CVPR JUNE 3-7, 2026 DENVER COLORADO

Reviving ConvNeXt for Efficient Convolutional Diffusion Models

Taesung Kwon, Lorenzo Bianchi, Lennart Wittke, Felix Watine, Fabio Carrara, Jong Chul Ye, Romann Weber, Vinicius Azevedo

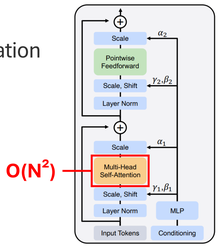
KAIST · ETH Zürich · ISTI-CNR · University of Pisa

Can we generate high-quality images using only convolutions? $O(N^2) \rightarrow O(N)$

Motivation

- Diffusion Transformers (DiTs) dominate image generation
- Self-attention is $O(N^2)$ — heavy compute cost
- **Goal: $O(N^2) \rightarrow O(N)$ while keeping quality**

DiT Block



Background: ConvNeXt

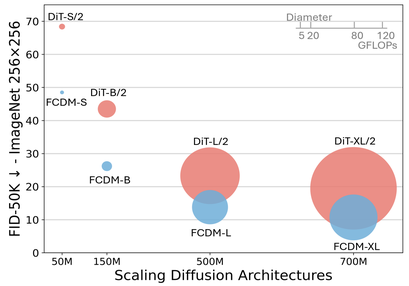
- *ConvNeXts* proved that modernized ConvNets can match ViT in classification
- High performance is unlocked via key designs: inverted bottlenecks, large 7x7 separable convolutions, and global response normalization
- **Can ConvNeXt's efficiency and strengths transfer to generative diffusion?**

Key Contributions

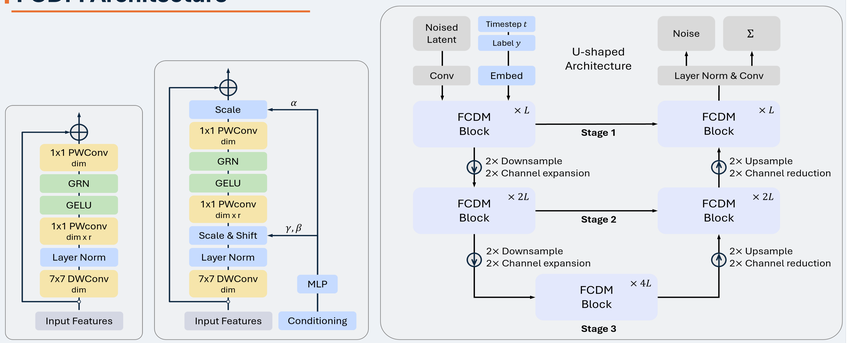
1. First fully convolutional diffusion backbone (FCDM) based on ConvNeXt architecture
2. Competitive FID with DiT-XL/2 at 50% compute cost (FLOPs)
3. 7× faster FID convergence demonstrating training efficiency

Superior Efficiency

50%	7×	4 - 4090 GPUs
Fewer FLOPs	Faster Training	Only Needed



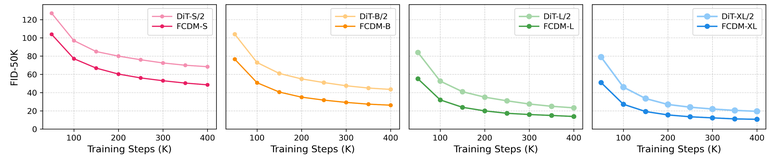
FCDM Architecture



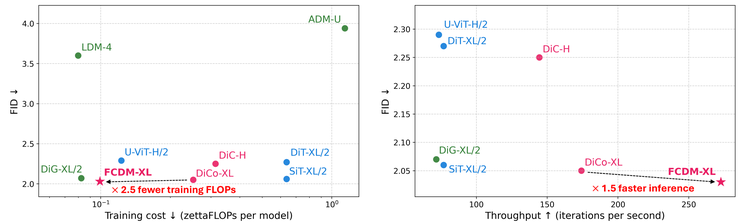
(a) ConvNeXt block (b) FCDM block (ConvNext block with adaLN) (c) Overall architecture of FCDM

- ConvNeXt block + Adaptive LayerNorm for conditioning
- Easy scalable U-shaped design from S to XL variants
- Fully convolutional — no self-attention layers

Training Convergence

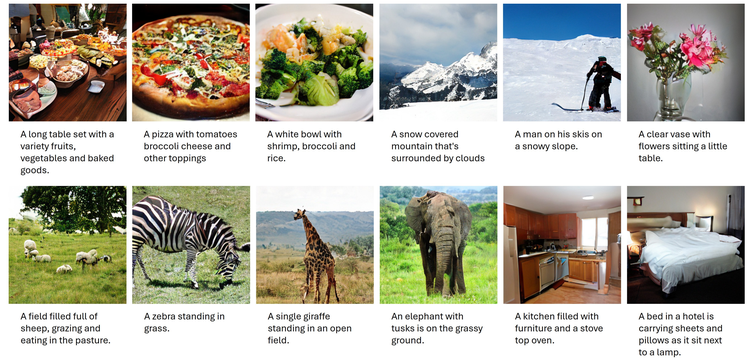


Efficiency Comparison



Generated Samples (ImageNet 512×512 & 256×256)

Text-to-image (MS-COCO 256x256)



A long table set with a variety fruits, vegetables and baked goods.

A pizza with tomatoes broccoli cheese and other toppings

A white bowl with shrimp, broccoli and rice.

A snow covered mountain that's surrounded by clouds

A man on his skis on a snowy slope.

A clear vase with flowers sitting a little table.

A field filled full of sheep, grazing and eating in the pasture.

A zebra standing in grass.

A single giraffe standing in an open field.

An elephant with tusks is on the grassy ground.

A kitchen filled with furniture and a stove top oven.

A bed in a hotel is carrying sheets and pillows as it sit next to a lamp.