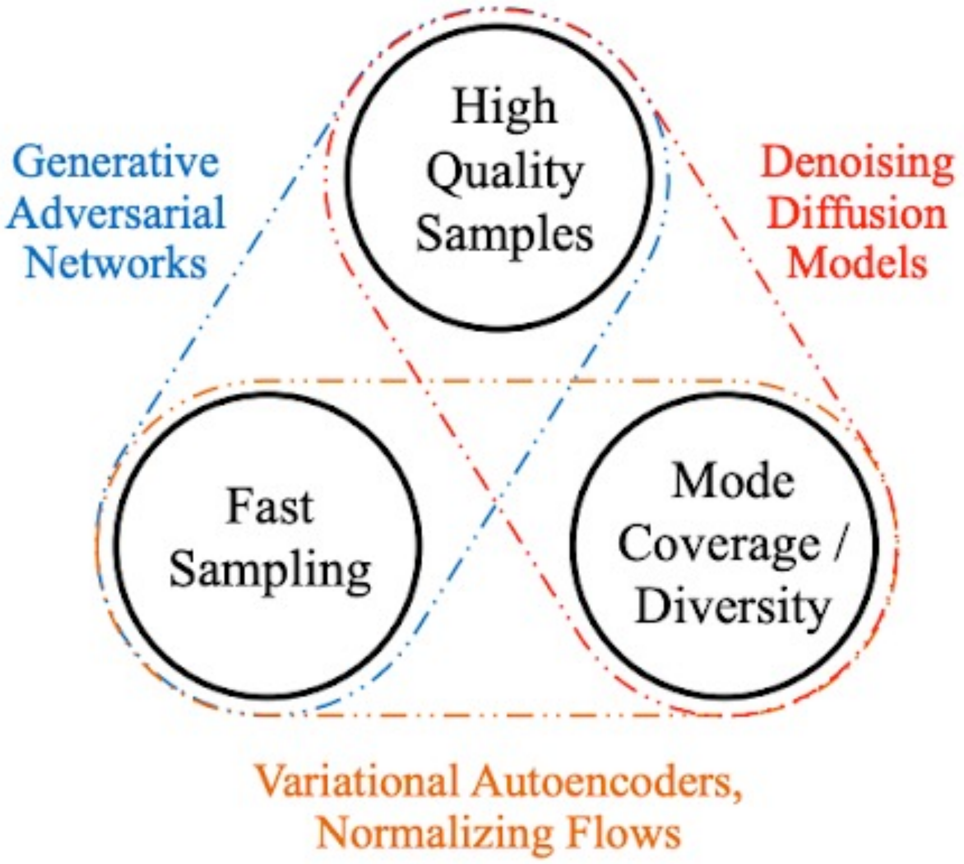
Diffusion Probabilistic Model Made Slim [CVPR 2023]

Xingyi Yang, Daquan Zhou, Jiashi Feng, Xinchao Wang

National University of Singapore, Bytedance

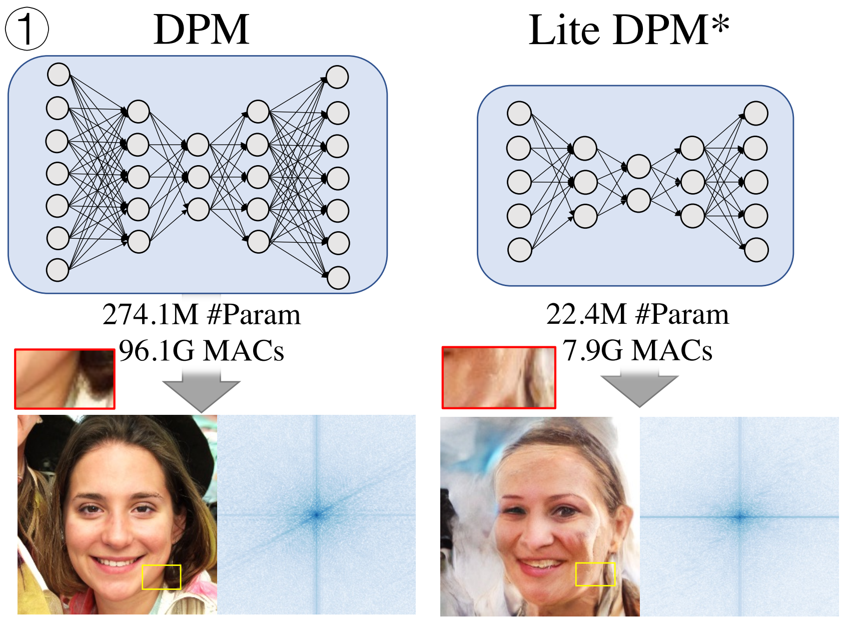
Challenge: DPM Lacks Efficiency



LEARNING AND VISION

Method	#Param	FID↓	Low-freq Error↓	High-freq Error↓
LDM	274.1M	5.0	0.11	0.75
Lite-LDM	22.4M	17.3	0.28(+0.17)	3.35(+2.17)

Table 1. Low-freq and High-freq error for different model size.



Challenge 2: Small Diffusion High-frequency Deficiency

Frequency Analysis

1. **Spectrum Evolution**
   - Low to high recovery
2. **Frequency Bias**
   - Deficiency on Long-tail patterns

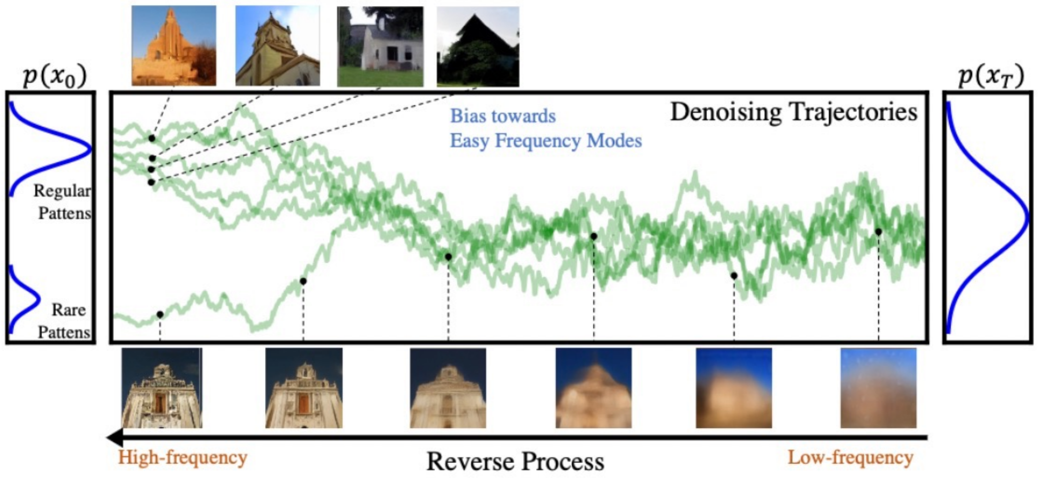


Figure 2. **Illustration of the Frequency Evolution and Bias for Diffusion Models.** In the reverse process, the optimal filters recover low-frequency components first and add on the details at the end. The predicted score functions may be incorrect for rare patterns, thus failing to recover complex and fine-grained textures.

LEARNING AND VISION

I. Spectrum Evolution

Simplified Assumption: Linear Filter, additive Gaussian, wide-sense stationary signal

Weiner Filter

Proposition 1. *Assume* $\mathbf{x}_0$ *is a wide-sense stationary signal and* $\boldsymbol{\epsilon}$ *is white noise of variance* $\sigma^2 = 1$. *For* $\mathbf{x}_t = \sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}$, *the optimal linear denoising filter* h_t *at time t that minimize* $J_t = \|h_t * \mathbf{x}_t - \boldsymbol{\epsilon}\|^2$ *has a closed-form solution*

$$\mathcal{H}_t^*(f) = \frac{1}{\bar{\alpha}|\mathcal{X}_0(f)|^2 + 1 - \bar{\alpha}} \tag{6}$$

where $|\mathcal{X}_0(f)|^2$ *is the power spectrum of* $\mathbf{x}_0$ *and* $\mathcal{H}_t^*(f)$ *is the frequency response of* h_t^*.

LEARNING AND VISION

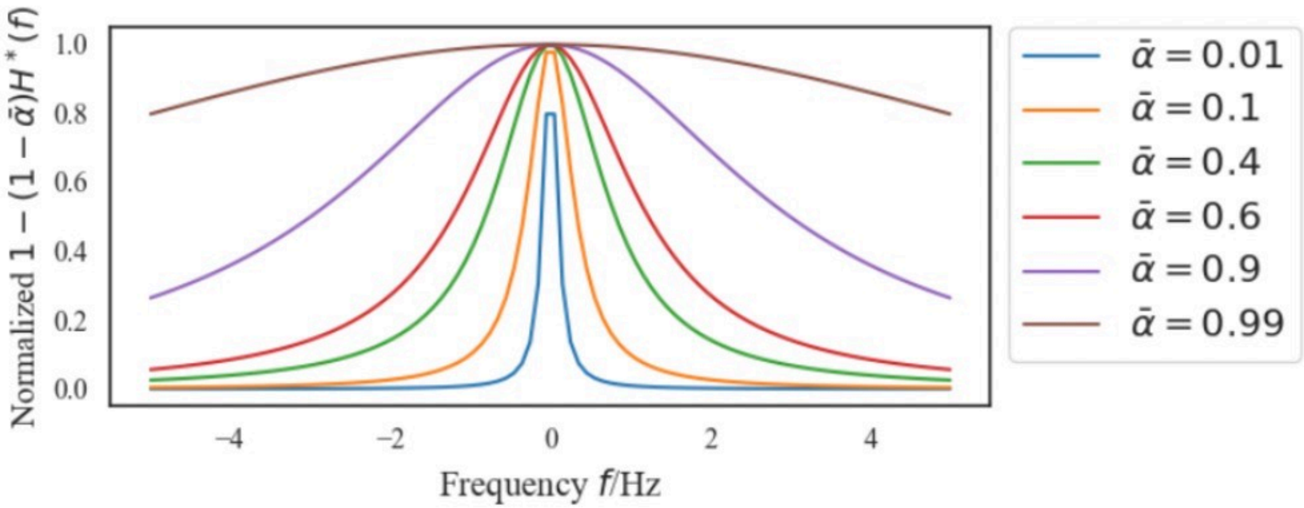
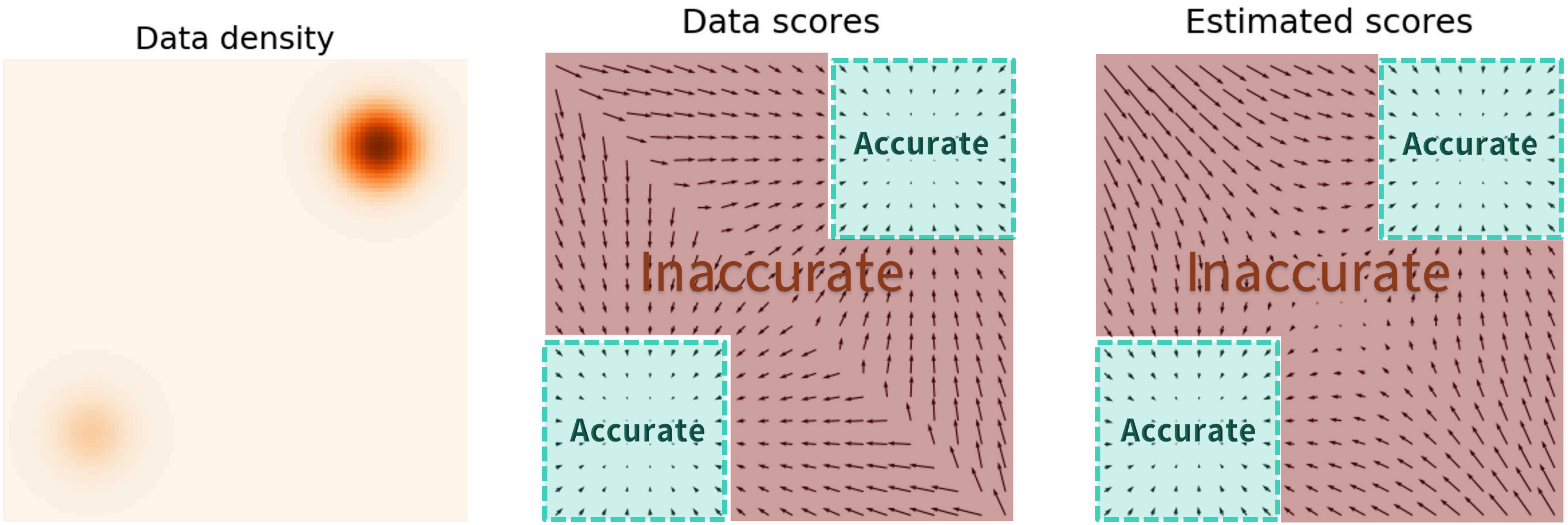


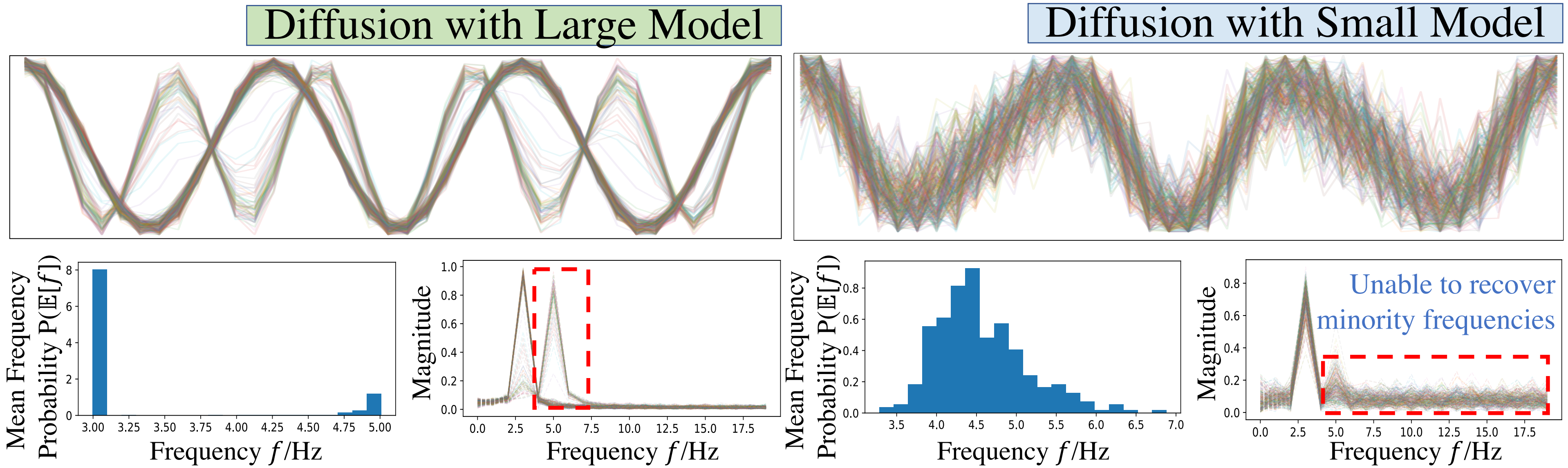
Figure 3. $1-(1-\bar{\alpha})|H^*(f)|^2$ of the optimal linear denoising filter with different $\bar{\alpha}$.

I. Spectrum Evolution

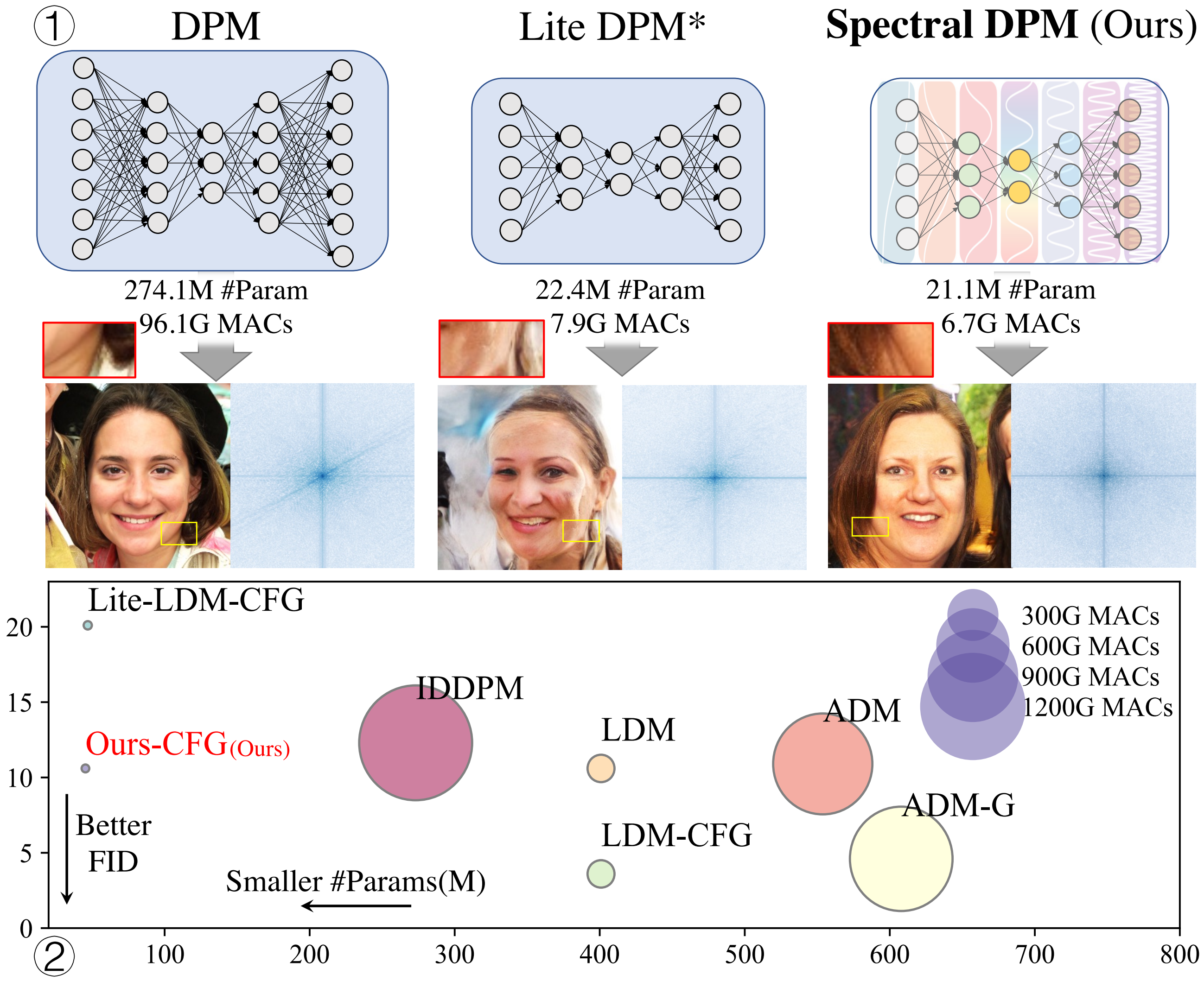
II. Frequency Deficiency



II. Frequency Deficiency



Our solution: Spectral Diffusion



Module 1 Wavelet Gating

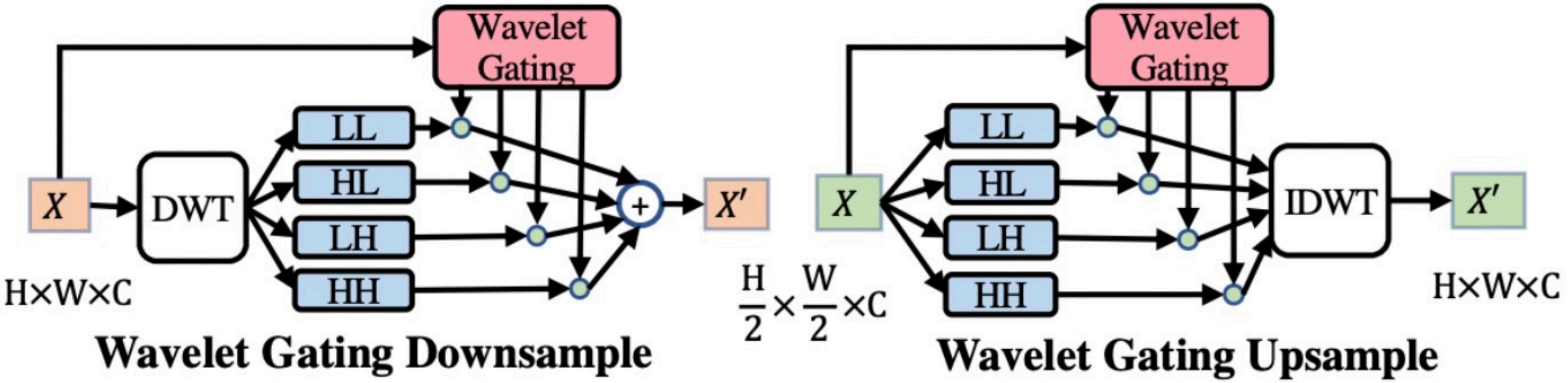
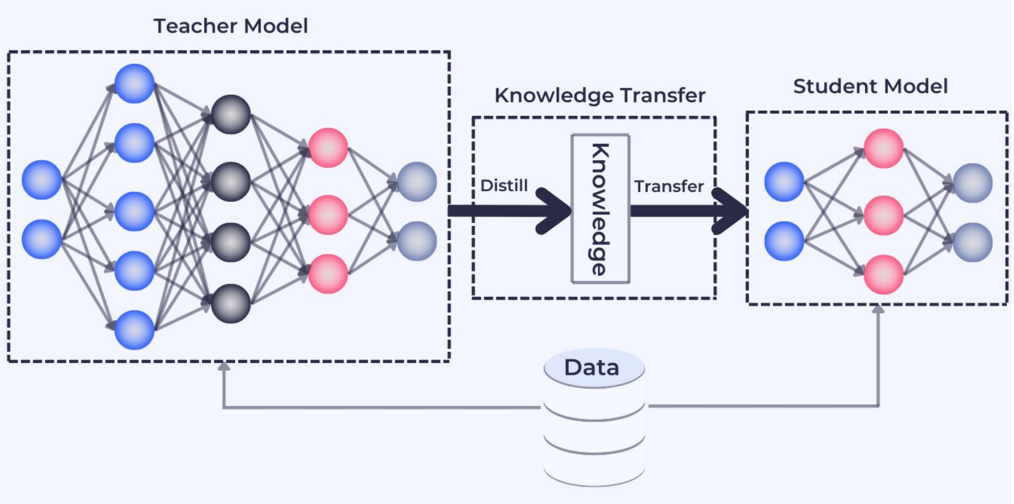


Figure 5. WG−Down and WG−Up with wavelet gating.

Make Diffusion Dynamic, to tackle Challenge 1

Module 2 Distill High-Frequency



$$\mathcal{X}_T^{(i)} = \mathcal{F}[\mathbf{X}_T^{(i)}], \mathcal{X}_S^{(i)} = \mathcal{F}[\mathbf{X}_S^{(i)}], \mathcal{X}^{(i)} = \mathcal{F}[\text{Resize}(\mathbf{x}_0)]$$

$$\mathcal{L}_{\text{freq}} = \sum_i \omega_i \| \mathcal{X}_T^{(i)} - \mathcal{X}_S^{(j)} \|_2^2, \text{where } \omega = |\mathcal{X}^{(i)}|^{\alpha}$$

Boost High-Freq, to tackle Challenge 2

Quantitative Results

FFHQ 256 × 256			
Model	#Param	MACs	FID↓
DDPM [18]	113.7M	248.7G	8.4
P2 [6]	113.7M	248.7G	7.0
LDM [48]	274.1M	96.1G	5.0
Lite-LDM	22.4M(12.2×)	7.9G(12.2×)	17.3(−12.3)
Ours	21.1M(13.0×)	6.7G(14.3×)	10.5(−5.5)

CelebA-HQ 256 × 256			
Model	#Param	MACs	FID↓
Score SDE [59]	65.57M	266.4G	7.2
DDGAN [62]	39.73M	69.9G	7.6
LDM [48]	274.1M	96.1G	5.1
Lite-LDM	22.4M(12.2×)	7.9G(12.2×)	14.3(−9.2)
Ours	21.1M(13.0×)	6.7G(14.3×)	9.3(−4.2)

LSUN-Bedroom 256 × 256			
Model	#Param	MACs	FID↓
DDPM [18]	113.7M	248.7G	4.9
IDDPM [42]	113.7M	248.6G	4.2
ADM [8]	552.8M	1114.2G	1.9
LDM [48]	274.1M	96.1G	3.0
Lite-LDM	22.4M(12.2×)	7.9G(12.2×)	10.9(−7.9)
Ours	21.1M(13.0×)	6.7G(14.3×)	5.2(−2.2)

LSUN-Church 256 × 256			
Model	#Param	MACs	FID↓
DDPM [18]	113.7M	248.7G	4.9
IDDPM [42]	113.7M	248.6G	4.3
ADM [8]	552.8M	1114.2G	1.9
LDM [48]	295.0M	18.7G	4.0
Lite-LDM	32.8M(9.0×)	2.1G(8.9×)	13.6(−9.6)
Ours	33.8M(8.7×)	2.1G(8.9×)	8.4(−4.4)

Table 2. Unconditional generation results comparison to prior DPMs. The results are taken from the original paper, except that DDPM is take from the [6].

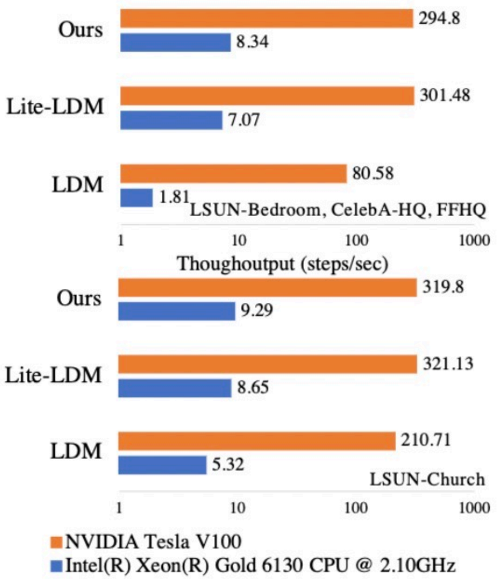


Figure 6. Throughput for unconditional image generation.

Quantitative Results

Method	#Param	MACs	FID↓
IDDPM [42]	273.1M	1416.3G	12.3
ADM [8]	553.8M	1114.2G	10.9
LDM [48]	400.9M	99.8G	10.6
ADM-G [8]	553.8+54.1M	1114.2+72.2G	4.6
LDM-CFG [48]	400.9M	99.8G	3.6
Lite-LDM-CFG	47.0M(8.5×)	11.1G (9.0×)	20.1(−16.5)
Ours-CFG	45.4M(8.8×)	9.9G (10.1×)	10.6(−7.0)

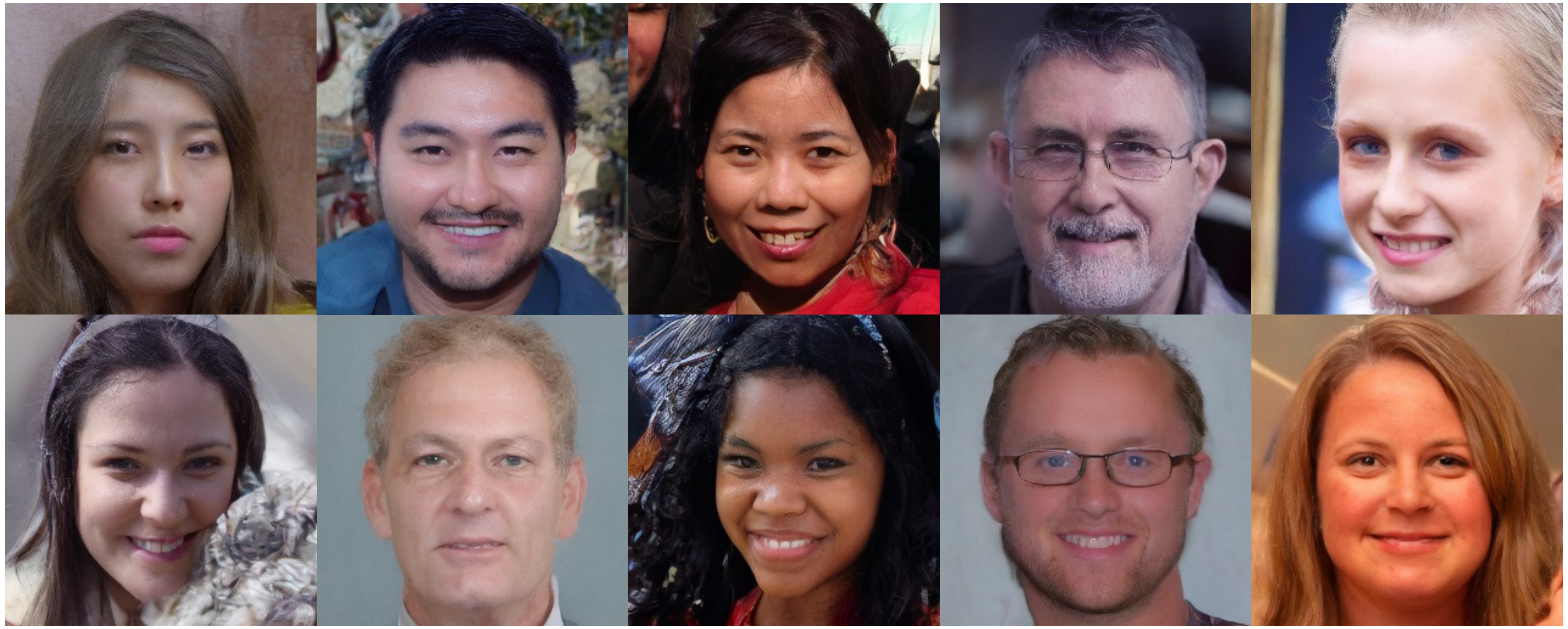
Table 3. Comparison of class-conditional image generation methods on ImageNet [7] with recent state-of-the-art methods. "G" stands for the classifier guidance and "CFG" refers to the classifer-free guidance for conditional image generation.

Method	#Param	FID↓
GLIDE [41]	5.0B	12.24
DALLE2 [45]	5.5B	10.39
Imagen [51]	3.0B	7.27
LDM [48]	1.45B	12.63
Ours	77.6M(18.7×)	18.87

Table 4. Zero-Shot FID on MS-COCO text-to-image generation.

Visualizations: CelebA-HQ

Visualizations: FFHQ



Visualizations: ImageNet

Ablation Study

Method	**FFHQ** 256 × 256							
+ Wavelet Gating		✓			✓		✓	✓
+ Spatial Distill			✓		✓	✓		✓
+ Freq Distill				✓		✓	✓	✓
FID↓	17.3	14.7	16.6	15.3	12.3	12.4	11.4	10.5

Table 5. Ablation study on FFHQ dataset.

Ablation Study

- Lite-LDM lacks recovery for high-freq
- Our SD gets better high-freq reconstruction

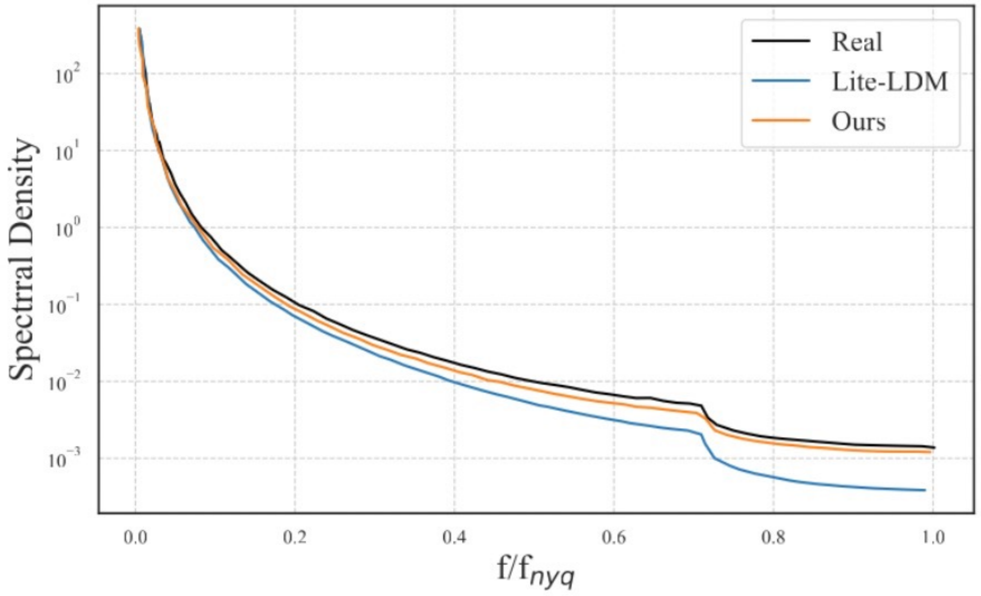


Figure 3. Mean reduced spectrum from real and generated images.

Thanks for Listening