NeurIPS 2026

RePiD

Efficient Recursive Pixel-Space Diffusion via Hierarchical Patch Denoising



ImageNet 256×256 samples generated by RePiD

Duoyou Chen, Yaxue Guo, Can Zhang, Cheng Chen

University of Science and Technology Beijing

Two routes to diffusion image generation

Latent diffusion

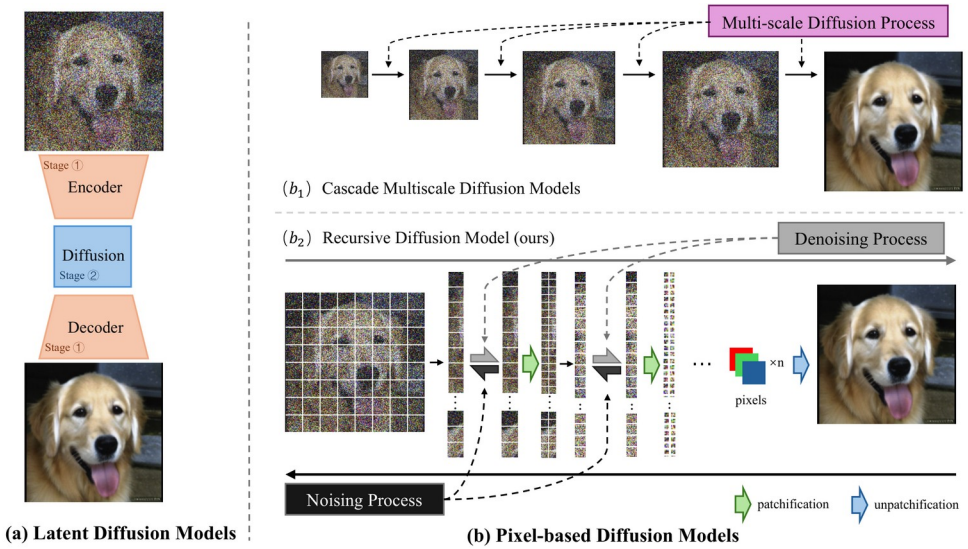
LDM · DiT · SiT

Diffuse in a compressed latent space. Efficient, but the autoencoder caps fidelity and loses fine details

Pixel diffusion

ADM · RDM · PixelFlow · JiT · PixelDiT

Diffuse on raw pixels. No reconstruction bottleneck, but every step touches the full-resolution image



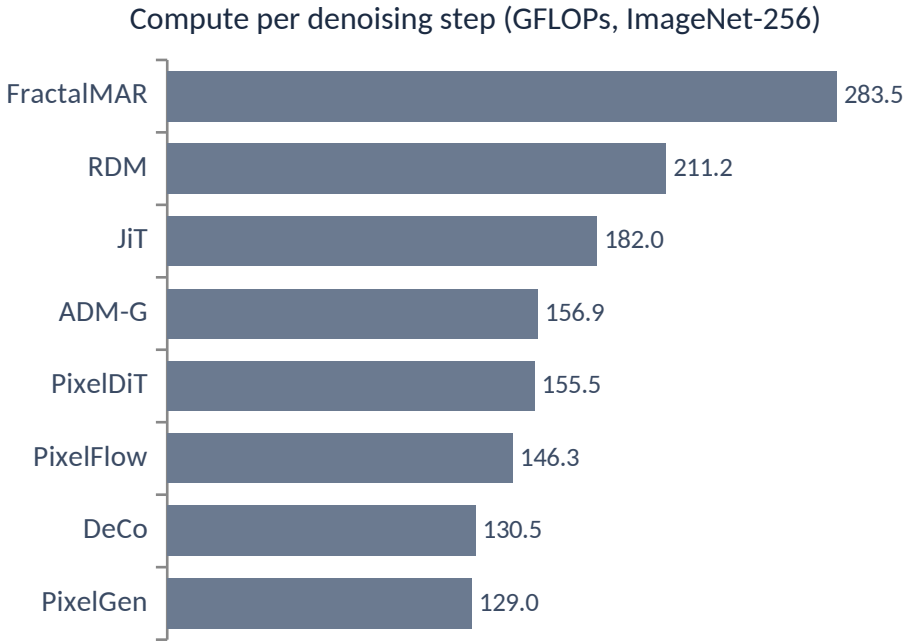
Paradigm comparison: (a) latent diffusion, (b1) cascaded multi-scale, (b2) RePiD

Existing pixel models pay for full resolution

Cascaded multi-scale *RDM, PixelFlow*
Relay diffusion across growing resolutions

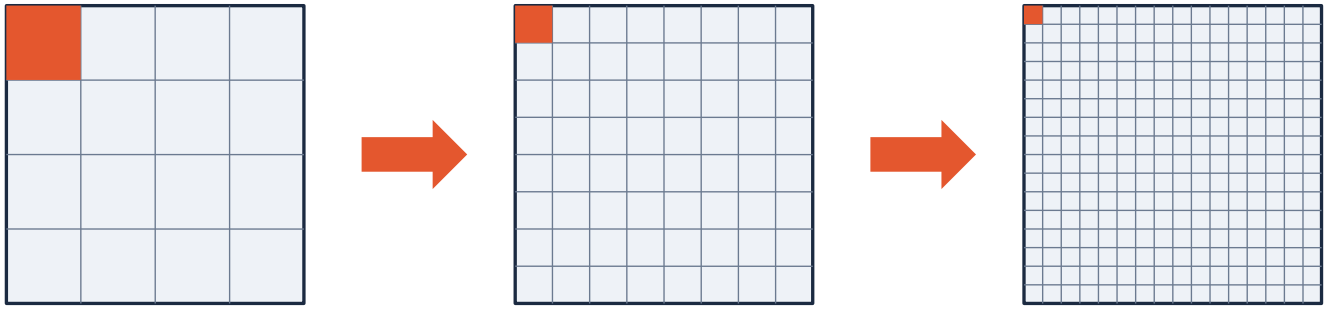
Recursive autoregressive *FractalMAR*
Compose generators recursively, token by token

Single-scale pixel DiTs *JiT, PixelDiT, DeCo, PixelGen*
Large Transformers over the full image



OUR IDEA

Shrink the problem, not the image



Stage 1
16×16 blocks
256 blocks

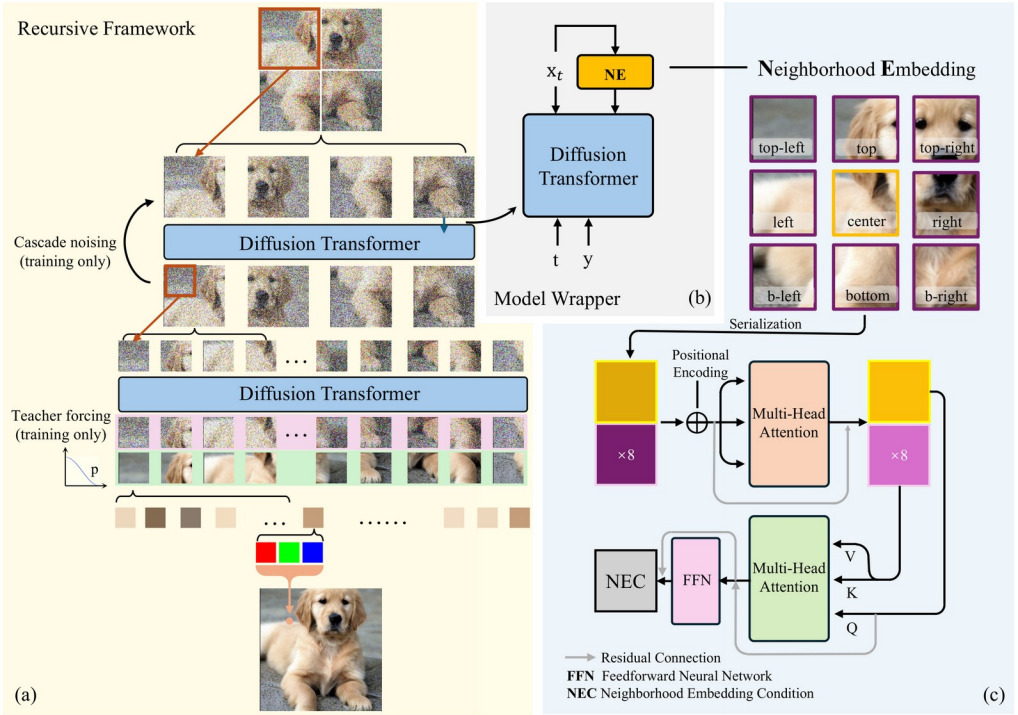
Stage 2
4×4 blocks
4,096 blocks

Stage 3
single pixels
65,536 blocks

≈1/240
attention cost vs. full image

The resolution stays fixed; only the block size shrinks along the denoising trajectory, coupling diffusion time with spatial granularity.

Parallel block denoising with neighbor context



1 **One small DiT per stage**

Denoises every block in parallel, never the whole image

2 **Neighborhood embedding**

Each block attends to its 8 neighbors to keep seams consistent

3 **Teacher forcing in training**

A cosine schedule stabilizes training across stages

RESULTS

ImageNet-256: competitive FID with far less compute

Model	Type	Params	FID ↓	IS ↑	GFLOPs ↓	Time (s) ↓
DiT-XL/2	Latent	86M+675M	2.27	278.2	118.6	5.53
SiT-XL/2	Latent	86M+675M	2.06	277.5	118.6	5.03
RDM	Pixel	553M+553M	1.99	260.4	211.2	16.93
PixelFlow-XL/4	Pixel	677M	1.98	282.1	146.3	12.53
PixelDiT-XL/16	Pixel	797M	1.61	292.7	155.5	23.90
JiT-H/16	Pixel	953M	1.86	303.4	182.0	31.20
DeCo-XL/16	Pixel	682M	1.69	304.0	130.5	23.20
PixelGen-XL/16	Pixel	676M	1.83	293.6	129.0	17.80
RePiD (ours)	**Pixel**	**280M**	**1.87**	**341.6**	**43.3**	**3.43**

Recent pixel models use official results; latencies normalized to one RTX 4090. Full table in the paper.

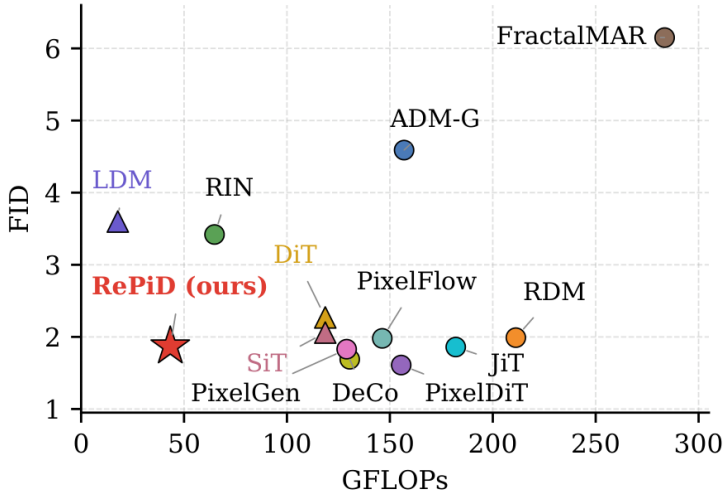
A favorable quality–efficiency trade-off

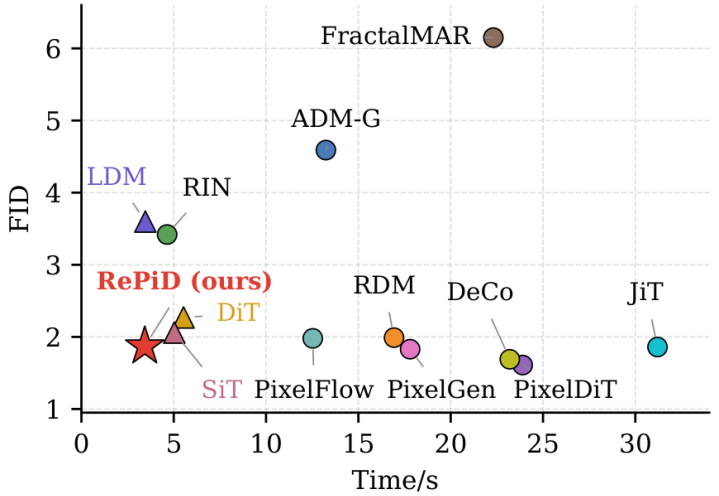
3.4×
fewer parameters than JiT

4.2×
fewer GFLOPs than JiT

3.43 s
per image on one RTX 4090

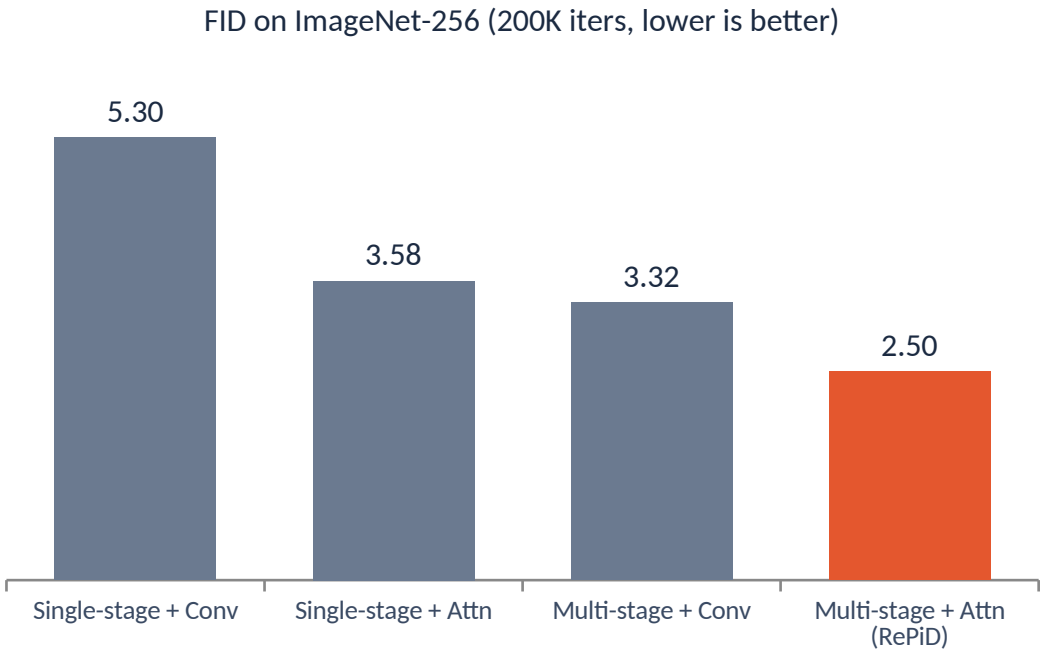
1 GPU
to train (91 GPU-hours)





ABLATION

Recursion and neighbor attention both matter



5.30 → 3.32
Recursion alone (conv)

3.32 → 2.50
Adding neighbor attention

3.31% / 1.85%
Boundary-artifact deviation from real images

Takeaway

Recursive patch-level denoising is a practical path to efficient pixel-space generation



Thank you!