Learning Heterogeneous Global-Local Frequency Dependencies in Diffusion-Based Image Compression

YuBing Luo[1] , **Zekai Ji**[1] , **Jia Qin**[1,†] , **Zhihang Chen**[1] , **Tengyue Guo**[2] , **Pinle Qin**[1,†] , **Rui Chai**[1] , **Jianchao Zeng**[1]

[1]School of Computer Science and Technology, North University of China, Taiyuan, China
[2]School of Automation, Nanjing University of Information Science and Technology, Nanjing, China

Abstract

Diffusion-based image compression has exhibited robust performance. However, most existing methods primarily emphasize spatial domain, causing frequency dependencies to be learned only implicitly. We revisit this problem from a frequency perspective, and observe pronounced heterogeneity between global and local frequency dependencies. Global frequency patterns describe the energy distribution over the entire image, whereas patch level frequency relations govern the coupling of local details. These two forms of dependency differ substantially in both scale and semantic level, yet most methods overlook this distinction. To address this issue, we propose a learning heterogeneous global-local frequency dependencies in diffusion-based image compression which uses Fourier and Mamba jointly models both global and local frequency correlations(FMDiff). The core of FMDiff is the dual branch FMBlock. In the frequency branch, features are decomposed into magnitude and phase, then fed into Mamba separately. Its scanning mechanism captures cross frequency coupling within each patch while aggregating long range frequency context along the sequence. Magnitude phase interactive modulation is then used to explicitly restore spectral information corrupted by compression. The spatial branch provides high level semantic constraints and is fused with the frequency branch during diffusion reconstruction. Extensive experiments show that FMDiff consistently improves performance across multiple datasets.

1 Introduction

Image compression aims to minimize redundancy in data representation while preserving the key information contained in images as much as possible, thereby enabling high fidelity reconstruction [Bellard, 2014; He *et al.*, 2022; Mentzer *et al.*, 2020]. In recent years, diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020] have achieved substantial progress in tasks such as super resolution reconstruction [Saharia *et al.*, 2022; Ma *et al.*, 2023] and image inpainting [Xia *et al.*, 2023]. However, during image compression, diffusion models often prioritize perceptual consistency over pixel level accuracy [Li *et al.*, 2024d; Lei *et al.*, 2023], which may lead to suboptimal fidelity at the patch level, particularly in terms of frequency dependency. Therefore, achieving a better trade-off among bitrate, distortion, and perceptual quality remains a key research problem.

Unifying low distortion and high perceptual quality is primarily challenged by the stable reconstruction of high frequency details after compression. Frequency dependency describes structured correlations in the frequency domain, indicating that they are not independent in distribution or structure [Li *et al.*, 2024a; Luan *et al.*, 2025]. Specifically, global frequency dependency characterizes the energy distribution of the entire image. Local frequency dependency reflects how frequency components within a patch jointly determine edges, textures, and their coupling with content semantics. In compression, high frequency energy is low and is easily suppressed by quantization [Wallace, 1991]. Without explicit constraints on frequency dependency, reconstruction can produce inconsistent cross frequency compositions. This shifts the global energy distribution and damages local structural consistency.

In this paper, we analyze the frequency components of natural images and obtain two key observations. First, as shown in Fig. 1(a) and Fig. 1(c), the Fourier transform exhibits strong separability between magnitude and phase. Exchanging the phase between different images markedly changes structural details, while changing the magnitude mainly affects overall brightness and contrast. This phenomenon suggests that compression degradation more readily perturbs phase(structure), while magnitude remains relatively stable. Therefore, frequency restoration should be modeled from two complementary perspectives, intensity and structure. The intensity prior can be used to guide the recovery of structural details. Second, as shown in Fig. 1(b), cross frequency correlations exhibit pronounced heterogeneity across global and local scales. At the whole image level, frequency relationships tend to be weakly correlated. In contrast, within local patches, the coupling strength across frequencies varies with content. It becomes more evident in texture rich regions and is weak or unstable in flat areas. Existing reconstruc-

†Corresponding authors.

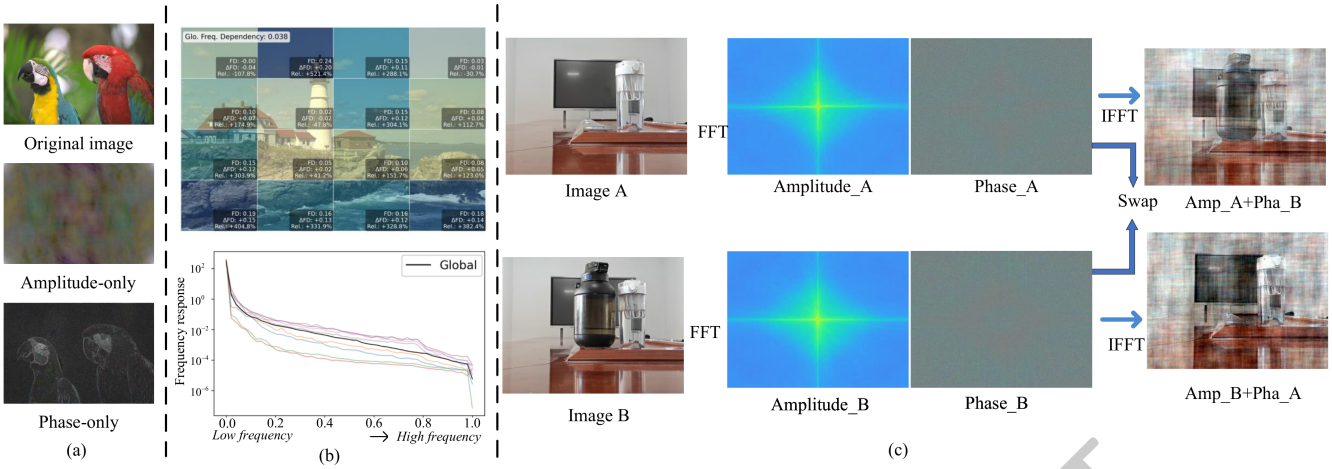


Figure 1: (a) and (c) highlight the semantic difference between magnitude and phase, magnitude mainly reflects the energy distribution of the image, whereas phase is more decisive for structure and texture details. (b) shows the heterogeneity between global and patch level frequency dependencies. The top heatmap visualizes heterogeneity strength (darker indicates stronger), with texture rich regions exhibiting more pronounced heterogeneity and being harder to restore. The bottom plots compare global and local cross frequency relationships via frequency response curves, providing a spectral level view of this heterogeneity. In (b), FD denotes patch level frequency dependency, ΔFD denotes the absolute deviation from the global frequency dependency, and Rel denotes the relative deviation in percentage.

tion methods often focus on global context modeling, or only small scale details. They implicitly assume that different regions share similar frequency relationships. This prior assumption can lead to spectral energy shifts during texture and structure reconstruction, which ultimately produces artificial textures or distortions.

Based on these observations, we propose FMDiff, a Fourier Mamba guided diffusion-based image compression framework that explicitly models global spectral energy distribution and patch level cross frequency coupling. At the core of FMDiff is FMBlock, which performs cooperative modeling with a dual branch design in the spatial and frequency domains. In the frequency branch, we transform features into the Fourier domain and decouple the complex spectrum into magnitude and phase. Magnitude learns global spectral energy relationships and suppresses compression induced energy shifts. Phase encodes structural representations. We then feed them into Mamba separately. Its scanning mechanism captures cross frequency coupling within each patch and aggregates global frequency context along the scanning order. This enables unified encoding of local and global dependencies. We further apply magnitude phase interactive modulation so that the intensity prior can adaptively guide the recovery of global spectral energy and local structural details. This improves the realism and stability of textures. Meanwhile, the spatial branch leverages Mamba long range modeling to extract global semantic context. This context is injected into the frequency branch as conditional information, providing controllable semantic constraints for high frequency restoration. In addition, we introduce learnable scaling factors at skip connections to adaptively balance local details and global consistency. Finally, features from both domains are unified during diffusion denoising, leading to a better trade-off between distortion and perceptual quality.

In summary, our contributions are as follows.

- We analyze dependencies among different frequency components and show that they exhibit pronounced global and local heterogeneity. Global frequency dependency constrains the energy distribution over the entire image. Local dependency determines the consistency of internal textures and semantics. Ignoring this difference can cause spectral energy shifts and local structural imbalance.
- We design a Fourier state space module that comprehensively captures both local and global frequency dependencies. The spatial branch analyzes complex spatial details while also capturing global semantics. The frequency branch decouples Fourier representations and modulates, then interacts them to recover degraded information across the full frequency band.
- Extensive results show that our method achieves better perceptual quality while also attaining lower distortion.

2 Related Work

2.1 Lossy Image Compression

Image compression is essential for storage and transmission. Traditional codecs such as VVC [Bross *et al.*, 2021] remain widely used, but their block-based designs often overlook spatial dependencies. Balle[Ballé *et al.*, 2016] introduced end to end learned compression with hyperpriors, while later works enhance coding and context modeling, e.g., ELIC [He *et al.*, 2022], LIC-TCM [Liu *et al.*, 2023], CCA [Han *et al.*, 2024], and FTIC [Li *et al.*, 2024b]. However, most of these methods primarily optimize rate and distortion, which can lead to perceptual artifacts and a distribution shift from natural images. GAN-based approaches [Mentzer *et al.*, 2020; Muckley *et al.*, 2023] enhance perceptual quality but may sacrifice fidelity. More recently, diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020] have shown strong

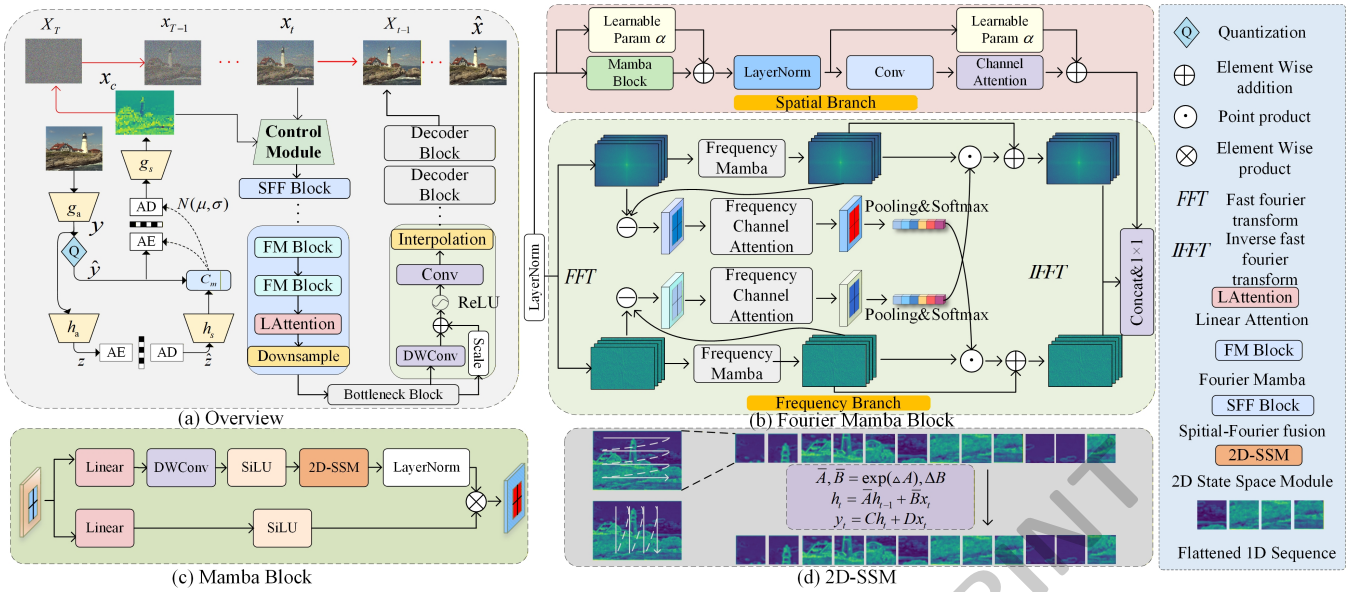


Figure 2: Overview of our FMDiff. Our FMDiff consists of the entire process of diffusion-based image compression, including FMBlock. The core modules of FMBlock are the spatial branch and frequency branch.

perceptual performance via iterative denoising. CDC [Yang and Mandt, 2023] replaces conventional decoders with conditional diffusion models for perceptual reconstruction, while Text+Sketch [Lei *et al.*, 2023] and DiffEIC [Li *et al.*, 2024d] leverage Stable Diffusion for conditional or pretrained generation from compressed representations. However, these methods largely operate in the spatial domain and leave frequency dependencies underexplored, motivating us to revisit diffusion image compression from a frequency perspective.

2.2 Frequency Domain Learning

The Fourier transform is a classic technique in signal processing. Recent studies show that frequency domain representations have been widely used in low light enhancement [Huang *et al.*, 2022], medical image segmentation [Chen *et al.*, 2024a], image restoration and feature enhancement [Chen *et al.*, 2025; Dong *et al.*, 2022; Dong *et al.*, 2024], and network optimization [Yin *et al.*, 2019]. For example, DeepRFT [Mao *et al.*, 2023] performs frequency selection in the Fourier domain to capture global structure and restore details at multiple scales. FourLLIE [Wang *et al.*, 2023] and DMFourLLIE [Zhang *et al.*, 2024] further adopt multi stage strategies to jointly recover illumination and details. However, research that explicitly investigates dependencies among different frequency components remains relatively limited.

2.3 State Space Models

In recent years, state space models (SSMs) have gained attention in vision for their strong long range modeling ability and linear complexity. Mamba [Gu and Dao, 2024], enabled by selective scanning, has been applied to medical segmentation [Wu *et al.*, 2025], video understanding [Li *et al.*, 2024c], and object detection [Chen *et al.*, 2024b]. For low level vision, MambaIR [Guo *et al.*, 2024] addresses local forgetting and channel redundancy via local enhancement and channel attention, while UVM Net [Zheng and Wu, 2024] integrate Mamba with convolution or dual branch designs to capture both local and global cues. Compared with existing methods, our approach explicitly models both global spectral statistics and patch level cross frequency coupling. It further improves spectral correction and stabilizes texture recovery through magnitude phase interactive modulation.

3 Proposed Method

3.1 Overall Network Architecture

As shown in Fig. 2, we illustrate the overall architecture of the proposed FMDiff network in this chapter. First, the analysis transform g_a maps the input image x to a latent representation y with 256 channels. The model then applies a hyper analysis transform h_a to y to obtain the hyper latent representation z, which is fed into the hyper decoder. A joint auto regressive and hierarchical prior model is adopted to predict the entropy parameters (μ, σ) of the latent representation y, thereby modeling the probability distribution of the quantized latent $\hat{y}$ and performing entropy estimation. Finally, $\hat{y}$ is passed through the synthesis transform g_s to produce the conditional features x_c required by the diffusion model.

During the reverse diffusion stage, the noisy image x_t at the current timestep and the conditional features x_c are jointly fed into a control module for feature alignment and fusion. The encoder each stage is built upon a Dual branch fusion(DBF) module to simultaneously model global and local correlations in the frequency domain. Specifically, each DBF module contains two Fourier Mamba blocks, together with a linear attention layer and a downsampling layer. Finally, we employ DDIM sampling [Song *et al.*, 2020] to obtain the reconstructed output $\hat{x}$ by iteratively updating from the initial noise over multiple steps.

3.2 Frequency Branch

Existing studies based on the Fourier Mamba [Zou *et al.*, 2024; Li *et al.*, 2024a] implicitly assume that different regions share similar frequency relationships. However, we observe that frequency dependencies are highly heterogeneous. At the whole image level, frequency statistics are weakly correlated and easily averaged out. Within local patches, cross frequency coupling is stronger in textured regions and weaker in flat areas. To explicitly model spectral information and enhance high frequency recovery, we combine Mamba with the Fourier transform and build a Fourier Mamba module. It consists of a frequency branch and a spatial branch. As illustrated in Fig. 2, the network architecture of the frequency branch is shown in the middle right part. Given a feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we apply the fast Fourier transform to obtain its spectrum:

$$F = \mathcal{F}(X) \in \mathbb{C}^{C\times H\times W}, \quad A_x = |F|, \quad P_x = \angle(F)$$

Here, A_x denotes the magnitude, which represents the spectral intensity distribution. P_x denotes the phase, which directly determines structures and textures. The degradation introduced by compression can be viewed as a perturbation to F. It can be reflected as magnitude loss and phase shifts. Therefore, instead of learning an difficult mapping directly in the frequency domain, we model A_x and P_x separately. This design allows the model to learn with two explicit objectives, correcting energy imbalance and mitigating structural shifts. Furthermore, to capture cross frequency coupling within local regions, we partition and serialize A_x and P_x according to the frequency domain arrangement and feed them into Mamba Block separately. Specifically, we divide A_x and P_x into 8×8 patches with stride 8.

$$P'_x = \mathrm{Mamba}(P_x)$$
$$A'_x = \mathrm{Mamba}(A_x)$$

To mitigate the directional bias introduced by one dimensional sequence flattening, we encode the frequency domain token sequence using row and column directional serpentine scanning. Intuitively, Mamba's scanning mechanism can establish long range dependencies along the sequence. This allows it to encode the coordination among frequency components within each local patch and also absorb global spectral context across patches. As a result, it can capture inconsistencies between local and global frequency patterns. Next, reconstructing directly using only A'_x and P'_x can easily introduce irrelevant spectral changes. Here, we treat the frequency domain residual as the model's correction for compression degradation:

$$\Delta A_x = A'_x - A_x$$
$$\Delta P_x = P'_x - P_x$$

A larger residual magnitude indicates that the corresponding frequency component is more severely corrupted. Based on this intuition, we use channel attention and global pooling to compress the residual into a channel wise gating vector, which is used to produce adaptive restoration:

$$g_A = \sigma(\mathrm{pool}(\mathrm{CA}(\Delta A_x)))$$

$$g_P = \sigma(\mathrm{pool}(\mathrm{CA}(\Delta P_x)))$$

Here, pool denotes global average pooling, CA denotes channel attention, and σ denotes the Sigmoid. The g controls channel selection and modulation strength in the spectrum. Moreover, magnitude and phase are not independent. Recovering realistic textures often requires them to remain consistent in local frequency compositions. To this end, we adopt symmetric interactive modulation. The g_A modulates the phase, and the g_P modulates the magnitude. This yields a restoration mechanism in which the two branches mutually constrain and guide each other:

$$\hat{P}_x = P'_x + P'_x \odot g_A$$
$$\hat{A}_x = A'_x + A'_x \odot g_P$$

Here, $\odot$ denotes element wise multiplication. Finally, we apply the inverse Fourier transform to the restored magnitude and phase to obtain the output features.

3.3 Spatial Branch

Although Mamba provides efficient long range dependency modeling for sequences, directly flattening two dimensional features into a one dimensional sequence introduces a structural issue. Local neighborhoods in two dimensions no longer remain adjacent under one dimensional indexing. Pixels that are adjacent in 2D may be far apart in the 1D scanning order. This makes state propagation in the state space model favor correlations along the scan path, which weakens true local pixel coupling. This leads to a typical inconsistency, global context can be modeled, but the details required for local textures are insufficient. As a result, feature alignment between global semantics and local details becomes unstable.

To this end, we follow prior work such as [Zheng and Wu, 2024; Guo *et al.*, 2024], we introduce a learnable scaling factor in the spatial branch as a local consistency correction term. The goal is to retain Mamba's global modeling capability while adaptively adjusting the local information during skip fusion. Specifically, as illustrated in the upper part of Fig. 2(b), given a feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we first apply normalization and feed it into Mamba Block to obtain a global context representation:

$$G = \mathrm{LN}(X'), \quad X' = \mathrm{Mamba}(X)$$

Here, G mainly represents global semantics. It is used to provide semantic constraints for frequency domain restoration. Next, to explicitly compensate for the local neighborhood relationships, we generate a scaling gate S in the residual branch. It modulates the strength of local information injected at the skip fusion stage:

$$Y = G + S \odot X$$

Here, $S \odot X$ denotes adaptive reweighting of local representations. In textured regions, the gate tends to strengthen local information to mitigate structural distortion. In flat regions, it suppresses local responses to avoid noise. Finally, we introduce channel attention to recalibrate channel responses. The state space model first produces diverse candidate channel representations.

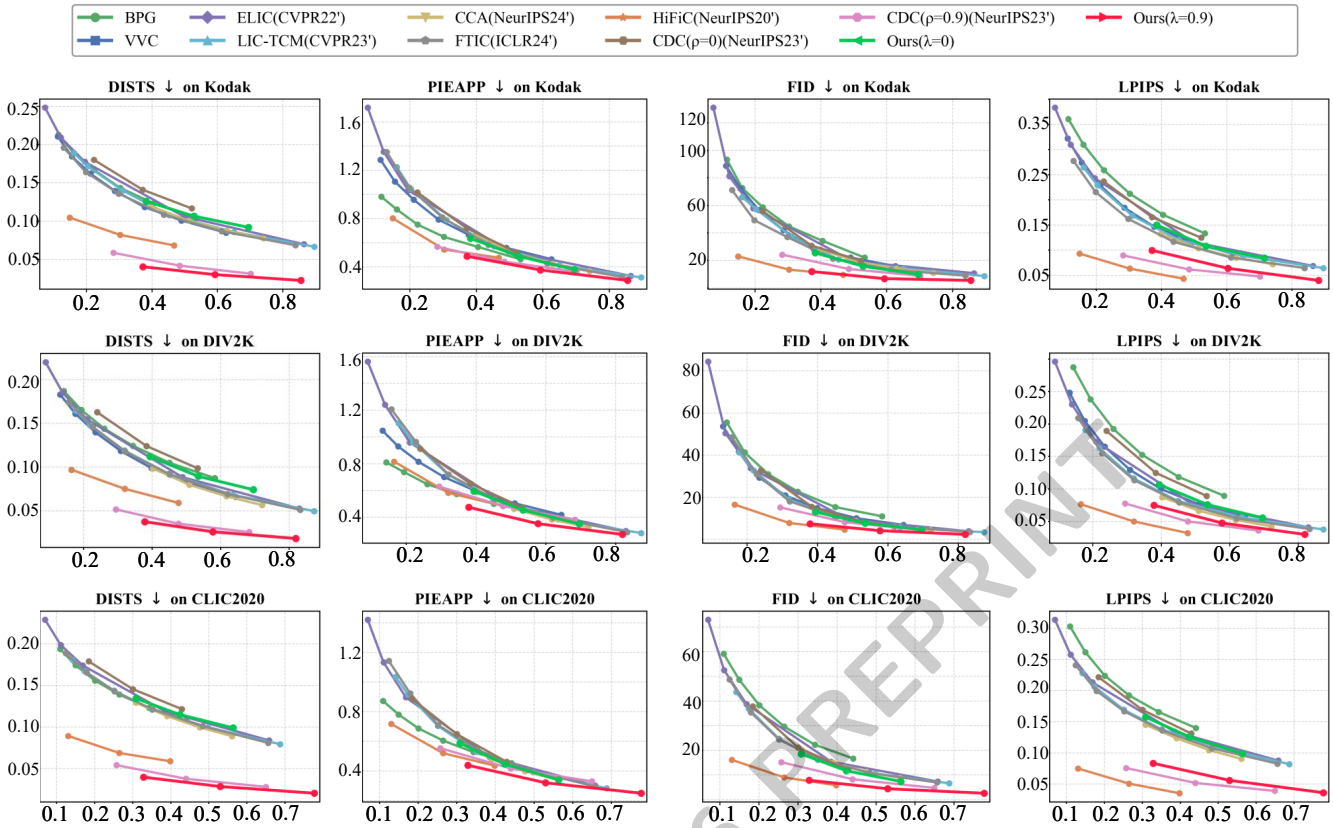


Figure 3: Quantitative comparisons with state-of-the-art methods in terms of perceptual quality on the Kodak, DIV2K, and CLIC2020.

4 Experiment and Analysis

4.1 Experimental Setup

Implementation details: In the warm up stage, the Lagrangian multiplier is fixed. To obtain different bitrates, we set β to 0.0032, 0.0064, 0.0128, 0.0512, 0.1024, and 0.2048. We warm up FMDiff for 500K steps with a learning rate of 10^{-5}, then train for another 1000K steps with exponential decay, reducing the learning rate by 20% every 100K steps. The batch size is 4. For diffusion, we use 8193 timesteps with a cosine variance schedule. At inference, we adopt DDIM sampling [Song *et al.*, 2020] with 65 steps.

Datasets: We train on the public Vimeo-90K dataset [Xue *et al.*, 2019] with 90,000 clips, each containing seven 448×256 frames. For evaluation, we use Kodak [Franzen, 2013], DIV2K [Agustsson and Timofte, 2017], and CLIC2020 [Toderici *et al.*, 2020]. Kodak includes 24 images at 768×512, while DIV2K and CLIC2020 contain 100 and 428 high resolution images, respectively.

Metrics: For quantitative evaluation, we measure perceptual quality using DISTS [Ding *et al.*, 2020], PIEAPP[Prashnani *et al.*, 2018], FID[Heusel *et al.*, 2017], and LPIPS[Zhang *et al.*, 2018]. We also report distortion metrics, including PSNR and MS-SSIM [Wang *et al.*, 2003], and use BPP to evaluate compression performance.

4.2 Comparisons With State-of-the-art Methods

To comprehensively evaluate the proposed FMDiff, we compare it with representative state-of-the-art image compression methods. This set includes conventional standard codecs such as BPG [Bellard, 2014] and VVC [Bross *et al.*, 2021], VAE-based methods such as ELIC [He *et al.*, 2022], LIC-TCM [Liu *et al.*, 2023], CCA [Han *et al.*, 2024], and FTIC [Li *et al.*, 2024b], as well as HiFiC [Mentzer *et al.*, 2020], which uses GAN-based generation to improve perceptual quality. In addition, we include CDC [Yang and Mandt, 2023], a diffusion-based image compression method, as an important baseline in the diffusion family.

1) Quantitative Comparisons: Figs. 3 and 4 show rate–distortion curves on Kodak, DIV2K, and CLIC2020. Fig. 3 reports perceptual metrics, while Fig. 4 reports fidelity metrics. Following CDC, we evaluate two settings: a distortion oriented FMDiff with $\lambda = 0$ and a perception oriented FMDiff with $\lambda = 0.9$, prioritizing fidelity and perceptual quality, respectively.

Fig. 3 plots rate-perception curves on the three datasets. FMDiff consistently outperforms BPG, VVC, ELIC, CCA, LIC-TCM, and FTIC across all perceptual metrics, and surpasses generative codecs such as HiFiC and CDC with lower DISTS, FID, and PIEAPP, indicating closer perceptual similarity to the originals.

Fig. 4 compares rate-distortion performance. Our distor-

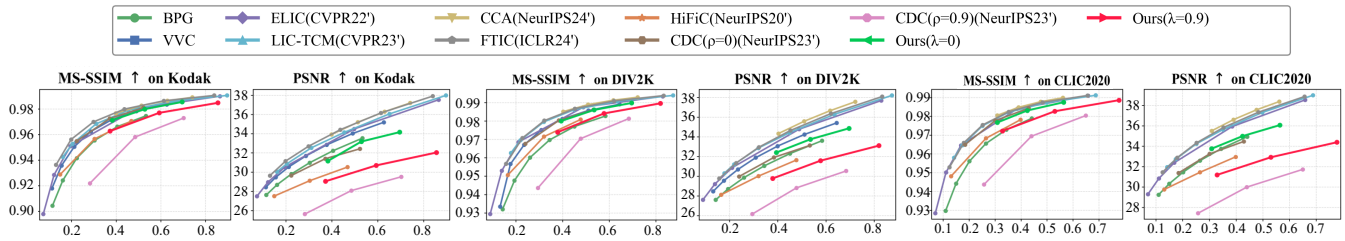


Figure 4: Quantitative comparisons with state-of-the-art methods in terms of fidelity quality on the Kodak, DIV2K, and CLIC2020.

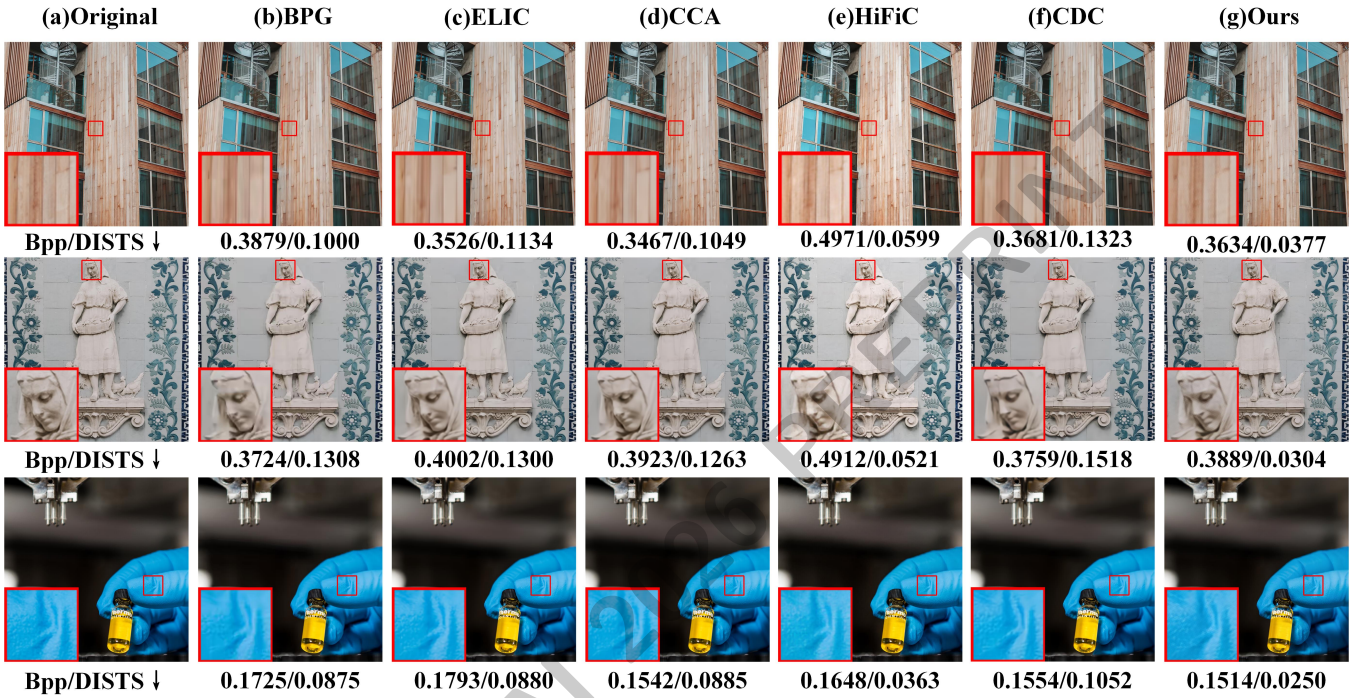


Figure 5: Visualization of the reconstructed images from Kodak dataset. The metrics are [BPP↓ DISTS↓]

tion oriented model achieves distortion comparable to BPG, VVC, ELIC, CCA, LIC-TCM, and FTIC. The perceptual oriented model emphasizes realistic detail synthesis via diffusion and thus trades off pixel level accuracy, but both variants outperform other generative codecs in fidelity; in particular, the perceptual model sacrifices some fidelity for improved realism.

2) Qualitative comparison: In Fig. 5, we visualize reconstructions produced by most of the compared methods. The reconstructions from BPG, ELIC, and CCA exhibit noticeable blur and artifacts. HiFiC and CDC produce sharper edges and textures. However, they also introduce spurious details that do not exist in the original images. Our method achieves strong semantic consistency and more faithful local details. It also attains better DISTS at a lower bitrate.

3) Complexity Comparisons: We further compare the complexity with state-of-the-art image compression methods. We report runtime under the available implementations or reported settings for reference. As shown in Table 1, we report the mean and standard deviation of the encoding and decoding time in seconds on the Kodak dataset for four categories of methods, including traditional codecs, VAE-based methods, GAN-based methods, and diffusion-based methods. The results indicate that diffusion-based methods generally incur higher encoding and decoding complexity than VAE-based and GAN-based approaches. Nevertheless, the encoder of FMDiff is notably faster than Text+Sketch [Lei *et al.*, 2023] and DiffEIC [Li *et al.*, 2024d].

4.3 Ablation Studies

To thoroughly analyze the role of key modules in FMDiff, we conducted ablation studies on FMBlock on the Kodak dataset. As shown in Table 2, the ablation experiments systematically investigated the impact of each component.

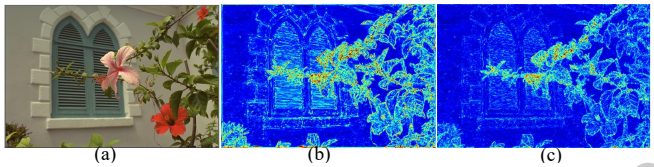
1) Effects of FMBlock. We ablate FMBlock by removing either the spatial or frequency branch and evaluate distortion and perceptual metrics (Table 2). For w/o Spat., the explicit spatial branch is removed while the cross attention condition for the frequency branch is retained. Both variants

Type	Method	Denoising Step	Encoding Speed	Decoding Speed	Platform
Traditional method	VVC	–	13.862 ± 9.821	0.066 ± 0.006	13th Core i9-13900K
VAE-based methods	ELIC	–	0.192 ± 0.019	0.113 ± 0.002	RTX4090
	CCA	–	0.117 ± 0.008	0.134 ± 0.010	RTX4090
GAN-based methods	HiFiC	–	0.038 ± 0.004	0.059 ± 0.004	RTX4090
Diffusion-based methods	Text+Sketch	25	62.045 ± 0.516	12.028 ± 0.413	RTX4090
	DiffEIC	50	0.128 ± 0.005	4.574 ± 0.006	RTX4090
	FMDiff(Ours)	25	0.010 ± 0.004	0.959 ± 0.002	RTX4090
	FMDiff(Ours)	65	0.016 ± 0.002	2.499 ± 0.004	RTX4090

Table 1: Encoding and decoding speed on the Kodak dataset in terms of seconds.

Methods	Frequency branch	Spatial branch	Relative change (%)	
			PSNR	LPIPS
Baseline	×	×	10.10	19.20
w/o Spat.	✓	×	1.43	3.51
w/o Freq.	×	✓	5.59	18.44
Full FMDiff	✓	✓	0	0

Table 2: Ablation study for investigating the components of FMBlock on the Kodak dataset.



(a) (b) (c)

Figure 6: Error map (b) is the difference between the reconstruction by Spatial branch and the ground truth. Error map (c) is the difference between our reconstruction result and GT with fewer huge error values. (a) Input. (b) Spat. (c) Ours.

show worse performance than the full model, indicating that the two branches are complementary.

We further visualize error maps for the variant without FMBlock in Fig. 6. The results suggest that FMBlock better captures joint spatial frequency features, and that the frequency branch in particular improves the recovery of textures and fine details.

2) Analysis of the Frequency Branch. Table 3 reports ablations of the frequency branch. Using both amplitude and phase outperforms either alone, confirming their complementarity. Adding cross modulation consistently improves both fidelity and perceptual quality. Finally, removing the Mamba block degrades perceptual performance, at comparable bitrate, enabling Mamba markedly lowers DISTS while keeping PSNR similar, suggesting it captures long range dependencies.

3) Analysis of the Spatial Branch. Table 4 reports ablations of the spatial branch. The full spatial branch achieves the highest PSNR, while the LPIPS values are close among different variants. This suggests that the spatial branch mainly improves fidelity, whereas perceptual quality is more strongly affected by the frequency branch.

Methods	bpp	PSNR↑	DISTS↓
Only amplitude	0.4425	32.1499	0.1355
Only phase	0.4563	32.1497	0.1390
Amp + Phase	0.4450	32.2320	0.1347
w/o cross modulation	0.4529	32.1898	0.1369
w/ cross modulation	0.4497	32.2538	0.1333
w/o Mamba	0.4523	32.2130	0.1383
w/ Mamba	0.4563	32.2882	0.1307

Table 3: Performance comparison of frequency branch variants.

Methods	bpp	PSNR↑	LPIPS↓
w/o Mamba	0.4619	<u>32.1810</u>	**0.1486**
w/o Scale	**0.4389**	32.1036	<u>0.1508</u>
Full Spatial Branch	<u>0.4448</u>	**32.2094**	0.1522

Table 4: Performance comparison of spatial branch variants.

5 Conclusion

We propose FMDiff, a Fourier Mamba reconstruction methods for diffusion-based image compression, designed to mitigate spectral energy shifts and local texture inconsistencies caused by high frequency degradation. Motivated by the global local heterogeneity of frequency dependencies, FMDiff uses the FMBlocks to explicitly model both, the Fourier branch decouples magnitude and phase, captures patch-wise cross frequency coupling and long range frequency context using Mamba scanning, and restores corrupted spectra via magnitude phase interactive modulation. The spatial branch leverages Mamba to provide semantic and spatial consistency, with a learnable scaling factor and cross domain fusion balancing global structure and local details. Extensive experiments show that FMDiff improves the distortion perception trade-off across multiple datasets, yielding more realistic textures and more stable reconstructions.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was partly supported by the Natural Science Foundation of China (62302466), Natural Science Basic Research Program of Shanxi Province (202303021212188).

References

[Agustsson and Timofte, 2017] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

[Ballé *et al.*, 2016] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[Bellard, 2014] Fabrice Bellard. Bpg image format, 2014.

[Bross *et al.*, 2021] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[Chen *et al.*, 2024a] Jinsha Chen, Gang Yang, Aiping Liu, Xun Chen, and Ji Liu. Sfe-net: Spatial-frequency enhancement network for robust nuclei segmentation in histopathology images. *Computers in Biology and Medicine*, 171:108131, 2024.

[Chen *et al.*, 2024b] Tianxiang Chen, Zi Ye, Zhentao Tan, Tao Gong, Yue Wu, Qi Chu, Bin Liu, Nenghai Yu, and Jieping Ye. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Chen *et al.*, 2025] Linwei Chen, Lin Gu, Liang Li, Chenggang Yan, and Ying Fu. Frequency dynamic convolution for dense image prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30178–30188, 2025.

[Ding *et al.*, 2020] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

[Dong *et al.*, 2022] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[Dong *et al.*, 2024] Jiahua Dong, Hongliu Li, Yang Cong, Gan Sun, Yulun Zhang, and Luc Van Gool. No one left behind: Real-world federated class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2054–2070, 2024.

[Franzen, 2013] Rich Franzen. True color kodak images. http://r0k.us/graphics/kodak/, 2013.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

[Guo *et al.*, 2024] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pages 222–241. Springer, 2024.

[Han *et al.*, 2024] Minghao Han, Shiyin Jiang, Shengxi Li, Xin Deng, Mai Xu, Ce Zhu, and Shuhang Gu. Causal context adjustment loss for learned image compression. *arXiv preprint arXiv:2410.04847*, 2024.

[He *et al.*, 2022] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5718–5727, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2022] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*, pages 163–180. Springer, 2022.

[Lei *et al.*, 2023] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text+ sketch: Image compression at ultra low rates. *arXiv preprint arXiv:2307.01944*, 2023.

[Li *et al.*, 2024a] Dong Li, Yidi Liu, Xueyang Fu, Senyan Xu, and Zheng-Jun Zha. Fouriermamba: Fourier learning integration with state space models for image deraining. *arXiv preprint arXiv:2405.19450*, 2024.

[Li *et al.*, 2024b] Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware transformer for learned image compression. iclr, 2024, 19. *arXiv preprint arXiv:2310.16387*, 2024.

[Li *et al.*, 2024c] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024.

[Li *et al.*, 2024d] Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Towards extreme image compression with latent feature guidance and diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Liu *et al.*, 2023] Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14388–14397, 2023.

[Luan *et al.*, 2025] Xin Luan, Huijie Fan, Qiang Wang, Nan Yang, Shiben Liu, Xiaofeng Li, and Yandong Tang. Fmambair: A hybrid state space model and frequency domain for image restoration. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[Ma *et al.*, 2023] Yiyang Ma, Huan Yang, Wenhan Yang, Jianlong Fu, and Jiaying Liu. Solving diffusion odes with optimal boundary conditions for better image super-resolution. *arXiv preprint arXiv:2305.15357*, 2023.

[Mao *et al.*, 2023] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1905–1913, 2023.

[Mentzer *et al.*, 2020] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020.

[Muckley *et al.*, 2023] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023.

[Prashnani *et al.*, 2018] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018.

[Saharia *et al.*, 2022] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Toderici *et al.*, 2020] George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Balle, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. Workshop and challenge on learned image compression (clic2020), 2020.

[Wallace, 1991] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[Wang *et al.*, 2003] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[Wang *et al.*, 2023] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7459–7469, 2023.

[Wu *et al.*, 2025] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *Neurocomputing*, 624:129447, 2025.

[Xia *et al.*, 2023] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023.

[Xue *et al.*, 2019] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019.

[Yang and Mandt, 2023] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36:64971–64995, 2023.

[Yin *et al.*, 2019] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2024] Tongshun Zhang, Pingping Liu, Ming Zhao, and Haotian Lv. Dmfourllie: dual-stage and multi-branch fourier network for low-light image enhancement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7434–7443, 2024.

[Zheng and Wu, 2024] Zhuoran Zheng and Chen Wu. U-shaped vision mamba for single image dehazing. *arXiv preprint arXiv:2402.04139*, 2024.

[Zou *et al.*, 2024] Zhen Zou, Hu Yu, Jie Huang, and Feng Zhao. Freqmamba: Viewing mamba from a frequency perspective for image deraining. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 1905–1914, 2024.