DiMSUM: Diffusion Mamba - A Scalable and Unified Spatial-Frequency Method for Image Generation

Hao Phung[*,1,3,†] Quan Dao[*,1,2,†] Trung Dao[1] Hoang Phan[4] Dimitris Metaxas[3] Anh Tran[1]





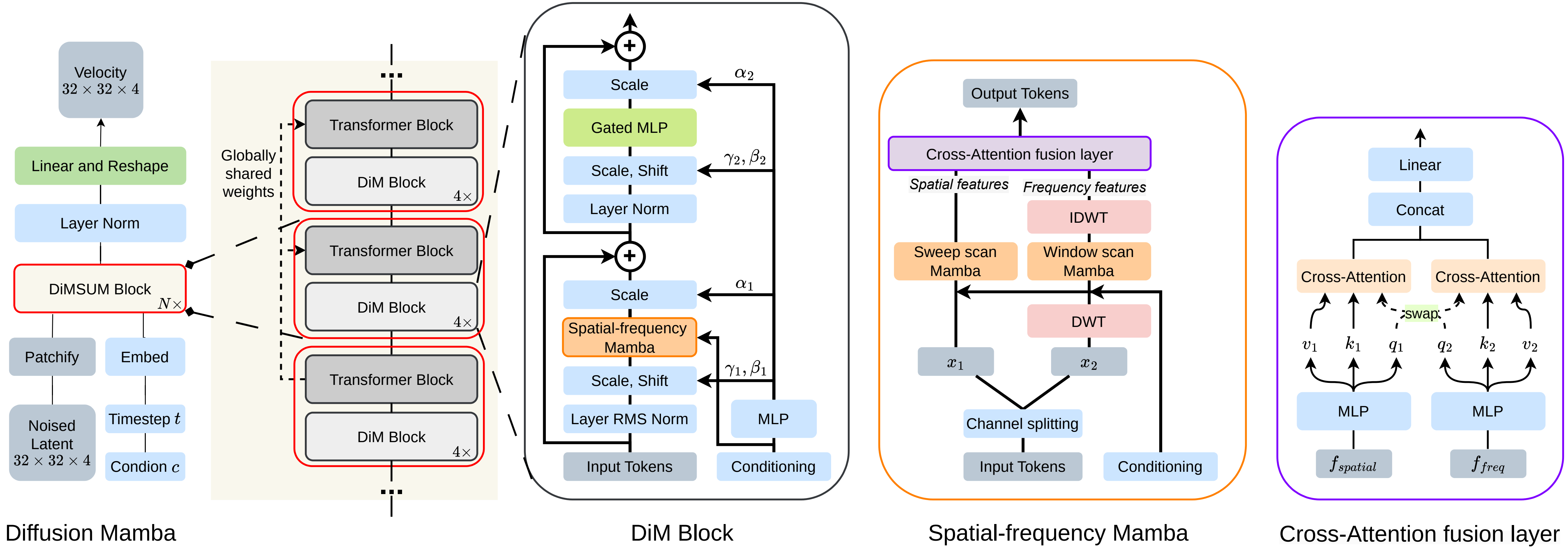

[1]VinAI Research [2]Rutgers University [3]Cornell University [4]New York University

[*]*Equal contribution* [†]*Work done while at VinAI Research*

Poster session: Wed 11 Dec 11 a.m. PST — 2 p.m. PST

DiMSUM

We propose DiMSUM, a hybrid Mamba-Transformer diffusion model that synergistically leverages both spatial and frequency information for high-quality image synthesis.



Highlight

- Outperforms both DiT and DIFFUSSM
- Requires less than a third training iterations than DiT and SiT.

Model	FID↓	Recall↑	Params	#Iters × Bs	Epoch
Ours	**8.61**	**0.67**	460M	936K × 704	510
Ours-G	**2.11**	**0.59**	460M	936K × 704	510
SSM-based					
DIFFUSSM-XL [65]	9.07	0.64	673M	2578K× 256	515
DIFFUSSM-XL-G	2.28	0.56	673M	2578K × 256	515
UNet-based					
LDM-4 [51]	10.56	0.62	400M	178K × 1200	200
LDM-4-G	3.60	0.48	400M	178K × 1200	200
Transformer-based					
DiT-L/2 [48]	23.33	-	458M	400K × 256	80
DiT-XL/2	9.62	0.67	675M	7000K × 256	1.4K
DiT-XL/2-G	2.27	0.57	675M	7000K × 256	1.4K
SiT-XL/2 [44]	9.40	-	675M	7000K × 256	1.4K
SiT-XL/2-G	2.15	0.59	675M	7000K × 256	1.4K
GAN model					
BigGan-deep [4]	6.95	0.28	160M	-	
StyleGAN-XL [54]	2.30	0.53	166M	25000K × 256	4K

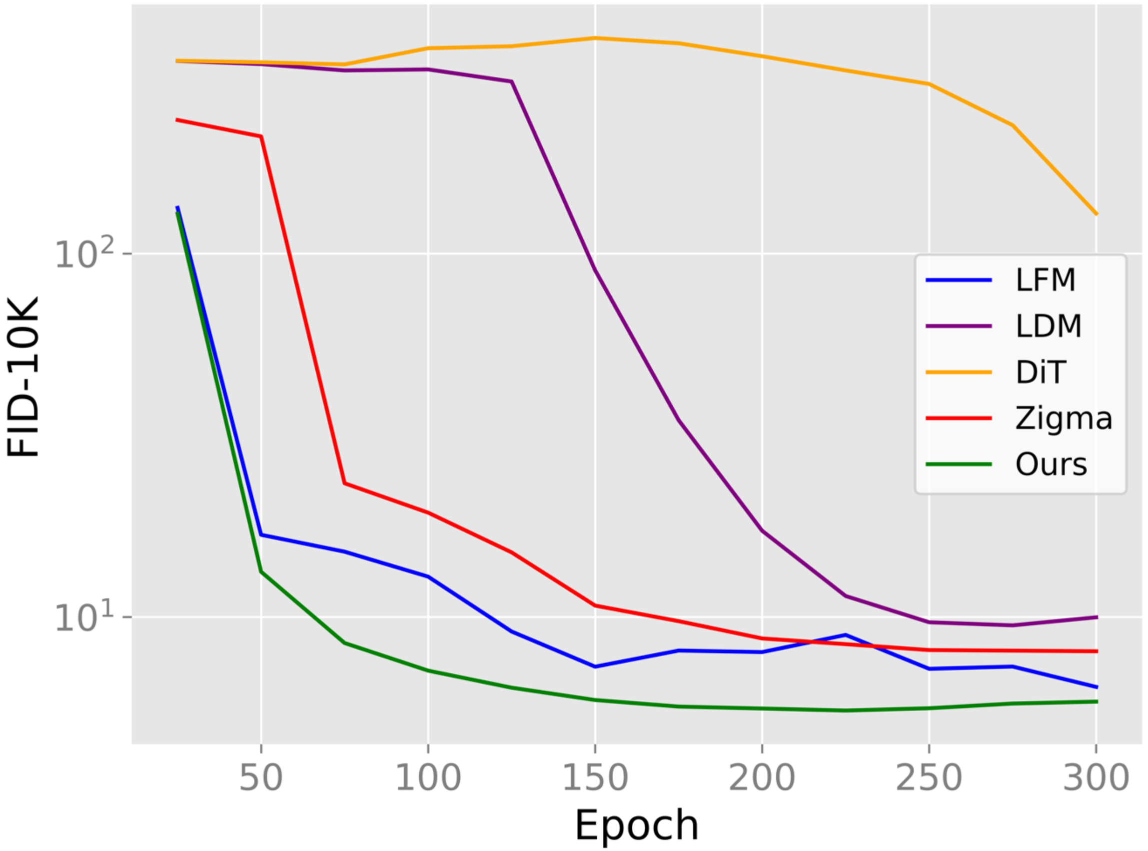
Conditional image generation on ImageNet-1K 256×256

Highlight

Model	FID↓	Recall↑	Params	#Iters × Bs	Epoch
Ours	**8.61**	**0.67**	460M	936K × 704	510
Ours-G	**2.11**	**0.59**	460M	936K × 704	510
SSM-based					
DIFFUSSM-XL [65]	9.07	0.64	673M	2578K× 256	515
DIFFUSSM-XL-G	2.28	0.56	673M	2578K × 256	515
UNet-based					
LDM-4 [51]	10.56	0.62	400M	178K × 1200	200
LDM-4-G	3.60	0.48	400M	178K × 1200	200
Transformer-based					
DiT-L/2 [48]	23.33	-	458M	400K × 256	80
DiT-XL/2	9.62	0.67	675M	7000K × 256	1.4K
DiT-XL/2-G	2.27	0.57	675M	7000K × 256	1.4K
SiT-XL/2 [44]	9.40	-	675M	7000K × 256	1.4K
SiT-XL/2-G	2.15	0.59	675M	7000K × 256	1.4K
GAN model					
BigGan-deep [4]	6.95	0.28	160M	-	
StyleGAN-XL [54]	2.30	0.53	166M	25000K × 256	4K

Conditional image generation on ImageNet-1K 256 × 256

Outperforms Zigma and other baselines in convergence rates.



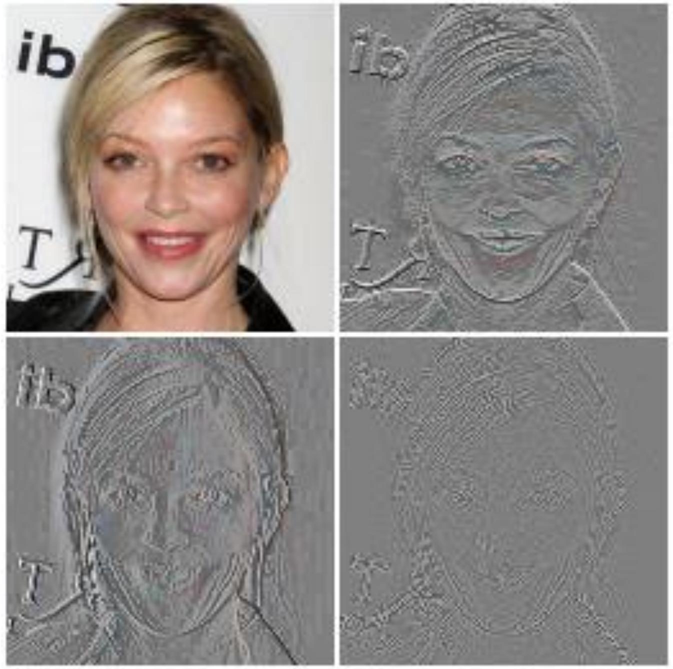
Training convergence curves on CelebA

First, let’s take a look at the wavelet transformation

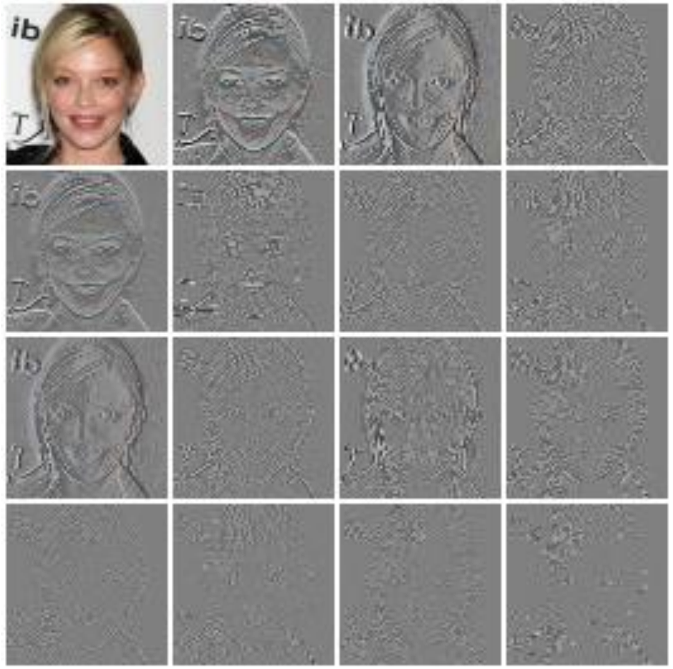
Decompose image with wavelet transformation

Input Image

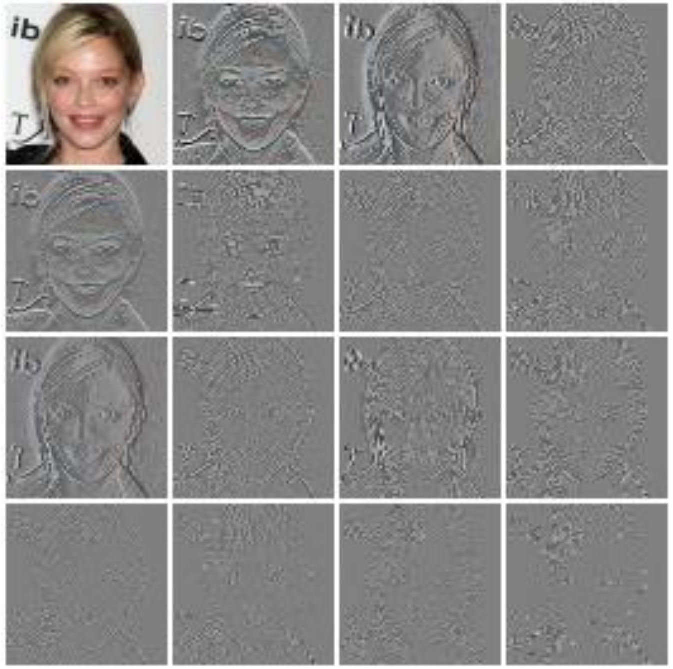
Wavelet level 1



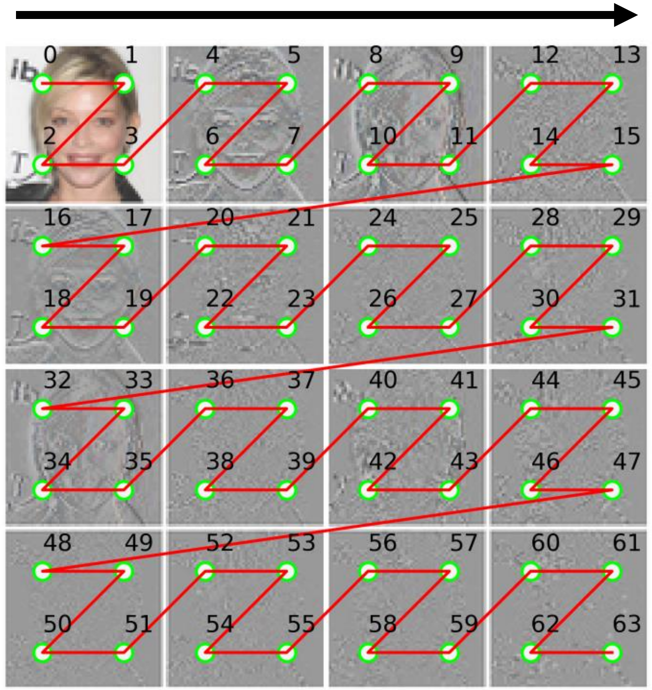
Wavelet level 2

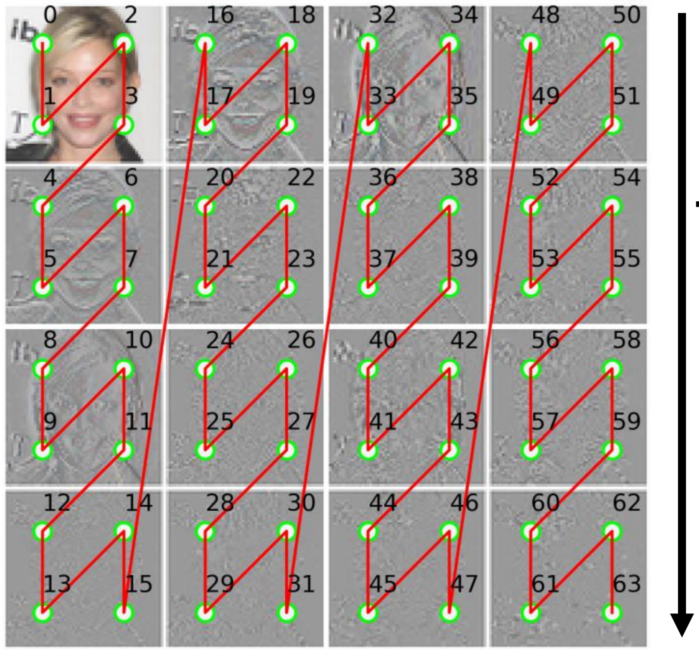


Perform scanning in frequency space



Left to Right





Top to Bottom

**Each window corresponds to a whole wavelet subband, which means our method can ____*

- Better capture long-range frequencies within wavelet subband
- Preserve the structure consistency of frequencies across subbands

Integration of condition input into Mamba

Mamba has no attention mechanism like Transformers. How can we incorporate conditional context into its process?

The answer is really simple, it lies in the recurrent process of Mamba.

Integration of condition input into Mamba

Recurrent process of Mamba

$$\begin{cases} h_0 = \overline{\boldsymbol{B}} x_0 \\ y_0 = \boldsymbol{C} h_0 \end{cases}$$

$$\begin{cases} h_t = \overline{\boldsymbol{A}} \boxed{h_{t-1}} + \overline{\boldsymbol{B}} x_t \\ y_t = \boldsymbol{C} h_t \end{cases}$$

where $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C}$ are the parameter triplet of Mamba

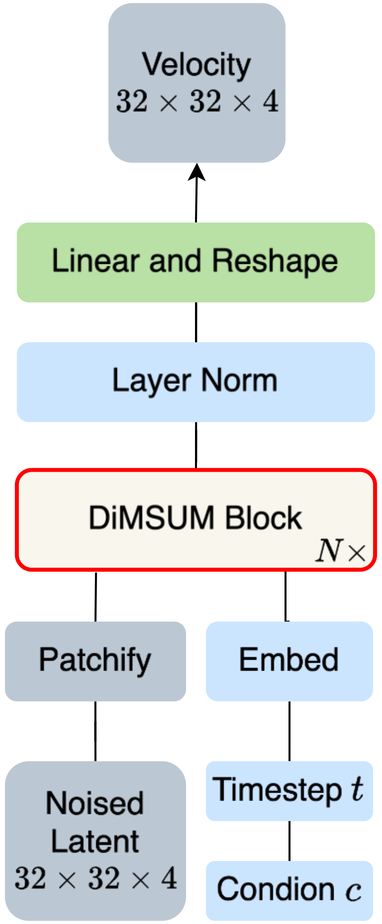
Inject context condition c

$$\begin{cases} h_0 = \boxed{\overline{\boldsymbol{A}} h_{-1}} + \overline{\boldsymbol{B}} x_0 \\ y_0 = \boldsymbol{C} h_0 \end{cases}$$

where $h_{-1} = Linear_{\mathrm{D}}(\mathrm{c})$ is a linear projection of context input c

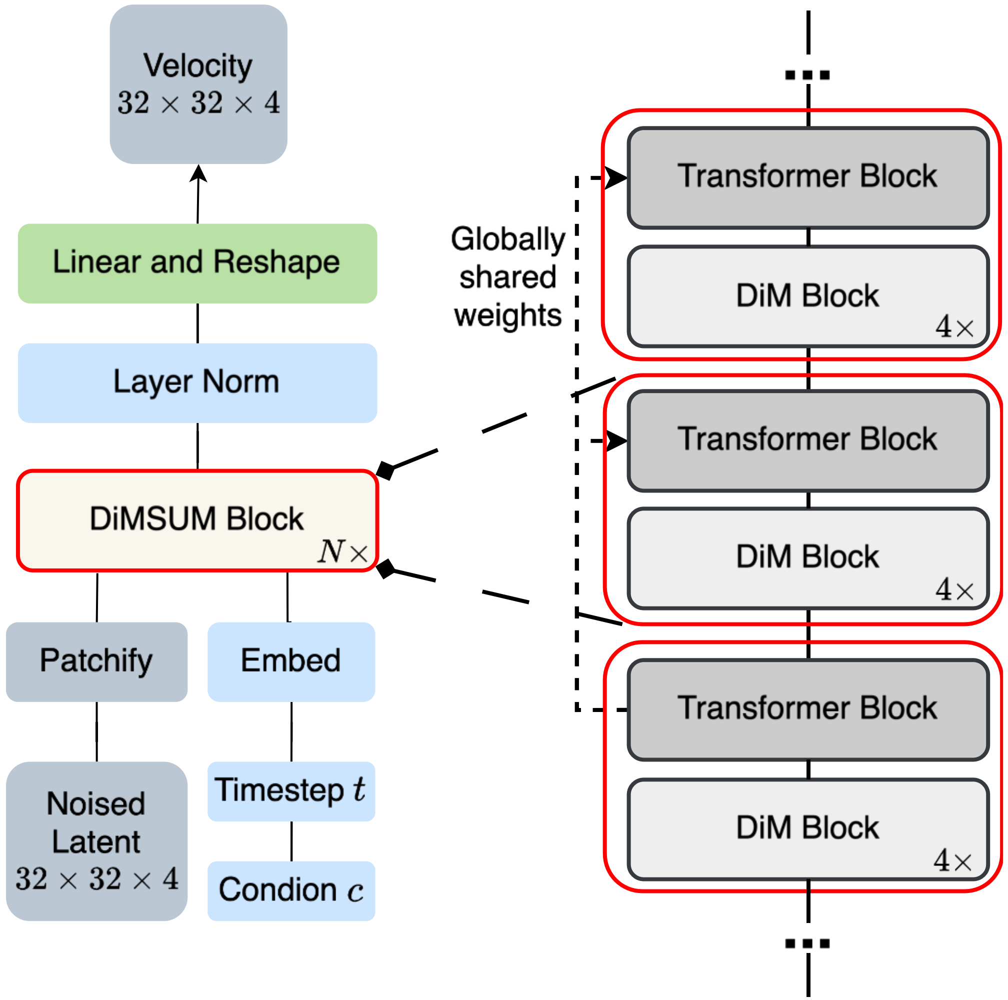
Now, let’s dive into the architecture

DiMSUM Architecture



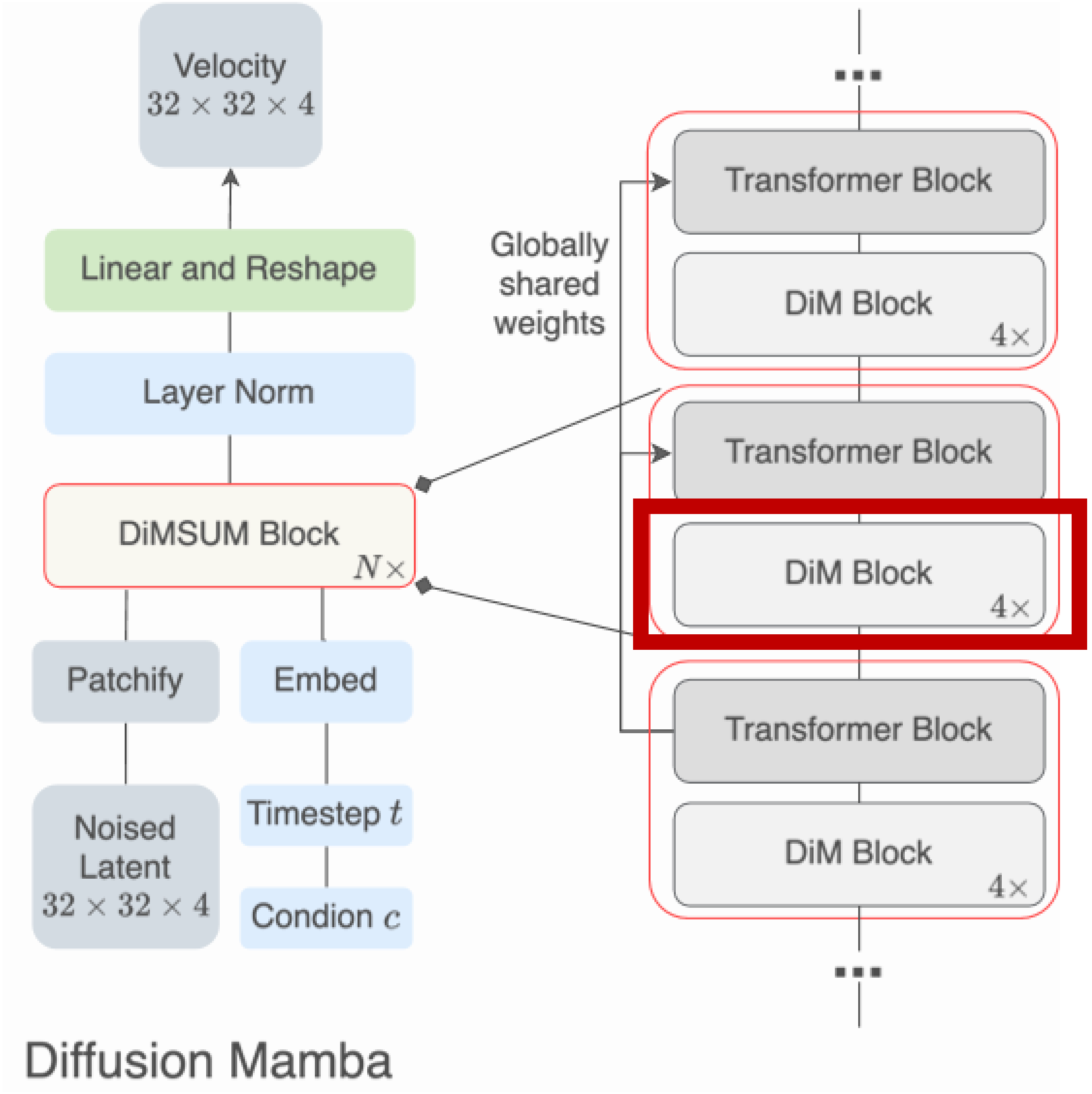
Diffusion Mamba

Inherit the simplicity in design of DiT architecture



Diffusion Mamba

- Only a single transformer block is added which is shared throughout the architecture.
- #Params is still comparable with DiT-L (460M vs 458M)
- Transformer block is inserted after every 4 DiM Block, resulting in only marginal compute overhead.


Velocity
$32 \times 32 \times 4$
Linear and Reshape
Layer Norm
DiMSUM Block
$N\times$
Patchify
Embed
Noised Latent
$32 \times 32 \times 4$
Timestep t
Condion c
Globally shared weights
Transformer Block
DiM Block
$4\times$
Transformer Block
DiM Block
$4\times$
Transformer Block
DiM Block
$4\times$
Diffusion Mamba

DiM Block design

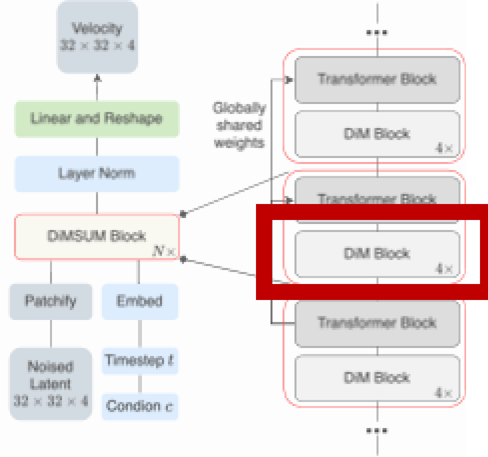
We introduce two key components:

(1) Spatial-Frequency Mamba

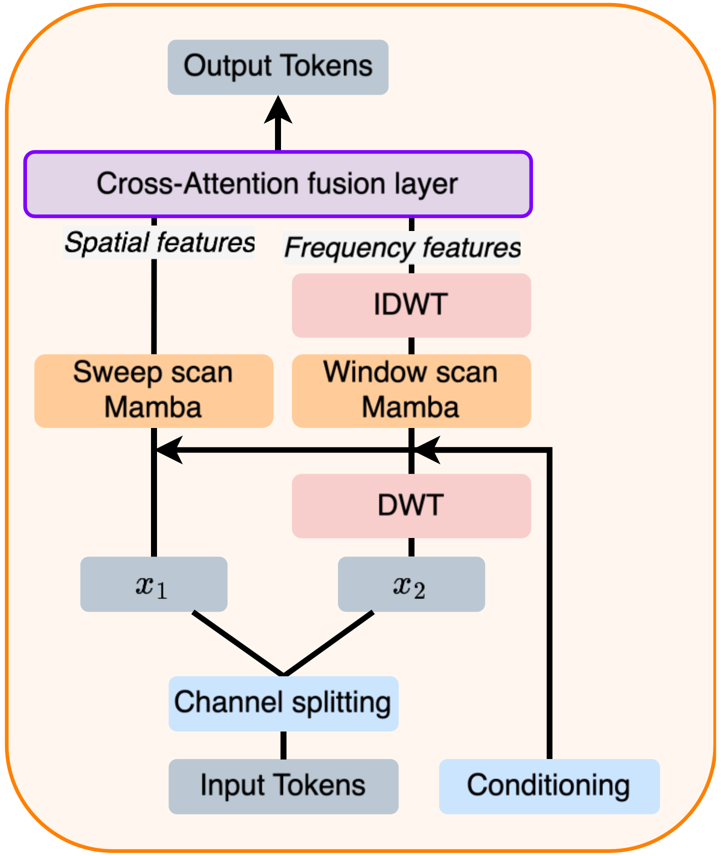
Input features are split by channels, with one half assigned to the spatial branch and the other to the frequency branch.

Spatial branch is just a Mamba Block with sweep scan

Frequency branch is our proposed Wavelet Mamba Block



Diffusion Mamba



Spatial-frequency Mamba

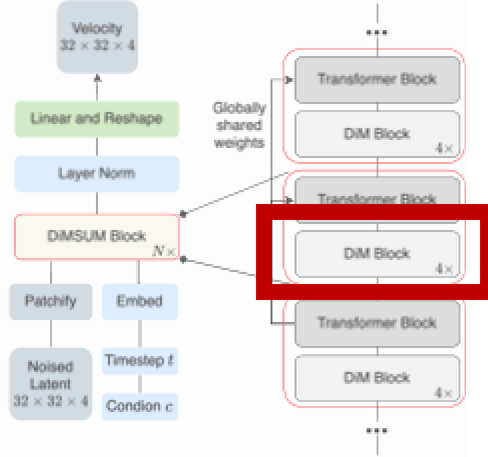
DiM Block design

We introduce two key components:

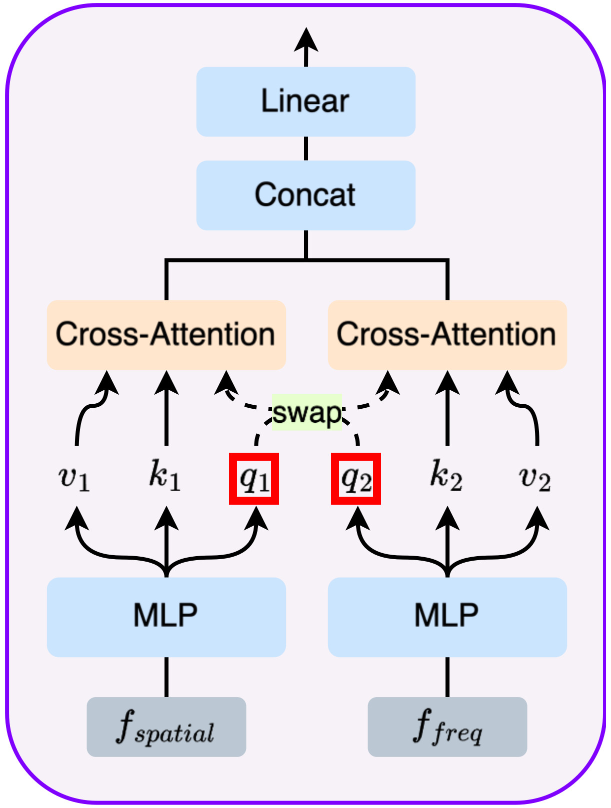
(1) Spatial-Frequency Mamba

(2) Cross-Attention fusion layer

Fusing spatial branch & frequency branch by simply swapping their queries

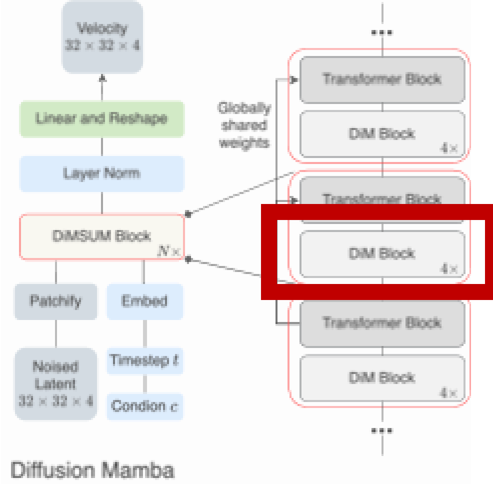


Diffusion Mamba



Cross-Attention fusion layer

DiM Block design

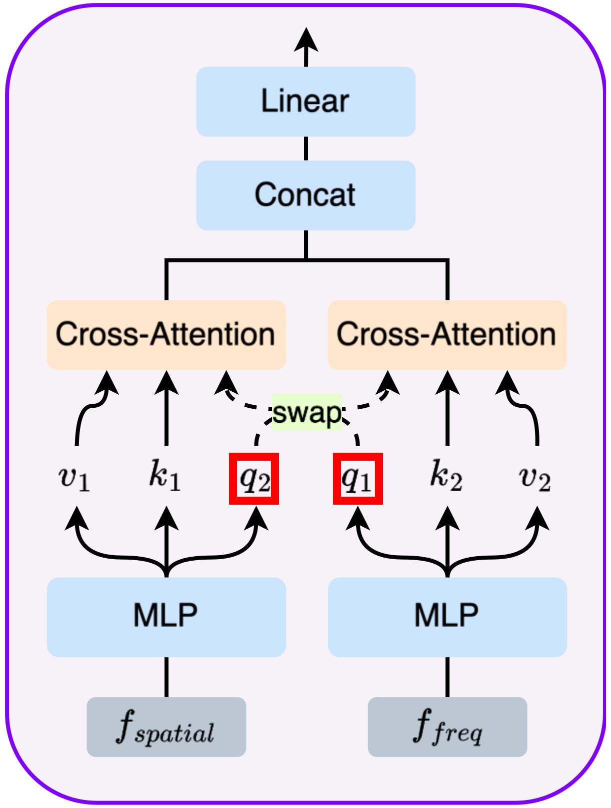


We introduce two key components:

(1) Spatial-Frequency Mamba

(2) Cross-Attention fusion layer

Fusing spatial branch & frequency branch by simply swapping their queries



Cross-Attention fusion layer

Ablation studies on CelebA

(a) Scanning

Order	FID↓	Recall↑	iters/s ↑
Conditional Mamba Only			
Bi	6.39	0.44	2.06
Sweep-4	5.27	0.49	2.06
Sweep-8	5.53	0.48	1.97
Zigzag-8	6.17	0.46	1.97
Jpeg-8	6.26	0.45	1.97
Window	10.88	0.36	2.05
Spatial-frequency Mamba			
Sweep-4 | Sweep-4	5.41	0.49	1.54
Sweep-4 | Window	4.92	0.50	1.54

(b) Components

	FID↓	Recall↑	Params	GFLOPs
Baseline	6.19	0.46	413M	51.65
+ Conditional Mamba	5.27	0.49	446M	51.69
+ Wavelet Mamba (w/ concat)	5.87	0.47	394M	56.54
+ Cross-Attention fusion layer	4.92	0.50	436M	62.42
+ Shared transformer block	4.65	0.52	459M	84.49

(c) Frequency types

	FID↓	Recall↑	Params	GFLOPs
DCT	5.53	0.50	436M	67.33
EinFFT	5.63	0.48	371M	66.96
Wavelet	4.92	0.50	436M	62.42

Generated examples

Unconditional CelebA

Class-Conditional ImageNet

Sampling speed

Method	Time	MEM	Params	GFlops
256 (latent size: 32×32)				
Ours-L/2	2.20s	2.42G	460M	84.49
DiT-L/2	3.80s	2.30G	458M	80.74
512 (latent size: 64×64)				
Ours-L/2	2.86s	2.46G	461M	337.48
DiT-L/2	4.78s	2.34G	459M	361.14

Thank you for watching!

DiMSUM

Please scan this for more information