MBMAMBA: WHEN MEMORY BUFFER MEETS MAMBA FOR STRUCTURE-AWARE IMAGE DEBLURRING

Anonymous authors
Paper under double-blind review

ABSTRACT

The Mamba architecture has emerged as a promising alternative to CNNs and Transformers for image deblurring. However, its flatten-and-scan strategy often results in local pixel forgetting and channel redundancy, limiting its ability to effectively aggregate 2D spatial information. Although existing methods mitigate this by modifying the scan strategy or incorporating local feature modules, it increase computational complexity and hinder real-time performance. In this paper, we propose a structure-aware image deblurring network without changing the original Mamba architecture. Specifically, we design a memory buffer mechanism to preserve historical information for later fusion, enabling reliable modeling of relevance between adjacent features. Additionally, we introduce an Ising-inspired regularization loss that simulates the energy minimization of the physical system's "mutual attraction" between pixels, helping to maintain image structure and coherence. Building on this, we develop MBMamba. Experimental results show that our method outperforms state-of-the-art approaches. Our code is available at `https://anonymous.4open.science/r/MBMamba-C83B`

1 INTRODUCTION

Image deblurring seeks to restore a sharp latent image from a blurred observation. Given the ill-posed nature of this inverse problem, traditional methods Karaali & Jung (2017); Dong et al. (2011) often incorporate handcrafted features or explicit priors to narrow the solution space toward natural images. However, designing such priors is not only challenging but also lacks generalizability, making them less effective in real-world applications.

Benefiting from the rapid progress of deep learning in high-level vision tasks, numerous data-driven methods have adopted CNNs as backbone architectures Kuo et al. (2025); Cui et al. (2024). Although convolutions are effective at capturing local patterns, their inherent limitations—such as restricted receptive fields and content-agnostic operations—hinder their ability to model long-range dependencies. To address these issues, several approaches Xu et al. (2025); Feng et al. (2023) have introduced transformers into image deblurring, achieving superior performance over CNN-based methods by leveraging attention mechanisms within spatial windows or across channel dimensions. Nonetheless, these methods still face challenges: they either struggle to fully utilize spatial details or are constrained by coarse partitioning strategies that limit the extraction of fine-grained features within individual windows.

State space models Dao & Gu (2024); Mehta et al. (2022), especially the enhanced Mamba variant, have recently attracted considerable attention for their ability to model long-range dependencies with linear complexity. However, the widely adopted flatten-and-scan strategy often leads to the loss of local pixel information and redundant channel features, limiting the model's capacity to effectively capture 2D spatial structures. Given the importance of local detail and channel-wise cues in image deblurring, directly applying state space models typically results in subpar performance. To overcome these limitations, recent studies have proposed a variety of scanning strategies, such as four-directional scanning Guo et al. (2024), shortest path traversal Zhou et al. (2025), and slice-and-scan Liu et al. (2025), as well as the integration of local feature enhancement modules. However,

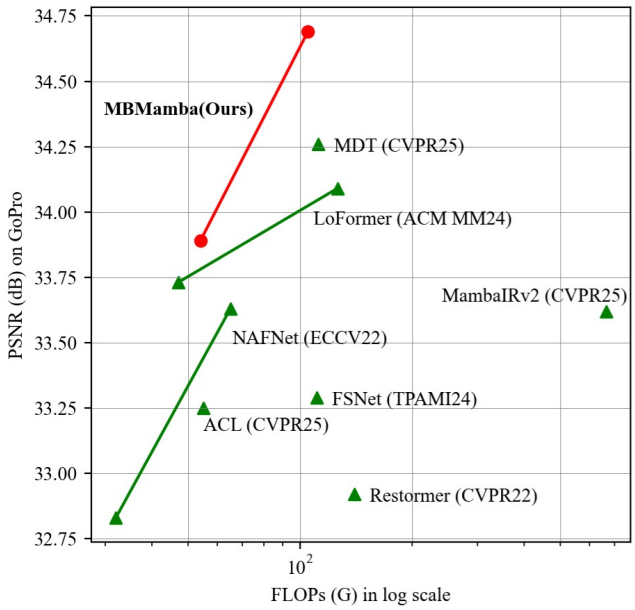

Figure 1: Computational cost vs. PSNR of models on the GoPro dataset Nah et al. (2016). Our MBMamba achieve the SOTA performance while simultaneously reducing computational costs.

these solutions often require multiple scans or additional components, which inevitably increase computational overhead and compromise real-time efficiency.

Based on the above analysis, a natural question arises: can we design a structure-aware image deblurring network that efficiently integrates both local and global features without increasing the number of scans or adding extra local modules? To address this, we propose MBMamba, which incorporates several key components. Specifically, we design a memory buffering mechanism that stores historical information, which is then fused with the current features via a cross-attention mechanism, enabling more robust modeling of dependencies between adjacent features. In addition, we present an Ising-inspired regularization loss that mimics the energy minimization process of physical systems, capturing the "mutual attraction" between pixels to better preserve image structure and coherence. Furthermore, to fully exploit the decoder's potential, we utilize a pre-trained encoder and implement multiple sub-decoders alongside multi-scale output designs, thereby easing the training process. As demonstrated in Figure 1, MBMamba achieves state-of-the-art results while maintaining computational efficiency compared to existing approaches.

The main contributions of this work are:

1. We propose MBMamba, a structure-aware image deblurring network that efficiently integrates both local and global features.
2. We design a memory buffer mechanism that preserves historical information for subsequent fusion, enabling reliable modeling of relationships between adjacent features.
3. We present an Ising-inspired regularization loss to capture the "mutual attraction" between pixels to better maintain image structure and coherence.
4. Extensive experiments show that MBMamba achieves competitive performance compared to state-of-the-art methods.

2 Related Work

2.1 Hand-crafted prior-based methods

Given the inherently ill-posed nature of image deblurring, early methods Karaali & Jung (2017); Chen et al. (2020); Wen et al. (2021) often relied on manually designed priors to narrow the solution space. Although some approaches have attempted to incorporate additional sensor data to estimate

blur kernels more accurately Rong et al. (2024), these prior-driven strategies generally struggle to model the complex degradation process and often lack robustness and general applicability.

2.2 CNN-BASED METHODS

With the rapid progress of deep learning, many approaches have shifted from manually crafting image priors to developing various CNN-based models for image deblurring. To effectively balance spatial detail preservation and contextual understanding, MPRNet Zamir et al. (2021) introduces cross-stage feature fusion to utilize features from multiple processing stages. MIRNet-V2 Zamir et al. (2022b) adopts a multi-scale design to extract richer features for restoration tasks, while IRNeXt Cui et al. (2023a) reconsiders CNN architecture to build a more efficient and effective network. NAFNet Chen et al. (2022) simplifies the model structure by analyzing and refining baseline components, removing or replacing non-linear activations. SFNet Cui et al. (2023b) and FSNet Cui et al. (2024) propose dynamic and compact frequency selection modules that identify the most informative components for restoration. ELEDNet Kim et al. (2025) combines cross-modal information with low-pass filtering to suppress noise while retaining structure. MR-VNet Roheda et al. (2024) leverages Volterra layers for efficient blur removal. DSDNe Kuo et al. (2025) formulates the deblurring task as separate data and regularization sub-problems to improve speed and accuracy. While these CNN-based methods outperform traditional prior-driven approaches, their reliance on local receptive fields inherently limits their ability to address long-range degradation patterns effectively.

2.3 TRANSFORMER-BASED METHODS

Thanks to their content-aware global receptive fields, Transformer architectures Vaswani et al. (2017) have recently become increasingly popular in image restoration, consistently outperforming traditional CNN-based methods. However, image deblurring often involves high-resolution inputs, where the quadratic computational complexity of standard attention mechanisms leads to heavy processing overhead. To mitigate this, models like Uformer Wang et al. (2022), SwinIR Liang et al. (2021), and U^{2}former Feng et al. (2023) adopt window-based self-attention to localize computation. Yet, this strategy limits the capacity to capture full contextual information within each patch. To improve efficiency, Restormer Zamir et al. (2022a) and MRLPFNet Dong et al. (2023) shift attention computation to the channel dimension, achieving linear complexity. Nevertheless, this design compromises spatial feature modeling. FFTformer Kong et al. (2023) explores attention computation in the frequency domain using Fourier transforms, but requires inverse operations that introduce additional cost. MAT Xu et al. (2025) proposes a motion-adaptive Transformer, leveraging motion cues to build more robust global dependencies. For more realistic deblurring scenarios, HI-Diff Chen et al. (2023) incorporates diffusion models to generate informative priors, which are then hierarchically integrated to enhance the deblurring process.

2.4 STATE SPACES MODEL

State space models Gu et al. (2021); Smith et al. (2023) have recently gained prominence for their efficiency in modeling long-range dependencies with linear computational complexity. Mamba Dao & Gu (2024) enhances this framework with a selective mechanism and a hardware-friendly parallel algorithm. Nonetheless, when applied to image restoration tasks, the standard Mamba still suffers from issues such as local pixel information loss and redundant channel features. To overcome these limitations, several recent approaches have explored alternative scanning strategies and incorporated modules aimed at enhancing local feature representation. MambaIR Guo et al. (2024) introduces a four-directional unfolding method combined with channel attention. ALGNet Gao et al. (2024) leverages a fusion module that jointly captures global and local information for more accurate feature extraction. LoFormer Mao et al. (2024) applies local channel-wise self-attention in the frequency domain to model cross-covariance within low- and high-frequency local regions. XYScanNet Liu et al. (2025) proposes an alternating slice-and-scan strategy along intra- and inter-slice directions. MambaIRV2 Guo et al. (2025) further enhances Mamba by introducing non-causal modeling similar to ViTs, enabling a more attentive state space representation.

However, these enhancements typically come at the cost of increased complexity, requiring additional scanning passes or supplementary modules, which hinders real-time performance. To address

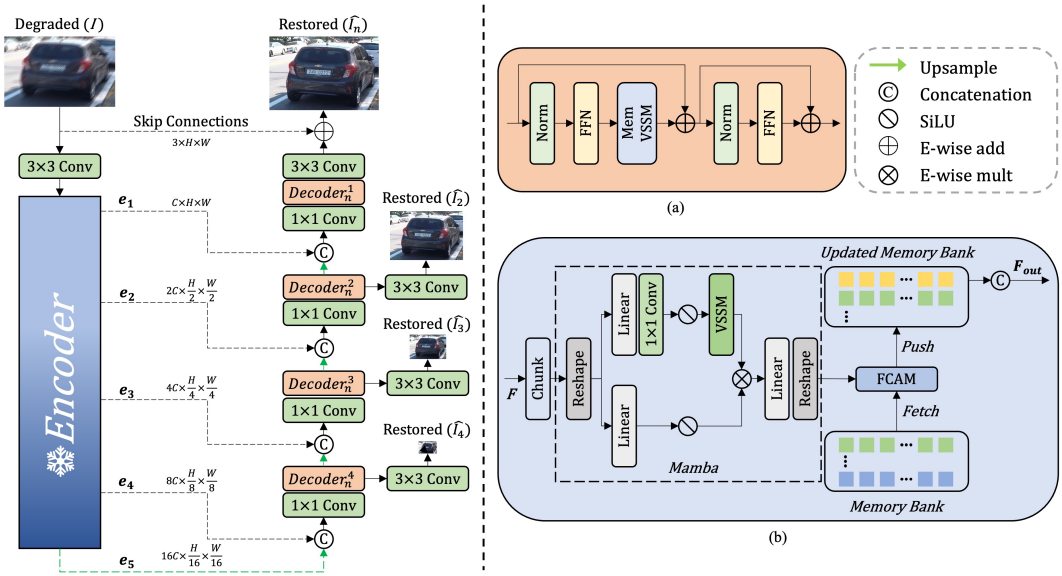

Figure 2: The overall architecture of the proposed MBMamba: (a) The decoder is composed of vision state space models equipped with a memory buffering mechanism (MemVSSM); (b) The internal structure of MemVSSM.

these challenges, we propose MBMamba to effectively integrates both local and global features for image deblurring—without incurring the overhead of multiple scans or added modules.

3 METHOD

In this section, we first provide an overview of the entire MBMamba pipeline. We then dive into the details of the proposed decoder, which comprises vision state space models equipped with a memory buffering mechanism (MemVSSM). Lastly, we present the Ising-inspired regularization loss.

3.1 OVERALL PIPELINE

Our proposed MBMamba, illustrated in Figure 2, consists of a frozen encoder and n sub-decoders, each comprising four decoding stages. Given a degraded image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, MBMamba first uses a convolutional layer to extract shallow features $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels of the feature map, respectively. These shallow features are passed through a pre-trained encoder to obtain multi-scale encoder features e_i (where $i = 1, 2, 3, 4, 5$) at different scales. The encoder features are then fed into the decoder, which generates decoder features d_n^i at different scales, progressively restoring them to their original size. Notably, since MBMamba incorporates multiple sub-decoders, the input to each subsequent sub-decoder is the output of the previous one. Finally, a convolutional layer is applied to the refined features to generate the residual image $\mathbf{X_n} \in \mathbb{R}^{H\times W\times 3}$ for n_{th} sub-decoder. This residual image is added to the degraded image to produce the restored output: $\hat{\mathbf{I}}_\mathbf{n} = \mathbf{X_n} + \mathbf{I}$.

3.2 MEMVSSM

The Mamba architecture has recently emerged as a promising alternative to CNNs and Transformers for image deblurring. However, its flatten-and-scan strategy tends to cause local pixel forgetting and channel redundancy, which undermines its ability to effectively capture 2D spatial information. While some existing approaches address this issue by altering the scanning strategy or adding local feature modules, these modifications often lead to increased computational cost and reduced real-time performance. To better capture both local and global features without increasing scan frequency or introducing additional modules, we propose a vision state space models equipped with a memory buffering mechanism (MemVSSM) in the decoder. As shown in Figure 2(a), given the input features

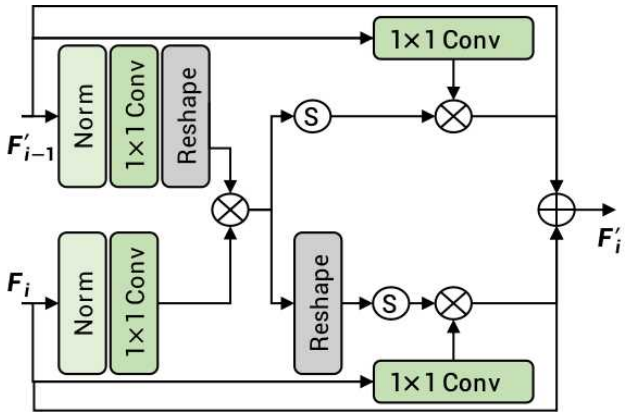

Figure 3: The structure of feature cross-attention mechanism (FCAM).

at the $(l-1)_{th}$ block X_{l-1}, the procedures of decoder can be defined as:

$$\begin{aligned} X_l^{'} &= MemVSSM(FFN(Norm(X_{l-1}))) \oplus X_{l-1} \\ X_l &= FFN(Norm(X_l^{'})) \oplus X_l^{'} \end{aligned} \tag{1}$$

Previous Mamba-based methods Liu et al. (2025); Guo et al. (2025) input image features into the VSSM at once, which necessitates additional scan iterations or the introduction of local feature modules to mitigate local pixel forgetting and channel redundancy caused by an excessive number of hidden units. In contrast, as illustrated in Figure 2(b), our MemVSSM first divides the input feature F into N channel-wise chunks.

$$F_1^d, ..., F_i^d, ..., F_N^d = Chunk(F) \tag{2}$$

Each chunk is sequentially processed by Mamba to produce enriched feature representations F_i, where i denotes the chunk index. A memory bank is then introduced to retain historical information for subsequent fusion. Specifically, we define the memory bank with the depth of K, which indicates the number of stored historical feature sets. The memory bank operates under a first-in, first-out (FIFO) policy for feature updating. In detail, during the processing of the i-th chunk, we first retrieve K previous outputs $[F'_{i-1}, F'_{i-2}, ..., F'_{i-K}]$, and fuse them with the current feature F_i to obtain an enhanced representation F'_i enriched with local contextual information. Then, the current feature F'_i is pushed into the memory bank, and the oldest entry F'_{i-K} is removed to maintain the buffer size. Finally, we concatenate all the features in memory bank to get the final output feature F_{out}. Formally, each divided feature segment F_i^d is first passed through Mamba to produce the output F_i:

$$\begin{aligned} F_i &= Mamba(F_i^d) = Reshape(Linear(F_i^t \otimes F_i^b)) \\ F_i^t &= VSSM(SiLU(f_{1\times1}^c(Linear(Reshape(F_i^d))) \\ F_i^b &= SiLU(Linear(Reshape(F_i^d))) \end{aligned} \tag{3}$$

where $f_{1\times1}^c$ represents 1×1 convolution.

Then we fuse the current feature F_i with the historical information $[F'_{i-1}, F'_{i-2}, ..., F'_{i-K}]$ in the memory bank to obtain the feature F'_i with enhanced local information as follows:

$$\begin{aligned} F'_i &= FCAM(F_i, F'_{i-1}, F'_{i-2}, ..., F'_{i-K}) \\ F'_{i-1}, F'_{i-2}, ..., F'_{i-K} &= fetch(MemoryBank) \\ UpdatesMemoryBank &= POP(F'_{i-K})\&PUSH(F'_i) \end{aligned} \tag{4}$$

where $FCAM(\cdot)$ denotes the feature cross-attention mechanism, which enhances the current features by integrating information from the memory bank.

As illustrated in Figure 3, for simplicity, we take the example that the memory bank has only one storage size. The fusion process follows a similar approach to mainstream feature fusion methods,

leveraging an attention mechanism to retain the most informative content through mutual comparison of features. The key distinction from existing methods Cui et al. (2024) lies in our objective: enhancing the local information within the current feature F_i. By integrating historical information from the memory bank into the current feature, the resulting representation not only preserves the global context captured by Mamba but also incorporates enriched local details. Specifically, given the current feature F_i and historical information F'_{i-1}, we first obtain the Q, K, V matrices used to perform the attention computation as follows:

$$\begin{aligned} Q_i &= K_{i-1} = f^c_{1\times1}(Norm(F_i)) \\ Q_{i-1} &= K_i = Reshape(f^c_{1\times1}(Norm(F'_{i-1}))) \\ V_i &= f^c_{1\times1}(F_i) \\ V_{i-1} &= f^c_{1\times1}(F'_{i-1}) \end{aligned} \tag{5}$$

Next up, we obtain the final enhanced feature F'_i by performing the attention calculation and feature fusion as follows:

$$\begin{aligned} F'_i &= \alpha Att_{i\to i-1} \oplus \beta Att_{i-1\to i} \oplus F_i \oplus F'_{i-1} \\ Att_{i\to i-1} &= SoftMax(Reshape(Q_i \otimes K_i)) \otimes V_i \\ Att_{i-1\to i} &= SoftMax(Q_{i-1} \otimes K_{i-1}) \otimes V_{i-1} \end{aligned} \tag{6}$$

where α and β are learnable cross-attention modulation parameters initialized to zero. Finally, we concatenate all the features stored in the memory bank to obtain the final output feature of MemVSSM, denoted as F_{out}:

$$F_{out} = concatenate(F'_i, F'_{i-1}, ..., F'_{i-K-1}) \tag{7}$$

Algorithm 1 Calculation of Ising Loss

Require: Predicted image $\hat{I}$ with shape $C \times H \times W$
1: Initialize loss $\mathcal{L}_{\text{Ising}} = 0$
2: **for** each pixel (x, y) in $\hat{I}$ **do**
3: **for** each neighboring pixel (x', y') of (x, y) **do**
4: **for** channel $c = 1$ to C **do**
5: $\mathcal{L}_{\text{Ising}} \mathrel{+}= \left|\hat{I}_c(x, y) - \hat{I}_c(x', y')\right|$
6: **end for**
7: **end for**
8: **end for**
9: Normalize: $\mathcal{L}_{\text{Ising}} \mathrel{/}= C \times H \times W$
10: **return** $\mathcal{L}_{\text{Ising}}$

3.3 Ising-inspired Regularization Loss

Since the input is processed in chunks for Mamba, the final restoration results often suffer from a lack of spatial coherence when relying solely on standard reconstruction loss, resulting in oversmoothed textures or fragmented structures. Inspired by the Ising model from statistical physics,which encourages local consistency among adjacent elements. We propose a novel Ising Loss to enhance spatial continuity and alleviate artifacts introduced by chunk-wise processing in Mamba-based architectures. This loss promotes similarity among neighboring pixel representations, thereby improving spatial coherence and enabling structure-aware image deblurring.

Specifically, given the predicted image $\hat{I}$, we define the Ising loss as the sum of absolute differences between a pixel and its neighboring pixels:

$$\mathcal{L}_{\text{Ising}} = \frac{1}{Z} \sum_{c=1}^{C} \sum_{x=1}^{H} \sum_{y=1}^{W} \sum_{(x',y')\in\mathcal{N}(x,y)} \left|\hat{I}_c(x, y) - \hat{I}_c(x', y')\right|, \tag{8}$$

where $\mathcal{N}(x, y)$ denotes the local neighborhood of pixel (x, y) (e.g., four- or eight-connected neighbors), and $Z = C \times H \times W$ is a normalization factor to make the loss invariant to image size.

Table 1: Quantitative evaluations of the MBMamba against state-of-the-art deblurring methods.

Methods	GoPro		HIDE	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
MPRNet Zamir et al. (2021)	32.66	0.959	30.96	0.939
FSNet Cui et al. (2024)	33.29	0.963	31.05	0.941
MambaIR Guo et al. (2024)	33.21	0.962	31.01	0.939
LoFormer Mao et al. (2024)	34.09	<u>0.969</u>	31.86	0.949
XYScanNet Liu et al. (2025)	33.91	0.968	31.74	0.947
MDT Chen et al. (2025)	34.26	<u>0.969</u>	31.84	0.948
ACL Gu et al. (2025)	33.25	0.964	-	-
Omni-Deblurring Li et al. (2025)	33.29	0.963	31.65	0.947
MambaIRv2 Guo et al. (2025)	33.62	0.967	31.63	0.948
MBMamba-S(Ours)	33.89	0.968	31.72	0.947
MBMamba-B(Ours)	<u>34.33</u>	<u>0.969</u>	<u>31.89</u>	0.949
MBMamba-L(Ours)	**34.68**	**0.972**	**32.22**	**0.953**

The computation of $\mathcal{L}_{\text{Ising}}$ is outlined in Algorithm 1. At each pixel location, we compute the sum of absolute differences with its neighbors and accumulate the result over all channels and spatial locations. Finally, the aggregated loss is normalized. To form the total training objective, we combine the Ising loss with standard reconstruction loss:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_c(\hat{I}, \overline{I}) + \delta\mathcal{L}_e(\hat{I}, \overline{I}) + \lambda\mathcal{L}_f(\hat{I}, \overline{I})) + \kappa\mathcal{L}_{\text{Ising}} \\ &= \sqrt{||\hat{I} - \overline{I}||^2 + \epsilon^2} + \delta\sqrt{||\triangle\hat{I} - \triangle\overline{I}||^2 + \epsilon^2} + +\lambda||\mathcal{F}(\hat{I}) - \mathcal{F}(\overline{I})||_1 + \kappa\mathcal{L}_{\text{Ising}}\end{aligned} \tag{9}$$

where $\overline{I}$ denotes the target images and $\mathcal{L}_c$ is the Charbonnier loss with constant $\epsilon = 0.001$. $\mathcal{L}_e$ is the edge loss, where $\triangle$ represents the Laplacian operator. $\mathcal{L}_f$ denotes the frequency domains loss, and $\mathcal{F}$ represents fast Fourier transform. To control the relative importance of loss terms, we set the parameters $\lambda = 0.1$, $\delta = 0.05$ as in Zamir et al. (2021); Cui et al. (2024) and $\kappa = 0.001$.

4 Experiments

We first present the experimental setup, followed by qualitative and quantitative comparisons, and ablation studies to validate our approach. **Additional experiments are provided in Appendix A.**

4.1 Experimental Settings

We adopt the Adam optimizer Kingma & Ba (2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set to 5×10^{-4} and decayed to 1×10^{-7} using a cosine annealing schedule Loshchilov & Hutter (2016). Training is conducted on 256×256 image patches with a batch size of 32 over 4×10^5 iterations. Data augmentation is performed through horizontal and vertical flipping. Furthermore, we construct three variants of MBMamba by adjusting the number of sub-decoders n (as shown in Figure 2(a)): MBMamba-S with 1 sub-decoder, MBMamba-B with 2 sub-decoders, and MBMamba-L with 4 sub-decoders.

4.2 Experimental Results

4.2.1 Evaluations on the synthetic dataset

Tables 1 present the performance of various image deblurring methods on the synthetic GoPro Nah et al. (2016) and HIDE Shen et al. (2019) datasets. Compared to the previous state-of-the-art MDT Chen et al. (2025), our MBMamba-L achieves a 0.42 dB gain on the GoPro dataset. Against other Mamba-based approaches, MBMamba-S surpasses MambaIRv2 Guo et al. (2025) by 0.27 dB, while MBMamba-L delivers a notable improvement of 1.06 dB. Furthermore, relative to the best-performing Mamba variant LoFormer Mao et al. (2024), MBMamba-L yields a significant enhancement of 0.57 dB. As shown in Figure 1, MBMamba's performance scales favorably with increased

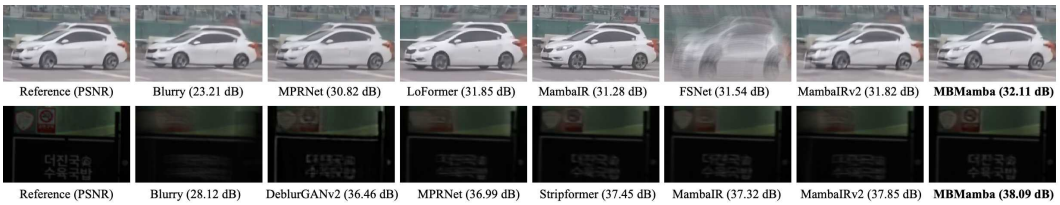

Figure 4: Image deblurring comparisons on the synthetic dataset Nah et al. (2016)(Top) and real-world dataset Rim et al. (2020)(Bottom).

Table 2: Quantitative real-world deblurring results.

Methods	RealBlur-R		RealBlur-J	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DeblurGAN-v2 Kupyn et al. (2019)	36.44	0.935	29.69	0.870
MPRNet Zamir et al. (2021)	39.31	0.972	31.76	0.922
Stripformer Tsai et al. (2022)	39.84	0.975	32.48	0.929
MRLPFNet Dong et al. (2023)	40.92	0.975	33.19	0.936
MambaIR Guo et al. (2024)	39.92	0.972	32.44	0.928
ALGNet Gao et al. (2024)	<u>41.16</u>	**0.981**	32.94	0.946
LoFormer Mao et al. (2024)	40.23	0.974	32.90	0.933
MambaIRv2 Guo et al. (2025)	40.36	0.979	32.91	0.940
MBMamba-S(Ours)	41.08	0.977	32.88	0.946
MBMamba-B(Ours)	41.12	<u>0.980</u>	<u>33.23</u>	<u>0.951</u>
MBMamba-L(Ours)	**41.21**	**0.981**	**33.41**	**0.953**

model size, highlighting its strong scalability. Although trained exclusively on the GoPro dataset, our model achieves a 0.36 dB PSNR improvement over LoFormer on the HIDE dataset, demonstrating strong generalization capability. Visual comparisons in Figure 4 further confirm that our method produces more visually appealing results.

4.2.2 Evaluations on the real-world dataset

We further evaluate MBMamba on real-world images from the RealBlur dataset Rim et al. (2020). As shown in Table 2, our method achieves higher PSNR and SSIM scores than previous approaches. In particular, while the improvement over the prior best method, ALGNet Gao et al. (2024), is modest on the RealBlur-R dataset, it is more pronounced on the RealBlur-J dataset, with a PSNR gain of 0.47 dB. Visual comparisons in Figure 4 further demonstrate that our model produces images with sharper details and better resemblance to the ground truth than competing methods.

4.3 Ablation Studies

We perform ablation studies to assess the effectiveness and scalability of our proposed method on the GoPro dataset Nah et al. (2016), with the results summarized in Table 3. Using NAFNet Chen et al. (2022) as the baseline, we gradually introduce our proposed modules to analyze their individual and combined contributions. As shown in Table 3, replacing the baseline with MemVSSM significantly enhances the model's ability to capture global information, leading to a notable improvement of approximately 0.81 dB. Additionally, introducing the Ising loss alone on the CNN-based baseline yields minimal performance gain, as the baseline already effectively models local information and benefits less from the smoothness regularization provided by the Ising loss. However, when

Table 3: Ablation study on individual components of the proposed MBMamba.

Method	PSNR
Baseline	32.83
Baseline replace with MemVSSM	33.64
Baseline + Ising loss	32.85
Baseline replace with MemVSSM + Ising loss	33.89

Table 4: Impact of MemVSSM design choices on model performance.

Net	PSNR	△ PSNR
VSSM Guo et al. (2024)	33.37	-
ASSM Guo et al. (2025)	33.65	+0.28
MemVSSM(Ours)	33.89	+0.52

Table 5: Effect of the pre-trained models.

Net	Pre-trained	Trainable	PSNR	△ PSNR
(a)		✔	33.74	-
(b)	✔	✔	33.86	+0.12
(c)	✔		33.89	+0.15

combined with MemVSSM, the use of Ising loss noticeably boosts the model's sensitivity to local structures, resulting in a further performance increase of about 0.25 dB. These results demonstrate the strong complementarity between our proposed modules.

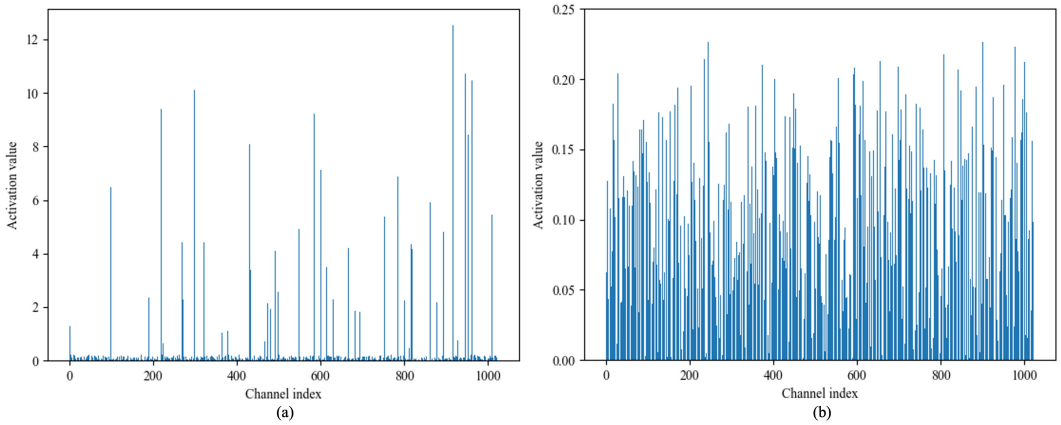

Figure 5: Following Guo et al. (2024), we use ReLU activation followed by global average pooling to compute channel activation values. (a) For the standard SSM, a significant portion of channels remain inactive, revealing channel redundancy. (b)The outputs from our MemVSSM module.

To further assess the effectiveness of our MemVSSM module, we replace it with existing VSSM methods Guo et al. (2024; 2025). As shown in Table 4, our MemVSSM achieves the best performance among all variants. Additionally, we apply ReLU activation followed by global average pooling on the MemVSSM outputs to compute channel activation values (see Figure 5). The results clearly indicate that MemVSSM effectively mitigates the problem of channel redundancy caused by an excessive number of hidden states in the state space model.

Since our MBMamba employs an existing pre-trained model as the encoder, we assess the influence of the pre-trained weights on overall performance. As shown in Table 5, leveraging pre-trained models clearly enhances performance. Furthermore, we investigate the effect of freezing encoder parameters and find that it has minimal impact on the results. Therefore, to conserve computational resources, we choose to freeze the encoder and train only the decoder.

5 Conclusion

In this paper, we present a structure-aware image deblurring network that effectively integrates both local and global features, without incurring the overhead of multiple scans or added modules. Specifically, we design a memory buffer mechanism to store and reuse historical information, facilitating more reliable modeling of the relevance between adjacent features. In addition, we introduce an Ising-inspired regularization loss inspired by physical systems, which simulates the "mutual attraction" between neighboring pixels through energy minimization. This loss encourages structural consistency and promotes smoother, more coherent image restoration. Extensive experimental results demonstrate that our proposed method achieves superior performance compared to state-of-the-art approaches.

REFERENCES

Duosheng Chen, Shihao Zhou, Jinshan Pan, Jinglei Shi, Lishen Qu, and Jufeng Yang. A polarization-aided transformer for image deblurring via motion vector decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 28061–28070, June 2025.

Liang Chen, Faming Fang, Shen Lei, Fang Li, and Guixu Zhang. Enhanced sparse model for blind deblurring. In *Proceedings of the European Conference on Computer Vision*, pp. 631–646, 2020. ISBN 978-3-030-58594-5.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *ECCV*, 2022.

Zheng Chen, Yulun Zhang, Ding Liu, bin xia, Jinjin Gu, Linghe Kong, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 29114–29125, 2023.

Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. Irnext: Rethinking convolutional network design for image restoration. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *The Eleventh International Conference on Learning Representations*, 2023b.

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1093–1108, 2024. doi: 10.1109/TPAMI.2023.3330416.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

J. Dong, J. Pan, Z. Yang, and J. Tang. Multi-scale residual low-pass filter network for image deblurring. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12311–12320, 2023.

Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

Xin Feng, Haobo Ji, Wenjie Pei, Jinxing Li, Guangming Lu, and David Zhang. U2-former: Nested u-shaped transformer for image restoration via multi-view contrastive learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023. doi: 10.1109/TCSVT.2023.3286405.

Hu Gao, Bowen Ma, Ying Zhang, Jingfan Yang, Jing Yang, and Depeng Dang. Learning enriched features via selective state spaces model for efficient image deblurring. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

Yubin Gu, Yuan Meng, Jiayi Ji, and Xiaoshuai Sun. Acl: Activating capability of linear attention for image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17913–17923, June 2025.

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. *arXiv preprint arXiv:2402.15648*, 2024.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28124–28133, 2025.

Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3):1126–1137, 2017.

Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 433–451, 2025.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5886–5895, 2023.

Pin-Hung Kuo, Jinshan Pan, Shao-Yi Chien, and Ming-Hsuan Yang. Efficient non-blind image deblurring with discriminative shrinkage deep networks. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

Orest Kupyn, T. Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8877–8886, 2019.

Yaowei Li, Hang An, Tong Zhang, Xiaoxuan Chen, Bo Jiang, and Jinshan Pan. Omni-deblurring: Capturing omni-range context for image deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

Hanzhou Liu, Chengkai Liu, Jiacong Xu, Peng Jiang, and Mi Lu. Xyscannet: An interpretable state space model for perceptual image deblurring. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 779–789, 2025.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2016.

Xintian Mao, Jiansheng Wang, Xingran Xie, Qingli Li, and Yan Wang. Loformer: Local frequency transformer for image deblurring. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*, 2022.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 257–265, 2016.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Siddharth Roheda, Amit Unde, and Loay Rashid. Mr-vnet: Media restoration using volterra networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6098–6107, June 2024.

Jianxiang Rong, Hua Huang, and Jia Li. Imu-assisted accurate blur kernel re-estimation in non-uniform camera shake deblurring. *IEEE Transactions on Image Processing*, 33:3823–3838, 2024.

Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5571–5580, 2019.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *ICLR*, 2023.

Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv*, 2017.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17683–17693, June 2022.

Fei Wen, Rendong Ying, Yipeng Liu, Peilin Liu, and Trieu-Kien Truong. A simple local minimal intensity prior and an improved algorithm for blind image deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):2923–2937, 2021.

Senyan Xu, Zhijing Sun, Mingchen Zhong, Chengzhi Cao, Yidi Liu, Xueyang Fu, and Yan Chen. Motion-adaptive transformer for event-based image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8942–8950, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022a.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022b.

Kun Zhou, Xinyu Lin, and Jiangbo Lu. Tsp-mamba: The travelling salesman problem meets mamba for image super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28134–28143, June 2025.

A Appendix

A.1 Overview

Mathematical Interpretation of MemVSSM A.2

Dataset A.3

More Ablation Studies A.4

Resource Efficient A.5

Additional Visual Results A.6

A.2 Mathematical Interpretation of MemVSSM

Consider one MemVSSM processing channel-chunk N. Let $x^{(n)} \in \mathbb{R}^{B\times d\times H\times W}$ be the input chunk and $T = H \cdot W$ the number of spatial tokens per chunk. Let $M(\cdot)$ denote the Mamba mapping applied to the chunk tokens, and define the Mamba output

$$o^{(n)} = M\big(x^{(n)}\big) \in \mathbb{R}^{B\times d\times H\times W}.$$

The implementation stores a fused memory $m^{(k)}$ after processing chunk k as

$$m^{(k)} := \mathrm{detach}\left(y^{(n)}\right),$$

where $y^{(n)}$ is the fused output produced for chunk n (see below); 'detach' indicates stopping gradients.

We also let C be the total channels and N the number of chunks so that the per-chunk channel dimension is

$$d = \frac{C}{N}.$$

We denote tokenized representations by indices $i, j \in \{1, \ldots, T\}$. FCAM computes scaled dot-product attention between current chunk queries (from $o^{(n)}$) and memory keys (from $m^{(k-1)}$):

$$a_{ij} = \frac{Q_\ell[i] \cdot Q_m[j]}{\sqrt{d}}, \tag{10}$$

$$S_{ij} = \mathrm{softmax}_j(a_{ij}), \qquad \sum_j S_{ij} = 1, \tag{11}$$

and two attention-propagated contributions (written per-token and then reshaped):

$$G_\ell[i] = \beta \sum_{j=1}^{T} S_{ij}\, V_m[j], \tag{12}$$

$$G_m[j] = \gamma \sum_{i=1}^{T} S'_{ji}\, V_\ell[i], \tag{13}$$

where $S' = \mathrm{softmax}(a^\top)$, and $\beta, \gamma \in \mathbb{R}^d$ are the per-channel learnable scales (initialized at zero in code). The fused forward output for chunk n is therefore

$$y^{(n)} \;=\; o^{(n)} \;+\; G_\ell \;+\; m^{(k-1)} \;+\; G_m, \tag{14}$$

with the implementation detail that $m^{(k-1)} = \mathrm{detach}(y^{(n-1)})$.

The baseline (no fusion) per-chunk update is $y^{(n)}_{\text{base}} = x^{(n)} + o^{(n)} +$ (FFN residual). With fusion, the Mamba output $o^{(n)}$ is additively augmented by attention-weighted transforms of the previous chunk's memory and by a direct additive memory term (Eq. 14). Thus fusion changes the forward dynamics seen by subsequent model components: information from chunk $n-1$ influences the activations of chunk n both through attention (G_ℓ) and direct addition ($m^{(k-1)}$). Concretely, fusion yields richer cross-channel / cross-spatial coupling in the activations received by downstream layers.

Because the implementation writes memory via $m^{(k-1)} = \mathrm{detach}(y^{(n-1)})$, the gradient path through memory is blocked. Using chain-rule notation,

$$\frac{\partial L}{\partial y^{(n-1)}} \;\not\ni\; \frac{\partial y^{(n)}}{\partial m^{(k-1)}} \frac{\partial L}{\partial y^{(n)}},$$

since $\partial m^{(k-1)} / \partial y^{(n-1)} = 0$. Therefore, although later chunks receive activations influenced by earlier chunks, the gradients do not accumulate recursively through the memory. This prevents the long-horizon gradient explosion or saturation typically observed in fully differentiable recurrent state updates. Gradients from the loss at chunk n still update the FCAM parameters (including the projection layers and β, γ), but they do not backpropagate into the Mamba parameters that generated $m^{(k-1)}$. In this design, the fusion modifies the forward dynamics by strengthening cross-chunk coupling in the activations, yet it keeps the backward dynamics local to each chunk. This locality is the key reason the implementation avoids gradient explosion or saturation when processing many chunks.

Because S_{ij} is a row-normalized softmax,

$$\|G_\ell[i]\| \leqslant \|\beta\|_\infty \sum_j S_{ij} \|V_m[j]\| \leqslant \|\beta\|_\infty \max_j \|V_m[j]\|. \tag{15}$$

Since β and γ are initialized to zero and learned gradually, the early-stage contribution of fusion is small, providing numerical stability. Combined with the softmax normalization, Eqs. 12–15 show fusion cannot produce unbounded outputs unless the learned scales diverge.

The memory fusion in MemVSSM effectively enhances Mamba with local historical context while preserving stable gradient flow. The behavior of the mechanism depends smoothly on K and N offering a controllable trade-off between local detail reinforcement and computational cost.

A.3 Datasets

In line with recent approaches Cui et al. (2024); Zamir et al. (2021), we train MBMamba on the GoPro dataset Nah et al. (2016), which consists of 2,103 training image pairs and 1,111 evaluation

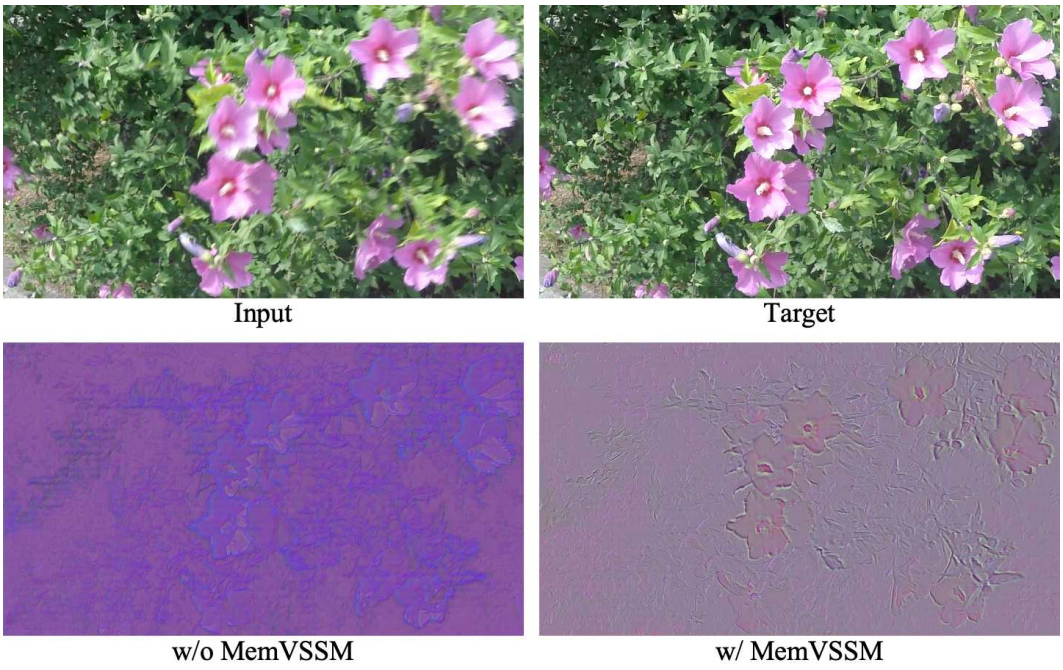

Figure 6: Effect of MemVSSM.

Table 6: The impact of hyperparameters.

N	K	κ	PSNR	$\triangle$ PSNR
4	4	0.001	33.89	-
4	4	0.005	33.75	-0.14
2	2	0.001	33.53	-0.36
6	4	0.001	33.77	-0.12
6	4	0.005	33.86	-0.03
4	2	0.001	33.87	-0.02

pairs. To evaluate the model's generalization capability, we directly test the GoPro-trained model on the HIDE Shen et al. (2019) and RealBlur Rim et al. (2020) datasets. The HIDE dataset, designed for human-centric motion deblurring, comprises 2,025 images. While both GoPro and HIDE are synthetically generated, RealBlur contains real-world image pairs, divided into two subsets: RealBlur-J and RealBlur-R.

A.4 More Ablation Studies

We present the visualization of feature maps in Figure 6 to emphasize the advantages of our proposed MemVSSM. As shown, the results clearly indicate that the features extracted with MemVSSM capture finer details and richer structural information compared to the baseline.

In addition to evaluating each hyperparameter individually, we analyze the interactions between the number of chunks N, memory bank depth K, and Ising loss weight κ. As shown in Table 6, chunk number N has the most significant influence, but the combination of N, K, and κ jointly determines optimal performance. For larger chunk numbers ($N = 6$), increasing κ slightly improves PSNR (33.81 $\rightarrow$ 33.86). This suggests that the optimal Ising loss weight depends on N. Larger N benefits from a slightly higher κ, while smaller N prefers a lower κ to avoid over-regularization. Since we only need the previous element in the memory bank to be fused with the current element, its depth does not have a big impact on the model performance. To facilitate the subsequent MemVSSM output, we set its depth to the same size as the chunk size.

Table 7: The impact of Ising loss.

Loss	PSNR	△ PSNR
Ising loss	33.89	-
Laplacian loss	33.74	-0.15

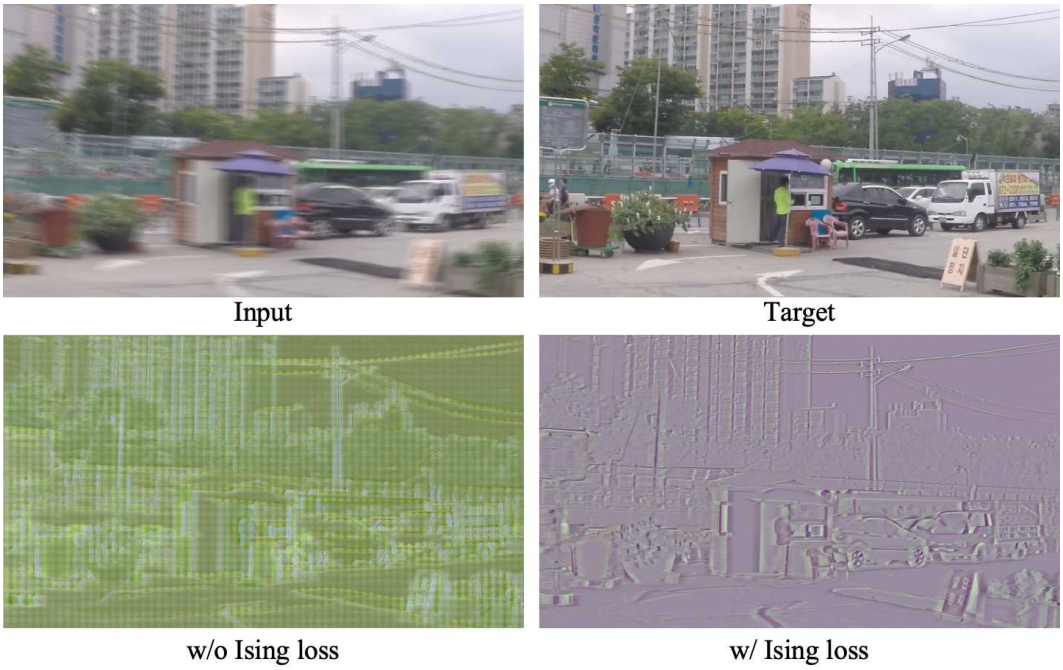

Figure 7: Effect of Ising loss.

To further demonstrate the effectiveness of Ising loss, we visualize its feature maps in Figure 7. In our MemVSSM, inputs are processed in chunks, and relying solely on standard reconstruction loss often leads to oversmoothed textures and fragmented structures due to poor spatial coherence. Ising loss enhances spatial continuity, alleviates artifacts caused by chunk-wise processing, and promotes similarity among neighboring pixel representations, thereby improving spatial coherence and enabling structure-aware image deblurring.

We also replace the Ising loss with Laplacian regularization. While Laplacian loss penalize local intensity differences uniformly to encourage smoothness, it tend to oversmooth fine structures and edges, which can degrade perceptual quality in image restoration tasks. In contrast, the Ising loss leverages a pairwise spin interaction model to selectively encourage piecewise-constant regions while preserving sharp boundaries. This leads to two distinct advantages: (i) enhanced preservation of high-frequency structures such as edges and textures, and (ii) stronger local consistency without introducing excessive blurring. Empirical results in Table 7 confirm that Ising regularization achieves superior reconstruction quality.

A.5 Resource Efficient

To further demonstrate the resource efficiency of our approach, we evaluate the model complexity and compare it with state-of-the-art methods in terms of inference time and FLOPs. Although our MemVSSM processes input features sequentially—thus not fully leveraging the parallelism inherent in the original SSM—it nonetheless achieves impressive efficiency. As shown in Table 8 and Figure 1, our MBMamba model not only delivers state-of-the-art performance but also substantially reduces computational overhead. Specifically, MBMamba-L surpasses the previous best-performing method, MambaIRv2 Guo et al. (2025), by 1.06 dB, while cutting computational cost by up to 84.2%

Table 8: The evaluation of model computational complexity on the GoPro dataset Nah et al. (2016).

Method	Time(s)	FLOPs(G)	PSNR	SSIM
MPRNet Zamir et al. (2021)	1.148	777	32.66	0.959
Restormer Zamir et al. (2022a)	1.218	140	32.92	0.961
IRNeXt Cui et al. (2023a)	<u>0.255</u>	114	33.16	0.962
FSNet Cui et al. (2024)	0.362	111	33.29	0.963
MambaIR Guo et al. (2024)	0.743	439	33.21	0.962
MambaIRv2 Guo et al. (2025)	0.743	664	33.62	0.967
MBMamba-S(Ours)	**0.249**	**54**	33.89	0.968
MBMamba-B(Ours)	0.283	<u>74</u>	<u>34.33</u>	<u>0.969</u>
MBMamba-L(Ours)	0.442	105	**34.68**	**0.972**

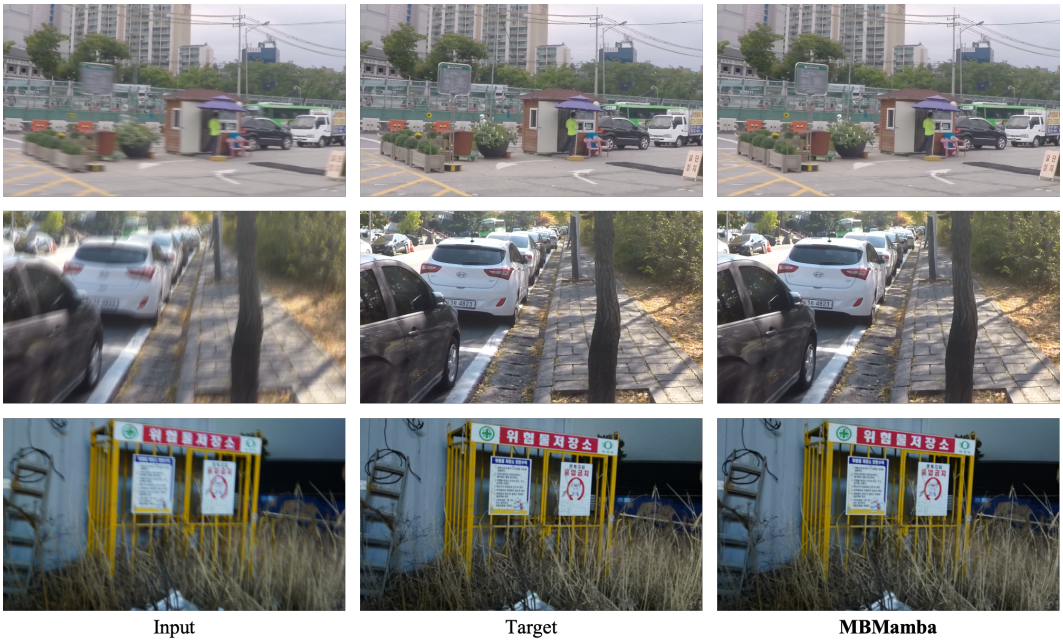

Figure 8: Visualization results on synthetic and real datasets.

and achieving nearly 1.5× faster inference. These results clearly demonstrate the effectiveness and efficiency of our method in balancing performance with resource usage.

A.6 Additional Visual Results

In this section, we present additional visual results to highlight the effectiveness of our proposed approach, as shown in Figures 8. It is clear that our model produces more visually appealing outputs for both synthetic and real-world motion deblurring.