RamIR: Reasoning and Action Prompting with Mamba for All-in-One Image Restoration

Aiqiang Tang, Yan Wu*, Yuwei Zhang

College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China.

*Corresponding author(s). E-mail(s): yanwu@tongji.edu.cn;
Contributing authors: aiqiangtang@tongji.edu.cn;
yuweizhang@tongji.edu.cn;

Abstract

All-in-one image restoration aims to recover various degraded images using a unified model. To adaptively reconstruct high-quality images, recent prevalent CNN and Transformer based models incorporate learnable prompts to dynamically acquire degradation-specific knowledge for different degraded images, achieving state-of-the-art restoration performance. However, existing methods exhibit limitations, including high computational burden and inadequate modeling of long-range dependencies. To address these issues, we propose a reasoning and action prompt-driven Mamba-based image restoration model, namely RamIR. Specifically, RamIR employs the Mamba block for long-range dependencies modeling with linear computational complexity relative to the feature map size. Inspired by Chain-of-Thought (CoT) prompting, we integrate Reasoning and Action (ReAct) prompts within the Mamba block. Hence, we utilize the capability of pretrained vision language (PVL) models to generate textual reasoning prompts describing the type and severity of degradations. Simultaneously, another output from PVL acts as action prompt representing the clean image caption. These prompts, employed in a CoT manner, enhance the network's sensitivity to degradation and elicit targeted recovery actions tailored to different reasoning prompts. Additionally, we explore the seamless interaction between Mamba blocks and prompts, introducing a novel prompt-driven module (PDM) to facilitate prompt utilization. Extensive experimental results demonstrate the superior performance of RamIR, highlighting its advantages in terms of input scaling efficiency over existing benchmark models for all-in-one image restoration.

Keywords: All-in-one Image Restoration, Image Restoration, Vision Mamba, Chain-of-thought

1 Introduction

Image quality is significantly diminished by degradations (e.g., rain, noise, haze, blur) that are often induced by unsuitable ambient conditions. Image restoration models learn to generate high-quality images pixel-by-pixel from their degraded counterparts. However, tackling each type of degradation with an independent model exposes challenges such as high training costs and poor generalization across various degradation types. Consequently, recent research has focused on various types and levels of degradation restoration, termed the all-in-one model, where the model implicitly classifies the type of degradation using shared knowledge. All-in-one models have broad practical applications.

Formally, for representing high-level patterns in all-in-one image restoration tasks, Convolutional Neural Networks (CNN) [1–3] and Vision Transformers (ViT) [4–6] have been extensively employed as two primary restoration backbones. CNN-based restoration networks have practical usage on resource-constrained devices due to their favorable efficiency in parallel convolution operations, despite extracting local features with limited receptive fields. On the other hand, ViT-based approaches incorporate the self-attention mechanism to capture long-range contextual information in images which realized a global receptive field. The main drawback of ViT is its quadratic growth in complexity as the image size increases, resulting in significant computational overhead.

Recent advancements in the natural language processing (NLP) domain, structured state-space sequence models [7] and Mamba [8] address the computational inefficiencies inherent in Transformers while retaining their global contextual modeling capabilities. Moreover, by leveraging a parallel scan algorithm [8], Mamba-based methods handle long sequence tokens in a parallel fashion, fully utilizing modern parallel computing devices, e.g., GPU. When comes to the computer vision domain, vision Mamba has emerged as favorable methods [9, 10] to obtain global receptive field at the cost of linear complexity. The aforementioned outstanding researches indicate that Mamba-based networks for image restoration can naturally activate more pixels to reconstruct high-quality images.

To address the generalization capacity of the all-in-one models, several works [2, 5, 11] utilize learned guidance vectors, so-called "prompt embeddings", help network learning degradation specific priors. Moreover, PVL models [11–14] are introduced into the restoration networks, where PVL models extract textual information serves as prompt fed into the all-in-one network and dynamically guiding reconstruction process, achieving state-of-the-art performance in unified image restoration tasks. PromptIR [5] is the most similar work as ours, however, utilizing prompts independently ignores the progressive logic within those prompts. With those observations, we question: 1) How does the prompt-driven vision Mamba-based network impact all-in-one image restoration performance? 2) What roles do those prompts play in the restoration process?

In this paper, we propose a ReAct prompt-driven Mamba-based network to perform all-in-one image restoration. We utilize Mamba blocks within our architecture to capture long-range dependencies, ensuring a comprehensive understanding of global context in images, while maintaining linear computational complexity. Experimental

results show vision Mamba facilitating restoration network modeling capability, outperforming current CNN-based and ViT-based restoration networks. Moreover, aiming at addressing network degradation awareness, we utilize the PVL model as a degraded image priors extractor, the obtained preliminary priors describe degradation type and clean image caption in high-level textual semantics. We then unleash various degradations removal under the guidance of degradation priors, motivated by ReAct [15], we treat the degradation type prior as reasoning prompt to learn degradation-specific knowledge. Corresponding to reasoning prompts, clean image captions serve as action prompts that improve reconstruction performance in the clean image reconstruction stage. In addition, we introduce the PDM to handle the reasoning and action prompts effectively. The main contributions of our work can be summarized as below:

- We introduce the ReAct prompt-driven Mamba-based model for all-in-one image restoration. Mamba blocks enrich the network by providing global context, interacting with ReAct prompts generated by the PVL model, thereby further enhancing the generalization capability of the all-in-one image restoration model.
- We design the PDM as the key component of RamIR, specifically to handle and facilitate the model's interaction with ReAct prompts.
- Experimental results demonstrate that our method achieves promising all-in-one restoration performance on various benchmark datasets while maintaining linear computational complexity.

2 Related works

Image restoration network. Image restoration is a fundamental and long-standing computer vision problem which includes various tasks such as image deraining [4, 16], dehazing [17, 18], and denoising [1, 19]. With the explosive development of deep learning, image restoration has significantly advanced in the past few years. In contrast to single degradation task [1, 4, 20–22], an all-in-one network handles multiple image restoration tasks[2, 3, 5, 6, 11]. Li et al. [3] introduced a unified network for multiple degradation image restoration tasks including denoising, draining, and dehazing. This network designed an image encoder to model comprehensive degradation representations by contrastively learning from degrade-clean paired images. It further utilizes these pre-captured representations as priors to assist the restoration network in dynamically reconstructing clean images. This method requires additional data pairs and a pretraining stage, composed to challenge of training cost and efficiency. CNN-based networks also limited the performance in both prior encoder stages and image recovery stages. To address above mentioned shortage, PromptIR [5] proposed the learnable vision prompt to implicitly cluster inputs according to the degradation type, incorporated with Transformer blockers achieved impressive results. Transformer-based approaches, such as those proposed in [6, 23, 24], significantly expand the receptive field, enabling the network to encompass a broader pixel region. This enhanced capability is crucial for image restoration tasks, as it allows the network to capture more contextual information and facilitate accurate reconstruction. By activating more pixels thus the anchor pixel can refer to more relevant pixels and represent image structure accurately. However, the shortcomings or limitations of

Transformer blocks primarily revolve around their computational demands and their effectiveness in handling spatial information when compared with CNN.

Vision Mamba. Recent advancements, including Vim [9] and VMamba [10], showcase that State Space Models (SSMs) introduce a groundbreaking approach to capturing long-range dependencies with linear complexity. This strategy circumvents the computational bottlenecks of Transformers while preserving their capacity for global context modeling. Motivated by the promising properties of vision Mamba, Mambair [22] exploring the power of recent advanced SSMs. They construct an image restoration network that incorporates vision Mamba blocks for capturing global contextual cues, subsequently employing a channel attention module to streamline channel information and mitigate redundancy. Notably, they introduce the local enhancement to handle the neighborhood pixel forgetting which results from the flattening strategy. This serves as an alternative image restoration network for degradation-specific tasks (e.g., denoising), resolving the dilemma of the trade-off between efficient computation and global effective receptive field. Researchers from Stanford [21] are also delving into the capabilities of the Mamba-based image restoration network. The introduced Channel-Aware U-Shaped Mamba incorporates a newly designed Mamba block that consists of channel-wise SSMs and spatial SSMs to attain global receptive fields while preserving channel-specific features. When combined with a U-Net, it enhances the understanding and reconstruction of images. Mamba-based methods hold great potential for single image restoration. However, the generalization of degradation awareness for an all-in-one Mamba-based image restoration model and the interaction with textual prompt capabilities are still under exploration.

Prompting learning. A noteworthy technique for refining vision-language models is prompt learning [25, 26], which has garnered significant attention. Prompt learning for image restoration involves integrating prompts that function as degradation priors directly into the image restoration network, thereby enhancing its ability to handle various degradation types. [5] initialized a learnable prompt using a lightweight module, highlighting its significance in a unified image restoration model. The PVL models are gaining increasing attention in image restoration tasks. [27] introduced reasoning prompt to adverse weather removal, utilizing high-level textual prompts of degradation priors which were obtained by formulating a question prompt as a query to [12]. Luo et al. [11, 28, 29] pretrained a degradation-aware CLIP model, which integrates an innovative controller module that adapts the fixed CLIP image encoder [14]. This controller adeptly adjusts the encoder to generate high-quality feature embeddings, tailored to the specific corruptions present in the input. By outputting degradation features that accurately reflect the input's actual distortions, the controller serves as a seamless classifier for distinguishing various degradation types. Clean image captions [13] describe the content of high-quality image as another prompt that guides the reconstruction process. InstructIR [2] builds an all-in-one restoration network upon [1], using human-written instructions to guide the reconstruction process, setting a new benchmark for researches on prompt-driven image restoration and enhancement. All these approaches treat prompts independently of each other. CoT [30–33] reasoning traces and task-specific actions in an interleaved manner [15], first introduced in NLP, demonstrating how prompt learning benefits from handling prompts through multiple

logical steps to enhance reasoning capabilities and take corresponding actions. ReAct prompt learning shows a potential direction of image restoration.

3 Method

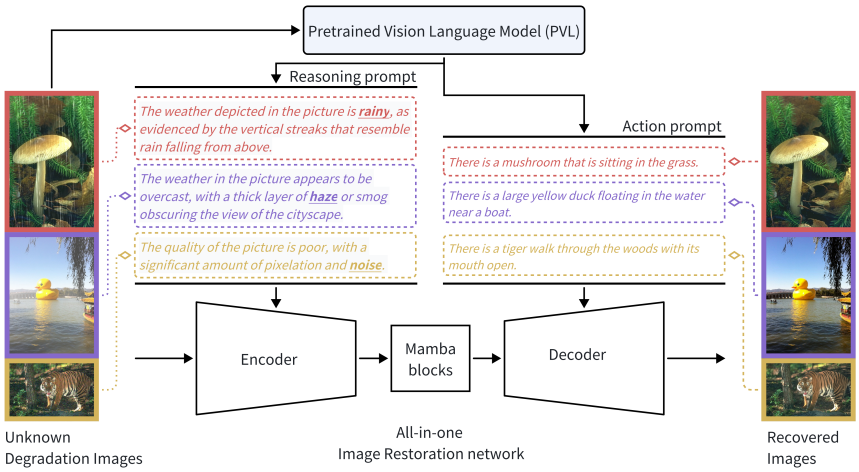

Fig. 1: Overview of our method RamIR. We propose a network based on Mamba, integrated with a PVL model, for performing image restoration across various types of degradation. The PVL model implicitly infers the reasoning prompt and action prompt from an input image with unknown degradation. These prompts are then injected into the Mamba-based image encoder and decoder to guide the removal of degradation and the reconstruction of a clean image, respectively.

Author: Enhance the Figure description.
For Review#3's 3rd comment.

We aim to develop an efficient network that focuses on modeling long-range dependencies with linear complexity for all-in-one image restoration. We propose RamIR, which integrates vision Mamba blocks into the U-Net architecture as the image encoder and decoder, guided by the reasoning prompt and dynamically taking recovery actions for various degradation tasks. As illustrated in Fig. 1, the Mamba-based image restoration network, which incorporates the PVL model, utilizes ReAct prompt to guide both the degradation removal stage and the clean image reconstruction stage. In this section, we first present the overall architecture of RamIR. Subsequently, we delve into its key components: the ReAct embedding generation module and the ReAct prompt-driven RAMamba block.

Author: Add the overall process description.
For Review#3's 3rd comment.

3.1 Overall architecture

The overall architecture of our method is illustrated in Fig. 2. Given unknown degradation image I_d, our method dynamically recovers the clean image I_c using the ReAct

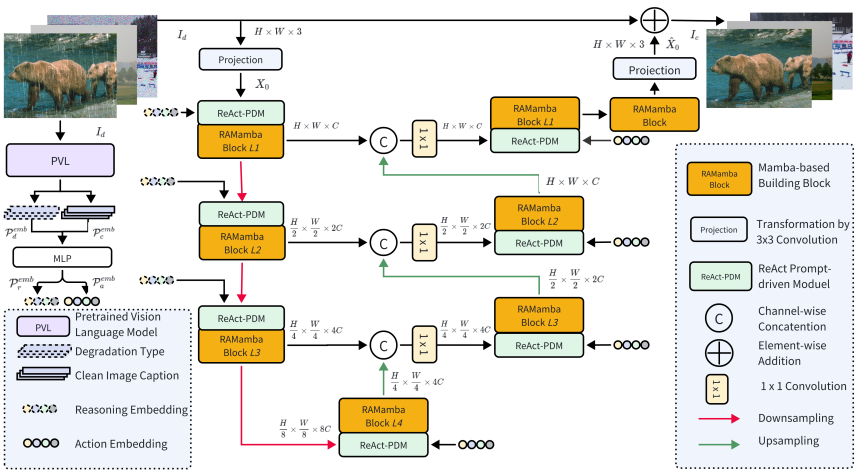

Fig. 2: The overall architecture of our approach. We use UNet-fashion network with Mamba blocks in the image encoding and decoding stages. Additionally, we preprocess the unknown degradation image using a PVL model to generate ReAct embedding. The core components of our framework, i.e., the PVL model, the ReAct prompt-driven module (PDM) and the RAMamba block. The PVL produces textual ReAct embedding, which are then injected into restoration network, the ReAct PDM dynamically adapts those embedding at various stages as ReAct prompt, while the RAMamba block interacts with the ReAct prompt to enrich the degradation-specific context, thereby enhancing the restoration process.

Author: Update caption of the figure. For Review#3's 2nd comment.

prompt-driven Mamba-based restoration network, where we generate degradation-aware ReAct embedding from PVL model [11]. The PVL is a per-trained CLIP model designed to produce high-quality image feature from a corrupted image while simultaneously predicting the degradation type. Specifically, in the preprocessing stage, the PVL model takes I_d as input, producing degradation priors $\mathcal{P}_d^{emb}$ and clean image content priors $\mathcal{P}_c^{emb}$. Subsequently, a dense network projects $\mathcal{P}_d^{emb}$ to $\mathcal{P}_r^{emb}$ as reasoning embedding and transforms $\mathcal{P}_c^{emd}$ to $\mathcal{P}_a^{emb}$ as action embedding. Meanwhile, in the restoration network, $I_d \in \mathbb{R}^{H\times W\times 3}$ initially passes through a 3×3 convolutional projection block to extract fundamental, low-level features $X_0 \in \mathbb{R}^{H\times W\times C}$; where $H \times W$ represents the spatial resolution and C denotes the channels. X_0 then undergoes a 4-level encoder-decoder, transforming into fine-grained, high-quality features $\hat{X}_0 \in \mathbb{R}^{H\times W\times 3}$. Specifically, the encoder comprises several reasoning prompt-driven RAMamba blocks and downsampling layers at each level l. The RAMamba block interacts with $\mathcal{P}_r^{emb}$ to drive diverse degradation feature encoding, the downsampling operator forms the feature map $X_l \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times 2^l C}$ by reducing spatial size and expanding the number of channels hierarchically, thereby yielding latent representations $X_4 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$ that will feed into the decoder. For feature reconstruction, we apply action prompt-driven RAMaba blocks and upsampling layers at each decoder level. We concatenate the corresponding encoder level features to the decoder level by

Author: Add decription of the PVL. For Review#3's 3rd comment.

Author: Grammar correct. For reviewer#4 3rd commet

utilizing a skip-connection. The decoded features are projected back to $\hat{X}_0$, yielding the final recovered high-quality image: $I_c = I_d + \hat{X}_0$.

3.2 ReAct embedding generation module

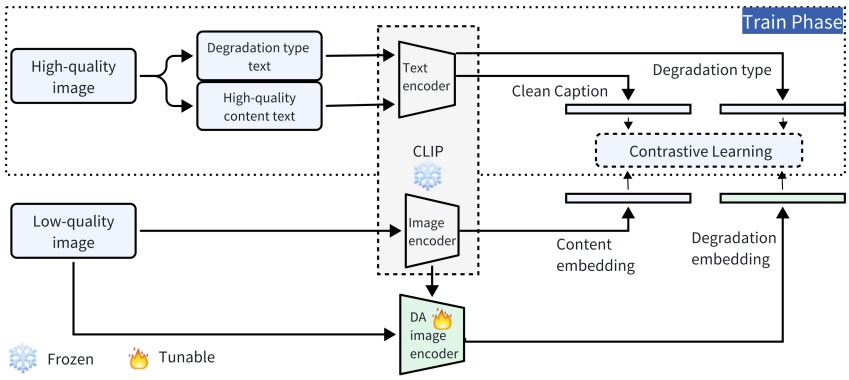

Fig. 3: The PVL [11] model overview. During PVL model train phase, freezes both text and image encoders of the CLIP but learns an additional degradation-adware(DA) image encoder with constrastive learning. The DA image encoder infers degradation embedding to match real corruption and then controls the image encoder to output clean content embedding.

Author: Add new figure of the PVL we used. Explain the core component.
For Review#3's 3rd comment.

We leverage the image understanding capability of the PVL model to preprocess degraded images. As depicted in Fig. 3, the pre-trained degradation-aware PVL generates both clean captions and degradation embeddings from a single degraded image. Inspired by CoT prompting, degradation embedding enables the restoration model to identify the type of degradation and take relevant reconstruction actions driven by the clean image caption embedding. As shown in Fig. 2, corrupted image I_d is directly fed into the PVL model for the generation of ReAct embedding. These embeddings are defined as Eq. 1

Author: Decription of the PVL.
For Review#3's 3rd comment.

Author: Improve the overall process description.
For Review#3's 3rd comment.

$$\mathcal{P}_d^{emb}, \quad \mathcal{P}_c^{emb} = PVL(I_d) \tag{1}$$

where PVL$(\cdot)$ is the PVL model, $\mathcal{P}_d^{emb}$ is the degradation embedding, and $\mathcal{P}_c^{emb}$ is the clean caption embedding. In Eq. 2, we apply a multilayer perceptron network, denoted as MLP$(\cdot)$, to the outputs of the PVL model. The refined embeddings, obtained through this process, are subsequently fed into the image restoration network.

$$\mathcal{P}_r^{emb} = MLP(\mathcal{P}_d^{emb}), \quad \mathcal{P}_a^{emb} = MLP(\mathcal{P}_c^{emb}) \tag{2}$$

In our restoration network, the $\mathcal{P}_r^{emb}$ and $\mathcal{P}_a^{emb}$ serve as the degradation reasoning embedding and restoration action embedding, respectively, enabling adaptive recovery of I_c.

3.3 ReAct Prompt-driven RAMamba block

We utilize an encoder-decoder structure as the image backbone network, with building blocks designed based on the vision Mamba block [10]. Additionally, we have integrated a ReAct PDM into the original vision Mamba block, see Fig. 4.

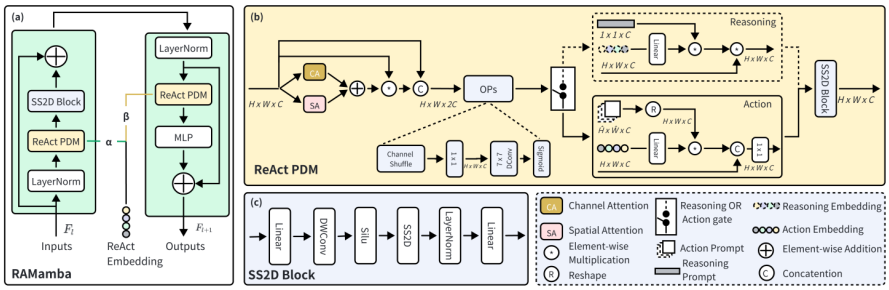

Fig. 4: Main components of RAMamba. (a) Mamba block with ReAct Prompt-driven Module. (b) Reasoning and Action Prompt-driven Module(ReAct PDM). (c) 2D Selective Scan Block of RAMamba.

RAMamba block: The block designed in the previous vision Mamba-based networks [10, 21, 22] mainly follow $Norm \rightarrow SS2D \rightarrow Norm \rightarrow MLP$ flow. Refer to Fig. 5, where the crucial module, 2D-selective-scan (SS2D), flattens the 2D image feature into a 1D sequence through cross-scanning in various distinct directions. Subsequently, the long-range dependency of each sequence is captured utilizing the discrete S6 block, as described in [10]. Finally, all sequences are merged by summation and then reshaped to restore the original 2D structure. Although SS2D can model long dependencies with linear computational complexity, simply integrating prompts into the block can only get sub-optimal guidance for image restoration. Therefore, The PDM will enhance network degradation awareness and reconstruction ability.

Author: Add SS2D moduel description. For Review#3's 3rd comment.

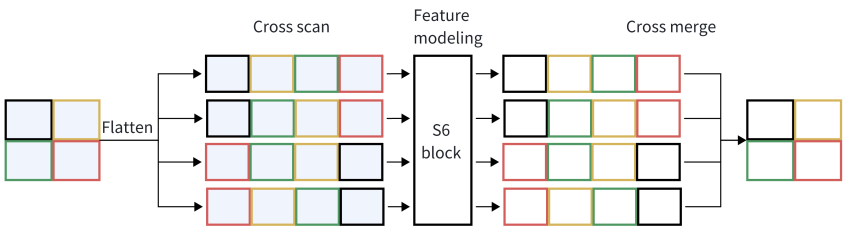

Fig. 5: Conceptual Visualization of the 2D Selective Scanning Approach

Author: Add Figure of SS2D. For Review#3's 3rd comment.

To this end, we propose the RaMamba block with a ReAct PDM to interact with ReAct embedding. As depicted in Fig. 4(a), given the input feature $F_l \in \mathbb{R}^{H \times W \times C}$

and the ReAct embedding $\mathcal{P}_{r|a}^{emb}$ which extracted from the PVL model. Firstly, we apply LayerNorm (LN) and SS2D block Fig. 4(c) with residual connection to capture the spatial long-range dependencies, and then the features are forwarded to LN and an MLP layer to enhance the model's representational capabilities. Moreover, we employ a ReAct PDM after each LN. Through experiments, we found the PDM provides better utilization of the ReAct embeddings. Additionally, we split the embedding into two parts by Eq. 3 and dynamically adapting to both PDMs with two learnable parameters α and β. The above process can be formulated as:

$$\mathcal{P}1_{r|a}^{emb}, \quad \mathcal{P}2_{r|a}^{emb} = Sp(\mathcal{P}_{r|a}^{emb}) \tag{3}$$

$$Z = SS2D(PDM(LN(F_l), \alpha \cdot \mathcal{P}1_{r|a}^{emb})) + F_l \tag{4}$$

$$F_{l+1} = MLP(PDM(LN(Z), \beta \cdot \mathcal{P}2_{r|a}^{emb})) + Z \tag{5}$$

where Sp is the split operation, LN is the LayerNorm, PDM and $SS2D$ refer to prompt-driven module and 2D-selective scan operation, respectively. From Eq. 4 global representations Z are obtained by SS2D operation with the driven of ReAct embedding $\mathcal{P}1_{r|a}^{emb}$, and then enhancing Z with another embedding $\mathcal{P}2_{r|a}^{emb}$ to get the final outputs F_{l+1}, shown in Eq. 5.

ReAct prompt-driven module: We design the ReAct PDM to facilitate interaction between the ReAct embedding and the normalized input feature F_l, enabling the model to take restoration actions based on degradation type, which is the reasoning embedding. To refine the F_l for better fusion with ReAct embedding, we apply attention mechanisms in both channel and spatial dimensions before sending to the reasoning OR action gate, which is designed to control reasoning or action prompt handle correspondingly. We utilize an additional SS2D block to obtain the prompt-driven feature.

The detailed procedure of the ReAct PDM is illustrated in Fig. 4 (b). F_l is preliminary re-weighted with Channel Attention (CA) and Spatial Attention (SA). It is then concatenated with F_l in the channel dimensions to better utilize the origin input feature. The weighted feature $F_l^w \in \mathbb{R}^{H \times W \times 2C}$ is defined as Eq. 6:

$$F_l^w = CAT[(CA(F_l) \oplus SA(F_l)) \odot F_l, F_l] \tag{6}$$

where $CAT[\cdot]$ is channel-wise concatenation, CA and SA denote channel and spatial attention operation, $\oplus$ is element-wise addition, and $\odot$ represents the element-wise multiplication. Next, we generate $\hat{F_l^w}$ by Eq. 7

$$\hat{F_l^w} = \sigma(DConv_{7\times7}(Conv_{1\times1}(CS(F_l^w)))) \tag{7}$$

where σ is the sigmoid function, $DConv_{k\times k}$ is $k \times k$ depth-wise separable convolution, $Conv_{k\times k}$ is the convolution operation with $k \times k$ kernel, and CS denotes channel shuffle operation. The fine-gained $\hat{F_l^w}$ then goes through the designed reasoning OR action gate to selectively interact with the ReAct embedding. Specifically, in the encoder stage, $\hat{F_l^w}$ applies element-wise multiplication with the reasoning embedding, which is projected to the length of C using linear operation. Additionally, we introduce

learnable prompts $p_r \in \mathbb{R}^{1\times1\times C}$ to dynamically handle various reasoning embedding, thereby enriching the encoder's awareness of degradation. As Eq. 8 reasoning embedding interaction is defined as:

$$P_R = ((Linear(\mathcal{P}_r^{emb}) \odot p_r) \odot \hat{F_l^w}) \tag{8}$$

where $Linear$ is the linear operation, $\mathcal{P}_r^{emb}$ is the reasoning embedding generated by Eq. 2. Different from the reasoning interaction, the action interaction happens in the decoder stage. Specifically, the projected action embedding is multiplied with the learnable prompt $p_a \in \mathbb{R}^{\hat{H}\times\hat{W}\times C}$, where $\hat{H}$, $\hat{W}$, C represent the height, width, and channels of the initial prompt. Since the network needs to recover images of different resolutions during inference, we re-scale p_a to match the size of the incoming degraded input image features. Action prompt P_A obtained by channel-wise concatenation of the weighted action embedding and $\hat{F_l^w}$. Then P_A is derived by Eq. 9

Author: Add explation for R() function, daynamic reshape the resoultion of the input images. For reviewer#4 5th question.

$$P_A = CAT[(Linear(\mathcal{P}_a^{emb}) \odot R(p_a), \hat{F_l^w}] \tag{9}$$

where $R(\cdot)$ denotes bilinear interpolation, re-scaling $p_a \in \mathbb{R}^{\hat{H}\times\hat{W}\times C}$ to the same size of $\hat{F_l^w} \in \mathbb{R}^{H\times W\times C}$, $\mathcal{P}_a^{emb}$ is the action embedding from Eq. 2. At end of the prompt interaction, we employ an additional refinement block to finalize the prompt-driven feature by Eq. 10

$$F_l^{'} = SS2D(P_A|P_R) \tag{10}$$

where $F_l^{'}$ represents the output feature of ReAct PDM, $SS2D(|)$ denotes applying 2D-selective scan to either reasoning prompts P_R or action prompts P_A.

4 Experiments and Analysis

To demonstrate the strong performance of our proposed RamIR, we experimentally evaluate our method on two different settings as shown in Table 1:

- **S3**: following [3–5], include three representative image restoration tasks: deraining, dehazing, denoising.
- **S5**: keep aligned with [2, 6], two more typical tasks: image deblurring and low-light enhancement involved.

Furthermore, we undertake comprehensive ablation studies through extensive experimentation, validating the efficacy and merit of our proposed method.

Author: Add a table to present the experiment settings. For Review#3's 5th comment.

Table 1: Two distinct experimental settings are utilized, with tasks marked with a checkmark (✓) indicating their inclusion.

Setting	Deraining	Dehazing	Denoising	Deblurring	Low-Light
S3	✓	✓	✓	✕	✕
S5	✓	✓	✓	✓	✓

4.1 Implementation Details

Parameter Settings: We train RamIR end-to-end. Following [11], the corrupt images pass to the PVL model for reasoning and action embedding generation. All inputs are normalized to the range of [0, 1] and resized to 224 × 224 using bicubic interpolation. Our RamIR backbone consists of a 4-level symmetrical encoder-decoder U-Net structure. The number of RAMamba blocks is set to [1, 2, 2, 4] and channel numbers are [48, 96, 192, 384] from level-1 to level-4, respectively. We employ the ReaAct PDM at the beginning of each level. In accordance with previous works [4, 5], we perform data augmentation by applying horizontal flips and random rotation and then crop the input images to a patch size of 128×128. To stabilize the training process, we use Adam optimizer ($\beta 1 = 0.9, \beta 2 = 0.999$). The initial learning rate is set to $2e^{-4}$ and gradually decreases to $1e^{-6}$ using the cosine annealing strategy. We adopt the L1 loss as the network loss function. With a batch size of 8, the training takes approximately 3 days for a total of 700k iterations on one NVIDIA 4090 GPU.

Datasets: We evaluate our method on various benchmark datasets. For image restoration in rainy condition, we utilize the Rain100L [34] dataset, which comprises 200 training and 100 testing clean-rainy image pairs. For dehazing, we adopt the SOTS dataset [35] with over 7,000 training images and 500 test images. For noise removal, we train the model on a combination of BSD400 [36] and WED [37] dataset. The BSD400 dataset contains 400 clean images and the WED dataset has 4,744 clean images. We generate noisy counterparts by adding Gaussian noise with varying levels of $\sigma \in \{15, 25, 50\}$ to these images. Similar noise addition is applied to BSD68 [38] and Urban100 [39] for testing. Following [2], we use the LOL [40] dataset for image low light enhancement and the GoPro [41] dataset for image deblurring. We combine all the above-mentioned datasets as joint all-in-one training and test datasets, respectively.

Evaluation Metrics: To assess the restoration quality of our method, we rely on two widely-accepted metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), both evaluated in the RGB color space. Higher values for both PSNR and SSIM indicate superior restoration performance.

4.2 All-in-one image restoration results

We report quantitative results on all-in-one setting across various datasets. Table 2 shows the proposed RamIR outperforms existing approaches on **S3** setting. Notably, our method yields a 0.6dB performance gain over the previous method PromptIR [5] and surpasses InstructID-3D [2] by 0.2 dB on averaged PSNR. The RAMamba block we have designed exhibits a remarkable capability to grasp the intricately structured information within image and effectively activate a greater number of pixels, resulting in a substantial boost in image recovery performance. Specifically, our approach surpasses PromptIR [5] by an impressive 1.79 dB in the challenging image deraining task, demonstrating a significant leap forward. Furthermore, for the dehazing task, our method achieves a notable 1.0 dB improvement over InstrctIR [2], underscoring its effectiveness and superiority in tackling diverse image restoration challenges. These results underscore the potential of our approach and its capability to revolutionize the field of image recovery. Visual results illustrated in Fig. 6, Our method efficiently

eliminates degradation, resulting in visually enhanced images that provide a clearer understanding of image structure and ultimately yield superior reconstructions. We

Table 2: Quantitative results on all-in-one **S3** setting. We respectively highlight the best and second-best methods with **bold** and underline text.

Author: Update caption of the table. For Review#3's 2nd comment.

Method	Deraining Rain100L [34] PSNR/SSIM	Dehazing SOTS [35] PSNR/SSIM	Denoising on BSD68 [38] σ=15 PSNR/SSIM	σ=25 PSNR/SSIM	σ=50 PSNR/SSIM	Average PSNR/SSIM
FDGAN [42]	29.89/0.933	24.71/0.924	30.25/ 0.910	28.81/0.868	26.43/0.776	28.02/0.883
MPRNet [43]	33.57/0.954	25.28/0.954	33.54/0.927	30.89/0.880	27.56/0.779	30.17/0.899
DL [44]	32.62/0.931	26.92/0.391	33.05/0.914	30.41/0.861	26.90/0.740	29.98/0.875
AirNet [3]	34.90/0.967	27.94/0.962	33.92/0.933	31.26/0.888	28.00/0.797	31.20/0.910
PromptIR[5]	36.37/0.972	30.58/0.974	33.98/0.933	31.31/0.888	28.06/0.799	32.06/0.913
DA-CLIP[11]	36.46/0.974	30.12/0.972	33.86/0.925	31.06/0.865	27.55/0.778	31.81/0.903
InstructIR-3D[2]	37.98/0.978	30.22/0.959	**34.15/0.933**	31.52/0.890	**28.30/0.804**	32.43/0.913
RamIR(Ours)	**38.16/0.981**	**31.29/0.977**	34.04/0.931	**31.61/0.891**	28.19/0.801	**32.65/0.916**

also report results on five challenging image restoration tasks in Table 3. It shows that our RamIR achieves average 0.63dB improvement over InstructIR [2]. Similar improvements are observed when more tasks are involved. For instance, when compared to InstructIR [2], our method yields performance gains of 0.7dB on the deraining task and 3.91 dB on dehazing task. Visual comparison in Fig. 7, our method is capable of reconstructing blur-free images, generating visually pleasant clean images. Qualitative examples are presented in Fig. 8, we compare with low-light enhancement methods, the RamIR provides better visual results. Based on these results, we can deduce that our method offers comprehensive all-in-one restoration capabilities, particularly excelling in deraining and dehazing images. In contrast, InstructIR [2], which secured the second-best outcome, benefits from its straightforward non-linear activation-free network design, enabling it to perform remarkably well in low-light image enhancement and denoising tasks.

Author: Emphasis the result vaues. For Review#2's 3rd comment.

Table 3: Comparisons under all-in-one **S5** setting for five challenging datasets.

Author: Update captions of the table. For Review#3's 2nd comment.

Method	Deraining Rain100L [34] PSNR/SSIM	Dehazing SOTS [35] PSNR/SSIM	Denoising BSD68 [38] PSNR/SSIM	Deblurring GoPro [41] PSNR/SSIM	Low-Light Enh. LOL [40] PSNR/SSIM	Average PSNR/SSIM
NAFNet [1]	35.56/0.967	25.23/0.939	31.02/0.883	26.53/0.808	20.49/0.809	27.76/0.881
HINet [45]	35.67/0.969	24.74/0.937	31.00/0.881	26.12/0.788	19.47/0.800	27.40/0.875
DGUNet [46]	36.62/0.971	24.78/0.940	31.10/0.883	27.25/0.837	21.87/0.823	28.32/0.891
MIRNetV2 [47]	33.89/0.954	24.03/0.927	30.97/0.881	26.30/0.799	21.52/0.815	27.34/0.875
SwinIR [23]	30.78/0.923	21.50/0.891	30.59/0.868	24.52/0.773	17.81/0.723	25.04/0.835
Restormer [4]	34.81/0.962	24.09/0.927	31.49/0.884	27.22/0.829	20.41/0.806	27.60/0.881
DL [44]	21.96/0.762	20.54/0.826	23.09/0.745	19.86/0.672	19.83/0.712	21.05/0.743
Transweather [48]	29.43/0.905	21.32/0.885	29.00/0.841	25.12/0.757	21.21/0.792	25.22/0.836
TAPE [49]	29.67/0.904	22.16/0.861	30.18/0.855	24.47/0.763	18.97/0.621	25.09/0.801
AirNet [3]	32.98/0.951	21.04/0.884	30.91/0.882	24.35/0.781	18.18/0.735	25.49/0.846
IDR [6]	35.63/0.965	25.24/0.943	**31.60/0.887**	27.87/0.846	21.34/0.826	28.34/0.893
InstructIR-5D [2]	36.84/0.973	27.10/0.956	31.40/0.887	**29.40/0.886**	**23.00/0.836**	29.55/0.907
RamIR(Ours)	**37.56/0.979**	**31.09/0.978**	31.44/0.886	28.82/0.878	22.02/0.828	**30.18/0.910**

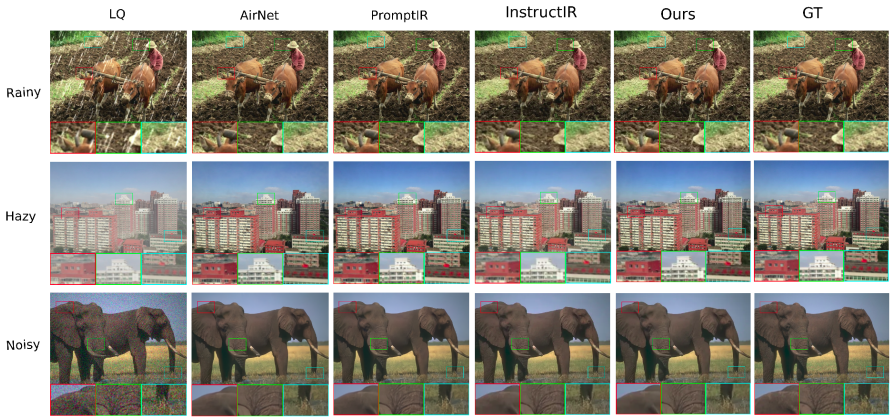

Fig. 6: Comparison of our method with other all-in-one approaches on **S3** setting. Zooming-in for a better visualization and comparisons for the areas of highlighted in colored rectangles.

Author: Update captions of the figure.
For Review#3's 2nd comment.

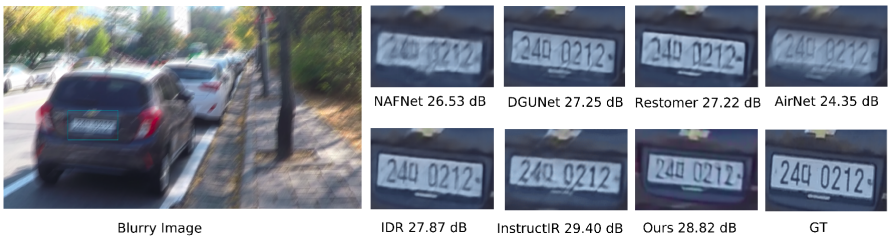

Fig. 7: Image deblurring results on **S5** setting. Comparison with other methods on GoPro [41].

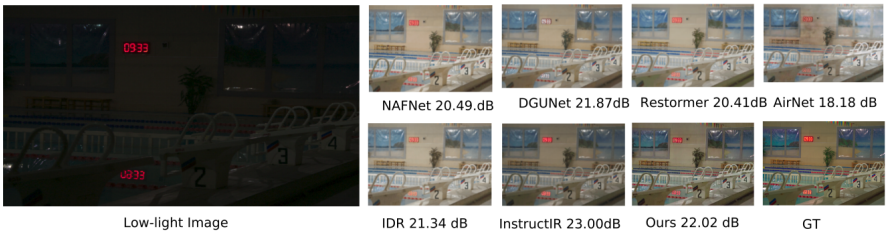

Fig. 8: Image low-light enhancement results on **S5** setting. Comparison with other methods on LOL [40].

4.3 Degradation-specific image restoration results

Our initial goal was to develop a powerful all-in-one image restoration model to handle various degradation types. Nevertheless, we also provide a quantitative comparison of degradation-specific tasks to show the ability of single degradation image restoration

which benefits from the ReAct prompt learning and Mamba block. Quantitative results for various degradation-specific tasks are presented in Appendix A.

4.4 Ablation Studies

We perform the ablation experiments on the **S3** setting to verify the effectiveness and scalability of our method. Firstly, we evaluate how the key ReAct prompt learning affects the restoration performance. As demonstrated in Table 4, employing ReAct prompt learning in a CoT manner significantly enhances the restoration quality of the network. Specifically, these prompts contribute to a notable 0.64 dB gain in performance. Next, Table 5 showcases the effectiveness of our well-designed PDM. By incorporating the PDM, the network achieves an additional 0.34 dB increase in PSNR, dynamically enhancing the interaction between the prompt and the backbone. The

Author: Emphasis the result vaues. For Review#2's 3rd comment.

Table 4: Impact of ReAct prompt.

Reasoning	Action	PSNR	SSIM
✗	✗	32.01	0.909
✓	✗	32.38	0.912
✗	✓	31.91	0.909
✓	✓	**32.65**	**0.916**

Table 5: Impact of PDM.

PDM	PSNR	SSIM
✗	32.31	0.915
✓	**32.65**	**0.916**

Fig. 9: tSnE plots of the degradation embeddings. The various colors represent different types of degradation. By incorporating ReAct prompt into the restoration network, we achieve superior clustering of embeddings for each specific task.

tSNE results illustrated in Fig. 9, the application of the ReAct prompt results in significantly enhanced clustering of various degradation types compared to its absence.

Author: Move the figure and description to experiement section. For Review#3's 1st comment.

This improved segregation underscores the effectiveness of the ReAct prompt in guiding the network to learn more discriminative degradation priors, thereby boosting the overall performance of the all-in-one image restoration process. Our method consistently produces appealing results across different degradation types. To evaluate the performance on unseen degradation image restoration, Appendix B offers the results on unseen degradation. Our method performs well on unknown scenarios over PromptIR [5].

4.5 Efficiency Analysis

We compare our model effectiveness with state-of-the-art restoration models and present model complexities and scales in Fig. 10, one can see that by incorporating RAMamba blocks to various degradation removal tasks, our method achieves high performance with relatively low model complexity. Specifically, RamIR achieves a PSNR that is 0.46dB higher than PromptIR [5], while using 34.2% fewer parameters and 39% more efficient in terms of computational complexity. This efficiency is largely due to our RAMamba block, which models long-distance dependencies with 2D selective scan mechanism and leverages parallel computing devices. .

Author: Emphasis the result vaues.
For Review#2's 3rd comment.

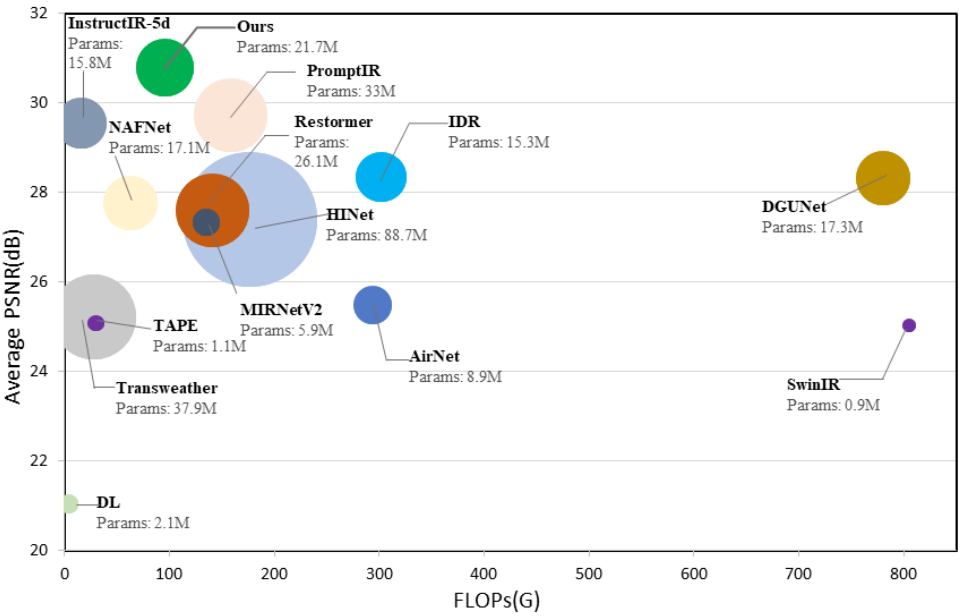

Fig. 10: Comparison of Parameter Scales and Model Performance in **S5** setting. FLOPs were computed on $256 \times 256 \times 3$ input, and the size of each bubble visually represents the number of parameters within each model.

Author: Update the efficency comparison.
For Review#2's 4th comment.

4.6 Discussion

Author: Update discussion part.
For Review#3's 7th comments.

Mamba-based image restoration network: In Fig. 11, as we progressively increase the input image resolution, our method demonstrates a linear growth in computational complexity, measured in FLOPs, compared to Transformer-based approaches such as PromptIR [5], Restormer [4], and IDR [6]. Leveraging the advantageous property of vision Mamba, which excels in processing long sequence inputs [9, 10], the RAMamba block effectively tackles the challenge of processing 2D image features with linear computational complexity. In Appendix C, we present quantitative results for Mamba-based networks in image restoration tasks, demonstrating the broad applicability of our method in this field. Fig. 12 illustrates the impact of training patch size on restoration performance, noting that our method achieves an average 0.5dB gain over InstructIR [2] when using input training image patches of 256×256, highlighting that activating more pixels significantly enhances recovery quality. Therefore, our proposed method demonstrates even greater superiority when handling larger input image resolutions. We believe that our substantial improvement in performance and efficiency for image processing tasks, which have the potential to serve as a fresh approach for future research.

Author: Emphasis the result vaues.
For Review#2's 3rd comment.

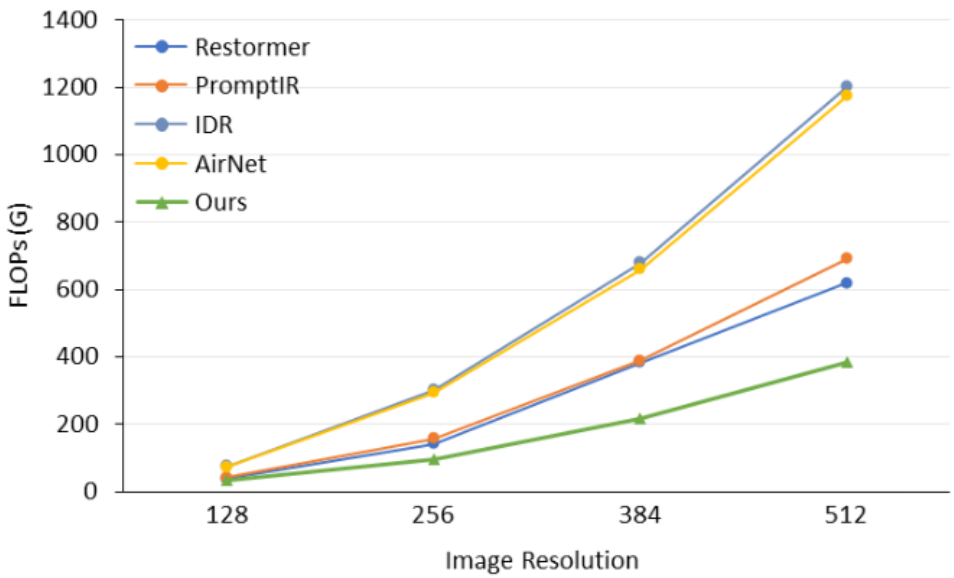

Fig. 11: FLOPs with progressively larger test image resolutions.

.

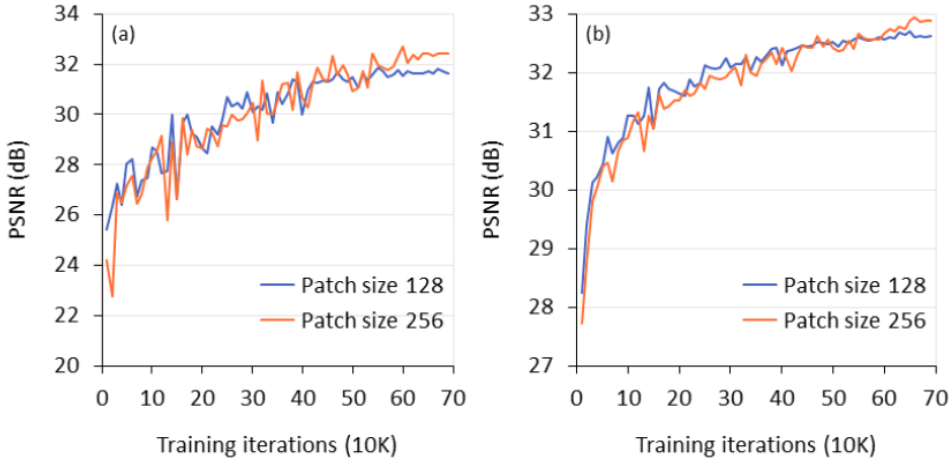

Fig. 12: Comparison of training curves (**S3**) with different patch sizes. (a) Image dehazing. (b) Averaged score across three tasks.

ReAct prompt: We visualize the t-SNE statistics of the clustered specific-task embeddings in Fig. 9. The learnable ReAct prompt significantly refresh the distribution of various degradation priors for the all-in-one task, potentially fostering further exploration of the interaction between the Mamba block and tunable ReAct prompt, such as through the cross-attention mechanism.

PVL model: Text-instructed image restoration approaches effectively steer the process of high-quality image reconstruction. Multi-modality models, such as PVL models, generate degradation-related embeddings that significantly bolster the model's capability to handle diverse degradation types within a unified framework. Furthermore, we utilize the degradation-aware PVL model [11] to output clean image embeddings, utilizing these as action prompts to further enhance the reconstruction process. The emerging advanced methodologies in NLP, such as CoT, prompt learning, and ReAct, have significantly accelerated the exploration of low-level computer vision tasks, including image restoration.

4.7 Limitations and Future work

Our approach delivers remarkable outcomes on conventional, demanding image restoration tasks, introducing a promising avenue for Chain-of-Thought (CoT) prompt learning to empower Mamba-driven recovery from unspecified degraded imagery. While promising, we recognize inherent limitations. Firstly, the reliability of our method relies crucially on the performance of the fine-tuned PVL model, which generates robust ReAct prompts. Secondly, similar to preceding approaches, RamIR encounters challenges in isolating and restoring a clean image from a single image exhibiting complex, mixed degradations. Although our approach achieves higher average PSNR scores, reflecting an implicit capability to categorize degradation types via

Author: Add limitations and our future work. For Review#2's 5th comment.

ReAct prompts, there is a loss of granularity in distinguishing between closely related degradations, limiting optimal results across diverse tasks.

In our future work, we will strive to 1) explore more PVL models to generate more robust high quality image representations and degradation embeddings. 2) explore works for closely or mixed degradation types removal, and increase the general performance across different degradation types.

5 Conclusion

We introduce RamIR, a robust all-in-one image restoration model that leverages the power of Prompt-based Vision Language (PVL) models. RamIR generates reasoning and action prompts to supervise the restoration network, enabling it to recover high-quality images from various degraded conditions. A key innovation lies in the introduction of the RAMamba building block, which interacts seamlessly with the ReAct prompt, significantly enhancing the model's recovery performance. This vision Mamba-based approach, guided by Chain-of-Thought (CoT) prompts, demonstrates competitive results across challenging image restoration tasks including deraining, dehazing, denoising, deblurring, and low-light enhancement.

Appendix A Degradation-specific task results

We re-train the model on each single degradation dataset. In Table A1, we compare with degradation-specific methods for deraining and dehazing, and for deblurring and low-light enhancement in image in Table A2. Also, extensive comparisons for image denoising shown in Table A3. Notably, our RamIR outperforms the PromptIR network by achieving favorable gains of 0.9dB and 1.34dB for deraining and dehazing tasks, respectively. This indicates that the Mamba-based network modeling of long-distance dependencies significantly enhances the recovery process. Furthermore, in the realm of image deblurring, our approach surpasses the CUMamba [21], a Mamba-based network, by achieving a gain of 0.2dB. The incorporation of the ReAct prompt further aids in improving noise removal capabilities.

Author: Emphasis the result vaues. For Review#2's 3rd comment.

Table A1: Deraining and Dehazing comparison: We compare with task-specific classical methods on benchmark datasets.

Deraining Rain100L [34]		Dehazing SOTS [35]	
Method	PSNR/SSIM	Method	PSNR/SSIM
UMR [50]	32.39/0.921	GFN [51]	21.55/0.844
LPNet [52]	33.61/0.958	GCANet [53]	19.98/0.704
AirNet [3]	33.90/0.977	AirNet [3]	23.18/0.900
Restormer [4]	36.74/0.978	Restormer [4]	30.87/0.969
PromptIR [5]	37.04/0.979	PromptIR [5]	31.31/0.973
RamIR	**38.43/0.982**	RamIR	**32.65/0.979**

Table A2: Quantitative comparison with state-of-the-art methods of debulrring and low-light enhancement.

Deblurring GoPro [41]		Low-light Enh. LOL [40]	
Method	PSNR/SSIM	Method	PSNR/SSIM
DeblurGAN-v2 [54]	Enlightengan [55]	17.61/0.653	
RNN [56]	29.19/0.931	URetinexNet [57]	21.32/0.931
Restormer [4]	32.92/0.961	LPNet [52]	21.32/0.835
MRL [58]	33.50/0.965	InstructIR [2]	22.83/0.836
CUMamba [21]	33.53/0.965	IAGC [59]	**24.53/0.842**
RamIR	**33.71/0.968**	RamIR	23.67/0.838

Table A3: Comparison on image denoising with advance methods, we report PSNR on benchmark datasets considering different σ noise levels.

Method	BSD68 [38]			Urban100 [39]		
	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$
IRCNN [60]	33.87	31.18	27.88	27.59	31.20	27.70
FFDNet [61]	33.87	31.21	27.96	33.83	31.40	28.05
DnCNN [19]	33.89	31.23	27.92	32.98	30.81	27.59
NAFNet [1]	33.67	31.02	27.73	33.14	30.94	27.32
HINet [45]	33.72	31.00	27.63	33.49	30.94	27.32
DGUNet [46]	33.85	31.10	27.92	33.67	31.27	27.94
AirNet [3]	34.14	31.48	28.23	34.40	32.10	28.88
Restormer [4]	34.29	31.64	28.41	34.67	32.41	29.31
PromptIR [5]	34.34	31.71	28.49	34.77	32.49	29.39
RamIR	**34.40**	**31.81**	**28.62**	**35.21**	**32.56**	**30.12**

Appendix B Experiments on out-of-distribution degradation

Thanks to our ReAct prompt learning, RamIR demonstrates promising "unified" image restoration capacity. We report results on BSD68 and Urban100 with noise level $\sigma = 100$, which are outside of the training dataset distribution, for evaluating degradation unseen tasks. In Table B4, our method demonstrates superior performance on this unseen dataset.

Table B4: Out of distribution dataset for all-in-one image restoration under **S3** setting of denoising level $\sigma = 100$. Our method generalize well compared to the AirNet [3] and PromptIR [5] baselines.

	Denoising $\sigma = 100$	
Method	BSD68 [38] PSNR	Urban100[39] PSNR
AirNet [3]	13.64	13.72
PromptIR [5]	21.03	20.05
RamIR	**21.06**	**20.11**

Appendix C Experiments of Mamba-based restoration network

Author: Add comparison table with mamba-based works.
For Review#3's 6th comment.

The power of the recently advanced state space model, specifically Mamba for image restoration, has been thoroughly explored to address the dilemma of balancing efficient computation with a global effective receptive field. Our method offers a novel perspective by integrating PVL with the Mamba-based restoration network. In Table C5, our approach outperforms the state-of-the-art Mamba-based image deraining network, showcasing immense potential in the field of image restoration.

Table C5: Quantitative comparison for image deraining. The results are reported the Y channel in YCbCr color space.

Method	Rain100H PSNR/SSIM	Rain100L PSNR/SSIM
MambaIR [22]	30.62/0.893	38.78/0.977
VMambaIR [62]	31.66/0.909	39.09/0.979
FreqMamba [63]	31.74/0.912	39.18/0.981
RamIR	**32.11/0.914**	**39.89/0.987**

Ethics declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

Ethical and informed consent for data used

This study relied on the openly accessible datasets, and cited the corresponding citations. As these datasets are publicly available and devoid of personally identifiable information, individual informed consent was not sought. Our utilization of these datasets aligns with ethical guidelines and legal obligations while honoring the provisions set by the data providers.

Data Availability

All data cited in this manuscript have been appropriately referenced. Detailed information regarding the cited datasets can be found in the references section with their respective citations. As such, all cited data are publicly available and accessible.

References

Author: Updated references. we keep the last five years advance works. For reviewer#1's 1st comment.

[1] Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, pp. 17–33. Springer, Berlin, Heidelberg (2022)

[2] Conde, M.V., Geigle, G., Timofte, R.: High-quality image restoration following human instructions. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)

[3] Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-In-One Image Restoration for Unknown Corruption. In: IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA (2022)

[4] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[5] Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.: Promptir: Prompting for all-in-one image restoration. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)

[6] Zhang, J., Huang, J., Yao, M., Yang, Z., Yu, H., Zhou, M., Zhao, F.: Ingredient-oriented multi-degradation learning for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5825–5835 (2023)

[7] Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: The International Conference on Learning Representations (ICLR) (2022)

[8] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

[9] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)

[10] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)

[11] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. arXiv preprint arXiv:2310.01018 (2023)

[12] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023)

[13] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)

[14] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)

[15] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)

[16] Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: A better and simpler baseline. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)

[17] Chen, Z., He, Z., Lu, Z.-M.: Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. IEEE Transactions on Image Processing **33**, 1002–1015 (2024)

[18] Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023)

[19] Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing **26**(7), 3142–3155 (2017)

[20] Tsai, F.-J., Peng, Y.-T., Tsai, C.-C., Lin, Y.-Y., Lin, C.-W.: Banet: A blur-aware attention network for dynamic scene deblurring. IEEE Transactions on Image Processing **31**, 6789–6799 (2022) https://doi.org/10.1109/TIP.2022.3216216

[21] Deng, R., Gu, T.: Cu-mamba: Selective state space models with channel learning for image restoration. arXiv preprint arXiv:2404.11778 (2024)

[22] Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.-T.: Mambair: A simple baseline for image restoration with state-space model. In: ECCV (2024)

[23] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. arXiv preprint arXiv:2108.10257 (2021)

[24] Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22367–22377 (2023)

[25] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[26] Zhou, K., Yang, J., Loy, C. Chen, Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) (2022)

[27] Yang, H., Pan, L., Yang, Y., Liang, W.: Language-driven all-in-one adverse weather removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24902–24912 (2024)

[28] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1680–1691 (2023)

[29] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. International Conference on Machine Learning (2023)

[30] Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Bras, R.L., Choi, Y., Hajishirzi, H.: Generated knowledge prompting for commonsense reasoning. arXiv preprint arXiv:2110.08387 (2022)

[31] Lu, Y., Hong, Y., Wang, Z., Zhou, G.: Enhancing reasoning capabilities by instruction learning and chain-of-thoughts for implicit discourse relation recognition. In: Conference on Empirical Methods in Natural Language Processing (2023)

[32] Liu, Y., Peng, X., Du, T., Yin, J., Liu, W., Zhang, X.: Era-cot: Improving chain-of-thought through entity relationship analysis. arXiv preprint arXiv:2403.06932 (2024)

[33] Zhang, Y., Wu, Y., Liu, Y., Peng, X.: Cpa-enhancer: Chain-of-thought prompted adaptive enhancer for object detection under unknown degradations. arXiv preprint arXiv:2403.11220 (2024)

[34] Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

[35] Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2019) https://doi.org/10.1109/TIP.2018.2867951

[36] Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 898–916 (2011) https://doi.org/10.1109/TPAMI.2010.161

[37] Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L.: Waterloo exploration database: New challenges for image quality assessment models. IEEE Transactions on Image Processing **26**, 1004–1016 (2017)

[38] Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 2, pp. 416–4232 (2001). https://doi.org/10.1109/ICCV.2001.937655

[39] Huang, J.-B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206 (2015). https://doi.org/10.1109/CVPR.2015.7299156

[40] Chen Wei, W.Y. Wenjing Wang, Liu, J.: Deep retinex decomposition for low-light enhancement. In: British Machine Vision Conference (2018). British Machine Vision Association

[41] Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

[42] Dong, Y., Liu, Y., Zhang, H., Chen, S., Qiao, Y.: FD-GAN: generative adversarial networks with fusion-discriminator for single image dehazing. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 10729–10736 (2020)

[43] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Multi-stage progressive image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

[44] Fan, Q., Chen, D., Yuan, L., Hua, G., Yu, N., Chen, B.: A general decoupled learning framework for parameterized image operators. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 33–47 (2021) https://doi.org/10.1109/TPAMI.2019.2925793

[45] Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 182–192 (2021)

[46] Mou, C., Wang, Q., Zhang, J.: Deep generalized unfolding networks for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition (CVPR) (2022)

[47] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.-H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV, pp. 492–511. Springer, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58595-2_30 . https://doi.org/10.1007/978-3-030-58595-2_30

[48] Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2343–2353 (2021)

[49] Liu, L., Xie, L., Zhang, X., Yuan, S., Chen, X., Zhou, W., Li, H., Tian, Q.: Tape: Task-agnostic prior embedding for image restoration. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, pp. 447–464. Springer, Cham (2022)

[50] Yasarla, R., Patel, V.M.: Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8397–8406 (2019). https://doi.org/10.1109/CVPR.2019.00860

[51] Ren, W., Ma, L., Zhang, J., Pan, J.-s., Cao, X., Liu, W., Yang, M.-H.: Gated fusion network for single image dehazing. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3253–3261 (2018)

[52] Li, J., Li, J., Fang, F., Li, F., Zhang, G.: Luminance-aware pyramid network for low-light image enhancement. IEEE Transactions on Multimedia **23**, 3153–3165 (2021) https://doi.org/10.1109/TMM.2020.3021243

[53] Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G.: Gated context aggregation network for image dehazing and deraining. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1375–1383 (2019). https://doi.org/10.1109/WACV.2019.00151

[54] Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: The IEEE International Conference on Computer Vision (ICCV) (2019)

[55] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing **30**, 2340–2349 (2021)

[56] Zhang, J., Pan, J., Ren, J., Song, Y., Bao, L., Lau, R.W.H., Yang, M.-H.: Dynamic scene deblurring using spatially variant recurrent neural networks. In:

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2521–2529 (2018). https://doi.org/10.1109/CVPR.2018.00267

[57] Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5891–5900 (2022). https://doi.org/10.1109/CVPR52688.2022.00581

[58] Dong, J., Pan, J., Yang, Z., Tang, J.: Multi-scale residual low-pass filter network for image deblurring. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12311–12320 (2023). https://doi.org/10.1109/ICCV51070.2023.01134

[59] Wang, Y., Liu, Z., Liu, J., Xu, S., Liu, S.: Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13082–13091 (2023). https://doi.org/10.1109/ICCV51070.2023.01207

[60] Yang, B., Qin, L., Liu, J., Liu, X.: Ircnn: An irregular-time-distanced recurrent convolutional neural network for change detection in satellite time series. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022) https://doi.org/10.1109/LGRS.2022.3154894

[61] Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing **27**(9), 4608–4622 (2018) https://doi.org/10.1109/TIP.2018.2839891

[62] Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., Yang, W.: Vmambair: Visual state space model for image restoration. ArXiv preprint arXiv:2403.11423 (2024)

Author: Add latest Mamba-baser IR works.
For reviewer#1 1st comment.

[63] Zhen, Z., Hu, Y., Feng, Z.: Freqmamba: Viewing mamba from a frequency perspective for image deraining. ArXiv preprint arXiv:2404.09476 (2024)