Vision Mamba Mender

Jiacong Hu[1,4], **Anda Cao**[1], **Zunlei Feng**[2,3,4*],
Shengxuming Zhang[2], **Yi Wang**[1], **Lingxiang Jia**[1], **Mingli Song**[1,3,4]

[1]College of Computer Science and Technology, Zhejiang University,
[2]School of Software Technology, Zhejiang University,
[3]State Key Laboratory of Blockchain and Data Security, Zhejiang University,
[4]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

{jiaconghu,caoanda,zunleifeng}@zju.edu.cn,
{zsxm1998,y_w,lingxiangjia,brooksong}@zju.edu.cn

Abstract

Mamba, a state-space model with selective mechanisms and hardware-aware architecture, has demonstrated outstanding performance in long sequence modeling tasks, particularly garnering widespread exploration and application in the field of computer vision. While existing works have mixed opinions of its application in visual tasks, the exploration of its internal workings and the optimization of its performance remain urgent and worthy research questions given its status as a novel model. Existing optimizations of the Mamba model, especially when applied in the visual domain, have primarily relied on predefined methods such as improving scanning mechanisms or integrating other architectures, often requiring strong priors and extensive trial and error. In contrast to these approaches, this paper proposes the Vision Mamba Mender, a systematic approach for understanding the workings of Mamba, identifying flaws within, and subsequently optimizing model performance. Specifically, we present methods for predictive correlation analysis of Mamba's hidden states from both internal and external perspectives, along with corresponding definitions of correlation scores, aimed at understanding the workings of Mamba in visual recognition tasks and identifying flaws therein. Additionally, tailored repair methods are proposed for identified external and internal state flaws to eliminate them and optimize model performance. Extensive experiments validate the efficacy of the proposed methods on prevalent Mamba architectures, significantly enhancing Mamba's performance. For more information, please visit https://vision-mamba-mender.github.io/.

1 Introduction

Deep learning has demonstrated outstanding performance in various fields of artificial intelligence, with deep neural networks such as convolutional neural networks (CNNs) [1, 2] and Vision Transformer [3, 4] dominating the field of computer vision. However, the limited receptive field [5, 6] of CNNs and the quadratic computational complexity [7, 8] of Transformers constrain their further development. To overcome these limitations, an increasing number of studies have attempted to propose more advanced models [9–12]. Recently, Mamba [13], based on the state space model [11], has become the focus of these research efforts. Mamba can maintain nearly linear computational complexity while achieving a global receptive field, leading to outstanding performance. Consequently, it has been widely adopted in the field of computer vision [14–16].

[*] Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To enhance the performance of Mamba in visual tasks, existing works mainly focus on improving the architecture of Mamba [17–21]. For instance, Zhu et al.[17] proposed Vision Mamba, which adds branches to the original Mamba to simultaneously process image sequences in both forward and backward directions. Nearly simultaneously, Liu et al.[18] introduced VMamba, which adopts a four-way scanning strategy on top of the original Mamba to achieve a more comprehensive global receptive field. Other improvements include PlainMamba [19], Mamba-ND [20], SiMBA [21], and MambaMixer [22], all of which are variants of the original Mamba framework.

However, the aforementioned methods that optimize the Mamba model by improving its architecture are predefined and require strong prior knowledge and extensive trial and error. Additionally, Yu et al. [23] recently pointed out in their latest research that the current improvements made to Mamba-based visual models are unnecessary for visual tasks, especially for visual recognition tasks. This opposing view underscores the necessity and urgency of further optimizing Mamba models in the field of computer vision. Hence, unlike the aforementioned pre-optimization methods, this paper attempts to analyze the working mechanism of Mamba from a post-perspective, identify the reasons of flaws leading to incorrect prediction results, and automatically rectify them to further enhance the performance of Mamba models. Moreover, this approach is applicable to all Mamba-like models.

Based on this idea, in this paper, we propose Vision Mamba Mender, a systematic approach to understanding the working mechanism of Mamba from a post-perspective, identifying flaws within it, and rectifying them to ultimately improve model performance. In understanding the operational mechanism of the Mamba model, we categorize the computational process of Mamba into external state interaction and internal state interaction.

Along these two perspectives of external and internal states, we introduce a state correlation analysis method tailored for Mamba to establish the correlation between hidden states and predicted results. Additionally, we define external state correlation scores and internal state correlation scores to quantitatively analyze differences in state correlations, revealing flaws existing respectively in the external and internal states. Specifically, external state flaws refer to instances where correct model predictions in certain states are predominantly associated with foreground regions, while incorrect predictions are primarily linked to background regions. Internal state flaws, on the other hand, pertain to cases where correct predictions within a class are correlated with the same regions within the state and exhibit low overall complexity, whereas incorrect predictions within the class focus on different internal regions of the state and demonstrate higher overall complexity.

Furthermore, we propose repair methods tailored for addressing both external state flaws and internal state flaws. Specifically, in the repair of external state flaws, we impose constraints on the external state correlations within certain modules of Mamba to increase their correlation with foreground information during prediction. On the other hand, in the repair of internal state flaws, we impose constraints on the internal state correlations within certain modules of Mamba to enhance their correlation with genuine class-specific internal information during prediction. Through extensive experimentation, we demonstrate that the proposed Vision Mamba Mender is applicable to state-of-the-art Vision Mamba architectures.

The contributions of this paper are summarized as follows:

- We propose a novel post-hoc optimization method named Vision Mamba Mender. This method is applicable to existing state-of-the-art Vision Mamba architectures, identifying and repairing flaws in the Mamba model's mechanisms for visual recognition tasks from a post-hoc perspective, ultimately enhancing the model's performance.
- We introduce a state correlation analysis method and correlation score definitions for Mamba from both external and internal hidden state perspectives. These methods are used to identify flaws. Additionally, we introduce a state correlation constraint method to rectify these flaws.
- Extensive experiments demonstrate that the proposed Vision Mamba Mender can effectively identify and repair flaws in the Mamba model without introducing additional parameters, significantly improving the performance of the Vision Mamba.

2 Preliminaries

Mamba [13] is a novel and efficient sequence modeling model composed of multiple uniformly stacked Mamba blocks. Each Mamba block is constructed based on a selective state space model

(SSM). Unlike the previous time-invariant SSM [11], it allows the parameters of the SSM to depend on the input, enhancing the model's expressive capacity. Furthermore, inspired by H3 [24] and Gated MLP [25], each Mamba block also incorporates modules such as causal convolution and gating mechanisms.

As illustrated in Figure 1, given the hidden state $h_n^{(\ell)-1}$ of the i-th token in the input of the ℓ-th Mamba block, the computation of the hidden state $h_n^{(\ell)}$ of the i-th token in the output of the ℓ-th Mamba block is as follows:

$$\begin{aligned} x_n^{(\ell)} &= \mathrm{SiLU}(h_n^{(\ell)-1} \cdot W_x^{(\ell)}), && (1) \\ c_1^{(\ell)}, c_2^{(\ell)}, \ldots, c_n^{(\ell)} &= \text{cacusal-Conv1D}(x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_n^{(\ell)}), && (2) \\ s_n^{(\ell)} &= \text{selective-SSM}(c_1^{(\ell)}, c_2^{(\ell)}, \ldots, c_n^{(\ell)}), && (3) \\ z_n^{(\ell)} &= \mathrm{SiLU}(h_n^{(\ell)-1} \cdot W_z^{(\ell)}), && (4) \\ y_n^{(\ell)} &= (s_n^{(\ell)} \odot z_n^{(\ell)}) \cdot W_y^{(\ell)}, && (5) \end{aligned}$$

where SiLU$(.)$, causal-Conv1D$(.)$, and selective-SSM$(.)$ denote the activation function, the casual 1D convolution, and the selective state model, respectively. $W_x^{(\ell)}$, $W_z^{(\ell)}$, and $W_y^{(\ell)}$ are the projection matrices for the linear operations in the Mamba block, while x, z, c, s, and y represent the intermediate states within the Mamba block. Finally, by applying a residual connection, the hidden state $h_n^{(\ell)}$ is obtained as follows:

$$h_n^{(\ell)} = h_n^{(\ell)-1} + y_n^{(\ell)}. \qquad (6)$$

In Eqn. (5), $\odot$ denotes the Hadamard product. The term $z_n^{(\ell)}$ in Eqn. (4) is computed through a separate pathway, serving as a gating mechanism to regulate the information flow in the main pathway.

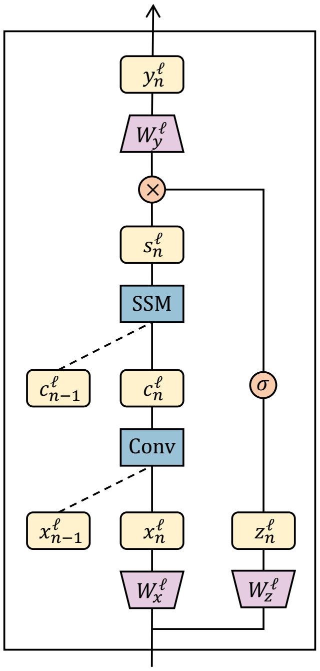

Figure 1: The computational process of a Mamba block.

However, the original Mamba block is designed for one-dimensional sequences and is not suitable for handling multidimensional visual data, particularly for vision tasks requiring spatial awareness. Existing vision Mamba architectures enhance the basic Mamba block to accommodate these requirements, such as ViM [17], VMamba [18], PMamba [19], Mamba-ND [20], and SiMBA [21]. Unlike the predefined optimizations, our approach analyzes the working mechanism of the Mamba model post-hoc, identifying flaws and making repairs to further enhance the model's performance.

3 Where Do Flaws Occur?

Identifying flaws in Mamba first requires an understanding of how Mamba operates. Some studies have provided empirical evidence to elucidate the mechanisms of Mamba models in the NLP domain, such as their contextual learning ability [26], factual recall capability [27], and interpretability [28]. However, elucidating the operational mechanisms of Mamba models in the visual domain remains a significant challenge. Ali et al. [29] established a connection between the selective SSM within the Mamba block and the self-attention mechanism in Transformers, allowing the selective SSM to represent the interaction process between any two states by constructing a self-attention matrix, which is utilized for image feature attribution. However, focusing solely on the selective SSM within the Mamba block is far from sufficient, as causal convolution, as shown in Eqn. (2), also participates in the interaction between states. Moreover, other computational modules within the Mamba block used for state interaction must also be considered.

In this section, we first investigate the working mechanisms of Mamba. We summarize the computational processes within Mamba as state interactions[1]. These interactions are categorized into *external state interactions* (where a state interacts with other states to form a new state, as shown in Eqn. (2) and (3)) and *internal state interactions* (where a state interacts only with its internal information to form a new state, as shown in Eqn. (1), (4), and (5)). We explore the operational mechanisms of Mamba from both the external and internal state interaction perspectives to identify flaws.

[1] In Mamba models used for visual recognition, each state corresponds to an image patch or a learnable classification token.

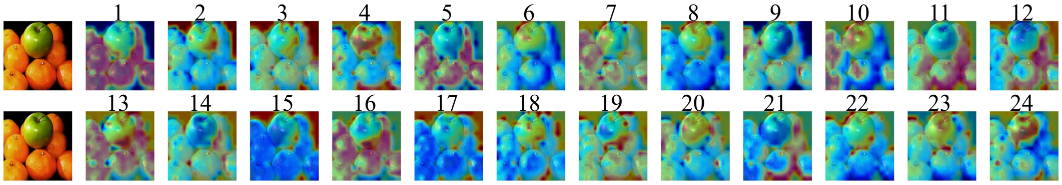

Figure 2: Visualization of the external state correlation $\mathbf{e}^{(\ell,s)}$ of the output $s_n^{(\ell)}$ from the selective-SSM module in different ViM [17] blocks. The depth of the ViM blocks increases from left to right.

3.1 External State Correlation Analysis

To understand the impact of external states on predictions within Mamba's working mechanisms, we propose the Grad-ESC method, inspired by Grad-CAM [30], which uses gradient and activation information to assess the importance of each neuron in decision-making. Grad-ESC calculates the correlation between external states and the model's prediction outcomes, termed external state correlation. Unlike the attention mechanism derived by Ali et al. [29], which applies only to the selective SSM module within the Mamba block, Grad-ESC allows for correlation analysis of the outputs from any module within the Mamba block.

Specifically, given the predictive distribution p output by the Mamba model and the true class k of the input sample, the calculation process for external state correlation $\mathbf{e}^{(\ell,s)} \in \mathbb{R}^{H \times W}$ (where H and W represent the height and width of the input image) of the states $\{s_n^{(\ell)}\}_{n=1}^{N}$ output by the selective SSM module (where $s_n^{(\ell)} \in \mathbb{R}^D$, N and D are the number and dimensionality of the states) is as follows:

$$\mathbf{e}^{(\ell,s)} = \mathcal{R}(\overline{s}_1^{(\ell)}, \overline{s}_2^{(\ell)}, \dots, \overline{s}_N^{(\ell)}),\ \overline{s}_n^{(\ell)} = \mathbb{E}_D(g^{(\ell,s)} \odot s_n^{(\ell)}),\ g^{(\ell,s)} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial p^k}{\partial s_n^{(\ell)}}, \tag{7}$$

where $g^{(\ell,s)} \in \mathbb{R}^D$ represents the weights used to weight the state $s_n^{(\ell)}$ along its dimensions based on gradient information, $\mathbb{E}_D(\cdot)$ denotes the average across all dimensions, $\hat{s}_n^{(\ell)} \in \mathbb{R}$ signifies the degree of correlation between the n-th state and the prediction after integrating gradient information, and $\mathcal{R}$ denotes the operation of reshaping and scaling to the original image size after retaining only the states corresponding to the image patches.

As illustrated in Figure 2, taking the state $s_n^{(\ell)}$ as an example, we visualize the external state correlations of $s_n^{(\ell)}$ in different blocks of ViM [17]. It can be observed that the external state correlations of $s_n^{(\ell)}$ vary across different blocks; some blocks exhibit correlations with foreground regions (such as the 1st layer), while others correlate with background regions (such as the 2nd layer). We consider correlations with the foreground to be interpretable, whereas correlations with the background are deemed uninterpretable. To quantify the interpretability of external state correlations, we propose the following definition of an external state correlation score based on commonly used interpretable evaluation methods such as perturbation test [29, 31] and segmentation test [32, 30, 33]:

Definition 1 (External Correlation Score). *Given a pre-trained Mamba model $\mathcal{F}(\cdot)$, an input image i, foreground annotations m of the input image, and external state correlation $\mathbf{e}^{(\ell,s)}$ computed through the proposed method, the external correlation score is defined as follows:*

$$\text{ECS}(\mathbf{e}^{(\ell,s)}) = \underbrace{\frac{\text{softmax}(\mathcal{F}(\mathbf{e}^{(\ell,s)+} \odot i))}{\text{softmax}(\mathcal{F}(\mathbf{e}^{(\ell,s)-} \odot i))}}_{\text{perturbation test}} \times \underbrace{\text{IoU}(\mathbf{e}^{(\ell,s)+}, m)}_{\text{segmentation test}}, \tag{8}$$

where $\mathbf{e}^{(\ell,s)+} = \mathbb{1}(\mathbf{e}^{(\ell,s)} \geq \alpha)$ denotes the retention of important regions (those greater than the threshold α) in $\mathbf{e}^{(\ell,s)}$, setting them to 1, and the removal of unimportant regions (those less than the threshold α), setting them to 0, for use in negative perturbation tests. Conversely, $\mathbf{e}^{(\ell,s)-} = \mathbb{1}(\mathbf{e}^{(\ell,s)} < \alpha)$ denotes the removal of important regions and the retention of unimportant regions, for use in positive perturbation tests. Additionally, $\text{ECS}(\mathbf{e}^{(\ell,s)}) \in [0, +\infty)$, where a higher value indicates a higher correlation score.

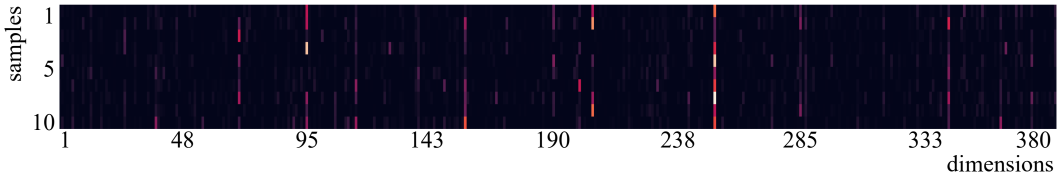

Figure 3: Visualization of the internal state correlations $\mathbf{i}_n^{(\ell,x)}$ of the output states $s_n^{(\ell)}$ from the linear mapping module $W_s^{(\ell)}$ in ViM [17] for samples of the same class. The horizontal axis represents the state dimensions, and the vertical axis represents the samples.

3.2 Internal State Correlation Analysis

To comprehend the influence of internal states on prediction outcomes within the operational mechanism of Mamba, we introduce the Grad-ISC method. This method is utilized for computing the degree of correlation between the internal states and model predictions, termed as the state-internal correlation.

Specifically, given the predicted distribution p outputted by the Mamba model and the true class k of the input sample, let's take the example of the n-th state $x_n^{(\ell)} \in \mathbb{R}^D$, which is the output of the linear mapping matrix $W_x^{(\ell)}$. The computation process for the corresponding internal state corelation $\mathbf{i}_n^{(\ell,x)} \in \mathbb{R}^D$ is as follows:

$$\mathbf{i}_n^{(\ell,x)} = g_n^{(\ell,x)} \odot x_n^{(\ell)},\, g_n^{(\ell,x)} = \frac{\partial p^k}{\partial x_n^{(\ell)}}, \tag{9}$$

where $g_n^{(\ell,x)} \in \mathbb{R}^D$ denotes the weights applied to the dimensions of state $x_n^{(\ell)}$ using gradient information. Similarly, the Grad-ISC method can perform internal state correlation analysis on the output of any module within the Mamba block.

Continuing with the example of state $x_n^{(\ell)}$, we visualize the internal state correlations $\mathbf{i}_n^{(\ell,x)}$ for samples belonging to the same class in ViM [17], as shown in Figure 3. It can be observed that for the same class, the regions of internal state correlation are relatively consistent. We posit that the more consistent and simpler the internal state correlation regions are for samples of the same class, the more interpretable they are. Conversely, the more inconsistent and complex the internal state correlation regions are for samples of the same class, the more difficult they are to interpret. To quantify the interpretability of internal state correlations, we propose a novel definition for the Internal Correlation Score:

Definition 2 (Internal Correlation Score). *Given J samples belonging to the same class and the internal state correlation $\mathbf{i}_n^{(\ell,x)} \in \mathbb{R}^D$ computed using the proposed method for a particular sample, the Internal Correlation Score is defined as follows:*

$$\text{ICS}(\mathbf{i}_n^{(\ell,x)}) = \underbrace{\mathbb{E}_D(\frac{1}{j}\sum_{j=1}^{J}\mathbf{i}_{n,j}^{(\ell,x)+})}_{\text{simplicity}} \times \underbrace{\mathbb{E}_D(\frac{\frac{1}{J}\sum_{j=1}^{J}\mathbf{i}_{n,j}^{(\ell,x)+}}{\mathbf{i}_{n,1}^{(\ell,x)+} \oplus \mathbf{i}_{n,2}^{(\ell,x)+} \oplus \cdots \oplus \mathbf{i}_{n,J}^{(\ell,x)+}})}_{\text{homogeneity}}, \tag{10}$$

where $\mathbf{i}_{n,j}^{(\ell,x)}$ represents the internal state correlation for the j-th sample, $\mathbf{i}_{n,j}^{(\ell,x)+} = \mathbb{1}(\mathbf{i}_{n,j}^{(\ell,x)} \geq \beta)$ denotes the binarized internal correlation (set to 1 if greater than β, otherwise set to 0), $\oplus$ denotes the XOR operation, and $\mathbb{E}_D(\cdot)$ denotes the average over all dimensions. Furthermore, $\text{ICS}(\mathbf{i}^{(\ell,x)}) \in [0, +\infty)$, where a higher value indicates a higher correlation score.

3.3 Identifying Flaw through Correlation Analysis

To analyze flaws within the Mamba model, we examined the external and internal state correlation scores under different conditions using the ViM model and the ImageNet-10 dataset. This analysis allowed us to observe variations in correlation relationships across different states.

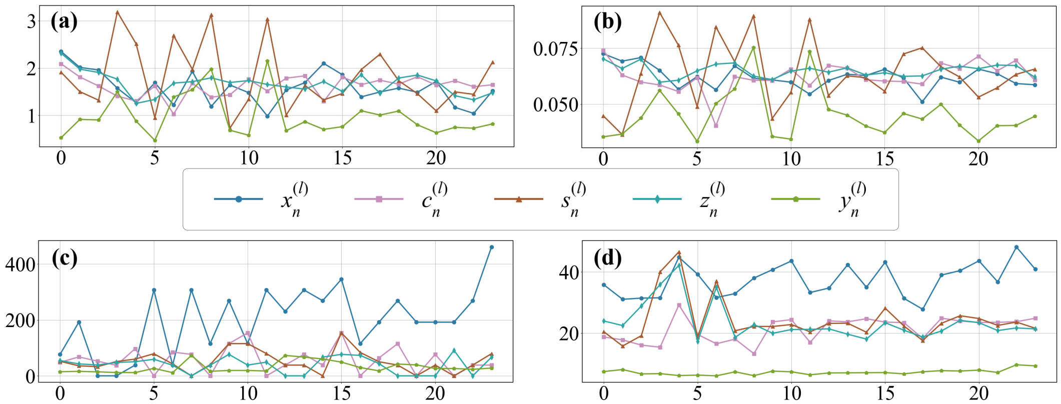

Figure 4: Comparison of external and internal state correlation scores across different blocks in the Mamba model between simple and difficult samples. **(a)** and **(b)** show the external state correlation scores for simple and difficult samples, respectively. **(c)** and **(d)** present the internal state correlation scores for simple and difficult samples, respectively.

Flaws in External State Correlation. To uncover flaws in the external correlations of states, we compare the scores of external state correlations in the Mamba model between simple and difficult samples, as illustrated in Fig. 4(a) and (b). For simple samples, it can be observed that across different blocks, the external correlation scores of states $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, and $z_n^{(\ell)}$ are all better than those of state $y_n^{(\ell)}$, especially in deeper blocks. Furthermore, by comparing simple samples (a) with difficult samples (b), it can be noted that the external correlation scores of all states in all blocks have decreased. This indicates that for difficult samples, the Mamba model tends to associate certain incomprehensible regions in the external states.

Flaws in Internal State Correlation. To unveil patterns in internal state correlations and detect flaws within them, we also conducted a comparative analysis of the internal correlation scores within the Mamba model, as depicted in Figure 4(c) and (d). For both simple and difficult samples, the internal correlation score of state $x_n^{(\ell)}$ is generally better than that of other states. Furthermore, simultaneous comparison of simple and difficult samples reveals a decrease in the internal correlation scores of all states, including state $x_n^{(\ell)}$. This suggests that for difficult samples, the Mamba model tends to correlate with some incomprehensible regions within the internal states.

4 How to Repair Flaws?

In principle, repairing flaws within the model's internals can enhance Mamba's decision-making process and improve its performance. However, there has been limited research specifically addressing such issues in Mamba models, particularly when applied in the domain of visual processing. Therefore, in this section, we investigate post-hoc flaw repair in the Mamba model from two perspectives: external state correlation and internal state correlation.

4.1 External State Correlation Repair

To repair flaws related to external state correlations, it is essential to identify the key components within the Mamba block that need fixing. In flaw identification regarding external state correlations, it is observed that the states $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, and $z_n^{(\ell)}$ exhibit flaws when predicting difficult samples. For these states, the external correlation scores for simple samples are higher than those for difficult samples. This implies that the model primarily correlates with foreground regions in correct predictions and with background regions in incorrect predictions. Furthermore, considering that external state interactions occur only within Conv and SSM modules (as indicated by Eqn.(2) and Eqn.(3)), we empirically suggest that the Conv and SSM components are crucial for influencing the model's anomalous decisions. Theoretically, it is also feasible to apply external flaw repair to other states.

We focus on repairing the external correlation flaws of the states $c_n^{(\ell)}$ and $s_n^{(\ell)}$ output by the Conv and SSM modules in the deeper blocks. Specifically, we first identify difficult samples from the training set and then constrain the external correlations of the hidden states $c_n^{(\ell)}$ and $s_n^{(\ell)}$ using foreground annotations m:

$$\text{Loss}_{\mathbf{e}} = \mathbb{E}_{HW}(\mathbf{e}^{(\ell,c)} \odot m) + \mathbb{E}_{HW}(\mathbf{e}^{(\ell,s)} \odot m), \tag{11}$$

where $\mathbb{E}_{HW}$ denotes the averaging operation over the two-dimensional matrix. It is important to note that during backpropagation, each term in the computational graph of $\text{Loss}_{\mathbf{e}}$ is differentiable, which involves second-order gradients as specified in Eqn. (7). The complete loss function, combining with the original task loss, is formulated as follows:

$$\text{Loss} = \text{Loss}_{\mathbf{ce}} + \lambda \text{Loss}_{\mathbf{e}}, \tag{12}$$

where $\text{Loss}_{\mathbf{ce}}$ denotes the cross-entropy loss for the image recognition task, and λ serves to balance the magnitudes of the respective loss components.

4.2 Internal State Correlation Repair

Similarly, when identifying flaws related to internal state correlations, it is observed that states in different blocks exhibit flaws when predicting difficult samples. For instance, the internal correlation scores of state $x_n^{(\ell)}$ for simple samples are higher than those for difficult samples. This suggests that in predictions biased toward correctness within a class, the model aligns with internally consistent regions of the state, characterized by lower overall complexity. Conversely, in predictions biased toward incorrectness within a class, the model tends to focus on internally inconsistent regions of the state with higher overall complexity. In our experiments, we consider the linear mapping $W_s^{(\ell)}$ within deeper blocks as a critical component influencing the model's predictions. However, theoretically, it is also feasible to apply internal flaw repair to other states.

We focus on repairing the internal correlation flaws of the states $x_n^{(\ell)}$ output by the linear mapping $W_x^{(\ell)}$ within deeper blocks. Specifically, we first select J simple samples for each class from the training set and create corresponding internal correlation templates $\hat{\mathbf{i}}_n^{(\ell,x)+} = 1 - (\frac{1}{J}\sum_{j=1}^{J} \mathbf{i}_{n,j}^{(\ell,x)+})$ for each class. We then utilize these templates to constrain the internal correlations of $x_n^{(\ell)}$:

$$\text{Loss}_{\mathbf{i}} = \mathbb{E}_D(\mathbf{i}_n^{(\ell,x)} \odot \hat{\mathbf{i}}_n^{(\ell,s)+}). \tag{13}$$

Similarly, during backpropagation, each term in the computational graph of $\text{Loss}_{\mathbf{i}}$ is also differentiable, which involves second-order gradients as specified in Eqn. (9). Combined with the loss for the original task, the complete loss is as follows:

$$\text{Loss} = \text{Loss}_{\mathbf{ce}} + \gamma \text{Loss}_{\mathbf{i}}, \tag{14}$$

where γ is a balancing factor to adjust the scale between loss functions.

It is important to note that external and internal state corrections represent two distinct perspectives, allowing these two methods to be used orthogonally. To maximize flaw repair, we recommend sequentially repairing external flaws (Eqn. (11)) followed by internal flaws (Eqn. (13)).

4.3 Results of Flaw Repair

We evaluated the proposed flaw repair method on five state-of-the-art Mamba models for visual tasks, including ViM-T [17], VMamba-T [18], SiMBA-S [21], EfficientVMamba-T (EMamba-T) [34], and LocalVim-T [35]. To ensure smooth operation of the Mamba models within limited computational resources, we modified certain model parameters, such as the number of blocks and the dimensionality of hidden states. Additionally, we conducted experiments on three scales of the ImageNet [36] dataset: ImageNet-50, ImageNet-300, and ImageNet-1K. To obtain the image foreground annotations m as defined in **Definition 1** from the ImageNet dataset, we utilized annotations from the ImageNet-S dataset [37]. These annotations correspond to those used during training, covering all 50 classes in ImageNet-50, 300 classes in ImageNet-300, and 919 classes in ImageNet-1K, with an average of 10 foreground-labeled samples per class. Further details on experimental settings are provided in Appendix B.

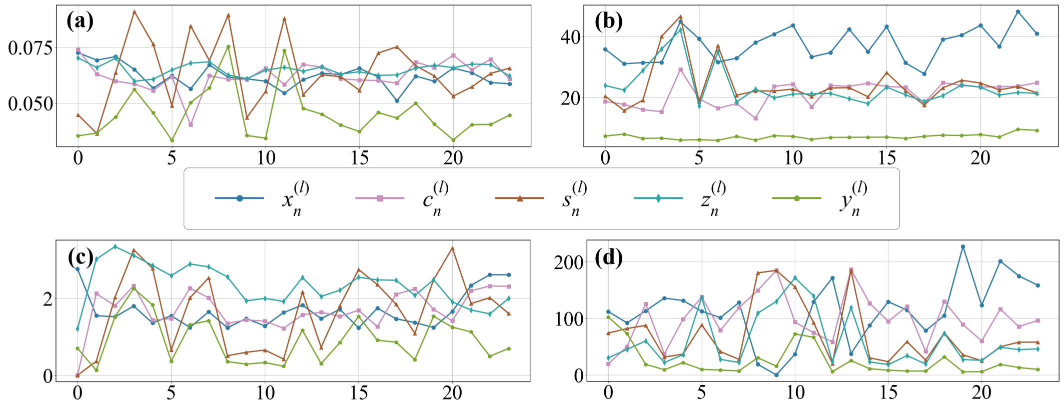

Figure 5: Comparison of external and internal state correlation scores across different blocks in the Mamba model before and after flaw repair. **(a)** and **(b)** show the external and internal state correlation scores before flaw repair, respectively. **(c)** and **(d)** present the external and internal state correlation scores after flaw repair, respectively.

Table 1: Comparison of model accuracy after external state flaw repair for different states. Taking '$x_n^{(\ell)}$' as an example, it represents the external flaw repair of state $x_n^{(\ell)}$. The results in the second row of the table correspond to external flaw repair, while the third row presents the outcomes of internal flaw repair. The experiment was conducted within the last block of the ViM model, using the ImageNet-50 dataset.

Base	$x_n^{(\ell)}$	$c_n^{(\ell)}$	$s_n^{(\ell)}$	$z_n^{(\ell)}$	$y_n^{(\ell)}$	$x_n^{(\ell)}$+$c_n^{(\ell)}$	$x_n^{(\ell)}$+$s_n^{(\ell)}$	$x_n^{(\ell)}$+$z_n^{(\ell)}$	$c_n^{(\ell)}$+$s_n^{(\ell)}$
76.44	78.40	78.44	78.28	77.96	78.28	78.56	78.04	78.04	78.48
76.44	78.20	79.16	78.88	78.32	78.32	77.72	78.76	78.44	78.42

Comparing State Correlation Scores Before and After Repair. To validate the effectiveness of the flaw repair methods, we compared the internal and external state correlation scores of difficult samples before and after flaw repair, as shown in Figure 5. It can be observed that the proposed flaw repair methods effectively enhance the correlation scores of the Mamba model, thereby improving the predicted correlation regions within both internal and external states. Specifically, for the ViM model, after external state flaw repair, the external correlation scores of the states $c_n^{(\ell)}$ are significantly higher than before repair. Similarly, after internal state flaw repair, the internal correlation scores of the states $x_n^{(\ell)}$ are also notably higher than before repair.

Repairing External and Internal Flaws in Different States. As shown in Table 1 (second row), we conducted experiments on external flaw repair for different states within the same block. The results indicate that simultaneously performing external flaw repair on states $c_n^{(\ell)}$ and $s_n^{(\ell)}$ leads to an improvement in model accuracy. This aligns with our findings in the main text regarding flaw detection, where states $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, and $z_n^{(\ell)}$ exhibit flaws when predicting challenging samples. Furthermore, since the Conv and SSM modules in each block facilitate external interactions, applying external flaw repair to states $c_n^{(\ell)}$ and $s_n^{(\ell)}$ is effective.

Similarly, as illustrated in Table 1 (third row), we performed experiments on internal flaw repair for different states within the same block. The results demonstrate that individually addressing internal flaws in state $x_n^{(\ell)}$ also results in improved model accuracy. This supports our earlier findings regarding flaw detection, where state $x_n^{(\ell)}$ shows flaws when predicting challenging samples, thus validating the effectiveness of internal flaw repair for state $x_n^{(\ell)}$.

Repairing SOTA Vision Mamba Models. As shown in the Table 2, we validated the proposed flaw repair methods on various mainstream Vision Mamba models. It can be observed that regardless of whether external or internal state flaw repair is performed, the accuracy of the repaired models exceeds

Table 2: Comparison of the accuracy of the SOTA Vision Mamba model after flaw repair. Base' denotes the original model, +Ext', +Int', and +All' represent models after external flaw repair, internal flaw repair, and simultaneous external and internal flaw repair, respectively.[2]

	ViM-T	**VMamba-T**	**SiMBA-S**	**EMamba-T**	**LocalVim-T**
ImageNet-50					
Base	76.44	79.96	81.32	81.44	75.08
+Ext	78.48+2.04	81.08+1.12	84.48+3.16	82.68+1.24	78.68+3.60
+Int	78.20+1.76	82.36+2.40	84.64+3.32	83.24+1.80	78.20+3.12
+All	79.68+3.24	82.28+2.32	86.52+5.20	83.32+1.88	80.56+5.48
ImageNet-300					
Base	75.11	75.04	68.08	74.67	70.39
+Ext	77.87+2.76	76.01+0.97	69.73+1.65	75.03+0.36	73.09+2.70
+Int	77.76+2.65	76.31+1.27	69.93+1.85	75.55+0.88	72.71+2.32
+All	79.53+4.42	76.75+1.71	70.58+2.50	75.66+0.99	74.84+4.45
ImageNet-1K					
Base	71.64	67.54	51.06	67.35	57.70
+Ext	73.02+1.38	68.34+0.80	52.12+1.06	67.62+0.27	59.30+1.60
+Int	72.79+1.15	68.67+1.13	52.23+1.17	67.89+0.54	59.90+2.20
+All	73.30+1.66	68.68+1.14	52.24+1.18	67.84+0.49	60.83+3.13

that of the original models. Specifically, for external state flaw repair, on ImageNet-50, the accuracy of VMamba increased by 1.12% after flaw repair, and the accuracy of SiMBA-S increased by 3.16%. For internal state flaw repair, on ImageNet-1K, the accuracy of ViM increased by 1.15% after flaw repair, and the accuracy of LocalViM increased by 2.20%. It is worth noting that external state flaw repair and internal state flaw repair can be orthogonal. For example, on ImageNet-50, simultaneously performing external state flaw repair and internal state flaw repair on ViM increased the accuracy by 2.04% and 1.76%, respectively, compared to performing each repair method individually. This resulted in an overall improvement of 3.24% compared to the original model.

5 Related Works

Model Optimization. Optimizing model performance through various prior designs or theoretical derivations has always been a pursuit in the field of artificial intelligence. Existing methods for optimizing the Mamba model, especially when applied to visual tasks, mainly involve adjusting the network architecture [17–22]. For instance, improvements have been made through the bidirectional scanning mechanism [17], cross-scanning mechanism [18], continuous 2D scanning mechanism [19], and incorporating new channel modeling techniques [21]. However, these enhancements are pre-designed and require extensive trial and error. For a more detailed understanding of Mamba's applications in computer vision, we recommend readers refer to recent surveys by Zhang et al. [14] and Xu et al. [16]. So far, there has been no research focusing on post-hoc optimization to correct internal flaws within the Mamba model to improve its performance. Optimizing models beyond Mamba, such as Convolutional Neural Networks [1, 38] and Transformers [3, 39], is also primarily achieved through architectural improvements [2, 40–42, 4], feature enhancement [43–46], or some post-hoc debugging methods [47–50]. However, these approaches either cannot be directly applied to a brand-new Mamba model or do not involve post-hoc optimization. In summary, unlike the aforementioned studies, the proposed Vision Mamba Mender is the first framework dedicated to post-hoc analysis and optimization of the Mamba model. Its aim is to rectify flaws within Mamba models in the domain of vision, thereby further enhancing the performance of Mamba models.

[2] We adjusted the experimental setup and refined methodological details to enhance the fairness and reproducibility of the comparisons. Detailed experimental settings can be found in Appendix B.

Model Explanation. Enhancing model transparency and trustworthiness through research on model explainability has been a major focus in the field of artificial intelligence. In the context of Mamba model explainability, most studies have primarily concentrated on the model's context learning capabilities [26], factual recall abilities [27], and comparisons of the explainability between Mamba and previous models [28]. These studies do not address the identification of flaws in the Mamba model's mechanisms and are focused on natural language processing (NLP), leaving a gap in their application to visual tasks. In the realm of computer vision, Ali et al.[29] established a connection between the selective SSM in the Mamba block and the self-attention mechanism in Transformers. They developed a feature attribution method for Mamba based on Attention-Rollout[51] and Transformer-Attribution [31], which can be used for image feature attribution. However, this method is limited to the selective SSM module and does not consider or apply to other modules within the Mamba model. In research on model explainability beyond the Mamba model, a plethora of methods have emerged, including activation-based methods [52–56], gradient-based methods [30, 57, 33], LRP-based methods [58, 59, 31], and perturbation-based methods [60–62]. However, most of these explainability methods are designed for specific network architectures and may not be directly applicable to Mamba. Inspired by Grad-CAM [30], this paper introduces Grad-ESC and Grad-ISC methods tailored for the Mamba's external and internal states, respectively. These methods effectively establish the correlation between the model's states and predictions, providing simple yet powerful analytical tools for identifying flaws in the Mamba model's mechanisms.

6 Discussion

As one of the most prominent models today, Mamba has found widespread application in the field of computer vision. While attitudes towards its utilization in visual tasks may vary, it does not impede the exploration of its novel internal operational mechanisms. On the contrary, it urges further research into optimizing the performance of Mamba models applied in visual tasks. Moreover, at this juncture, considering the potential changes in future architectures, devising a set of architecture-agnostic post-optimization methods becomes crucial.

However, in this work, we only analyzed the overall operational mechanisms of each module within Mamba, without delving into the internal details of each module. For instance, we hypothesize that there might be flaws within the SSM module itself during prediction, and further optimization of its internal details may potentially enhance the model's performance. Additionally, similar post-optimization paradigms should be further explored in non-visual tasks within Mamba or even applied in non-Mamba architectures.

7 Conclusion

Mamba incorporates intricate computational modules, making it non-trivial to delve into its operational mechanisms. In this paper, we have proposed a post-analysis and post-optimization approach to understand the workings of Mamba in visual recognition tasks and enhance its performance. Specifically, departing from existing pre-defined optimization methods, we have introduced predictive correlation analysis and correlation scoring definitions from both internal and external perspectives of Mamba's states. These methodologies aim to identify flaws within Mamba's operational mechanisms. Simultaneously, we have presented corresponding approaches for repairing internal and external flaws to optimize model performance. Extensive and comprehensive experiments have demonstrated the effectiveness of Vision Mamba Mender on mainstream Mamba architectures.

Acknowledgments and Disclosure of Funding

This research was supported by the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHZSD24F020001 and Ningbo Natural Science Foundation (2022J182).

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[5] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[8] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[9] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[10] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[11] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[12] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

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

[14] Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, and Zi Ye. A survey on visual mamba. *arXiv preprint arXiv:2404.15956*, 2024.

[15] Xiao Liu, Chenxu Zhang, and Lei Zhang. Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404*, 2024.

[16] Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*, 2024.

[17] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[18] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[19] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024.

[20] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. *arXiv preprint arXiv:2402.05892*, 2024.

[21] Badri N Patro and Vijay S Agneeswaran. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024.

[22] Ali Behrouz, Michele Santacatterina, and Ramin Zabih. Mambamixer: Efficient selective state space models with dual token and channel selection. *arXiv preprint arXiv:2403.19888*, 2024.

[23] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? *arXiv preprint arXiv:2405.07992*, 2024.

[24] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

[25] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *International conference on machine learning*, pages 9099–9117. PMLR, 2022.

[26] Riccardo Grazzi, Julien Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. Is mamba capable of in-context learning? *arXiv preprint arXiv:2402.03170*, 2024.

[27] Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in mamba. *arXiv preprint arXiv:2404.03646*, 2024.

[28] Gonçalo Paulo, Thomas Marshall, and Nora Belrose. Does transformer interpretability transfer to rnns? *arXiv preprint arXiv:2404.05971*, 2024.

[29] Ameen Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. *arXiv preprint arXiv:2403.01590*, 2024.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[31] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[32] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[34] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024.

[35] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[37] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. 2022.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[41] Shunyu Liu, Yihe Zhou, Jie Song, Tongya Zheng, Kaixuan Chen, Tongtian Zhu, Zunlei Feng, and Mingli Song. Contrastive identity-aware learning for multi-agent value decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11595–11603, 2023.

[42] Lin Chen, Zhijie Jia, Lechao Cheng, Yang Gao, Jie Lei, Yijun Bei, and Zunlei Feng. Vit-calibrator: Decision stream calibration for vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1147–1155, 2024.

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[44] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[45] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[46] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[47] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. Featureinsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 105–112. IEEE, 2015.

[48] Gabriel Cadamuro, Ran Gilad-Bachrach, and Xiaojin Zhu. Debugging machine learning models. In *ICML Workshop on Reliable Machine Learning in the Wild*, volume 103, 2016.

[49] Zunlei Feng, Jiacong Hu, Sai Wu, Xiaotian Yu, Jie Song, and Mingli Song. Model doctor: A simple gradient aggregation strategy for diagnosing and treating cnn classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 616–624, 2022.

[50] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5686–5697, 2016.

[51] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[53] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[54] Shunyu Liu, Jie Song, Yihe Zhou, Na Yu, Kaixuan Chen, Zunlei Feng, and Mingli Song. Interaction pattern disentangling for multi-agent reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[55] Shunyu Liu, Wei Luo, Yanzhen Zhou, Kaixuan Chen, Quan Zhang, Huating Xu, Qinglai Guo, and Mingli Song. Transmission interface power flow adjustment: A deep reinforcement learning approach based on multi-task attribution map. *IEEE Transactions on Power Systems*, 39(2):3324–3335, 2024.

[56] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022.

[57] Misha Denil, Alban Demiraj, and Nando De Freitas. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*, 2014.

[58] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[59] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[60] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[61] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[62] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[63] Xiao Wang, Shiao Wang, Yuhe Ding, Yuehang Li, Wentao Wu, Yao Rong, Weizhe Kong, Ju Huang, Shihao Li, Haoxiang Yang, et al. State space model for new-generation network alternative to transformers: A survey. *arXiv preprint arXiv:2404.09516*, 2024.

[64] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *arXiv preprint arXiv:2404.16112*, 2024.

[65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Vision Mamba Mender. The magnifying glass held in one hand symbolizes flaw detection, while the wrench held in the other hand symbolizes flaw repair.

Appendix for Vision Mamba Mender

To facilitate a better understanding of the value and significance of this work, as well as to thoroughly demonstrate the effectiveness and applicability of the proposed method, we have provided the algorithm code in supplementary material. This code will be made publicly available. Additionally, in the appendix, we have supplemented more information about the Mamba architecture, related work on Mamba interpretability, detailed experimental settings, additional results on state flaw identification, additional results on state flaw repair, and detailed ablation experiments, as follows.

A Origin of the Mamba Model

One of the core components of Mamba is the State Space Model (SSM). SSMs were initially inspired by continuous systems, which map a one-dimensional input signal $x(t) \in \mathbb{R}$ to a one-dimensional output signal $y(t) \in \mathbb{R}$ through an N-dimensional latent state $h(t) \in \mathbb{R}^N$. The general computational process can be defined as follows:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{15}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{B} \in \mathbb{R}^{N \times 1}$, and $\mathbf{C} \in \mathbb{R}^{1 \times N}$ are the state matrix, input matrix, and output matrix, respectively.

Unlike the continuity in Equation 15, Structured State Space Models (S4) [11] employ transformation methods, such as the zero-order hold (ZOH), to discretize it.

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}}h(t) + \overline{\mathbf{B}}x(t), \\ y(t) &= \overline{\mathbf{C}}h(t), \end{aligned} \tag{16}$$

where $\overline{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}$ are the discretized parameters, and Δ denotes the step size.

Building on this foundation, the Selective State Space Model (S6) creatively introduces selective scanning to overcome the limitations imposed by time-invariant parameters on context representation learning:

$$\mathbf{B} = \mathcal{S}_B(\mathbf{x}), \mathbf{C} = \mathcal{S}_C(\mathbf{x}), \Delta = \tau_\Delta(\Delta + S_\Delta(\mathbf{x})), \tag{17}$$

where S_B, S_C, and S_Δ are different linear mappings. Additionally, for more detailed information on Mamba's specifics and applications, we recommend readers consult several recent and well-regarded review articles [63, 64, 14–16].

B Detailed Experimental Settings

To facilitate the reproducibility of our results, we provide detailed experimental settings as follows. Throughout the entire experiment, we utilized 8 NVIDIA A40 GPU cards and a CPU with 24 cores and 500GB of memory.

Model Parameter Settings. To enable the Mamba model to train efficiently within limited computational resources, we adjusted certain parameters across the Mamba models. For instance, in the case of VMamba-T, we set the patch size to 16x16. For SiMBA-S, the model depth was adjusted to [2, 3, 3, 2], and we introduced a class token for classification in the last block. For EfficientVMamba-T, we similarly set the patch size to 16x16. In the case of LocalViM-T, we reduced the model depth from 20 to 9 and set the state dimensionality to 128. These parameter adjustments were made to validate the proposed methods under constrained resources. While this may somewhat reduce the baseline performance, it does not affect the fairness of comparing the Mamba model's performance before and after flaw identification and repair.

Model Training Settings. To ensure a fair comparison with limited resources, our training settings primarily followed the experimental setup of DeiT [43]. Specifically, we employed data augmentation techniques such as random cropping and random horizontal flipping. When training on 224x224 input images, we optimized the model using AdamW [65] with a momentum of 0.9, a total batch size of 128, and a weight decay of 0.1. We utilized a cosine learning rate schedule with an initial learning rate of 5e-4, training the Mamba model for 300 epochs. In particular during the baseline model training, these training strategies resulted in an exceedingly smooth training curve in the later stages, effectively optimizing the fit. During testing, we performed center cropping on the validation set to extract 224x224 images.

State Flaw Identification. For external state correlation analysis, the threshold α is set to 0.5 by default. For internal state correlation, the threshold β is set to 0.3 by default. The impact of thresholds on flaw detection can be found in the Appendices C and D. It is worth noting that for multi-branch Mamba, when conducting external flaw identification and repair, we merge all branches corresponding to the same state together. However, for internal flaw identification and repair, we consider each branch of the state separately.

State Flaw Repair. In our experiments, the balance weight λ for the loss function of external state flaw repair is set to $1e+7$ by default, and the balance weight γ for the loss function of internal state flaw repair is also set to $1e+7$ by default. Furthermore, based on the conclusions about flaw identification in the main text, external state flaw repair is applied by default to the first Mamba block, while internal state flaw repair is applied to the last Mamba block. The impact of repairing flaws in different blocks on model performance can be found in Appendices E and F.

C Impact of Threshold α on External State Correlation Scores

As shown in Figure 6, we compared the impact of different values of threshold α on external state correlation scores. By contrasting simple samples with difficult samples, we found that the value of α does not significantly affect the relative magnitudes of the external correlation scores. Specifically, regardless of whether α is set to 0.0, 0.2, or 0.8, the external correlation scores for states $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, $z_n^{(\ell)}$, and $y_n^{(\ell)}$ in difficult samples all decrease compared to those in simple samples. Additionally, when α is set between 0.4 and 0.6, more details can be observed; in different blocks, the external correlation scores of states $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, and $z_n^{(\ell)}$ are better than those of state $y_n^{(\ell)}$. Therefore, we recommend that the value of α should not be too extreme when calculating external state correlation scores. It can be set between 0.4 and 0.6; we use a default setting of $\alpha = 0.5$.

D Impact of Threshold β on Internal State Correlation Scores

We also examined the impact of different values of threshold β on internal state correlation scores, as shown in Fig. 7. Similarly, by comparing simple and difficult samples, it can be observed that the value of β does not affect the relative magnitudes of the internal correlation scores. Specifically,

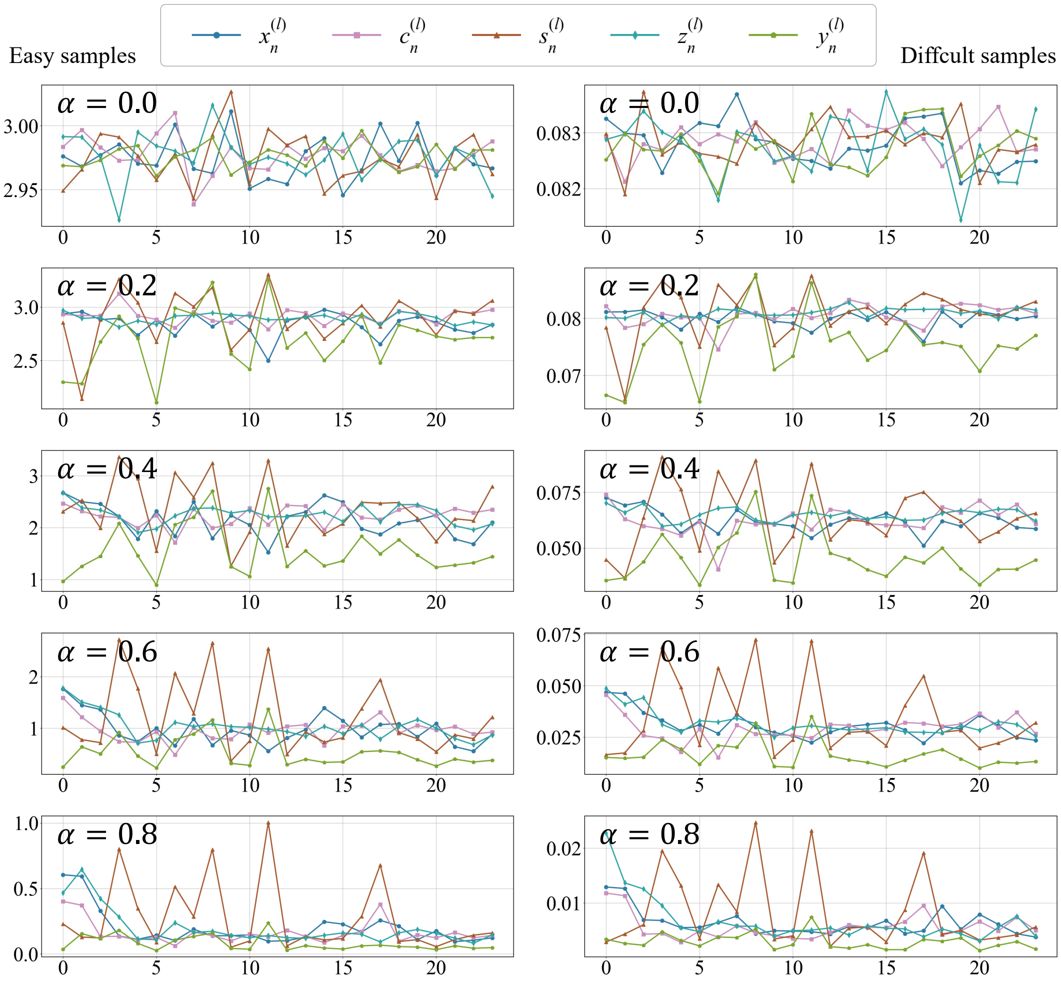

Figure 6: Comparison of external state correlation scores for simple and difficult samples under different threshold values of α. The left column of images shows the external state correlation scores for simple samples at various α values. The right column of images shows the external state correlation scores for difficult samples at various α values.

regardless of whether β is set to 0.0, 0.2, or 0.8, the internal correlation scores for states $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, $z_n^{(\ell)}$, and $y_n^{(\ell)}$ in difficult samples all decrease compared to those in simple samples. Additionally, when β is set between 0.4 and 0.6, more details can be observed; in different blocks, the internal correlation scores of state $x_n^{(\ell)}$ are better than those of states $c_n^{(\ell)}$, $s_n^{(\ell)}$, $z_n^{(\ell)}$, and $y_n^{(\ell)}$. Therefore, we recommend that the value of β should not be too extreme when calculating internal state correlation scores. It can be set between 0.4 and 0.6; we use a default setting of $\beta = 0.5$.

E External Flaw Repair in Different Blocks

As shown in Figure E(a), we conducted a study on repairing external flaws in different blocks, revealing a clear enhancement in model performance following external flaw repair. Specifically, when external flaws are repaired in the final block, the model's accuracy improves by 2.04% compared to the baseline. In contrast, while repairing external flaws in other blocks, such as the tenth block, also yields an accuracy increase, the improvement is only 1.72%. This discrepancy may be attributed to the presence of external flaws in the states across nearly every block, as illustrated in Figure 2, including states such as $x_n^{(\ell)}$, $c_n^{(\ell)}$, $s_n^{(\ell)}$, and $z_n^{(\ell)}$, which exhibit significant external flaws on difficult samples. However, due to structures like residual connections, the influence of external flaws on model predictions varies across different blocks.

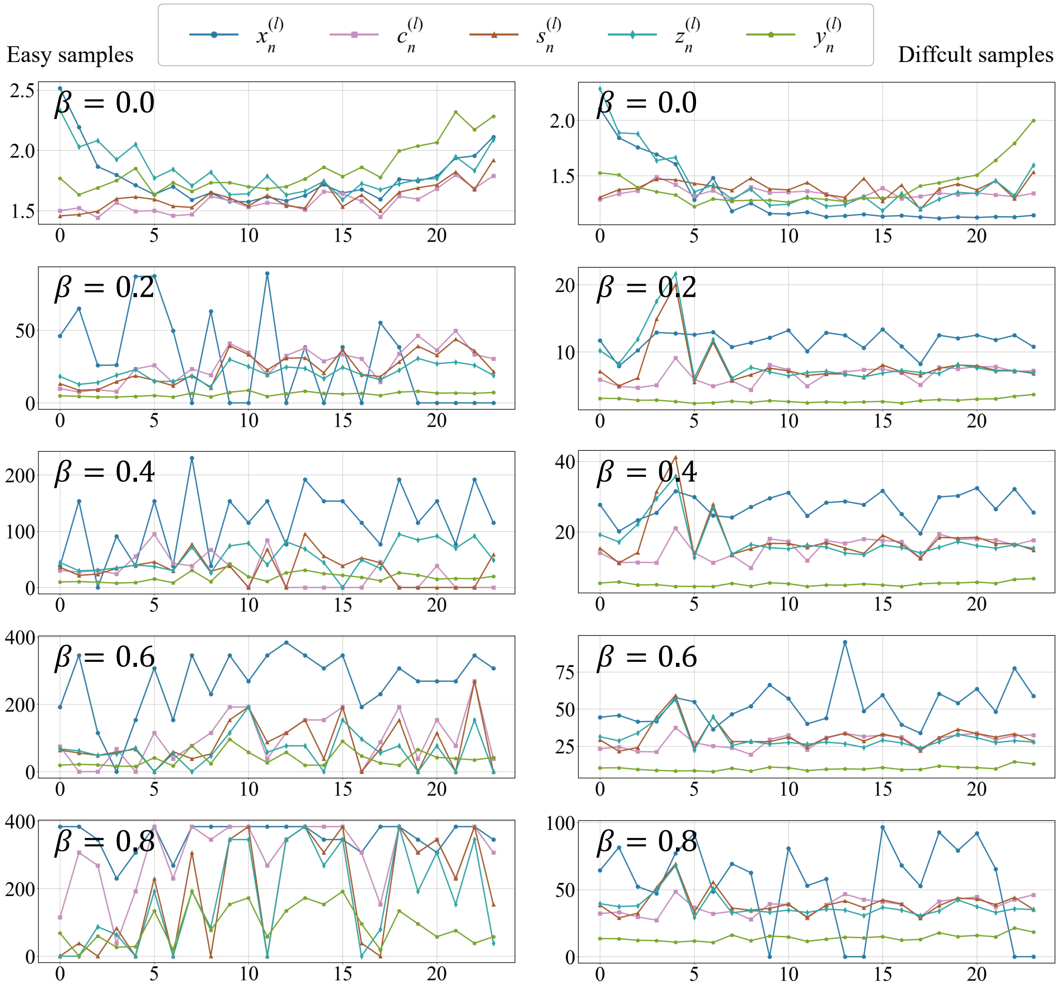

Figure 7: Comparison of internal state correlation scores for simple and difficult samples under different threshold values of β. The left column of images shows the internal state correlation scores for simple samples at various β values. The right column of images shows the internal state correlation scores for difficult samples at various β values.

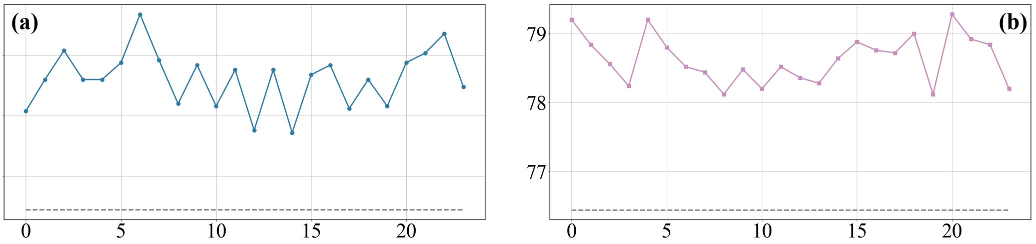

Figure 8: Comparison of model accuracy after state flaw repair in different blocks. (a) shows the results of external state flaw repair, conducted on states $c_n^{(\ell)}$ and $s_n^{(\ell)}$ in each block of the ViM model, using the ImageNet-50 dataset. (b) shows the results of internal state flaw repair, conducted on state $x_n^{(\ell)}$ in each block of the ViM model, also using the ImageNet-50 dataset.

F Internal Flaw Repair in Different Blocks

As shown in Figure E(b), we conducted a study on repairing internal flaws in different blocks, which also demonstrates a noticeable improvement in model performance. Specifically, when internal flaws are repaired in the final block, the model's accuracy increases by 1.76% compared to the baseline. In contrast, repairing internal flaws in other blocks, such as the second-to-last block, results in an accuracy improvement of 2.4%. Furthermore, repairing internal correlations in other blocks yields varying levels of accuracy enhancement. This variation may similarly stem from the model's internal structures, such as residual connections, which lead to different degrees of influence from internal flaws on model performance across the various blocks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The algorithm code is available in the supplementary material and will be publicly accessible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: To prevent the occurrence of accidental experimental results, we conducted each experiment at least twice to ensure the accuracy of the findings. The data presented in this paper are the averages of two or more experimental runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to the citations in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have provided the complete code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.