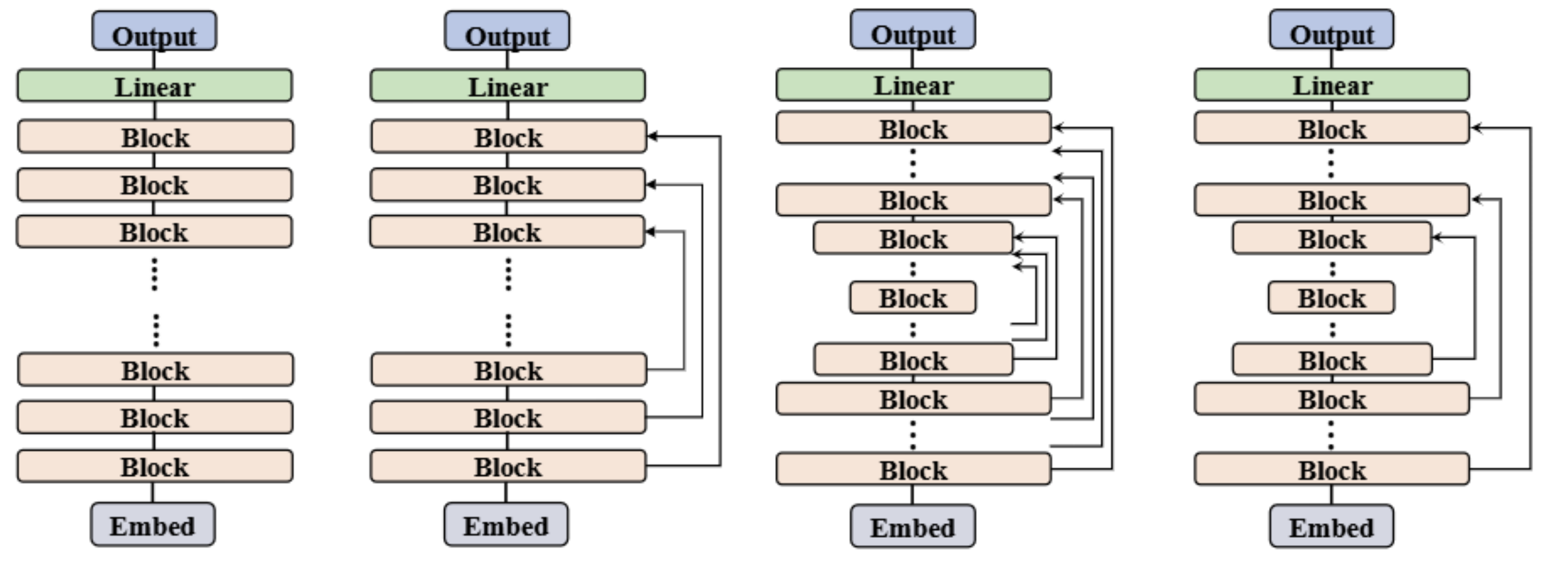
- Two **Conv3x3**s
- Residual Connection
- Removal of Attn



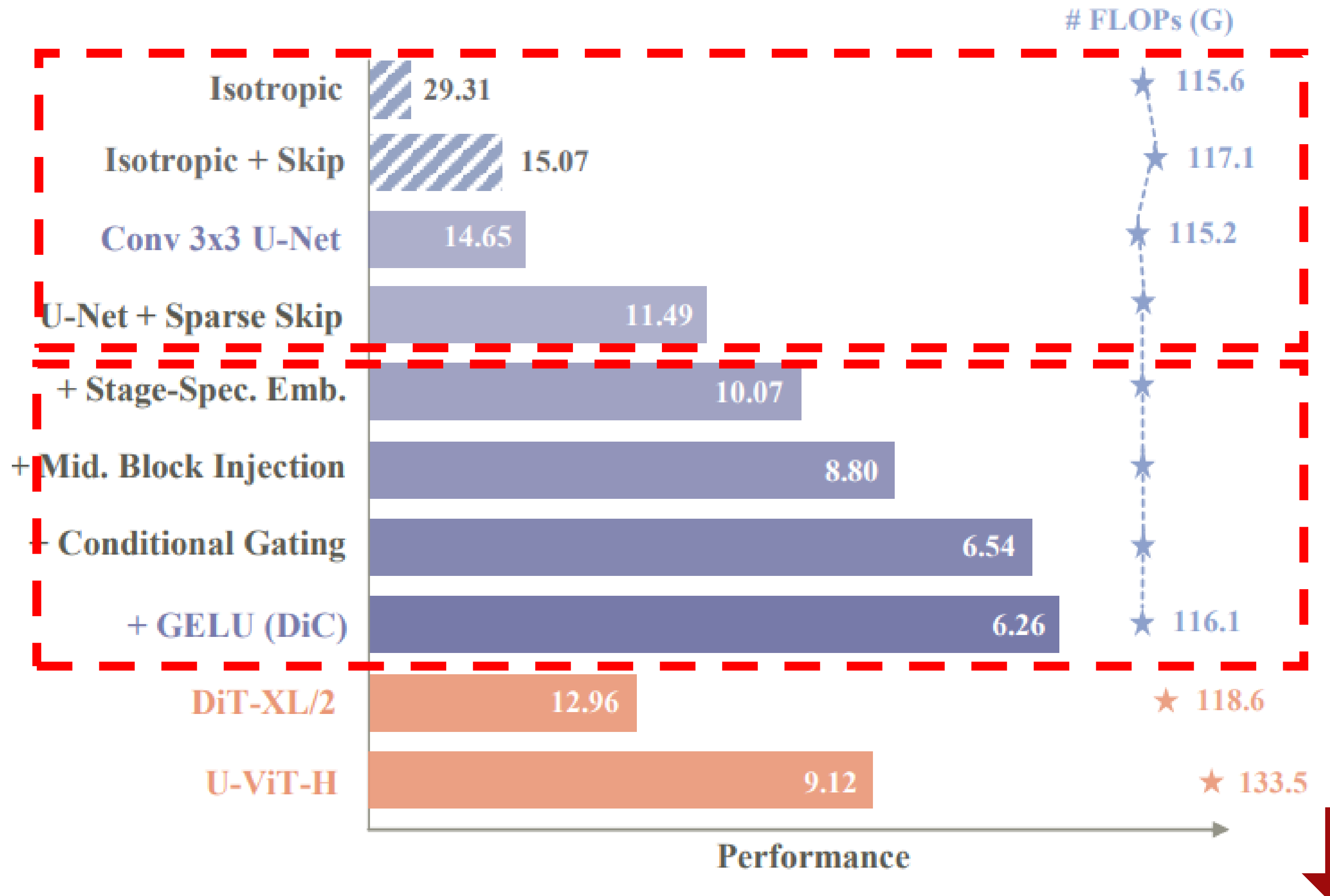
Macro Level

➢ *Architecture Arena*

U-Net arch performs the best



(a) Isotropic (b) Isotropic + Skip (c) U-Net (d) U-Net + Sparse Skip



➢ *Reduce the number of Skips*

Micro Level

➢ *Conditioning Improvements*

➢ *Activations*

Outstanding performance compared with DiT models!

Performance & Speed Advantages

⚡ High Throughput (TP)

ImageNet 256×256, 400K				
Model	FLOPs (G)	TP	FID↓	IS↑
U-ViT-XL [1]	113.0	72.6	18.35	76.59
DiT-XL/2 [30]	118.6	66.8	20.05	66.74
PixArt-α-XL/2 [2]	118.4	64.1	24.75	52.24
DiffiT-XL/2 [17]	118.5	64.1	36.86	35.39
DiT-LLaMA* [5]	118.6	65.2	20.22	70.10
DiC-XL (Ours)	116.1 (57.2)	**313.7**	13.11	100.15
DiC-H (Ours)	204.4 (97.2)	160.8	**11.36**	**106.52**

📉 Faster Convergence

ImageNet 256×256, Scale Up, w/o cfg			
Model	Training Steps	FID↓	IS↑
DiT-XL/2	2.4M	10.67	-
DiT-XL/2	7M	9.62	-
DiC-H	400K	11.36	106.52
DiC-H	600K	9.73	118.57
DiC-H	800K	**8.96**	**124.33**

🔍 Advantages on Larger Images

ImageNet 512×512, 3M, cfg=1.5				
Model	G FLOPs (Wino.)	TP	FID↓	IS↑
DiT-XL/2	524.7	16.2	3.04	240.82
DiC-XL	464.3 (228.7)	84.2	3.04	271.77
DiC-H	817.2 (388.4)	53.3	**2.96**	**293.54**

🔭 Good Potential with Advanced Training

ImageNet 256×256, Scale Up, w/ cfg			
Model	TP	BS×Iter	FID↓
DiT-XL/2	66.8	256×7M	2.27
U-ViT-H	63.9	1024×500K	2.29
DiC-H (Ours)	160.8	256×2M	**2.25**

ImageNet 256×256, REPA			
Model	Training iter	Sampling	FID↓
DiC-XL+U-REPA	1M	ODE	1.74
DiC-XL+U-REPA	1M	SDE	1.75

Takeaway: Convs, though neglected for long, are powerful diffusion archs.

ArXiv GitHub