# MambaIRv2: Attentive State Space Restoration

## Supplementary Material

## A. Efficiency Comparison on Large Inputs

Benefiting from the Mamba architecture, our proposed MambaIRv2 can achieve global pixel utilization. However, as an inevitable side effect, the global receptive field is usually accompanied by an increased computational cost since the model needs to process more tokens at once. Therefore, it is necessary to validate the efficiency on large resolution images. Here, we point out that, benefiting from our single-directional scan, our MambaIRv2 can in fact achieve a similar computational cost as the advanced Swin-Transformer [14] based method HAT [4]. In Tab. A.1, we give the MACs of our MambaIRv2 and HAT under varying input resolutions. As one can see, our MambaIRv2-B, which has a roughly similar number of parameters as HAT [4], is more efficient than HAT from resolution $64 \times 64$ to $1024 \times 1024$. For example, on the $256 \times 256$ resolution, which is a common inference patch size, our method achieves a 30% savings in computational cost metric MACs. At the high-resolution setting of $1024 \times 1024$, our method achieves fewer MACs than HAT. It is worth noting that in addition to this impressive efficiency, our MambaIRv2 still outperforms HAT by a noticeable margin, which has been extensively verified in the main paper.

Table A.1. The computational cost MACs with images of different resolutions. We compare our MambaIRv2-B and HAT [4]. We adopt the $4\times$ classical SR task and set the output size from $64\times$ to $1024 \times 1024$.

| models | $64 \times 64$ | $128 \times 128$ | $256 \times 256$ | $512 \times 512$ | $1024 \times 1024$ |
|---|---|---|---|---|---|
| HAT [4] | 26.05G | 58.62G | 162.85G | 527.63G | 1882.28G |
| MambaIRv2 | 7.12G | 28.49G | 113.97G | 455.89G | 1823.04G |

Table A.2. Ablation experiments on the hyper-parameters of the number of prompts $T$ in the prompt pool, and the inner rank $r$ in the semantic decoupling.

| $r \times T$ | Set14 | | Urban100 | | Manga109 | |
|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| $16 \times 64$ | 33.90 | 0.9205 | 32.96 | 0.9359 | 39.19 | 0.9781 |
| $16 \times 128$ | 33.91 | 0.9205 | 32.96 | 0.9354 | 39.22 | 0.9783 |
| $32 \times 128$ | **33.97** | 0.9210 | **32.97** | **0.9360** | 39.20 | 0.9783 |
| $32 \times 64$ | 33.95 | **0.9213** | **32.97** | 0.9355 | **39.24** | **0.9784** |

## B. More Ablation Results

### B.1. Ablation on Prompt Learning

In the proposed ASE, we use learnable prompts to absorb information of similar pixels across the whole image, which will be later inserted into the state space modeling to help the query pixel to see the unscanned tokens. The proposed prompt learning contains two key hyperparameters, namely the size of the prompt pool $T$, and the internal rank $r$ in the semantic decoupling. In this section, we perform hyperparameter ablation to investigate the impact of different $T$ and $r$ on the performance. As shown in Tab. A.2, when the $r$ is small, increasing the number of prompts $T$ can steadily improve performance. For example, when $r$ = 16, increasing $T$ from 64 to 128 can result in a 0.03dB improvement on Manga109. However, when $r$ is large, increasing the size of the prompt pool sometimes instead results in a slight performance drop. A similar observation also appears in the inner-rank $r$. In practice, we choose a moderate $T \times r = 32 \times 64$ considering the performance and efficiency trade-off.

### B.2. Visualization of Semantic Neighboring

In the proposed Semantic Guided Neighboring (SGN), we restructure the image so that semantically similar pixels are also spatially close in the unfolded 1D sequence. In this section, we visualize the learned restructured image in Fig. A.1 for better understanding. It can be seen that the previous distant pixels with similar semantics in the original feature map become spatially close after the restructuring of the SGN. By placing semantically similar pixels closer, the proposed SGN alleviates Mamba's long-range decay problem resulting from the causal modeling nature and thus facilitates better exploit those distant but similar pixels.

## C. Comparison on Receptive Field

As pointed out in previous work [10], a significant advantage of the Mamba architecture is the practical global receptive field, which helps the model activate more pixels to improve restoration performance. Here, we give visualization comparison results of LAM [9] and ERF [15] with other strong baselines. First, the Fig. A.2 gives the results of the LAM attribution map. One can see that our MambaIRv2 can activate more pixels than other state-of-the-art methods HAT [4] by presenting a wider LAM attribution and a higher DI, thus resulting in higher-quality restoration results. Second, the Fig. A.3 further gives the effective receptive filed visual comparison with other methods. Our MambaIRv2 exhibits darker colors across the entire image, indicating the global perception of the proposed method.

It is noteworthy that the ERF visualization in Fig. A.3 can also demonstrate the effectiveness of the proposed non-causal modeling strategy in our MambaIRv2. In detail, the ERF visualization of MambaIR [10] exhibits a clear criss-

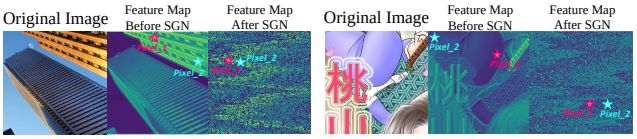


Figure A.1. Visualization of the effectiveness of the proposed SGN. Before SGN, the semantically similar pixels belonging to the same object in the original feature map are far apart. After SGN, these pixels are spatially close to each other, thus facilitating strong interactions in the unfolded 1D sequence.

crossing, which is a clear sign of the causal modeling property as the center pixel can only utilizes its previous pixels in the scanned 1D sequence. In contrast, our proposed MambaIRv2, which aims at eliminating causal modeling in Mamba for image restoration, does not exhibit such unfavorable crisscrossing, demonstrating the validity of our proposed non-causal modeling.

## D. More Implementation Details

We employ the DF2K [13, 20] dataset to train models on classic SR and use DIV2K only to train lightweight SR models. Moreover, we use Set5 [2], Set14 [21], B100 [17], Urban100 [11], and Manga109 [18] to evaluate the effectiveness of different SR methods. For Gaussian color image denoising and JPEG CAR, we utilize DIV2K [20], Flickr2K [13], BSD500 [1], and WED [16] as our training datasets. Our testing datasets for guassian color image denoising includes BSD68 [17], Kodak24 [7], McMaster [22], and Urban100 [11]. And we use Classic5 [6] and LIVE1 [19] datasets to evaluate the performance of the JPEG CAR task. The performance is evaluated using PSNR and SSIM on the Y channel from the YCbCr color space.

## E. Comparison to ATD

The Adaptive Token Dictionary (ATD) [23] which can generate input-specific tokens/prompts to help the query pixel see out of the window, appears close to our proposed MambaIRv2, with both including additional prompts for obtaining more information. Here, we summarize the main differences between them in the following aspects. First, the goals of introducing prompts in these two methods are clearly different. The ATD uses prompts to overcome the limited receptive field in the window attention, while our MambaIRv2 aims to mitigate the causal modeling of the Mamba. Second, the utilization of prompts for seeing beyond the scanned sequence in our MambaIRv2 is well-motivated. Specifically, we mathematically analyze the difference between state space and attention in the main paper, based on which we propose to add prompts to the $\mathbf{C}$ matrix in the state equation to attentively query relevant pixels across the image. In contrast, ATD adopts intuitive

Table A.3. Comparison with ATD [23] on 2× classic SR.

| methods | Set5 | Set14 | B100 | Urban100 | Manga109 |
|---|---|---|---|---|---|
| ATD [23] | 38.61 | **34.95** | **32.65** | **34.70** | 40.37 |
| MambaIRv2 | **38.65** | 34.89 | 32.62 | 34.49 | **40.42** |

cross-attention to incorporate prompts into the feature map. Third, the way in which the prompts are generated is different. Specifically, ATD uses the attention map to implicitly obtain the category of each pixel. However, attention maps are even not available in Mamba, and thus we propose to design separate routing modules to explicitly learn the category of each pixel. In Tab. A.3, we give the quantitative comparison of our MambaIRv2 against ATD, and it can be seen that our proposed method can achieve comparable performance to ATD. It should be noted that ATD [23] is a highly optimized Transformer-based method since Transformer has been introduced to image restoration for many years. Given the Mamba-based methods are still in their infancy since the introduction of MambaIR [10]. It is promising for the Mamba-based method to achieve further performance improvements over its transformer counterparts.

## F. Limitation and Future Works

Our MamabIRv2 can effectively alleviate the inherent causal nature of Mamba architecture [8] benefiting from the proposed attentive state space modeling. Nonetheless, our work can be further improved in the future in the following aspects. First, the Mamba architecture emerges as the third backbone option for image restoration, in addition to CNNs and ViTs, which provide more solutions for designing image restoration networks. Therefore, an in-depth interpretability analysis about what exactly Mamba or ViT has learned during the restoration of an image is important for further understanding and network design. Second, although this work follows existing works [12] to cover multiple image restoration tasks, some other tasks such as image deblurring, dehazing and deraining can also be explored in the future. The implementation of the U-shaped MambaIRv2 backbone for these tasks to achieve further performance improvement is also interesting and promising [3]. Finally, despite the promising results shown, we would like to point out that the Mamba-based image restoration network is still in its early stages. With the increasing research interest in Mamba, it will be promising to study the state-space models for low-level vision.

## G. Proof for Long-range Decay

As pointed out in the main paper, the causal property of Mamba leads to weak interactions between the query token and other remote tokens, *i.e.*, the long-range decay.

Here, given the condition of the causal modeling equation in Mamba, we attempt to derive the long-range decay as follows.

Formally, recall that the causal modeling of the state-space equation is given by:

$$\begin{aligned} h_i &= \overline{\mathbf{A}} h_{i-1} + \overline{\mathbf{B}} x_i, \\ y_i &= \mathbf{C} h_i + \mathbf{D} x_i. \end{aligned} \tag{A.1}$$

Then, we can continuously iterate Eq. (A.1) wit $i = 0, 1, \cdots k$. For example, setting $i = 0$ turns Eq. (A.1) into the following:

$$\begin{aligned} h_0 &= \overline{\mathbf{B}} x_0 \\ y_0 &= \mathbf{C} h_0 + \mathbf{D} x_0 = \mathbf{C}\overline{\mathbf{B}} x_0 + \mathbf{D} x_0 \end{aligned} \tag{A.2}$$

After that, we can further set $i = 1$ to obtain the following equation:

$$\begin{aligned} h_1 &= \overline{\mathbf{A}} h_0 + \overline{\mathbf{B}} x_1 = \overline{\mathbf{A}\mathbf{B}} x_0 + \overline{\mathbf{B}} x_1 \\ y_1 &= \mathbf{C} h_1 + \mathbf{D} x_1 = \mathbf{C}(\overline{\mathbf{A}\mathbf{B}} x_0 + \overline{\mathbf{B}} x_1) + \mathbf{D} x_1 \\ &= \mathbf{C}\overline{\mathbf{A}\mathbf{B}} x_0 + \mathbf{C}\overline{\mathbf{B}} x_1 + \mathbf{D} x_1 \end{aligned} \tag{A.3}$$

Set $i = 2$ gives the following:

$$\begin{aligned} h_2 &= \overline{\mathbf{A}} h_1 + \overline{\mathbf{B}} x_2 = \overline{\mathbf{A}}(\overline{\mathbf{A}\mathbf{B}} x_0 + \overline{\mathbf{B}} x_1) + \overline{\mathbf{B}} x_2 \\ &= \overline{\mathbf{A}}^2 \overline{\mathbf{B}} x_0 + \overline{\mathbf{A}\mathbf{B}} x_1 + \overline{\mathbf{B}} x_2 \\ y_2 &= \mathbf{C} h_2 + \mathbf{D} x_2 \\ &= \mathbf{C}(\overline{\mathbf{A}}^2 \overline{\mathbf{B}} x_0 + \overline{\mathbf{A}\mathbf{B}} x_1 + \overline{\mathbf{B}} x_2) + \mathbf{D} x_2 \\ &= \mathbf{C}\overline{\mathbf{A}}^2 \overline{\mathbf{B}} x_0 + \mathbf{C}\overline{\mathbf{A}\mathbf{B}} x_1 + \mathbf{C}\overline{\mathbf{B}} x_2 + \mathbf{D} x_2 \end{aligned} \tag{A.4}$$

By iterating continuously, we can generalize the output $y_k$ in the $k$-th time step being represented by $x_0 \cdots x_k$ as the following formula:

$$y_k = \mathbf{C}\overline{\mathbf{A}}^k \overline{\mathbf{B}} x_0 + \mathbf{C}\overline{\mathbf{A}}^{k-1} \overline{\mathbf{B}} x_1 + \cdots + \mathbf{C}\overline{\mathbf{B}} x_k + \mathbf{D} x_k \tag{A.5}$$

Eq. (A.5) actually quantitative the interaction between the $k$-th query token $x_k$ and all its previous $k$ tokens $x_0, x_1, \cdots, x_{k-1}$ in the causally scanned sequences to produce the $k$-th output $y_k$ of state-space model. It can be clearly seen in Eq. (A.1) that the contribution of $x_0$ to the generation of $y_k$ is weighted by $\mathbf{C}\overline{\mathbf{A}}^k \overline{\mathbf{B}}$, which is proportional to $\overline{\mathbf{A}}^k$. Since in the main paper we have empirically observed that the mean value of $\overline{\mathbf{A}}$ is statistically less than 1, as a result, when $k$ is large, *i.e.*, when the two pixels are distant, the contribution of $x_0$ to $x_k$ is small, *i.e.*, exhibiting long-range decay. If $x_0$ is very helpful to $x_k$, this decay can catastrophically impair the restoration of the $x_k$.

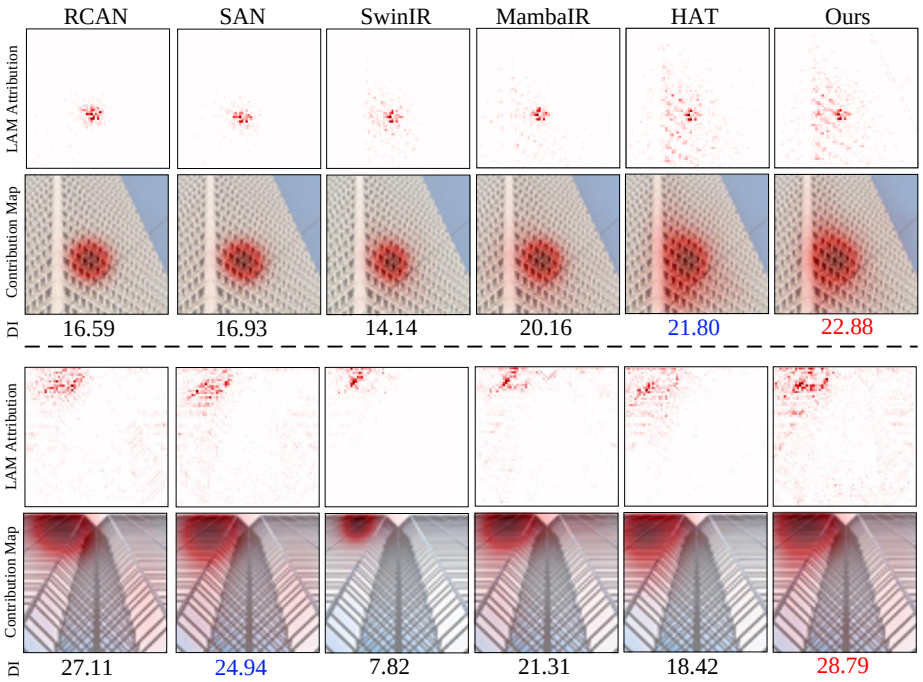


Figure A.2. The LAM visualization [9] comparison with different methods The diffusion index reflects the range of involved pixels. A higher DI represents a wider range of utilized pixels.

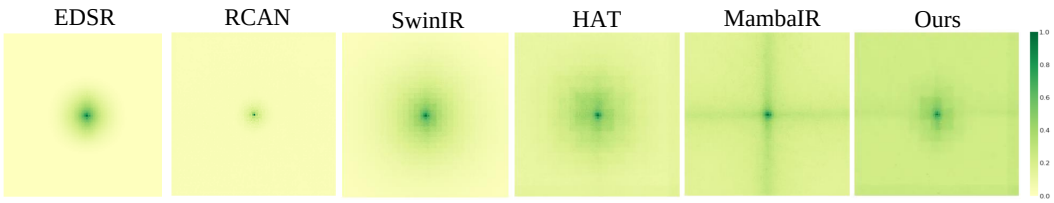


Figure A.3. The Effective Receptive Field (ERF) visualization [5, 15] for EDSR [13], RCAN [24], SwinIR [12], HAT [4], MambaIR [10], and the proposed MambaIRv2. A larger ERF is indicated by a more extensively distributed dark area. The proposed MambaIRv2 achieves a significant global effective receptive field.

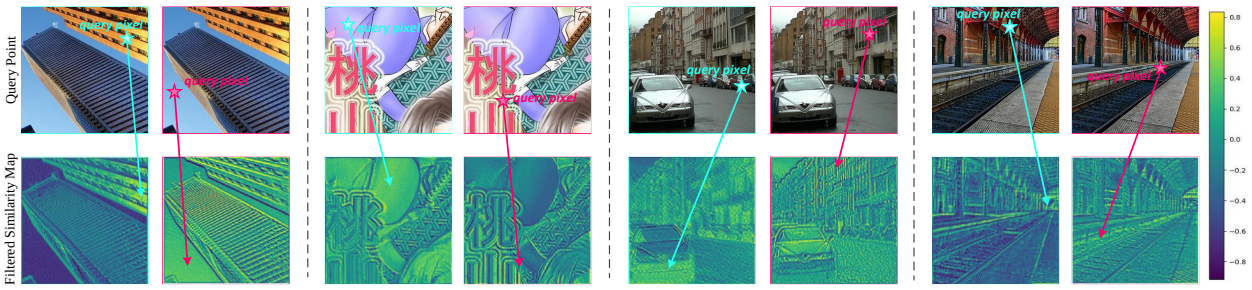


Figure A.4. More visualization results on the attentive state space modeling.

## References


[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2010. 2

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 2

[3] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. *arXiv preprint arXiv:2310.11881*, 2023. 2

[4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023. 1, 4

[5] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, pages 11963–11975, 2022. 4

[6] Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images. *IEEE TIP*, 16(5):1395–1411, 2007. 2

[7] Rich Franzen. Kodak lossless true color image suite. 2021. 2

[8] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[9] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *CVPR*, pages 9199–9208, 2021. 1, 4

[10] 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 *ECCV*, pages 222–241. Springer, 2025. 1, 2, 4

[11] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 2

[12] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 2, 4

[13] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 2, 4

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[15] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *NeurIPS*, 29, 2016. 1, 4

[16] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE TIP*, 26(2):1004–1016, 2016. 2

[17] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–423. IEEE, 2001. 2

[18] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using Manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 2

[19] H Sheikh. LIVE image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 2

[20] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, pages 114–125, 2017. 2

[21] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 2

[22] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2): 023016–023016, 2011. 2

[23] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *CVPR*, pages 2856–2865, 2024. 2

[24] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 4