LaMamba-Diff: Linear-Time High-Fidelity Diffusion Models Based on Local Attention and Mamba

Yunxiang Fu *Chaoqi Chen* *Yizhou Yu**
The University of Hong Kong, Shenzhen University

Motivation

Image Generation Requires:

- **Global context for scene understanding**
- **Fine-grained local details**

Self-attention in DiT

- **Quadratic complexity**
- **Insufficient local priors**

LaMamba-Diff

Novel architecture for diffusion models that use

- **Dynamic State Space Models (Mamba) to capture global context with linear complexity**
- **Local Attention to model fine-grained local details**

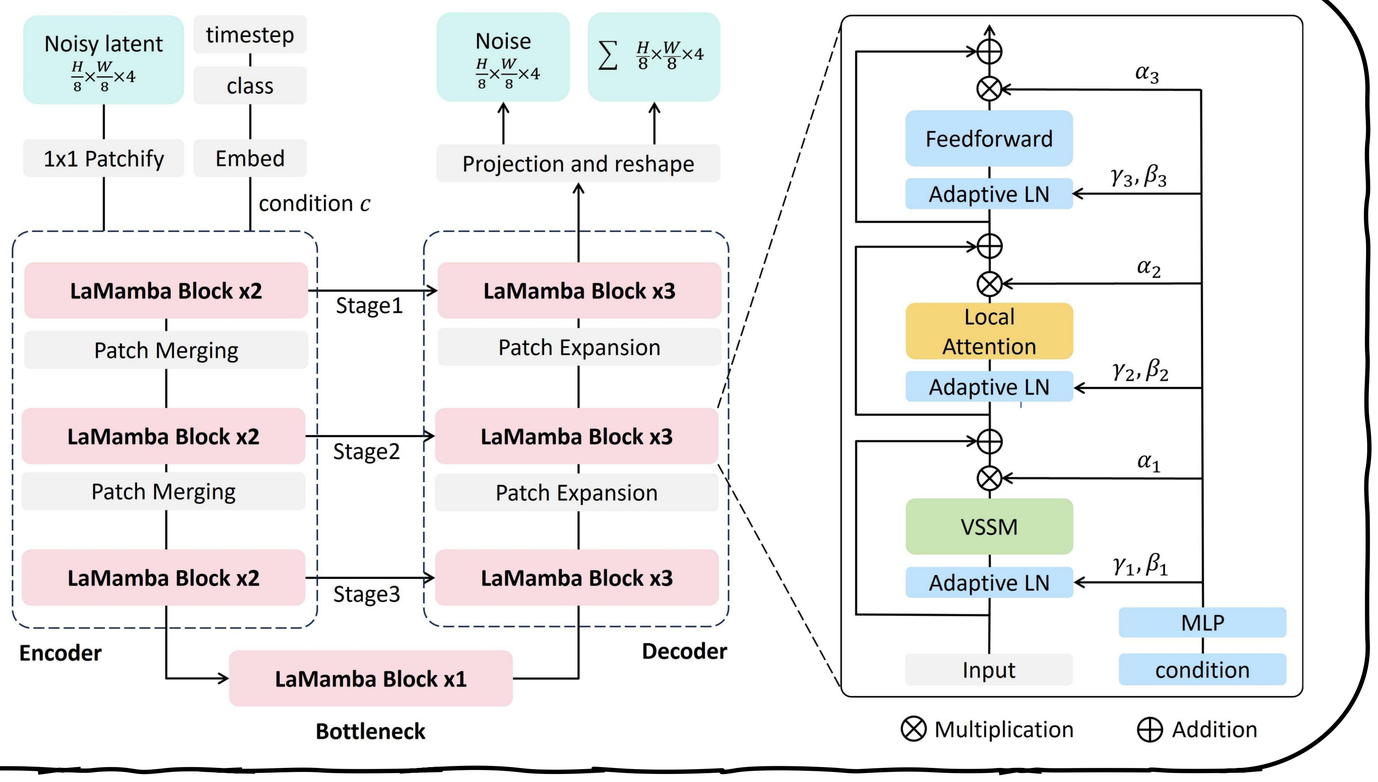
Method

Hybrid LaMamba Block: The Best of Both Worlds

- **Global Context via Mamba**: Integrates a Visual State Space Module (VSSM) to capture long-range dependencies with linear complexity (O(L)), ensuring efficient processing.
- **Local Details via Windowed Attention**: Augments with local self-attention within fixed windows, explicitly modeling fine-grained token interactions lost in Mamba's compression, critical for high fidelity.

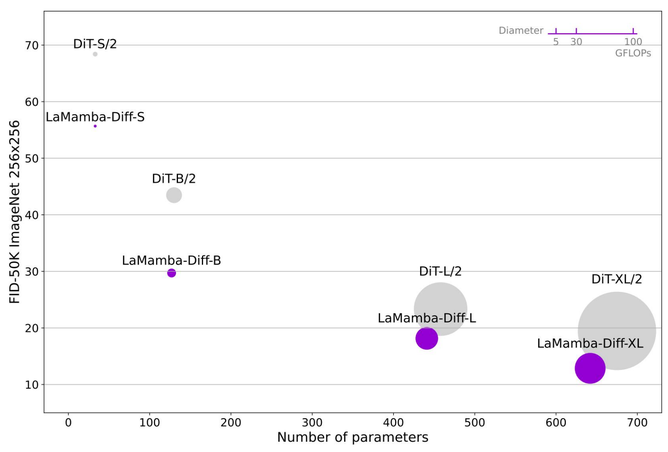
Fine-Grained Input Processing with 1x1 Patches

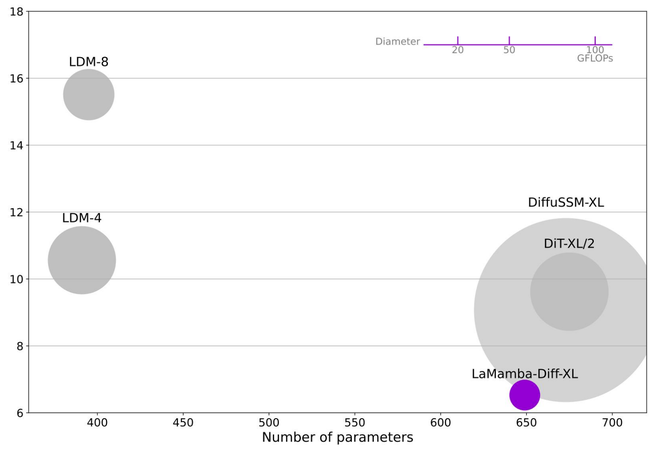
- The efficiency of LaMamba enables the use of a 1x1 patch size without prohibitive computational cost. This treats every latent pixel as a separate token, preventing the loss of high-frequency details that occurs with larger patches (e.g., 2x2 or 4x4), leading to sharper image generation.



Results

Unconditional Generation on ImageNet 256x256. Bubble area represents GFLOPs. (*Left*) FID-50K improves as model size (parameters/GFLOPs) increases. (*Right*) Our largest LaMamba-Diff model achieves better performance than DiT-XL/2 with significantly fewer GFLOPs.





256×256 samples generated

512×512 samples generated

Quantitative comparison of image generation quality on ImageNet 256x256 and 512x512. LaMamba-Diff-XL achieves competitive performance with substantially fewer GFLOPs.

ImageNet 256×256								
Model	Parameters(M)	GFLOPs	Training Steps(M)	FID↓	sFID↓	IS↑	Precision↑	Recall↑
				Unconditional				
ADM	554	1120	2.0	10.94	6.02	100.98	0.69	0.63
ADM-U	608	742	2.0	7.49	5.13	127.49	0.72	0.63
LDM-8	395	79	1.2	15.51	-	79.03	0.65	0.63
LDM-4	391	103	0.8	10.56	-	103.49	**0.71**	0.62
DiT-XL/2	675	118	7.0	9.62	6.85	121.50	0.67	**0.67**
DiffuSSM-XL	673	280	2.6	9.07	**5.52**	118.32	0.69	0.64
LaMamba-Diff-XL	656	50	2.0	**6.12**	5.58	**149.13**	**0.71**	0.64
				Classifier-free guidance				
ADM-G, ADM-U	673	761	2.0	3.94	6.14	215.84	0.83	0.53
LDM-8-G	395	79	1.2	7.76	-	209.52	0.84	0.35
LDM-4-G	391	103	0.8	3.60	-	247.67	**0.87**	0.48
U-ViT-H/2-G	501	133	2.0	2.29	-	247.67	**0.87**	0.48
DiffuSSM-XL-G	673	280	2.6	2.28	4.49	259.13	0.86	0.56
DiT-XL/2-G	675	118	7.0	2.27	4.60	278.24	0.83	**0.57**
DiS-H/2-G [7]	900	217	2.0	2.10	4.55	271.32	0.82	0.58
DiM-Huge	860	210	1.9	2.21	-	-	-	-
LaMamba-Diff-XL-G	656	50	2.0	**2.04**	**4.52**	**296.07**	0.84	0.56
ImageNet 512×512								
				Unconditional				
DiT-XL/2	675	524	3.0	12.03	7.12	105.25	0.75	**0.64**
LaMamba-Diff-XL	656	201	3.0	**7.76**	**6.89**	**122.41**	**0.81**	0.60
				Classifier-free guidance				
ADM-G, ADM-U	774	2834	1.9	3.85	5.86	221.72	0.84	0.53
U-ViT-H/4-G	501	133	2.0	4.05	8.44	261.13	0.84	0.48
DiffuSSM-XL-G	673	1066	1.2	3.41	5.84	255.06	0.85	0.49
DiT-XL/2-G	675	524	3.0	3.04	**5.02**	240.82	0.84	**0.54**
DiS-H/2-G [7]	900		4.0	2.88	4.74	272.33	0.84	0.56
DiM-Huge	860	708	-	3.78	-	-	-	-
LaMamba-Diff-XL-G	656	201	3.0	**3.01**	5.15	**277.19**	0.85	0.52

Comparison of GFLOPs and Images generated per second between LaMamba-Diff-XL and DiT-XL/2

Resolution	Model	GFLOPs	Image/s
256 × 256	DiT-XL/2	118.64	8.9
	LaMamba-Diff-XL	50.46	6.8
512 × 512	DiT-XL/2	524.60	1.9
	LaMamba-Diff-XL	201.20	2.0
1024 × 1024	DiT-XL/2	2910.30	0.3
	LaMamba-Diff-XL	804.18	0.5

Sample images generated by LaMamba-Diff-XL trained on ImageNet at 512×512 and 256×256 resolutions.