MambaLRP: Explaining Selective State Space Sequence Models

Farnoush Rezaei Jafari[1,2*]   **Grégoire Montavon**[3,2,1]   **Klaus-Robert Müller**[1,2,4,5,6]

Oliver Eberle[1,2*]

[1]Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany
[2]BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[3]Department of Mathematics and Computer Science, Freie Universität Berlin,
Arnimallee 14, 14195 Berlin, Germany
[4]Department of Artificial Intelligence, Korea University, Seoul 136-713, South Korea
[5]Max Planck Institute for Informatics, Stuhlsatzenhausweg 4, 66123 Saarbrücken, Germany
[6]Google DeepMind, Berlin, Germany

Abstract

Recent sequence modeling approaches using selective state space sequence models, referred to as Mamba models, have seen a surge of interest. These models allow efficient processing of long sequences in linear time and are rapidly being adopted in a wide range of applications such as language modeling, demonstrating promising performance. To foster their reliable use in real-world scenarios, it is crucial to augment their transparency. Our work bridges this critical gap by bringing explainability, particularly Layer-wise Relevance Propagation (LRP), to the Mamba architecture. Guided by the axiom of relevance conservation, we identify specific components in the Mamba architecture, which cause unfaithful explanations. To remedy this issue, we propose MambaLRP, a novel algorithm within the LRP framework, which ensures a more stable and reliable relevance propagation through these components. Our proposed method is theoretically sound and excels in achieving state-of-the-art explanation performance across a diverse range of models and datasets. Moreover, MambaLRP facilitates a deeper inspection of Mamba architectures, uncovering various biases and evaluating their significance. It also enables the analysis of previous speculations regarding the long-range capabilities of Mamba models.

1 Introduction

Sequence modeling has demonstrated its effectiveness and versatility across a wide variety of tasks and data types, including text, time series, genomics, audio, and computer vision [24, 82, 32, 9, 26]. Recently, there has been a surge of interest in a new class of sequence modeling architectures, known as structured state space sequence models (SSMs) [35, 66, 33]. This is due to their ability to process sequences in linear time, as opposed to quadratic time required by the more established Transformer architectures [69]. The recent Mamba architecture, a prominent and widely adopted instance of state space models, has demonstrated competitive predictive performance on a variety of sequence modeling tasks across domains and applications [33, 43, 83, 78, 70], while scaling linearly with sequence length.

---

*Correspondence to: rezaeijafari@campus.tu-berlin.de, oliver.eberle@tu-berlin.de

38th Conference on Neural Information Processing Systems (NeurIPS 2024).
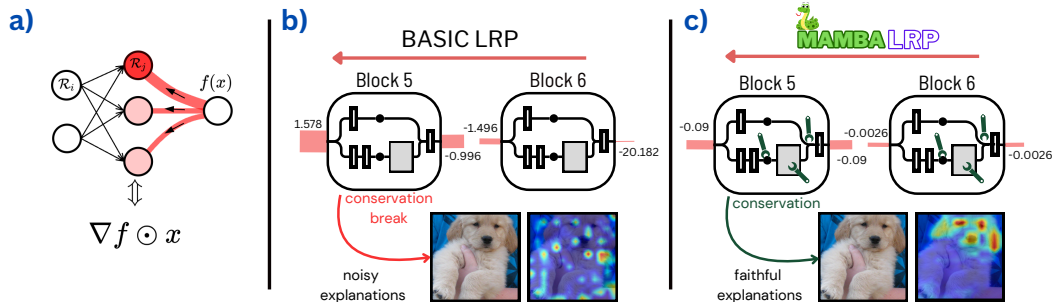
Figure 1: Conceptual steps involved in the design of MambaLRP. (a) Take as a starting point a basic LRP procedure, equivalent to Gradient × Input. (b) Analyze layers in which the conservation property is violated. (c) Rework the relevance propagation strategy at those layers to achieve conservation. The resulting MambaLRP method enables efficient and faithful explanations.

As Mamba models, and more generally SSMs, are rapidly being adopted into real-world applications, ensuring their transparency is crucial. This enables inspection beyond test set accuracy and uncovering various forms of biases, including 'Clever-Hans' effects [39]. It is particularly important in high-risk domains such as medicine, where the prediction behavior must be robust under real-world conditions and aligned with human understanding. The field of Explainable AI [48, 36, 8, 58] focuses on developing faithful model explanations that attribute predictions to relevant features and has shown success in explaining many highly nonlinear models such as convolutional networks [20], or attention-based Transformer models [3, 2].

Explaining the predictions of Mamba models is however challenging due to their highly non-linear and recurrent structure. A recent study [4] suggests viewing these models as attention-based models, enabling the use of attention-based explanation methods [1, 18]. Yet, the explanations produced by attention-based techniques are often unreliable and exposed to potential misalignment between input features and attention scores [75, 38]. As an alternative, Layer-wise Relevance Propagation (LRP) [10] decomposes the model function with the goal of explicitly identifying the relevance of input features by applying purposely designed propagation rules at each layer. A distinguishing feature of LRP is its adherence to a conservation axiom, which prevents the artificial amplification or suppression of feature relevance in the backward pass. LRP has been demonstrated to produce faithful explanations across various domains (e.g. [7, 62, 3, 20]). Nevertheless, the peculiarities of the Mamba architecture are not addressed by the existing LRP procedures, which may lead to the violation of the conservation property and result in unreliable explanations.

In this work, we present MambaLRP, a novel approach to integrate LRP into the Mamba architecture. By examining the relevance propagation process across Mamba layers through the lens of conservation, we pinpoint layers within the Mamba architecture that need to be addressed specifically. We propose a novel relevance propagation strategy for these layers, grounded in the conservation axiom, that is theoretically sound, straightforward to implement and computationally efficient. Through a number of quantitative evaluations, we show that the proposed MambaLRP approach allows to robustly deliver the desired high explanatory performance, exceeding by far the performance of various baseline explanation methods as well as a naive transposition of LRP to the Mamba architecture. We further demonstrate the usefulness of MambaLRP in several areas: gaining concrete insights into the model's prediction mechanism, uncovering undesired decision strategies in image classification, identifying gender bias in language models, and analyzing the long-range capabilities of Mamba. Our code is publicly available.[1]

2   Related Work

Structured State Space Sequence Models (SSMs).   Transformers [69] have emerged as the most widely used architectures for sequence modeling. However, their computational limitations, particularly with large sequence lengths, have restricted their applicability in modeling long sequences. Addressing these computational limitations, recent works [34, 35] have introduced structured state

---

[1]https://github.com/FarnoushRJ/MambaLRP

2

space sequence models (SSMs) as an alternative approach. SSMs are a class of sequence modeling methods, leveraging the strengths of recurrent, convolutional, and continuous-time methods, demonstrating promising performance across various domains, including language [30, 46], image [77, 13, 50], and video [71] processing, and beyond [59, 22, 42]. A recent advancement by Gu and Dao [33] introduced selective SSM, an enhanced data-dependent SSM with a selection mechanism that adjusts its parameters based on the input. Built on this dynamic selection, the Mamba architecture fuses the SSM components with multilayer perceptron (MLP) blocks. This fusion simplifies the architecture while improving its ability to handle various sequence modeling tasks, including applications in language processing [6, 53, 72], computer vision [41, 83, 78], medical imaging [44, 76, 31, 56, 40, 74, 73], and graphs [70, 14]. This fast adoption of SSMs and Mamba models underscores the need for reliable explanations of their predictions.

Explainable AI and SSMs. In efforts to explain Mamba models, [51] analyzed if the interpretability tools originally designed for Transformers can also be effectively applied to architectures such as Mamba. In this context, Ali et al. [4] and Zimerman et al. [84] recently proposed viewing the internal computations of Mamba models as an attention mechanism. This approach builds upon previous works that use attention signal as explanation, including Attention Rollout [1] and variants thereof [18, 17]. While these approaches can provide some insight, they inherit the limitations of using attention as an explanation [75, 38], including their inability to capture potential misalignment between tokens and attention scores, and the limited performance in empirical faithfulness evaluations. Alternative Explainable AI methods, not yet applied to Mamba models but in principle applicable to any model, include techniques using input perturbations [81, 85, 29] or leveraging gradient information [11, 64, 68, 65, 63]. Despite their wide applicability, these models have certain drawbacks, such as requiring multiple function evaluations for a single explanation or being susceptible to gradient noise, resulting in subpar performance, as our benchmark experiment will demonstrate. Alternatively, deriving tailored approaches that reflect the underlying model structure, has shown to be a promising direction in developing better attribution methods based on gradient analysis of the prediction function [7, 27, 62, 3]. In the Layer-wise Relevance Propagation framework, this necessitates suitable propagation rules, which are currently lacking for the Mamba architecture. To tackle these challenges, we introduce MambaLRP as an efficient solution for the computation of reliable and faithful explanations that are theoretically grounded in the axiom of relevance conservation.

3 Background

Before delving into the details of our proposed method, we begin with a brief overview of the selective SSM architecture, followed by an introduction to the LRP framework.

Selective SSMs (S6) An important component within the Mamba [33] architecture is the selective SSM. It is characterized by parameters, $\bar{A}$, $\bar{B}$, and C, and transforms a given input sequence $(x_t)_{t=1}^T$ into an output sequence of the same size $(y_t)_{t=1}^T$ via the following equations:

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t \tag{1}$$
$$y_t = C_t h_t \tag{2}$$

where the initial state $h_0 = 0$. What distinguishes the selective SSM from the original SSM (S4) [35] is that the evolution parameter, $\bar{A}_t$, and projection parameters, $\bar{B}_t$ and C_t, are functions of the input x_t. This enables dynamic adaptation of the SSM's parameters based on input. This dynamicity facilitates focusing on relevant information while ignoring irrelevant details when processing a sequence.

Layer-wise Relevance Propagation Layer-wise Relevance Propagation (LRP) [10] is an Explainable AI method that attributes the model's output to the input features through a single backward pass. This backward pass is specifically designed to identify neurons relevant to the prediction. LRP assigns relevance scores to neurons in a given layer and then propagates these scores to neurons in the preceding layer. The process continues layer by layer, starting from the network's output and terminating once the input features are reached. The LRP backward pass relies on an axiom called 'conservation' requiring that relevance scores are preserved across layers, avoiding to artificially amplify or suppress contributions. For example, let x and y be the input and output of some layer, respectively, and let $\mathcal{R}(x)$ and $\mathcal{R}(y)$ represent the sum of relevance scores in the respective layers. The conservation axiom requires that $\mathcal{R}(x) = \mathcal{R}(y)$ holds true.

3

4 LRP for Mamba

In this work, we bring explainability, particularly LRP, to Mamba models, following the conceptual design steps, shown in Fig. 1. We start by applying a basic LRP procedure, specifically one corresponding to Gradient × Input (GI), to the Mamba architecture. This serves as an effective initial step for identifying layers where certain desirable explanation properties, like relevance conservation, are violated. We analyze different layers of the Mamba architecture, derive relevance propagation equations and test the fulfillment of the conservation property. Our analysis reveals three components in the Mamba architecture where conservation breaks: the SiLU activation function, the selective SSM, and the multiplicative gating of the SSM's output. Leveraging the analysis above, we propose novel relevance propagation strategies for these three components, which lead to a robust, faithful and computationally efficient explanation approach, called MambaLRP.

4.1 Relevance propagation in SiLU layers

We start by examining the relevance propagation through Mamba's SiLU activation functions. This function is represented by the equation $y = x \cdot \sigma(x)$, where σ denotes the logistic sigmoid function.

Proposition 4.1 *Applying the standard gradient propagation equations yields the following result, which relates the relevance values before and after the activation layer:*

$$\underbrace{\frac{\partial f}{\partial x} x}_{\mathcal{R}(x)} = \underbrace{\frac{\partial f}{\partial y} y}_{\mathcal{R}(y)} + \underbrace{\frac{\partial f}{\partial y} \cdot \sigma'(x) \cdot x^2}_{\varepsilon} \tag{3}$$

The derivation for Eq. 3 can be found in Appendix A.1. We observe that the conservation property, *i.e.* $\mathcal{R}(x) = \mathcal{R}(y)$, is violated whenever the residual term ε is non-zero. We propose to restore the conservation property in the relevance propagation pass by locally expanding the SiLU activation function as:

$$y = x \cdot [\sigma(x)]_{\text{cst.}} \tag{4}$$

where $[\cdot]_{\text{cst.}}$ treats the given quantity as constant. This can be implemented e.g. in PyTorch using the `.detach()` function. Repeating the derivation above with this modification yields the desired conservation property, $\mathcal{R}(x) = \mathcal{R}(y)$. The explicit LRP rule associated to this LRP procedure is provided in Appendix B.

4.2 Relevance propagation in selective SSMs (S6)

The most crucial non-linear component of the Mamba architecture is its selective SSM component. It is designed to selectively retain or discard information throughout the sequence by adjusting its parameters based on the input, enabling dynamic adaptation to each token. To facilitate the analysis, we introduce an inconsequential modification to the original SSM by connecting C_t to h_t instead of x_t. To do so, we can redefine $\bar{A}_t$, $\bar{B}_t$, and C_t matrices as blockdiag($\bar{A}_t$, 0), ($\bar{B}_t$, I), and ($C_t \mid 0$) respectively, such that x_t becomes part of the state h_t without altering the overall functionality of the SSM.
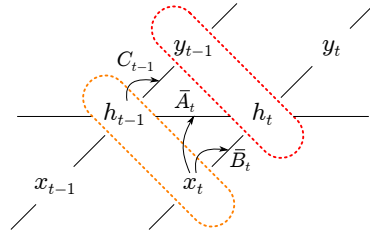
Figure 2: Unfolded view of SSM, highlighting two subsets of nodes, the relevance of which should be conserved throughout relevance propagation.

The unfolded SSM, with the aforementioned modification, is illustrated in Fig. 2. The complex relevance propagation procedure in the SSM component can be further simplified by considering two groups of units, illustrated in red and orange in Fig. 2. In these two groups, there are no connections within units of the same group, all the relevance propagation signals from the first group are directed towards the second group, and the second group receives no further incoming relevance propagation signal. With these properties, these two groups should, according to the principle of conservation, receive the same relevance scores.

4

Proposition 4.2 *Defining $\theta_t = (\bar{A}_t, \bar{B}_t, C_{t-1})$, and working out the propagation equations between these two groups yields the following relation:*

$$\underbrace{\frac{\partial f}{\partial x_t}x_t + \frac{\partial f}{\partial h_{t-1}}h_{t-1}}_{\mathcal{R}(x_t)+\mathcal{R}(h_{t-1})} = \underbrace{\frac{\partial f}{\partial h_t}h_t + \frac{\partial f}{\partial y_{t-1}}y_{t-1}}_{\mathcal{R}(h_t)+\mathcal{R}(y_{t-1})} + \underbrace{\frac{\partial f}{\partial \theta_t}\frac{\partial \theta_t}{\partial x_t}x_t + \frac{\partial f}{\partial \theta_t}\frac{\partial \theta_t}{\partial h_{t-1}}h_{t-1}}_{\varepsilon} \tag{5}$$

The derivation for Eq. 5 can be found in Appendix A.2. We note that the residual term ϵ, which is typically non-zero, violates conservation. Specifically, conservation fails due to the dependence of θ on the input. We propose to rewrite the state-space model at each step in a way that the parameters θ_t appear constant, *i.e.*:

$$h_t = [\bar{A}_t]_{\text{cst.}}h_{t-1} + [\bar{B}_t]_{\text{cst.}}x_t \tag{6}$$

$$y_t = [C_t]_{\text{cst.}}h_t \tag{7}$$

These equations can also be interpreted as viewing the selective SSM as a localized non-selective, *i.e.* standard, SSM. With this modification, conservation holds between the two groups, *i.e.* $\mathcal{R}(x_t) + \mathcal{R}(h_{t-1}) = \mathcal{R}(h_t) + \mathcal{R}(y_{t-1})$. By repeating the argument for each time step, conservation is also maintained between the input and output of the whole SSM component. Explicit LRP rules are provided in Appendix B.

4.3 Relevance propagation in multiplicative gates

In each block within the Mamba architecture, the SSM's output is multiplied by an input-dependent gate. In other words, $y = z_A \cdot z_B$, where $z_A = \text{SSM}(x)$ and $z_B = \text{SiLU}(\text{Linear}(x))$. Assume that the locally linear expansions introduced in Sections 4.1 and 4.2 are applied to the SSM components and SiLU activation functions, the mapping from x to y becomes quadratic.

Proposition 4.3 *Applying the standard gradient propagation equations establishes the following relation between the relevance values before and after the gating operation:*

$$\underbrace{\frac{\partial f}{\partial x}x}_{\mathcal{R}(x)} = \underbrace{\frac{\partial f}{\partial y}y}_{\mathcal{R}(y)} + \underbrace{\frac{\partial f}{\partial y}y}_{\varepsilon} \tag{8}$$

The derivation for Eq. 8 and explicit LRP rules can be found in Appendix A.3 and Appendix B, respectively. In this equation, we observe a spurious doubling of relevance in the backward pass. This can be addressed by treating half of the output as constant:

$$y = 0.5 \cdot (z_A \cdot z_B) + 0.5 \cdot [z_A \cdot z_B]_{\text{cst.}} \tag{9}$$

As for the previous examples, this ensures the conservation property $\mathcal{R}(x) = \mathcal{R}(y)$. An alternative would have been to make y linear by detaching only one of the terms in the product, as done for the SiLU activation or the SSM component. However, the strategy of Eq. 9 better maintains the directionality given by the gradient. We further compare these alternatives in an ablation study presented in Appendix C.5, demonstrating empirically that our proposed approach performs better.

4.4 Additional modifications and summary

The propagation strategies developed for the Mamba-specific components complement previously proposed approaches for other layers, including propagation through RMSNorm layers [3] and convolution layers via robust LRP-γ rules [49, 25] and their generalized variants. A summary of these additional enhancements is provided in Appendix C.2. Furthermore, our proposed propagation rules are generally applicable to other models that utilize similar components, such as multiplicative gates in recent architectures [54, 52, 45, 21].

A straightforward implementation of the propagation rules can be achieved by computing MambaLRP via Gradient $\times$ Input, where the gradient computations are modified to align with the proposed rules. The procedure consists of two main steps:

1. Perform the detach operations of Eqs. (4), (6), (7), and (9) (as well as similar operations for RMSNorm and convolutions).
2. Retrieve MambaLRP explanations by computing Gradient $\times$ Input on the detached model.

5

5 Experiments

To evaluate our proposed approach, we benchmark its effectiveness against various methods previously proposed in the literature for interpreting neural networks. We empirically evaluate our proposed methodology using Mamba-130M, Mamba-1.4B, and Mamba-2.8B language models [33], which are trained on diverse text datasets. The training details can be found in Appendix C.1. For the vision experiments, we use the Vim-S model [83]. Moreover, we perform several ablation studies to further investigate our proposed method.

Datasets In this study, we perform experiments on four text classification datasets, namely SST-2 [67], Medical BIOS [28], Emotion [60], and SNLI [16]. The SST-2 dataset encompasses around 70K English movie reviews, categorized into binary classes, representing positive and negative sentiments. The Medical BIOS dataset consists of short biographies (10K) with five specific medical occupations as targets. The SNLI corpus (version 1.0) comprises 570k English sentence pairs, with the labels entailment, contradiction, and neutral, used for the natural language inference (NLI) task. The Emotion dataset (20K) is a collection of English tweets, each labeled with one of six basic emotions. For the vision experiments, we use ImageNet dataset [23] with 1.3M images and 1K classes.

Baseline methods We compare our proposed method with several gradient-based, model-agnostic explanation techniques: Gradient $\times$ Input (GI) [11, 64], SmoothGrad [65], and Integrated Gradients [68]. Furthermore, we evaluate the performance of our proposed method against a naive implementation of LRP, *i.e.* LRP (LN-rule), where the LRP-0 rule is used in all linear and convolution layers, along with the LN-rule [3] in normalization layers.

We further compare the performance of our proposed method with two attention-based approaches, Attention Rollout (AttnRoll) and MambaAttr [4], which are recently proposed for Mamba models. Both methods are extensions of techniques originally developed for Transformer models: Attention Rollout [1] and Gradient$\times$Attention Rollout [18].

5.1 Conservation property

To verify the fulfillment of the conservation property, on which our method is based, we compare the network's output score with the sum of relevance scores attributed to the input features, for both the GI baseline and the proposed MambaLRP. The analysis is performed for Mamba-130M and Vim-S models trained on the SST-2 and ImageNet datasets, respectively. Full conservation is achieved if the output score equals the sum of relevance, as indicated by the blue line in Fig. 3. Our results show that conservation is severely violated by the GI baseline, and is addressed to a large extent by MambaLRP. Residual lack of conservation is due to the presence of biases in linear and convolution layers, which are typically non-attributable.
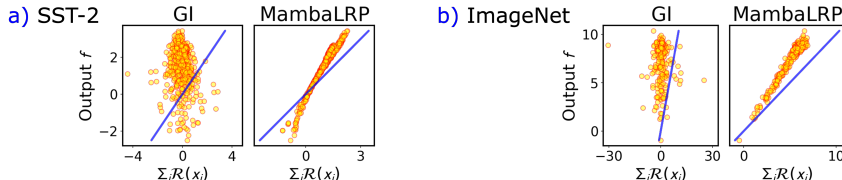
Figure 3: Conservation property. The x-axis represents the sum of relevance scores across the input features and the y-axis shows the network's output score. Each point corresponds to one example and its proximity to the blue identity line indicates the extent to which conservation is preserved, with closer alignment suggesting improved conservation.

5.2 Qualitative evaluation

In this section, we qualitatively examine the explanations produced by various explanation methods for Mamba-130M and Vim-S models. Fig. 4 illustrates the explanations generated to interpret the Mamba-130M model's prediction on a sentence from the SST-2 dataset with negative sentiment. We note that all of the explanation methods attribute positive scores to the word 'disgusting', which appears reasonable given the negative sentiment label. However, it is notable that the explanation

6

generated by MambaLRP is more sparse and focuses particularly on the terms 'so' and 'disgusting'. In contrast, the explanations produced by the gradient-based methods and AttnRoll appear to be quite noisy. Furthermore, we show the explanations produced to interpret the Vim-S model's predictions

MambaLRP	at least one scene is so disgusting that viewers may be hard pressed to retain their lunch .
GI	at least one scene is so disgusting that viewers may be hard pressed to retain their lunch .
SmoothGrad	at least one scene is so disgusting that viewers may be hard pressed to retain their lunch .
IG	at least one scene is so disgusting that viewers may be hard pressed to retain their lunch .
AttnRoll	at least one scene is so disgusting that viewers may be hard pressed to retain their lunch .
MambaAttr	at least one scene is so disgusting that viewers may be hard pressed to retain their lunch .

Figure 4: Explanations generated for a sentence of the SST-2 dataset. Shades of red represent words that positively influence the model's prediction. Conversely, shades of blue reflect negative contributions. The heatmaps of attention-based methods are constrained to non-negative values.

on images of the ImageNet dataset in Fig. 5. Purely gradient-based explanations tend to identify unspecific noisy features, while both attention-based approaches, AttnRoll and MambaAttr, are more effective at highlighting significant features. Among these methods, MambaLRP stands out for its ability to generate explanations that are particularly focused on key features used by the model to make a prediction. Take, for instance, the first image classified under the 'African elephant' category. We can see that the explanation generated by MambaLRP not only includes all occurrences of the 'African elephant' object but also highlights its distinctive features, such as the tusks. In the second image labeled 'wild boar', despite the presence of multiple objects in the image, MambaLRP's explanation remains focused on the 'wild boar' object, disregarding other objects. Moreover, in the third instance, MambaLRP uncovers a spurious correlation, the presence of a watermark in Chinese, influencing the model's prediction, a subtlety overlooked or not fully represented by other methods. Further qualitative results can be found in Appendix C.6.
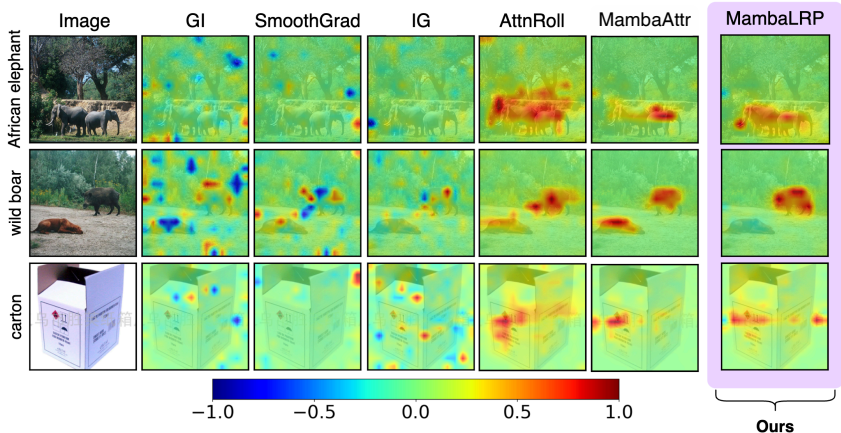
Figure 5: Explanations produced by different explanation methods for images of the ImageNet dataset. AttnRoll and MambaAttr are limited to non-negative heatmap values.

5.3 Quantitative evaluation

To quantitatively evaluate the faithfulness of explanation methods, we employ an input perturbation approach based on ranking input features by their importance [57], which can be done using either a Most Relevant First (MoRF) or Least Relevant First (LeRF) strategy. Ranked features are iteratively perturbed through a process known as *flipping*. We monitor the resulting changes in the output logit, f_c, for the predicted class c, and compute the area under the perturbation curve. The areas under the curves for LeRF and MoRF strategies are denoted by A_{LeRF}^F and A_{MoRF}^F, respectively. In contrast, the *insertion* method starts with a fully perturbed input and progressively restores important features. The areas under the curves for this method are indicated by A_{MoRF}^I and A_{LeRF}^I, for the MoRF and LeRF strategies, respectively. A reliable explanation method is characterized by low values of

7

A^F_{MoRF} or A^I_{LeRF}, and large values of A^F_{LeRF} or A^I_{MoRF}. In an effort to minimize the introduction of out-of-distribution manipulations, the recent study by Blücher et al. [15] advocates for harnessing both insights to derive a more resilient metric. Therefore, we follow the same strategy as [15, 2] to evaluate explanation methods. The evaluation metrics are defined as $\Delta A^F = A^F_{\text{LeRF}} - A^F_{\text{MoRF}}$ and $\Delta A^I = A^F_{\text{MoRF}} - A^F_{\text{LeRF}}$. For both metrics, a higher score is preferable, as it signifies a more accurate and reliable explanation method.

The outcomes of this analysis are represented in Table 1. MambaLRP consistently achieves highest faithfulness scores in comparison to other baseline methods. We observe that GI struggles with noisy attributions, leading to low faithfulness scores. However, methods like Integrated Gradients and MambaAttr have shown improvements in this regard. We note that LRP (LN-rule) outperforms most methods across the majority of the text classification tasks. Nevertheless, its performance is notably inferior compared to MambaLRP. Overall, we observe that MambaLRP significantly outperforms all other methods by a substantial margin. In both vision and NLP experiments, attention-based methods have shown superior performance compared to the purely gradient-based approaches.

Table 1: Evaluating explanation methods. Higher scores ΔA^F indicate more faithful explanations.

Methods	SST-2			Med-BIOS			SNLI			Emotion			ImageNet
	Mamba 130M	Mamba 1.4B	Mamba 2.8B	Mamba 130M	Mamba 1.4B	Mamba 2.8B	Mamba 130M	Mamba 1.4B	Mamba 2.8B	Mamba 130M	Mamba 1.4B	Mamba 2.8B	Vim-S
Random	-0.012	-0.106	0.007	0.044	-0.014	-0.037	0.010	0.002	0.000	-0.001	0.000	0.000	-0.001
GI [64]	0.078	-0.106	-0.043	0.200	-0.634	-1.434	-0.039	-0.039	0.083	-0.787	-0.409	-1.533	-0.018
SmoothGrad [65]	1.377	-0.383	-0.675	1.661	-2.300	-1.908	0.486	-0.687	-0.747	1.808	-1.852	-4.228	0.209
IG [68]	0.857	0.216	0.322	1.296	1.065	1.937	0.453	0.218	0.331	1.808	2.010	4.314	1.217
AttnRoll [4]	0.657	0.431	0.452	2.228	1.076	2.241	0.242	0.371	0.292	0.389	1.483	0.530	2.427
MambaAttr [4]	1.190	0.626	0.341	3.126	3.006	5.326	0.513	0.554	0.343	2.003	4.706	3.849	2.676
LRP (LN-rule, [3])	0.877	0.961	0.820	2.217	3.456	5.305	0.673	0.656	0.731	3.079	5.199	5.094	2.548
MambaLRP (ours)	**1.978**	**1.248**	**1.157**	**3.906**	**4.234**	**7.083**	**0.989**	**0.897**	**0.899**	**3.523**	**5.397**	**5.637**	**4.715**

Runtime comparison We report the runtimes of MambaLRP along with other methods used in this study in Appendix C.9. As shown in Table 10, our method's runtime is comparable to GI and can be implemented via a single forward and backward pass. Since approaches like Integrated Gradients require multiple function evaluations, their runtimes are considerably higher than MambaLRP.

Ablation study In Section 4, we proposed techniques for handling different non-linear components within the Mamba architecture. This ablation study aims to assess the significance of each technique by testing the effect of their exclusion on faithfulness. Table 2 shows that all three modifications are essential for achieving competitive explanation performance, with our proposed method for handling the SSM component being the most critical. Further experiments, comparing different strategies for handling the Mamba block's multiplicative gate, are detailed in Appendix C.5.

Table 2: Analyzing the impact of ablating the three proposed propagation rules on ΔA^F for the components in MambaLRP.

SiLU	SSM	Gate	SST-2	ImageNet
✓	✗	✓	0.577	0.144
✓	✓	✗	1.721	4.022
✗	✓	✓	1.943	4.618
✓	✓	✓	**1.978**	**4.715**

Table 3: Frequency of gendered words in explanations for 'Nurse' and 'Surgeon' classes of the Medical BIOS dataset across language models.

Models	Surgeon	Nurse
GPT2-base	0.14	0.24
T5-base	0.10	0.11
RoBERTa-base	0.01	0.06
Mamba-130M	0.009	0.058
Mamba-1.4B	0.001	0.042

6 Use cases

Uncovering gender bias in Mamba. Explanation methods serve as tools to uncover biases in pretrained vision and language models. Using our proposed method, we examine Mamba-130M and Mamba-1.4B models, trained on the Medical BIOS dataset, to investigate the potential presence of gender biases. Following the methodology in [28], we use MambaLRP to identify the top-5 tokens of highest importance and to quantify the prevalence of gendered words within these tokens. We find that the model exhibits a pronounced preference for female-gendered words in the 'Nurse' class

8
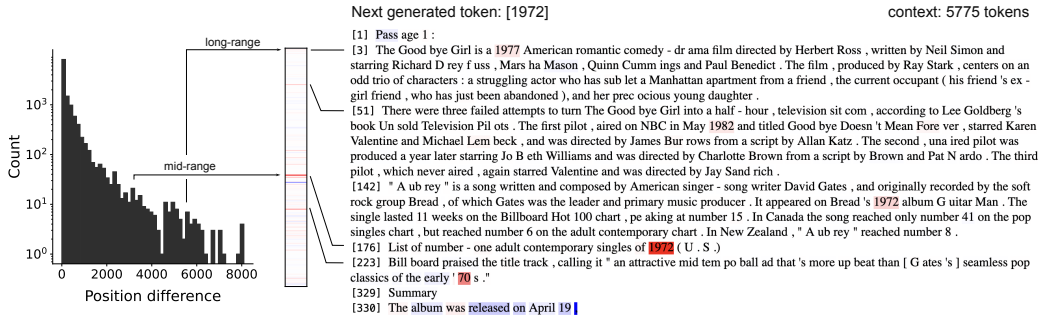
Figure 6: Analysis of the position of tokens relevant for next token generation. Left: Distribution of absolute position of the ten most relevant tokens for the prediction of the next word. Right: Long-range dependency between tokens of the input and the predicted next token (here: *1972*).
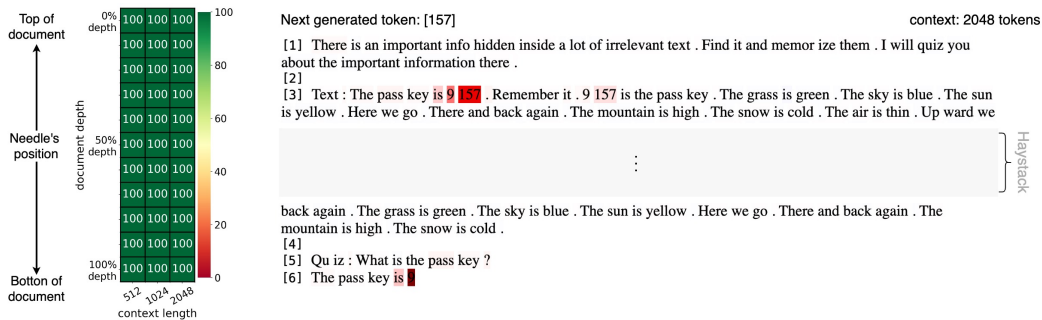
Figure 7: Explanation-based retrieval accuracy in the needle-in-a-haystack test verifying model reliance on relevant features for different context lengths.

(e.g. the proportion of gender-specific words is 0.058 for females, compared to 0.0 for males in Mamba-130M.). We also compare the results of our analysis with those achieved for the GPT2-base, T5-base, and RoBERTa-base models as mentioned in [28]. As shown in Table 3, both Mamba models are less dependent on gendered tokens compared to GPT2-base, T5-base, and RoBERTa-base models, with the Mamba-1.4B model showing a further decrease in bias compared to the Mamba-130M, suggesting improvements in reducing gender bias with increased model size.

Investigating long-range capabilities of Mamba. The ability of SSMs to model long-range dependencies is considered an important improvement over previous sequence models. In this use case, we analyze the extent to which the pretrained Mamba-130M model can use information from the entire context window. We use the HotpotQA [79] subset from the LongBench dataset [12], designed to test long context understanding. After selecting all 127 instances, containing sequences up to 8192 tokens, we prompt the model to summarize the full paragraph by generating ten additional tokens. Fig. 6 shows the distribution of the positional difference between a relevant token and the currently generated token. While we observe a pronounced pattern of attributing to the last few tokens, as seen in prior language generation studies [80, 61], the extracted explanations also identified relevant tokens across the entire context window, as presented for one example in Fig. 6 (right). This suggests that the model is indeed capable of retrieving long-range dependencies. We clearly see that in order to complete the sentence and assign a year to the album release date, the model analyzes previous occurrences of chronological information and MambaLRP identifies evidence supporting the decision for the date being '1972' as relevant. Our analysis demonstrates the previously speculated long-range abilities of the Mamba architecture [33], which we further explore in a comparison to Transformers in Appendix C.8.

Needle-in-a-haystack test. To assess the model's ability in retrieving relevant pieces of information from a broader context, we perform the needle-in-a-haystack test [47]. Our test involves extracting a single passkey (the 'needle') from a collection of repeated noise sentences (the 'haystack'), as described in [37]. We run this test at eleven different document depths with three different context

9

lengths. We use an instruction-finetuned Mamba-2.8B model in this experiment. To analyze the performance of the model, we introduce the explanation-based retrieval accuracy (XRA) metric. In this approach, we first identify the positions of the top-K relevant tokens by MambaLRP, and then, calculate the accuracy by comparing those positions to the needle's position. As shown in Fig. 7, MambaLRP accurately captures the information used by the model to retrieve the needle. In this case, the model could accurately retrieve the needle based on relevant information within the text. However, in more realistic and complex scenarios, the model may depend on irrelevant data yet still generate the correct token. This issue can be analyzed using XRA but cannot be evaluated by conventional retrieval accuracy metrics. Such cases and also further details about this experiment are shown in Appendix C.7.

7 Discussion and conclusion

Mamba models have emerged as an efficient alternative to Transformers. However, there are limited works addressing their interpretability [4, 84]. To address this issue, we proposed MambaLRP within the LRP framework, specifically tailored to the Mamba architecture and built upon the relevance conservation principle. Our evaluations across various models and datasets confirmed that MambaLRP adheres to the conservation property and provides faithful explanations that outperform other methods while being more computationally efficient. Moreover, we demonstrated how MambaLRP can help users debug state-of-the-art vision and language models while building trust in their predictions through various use cases. Future research can explore its potential across a broader range of applications and Mamba architectures, providing reliable insights into sequence models.

Limitations As a propagation-based explanation method, MambaLRP requires storing activations and gradients, leading to memory usage that depends on the model architecture and input sequence length. To reduce memory consumption, techniques such as gradient checkpointing can be utilized, which are applicable to other gradient-based methods as well. However, a limitation of these methods, including MambaLRP, is the potential inaccessibility of gradient information due to proprietary constraints. In such cases, approximating gradient information may offer a viable solution.

Acknowledgments and Disclosure of Funding

This work was funded by the German Ministry for Education and Research (refs. 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, 01IS18037A). K.R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

References

[1] S. Abnar and W. Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics.

[2] R. Achtibat, S. M. V. Hatefi, M. Dreyer, A. Jain, T. Wiegand, S. Lapuschkin, and W. Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. *arXiv:2402.05602*, 2024.

[3] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf. XAI for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR, 2022.

[4] A. Ali, I. Zimerman, and L. Wolf. The hidden attention of mamba models. *arXiv:2403.01590*, 2024.

[5] C. An, F. Huang, J. Zhang, S. Gong, X. Qiu, C. Zhou, and L. Kong. Training-free long-context scaling of large language models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1493–1510. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/an24b.html.

[6] Q. Anthony, Y. Tokpanov, P. Glorioso, and B. Millidge. BlackMamba: Mixture of experts for state-space models. *arXiv:2402.01771*, 2024.

10

[7] L. Arras, J. Arjona-Medina, M. Widrich, G. Montavon, M. Gillhofer, K.-R. Müller, S. Hochreiter, and W. Samek. Explaining and interpreting LSTMs. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 211–238, 2019.

[8] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115, 2020.

[9] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[11] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[12] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, and J. Li. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv:2308.14508*, 2023.

[13] E. Baron, I. Zimerman, and L. Wolf. 2-D SSM: A general spatial layer for visual transformers. *arXiv:2306.06635*, 2023.

[14] A. Behrouz and F. Hashemi. Graph Mamba: Towards learning on graphs with state space models. *arXiv:2402.08678*, 2024.

[15] S. Blücher, J. Vielhaben, and N. Strodthoff. Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks. *arXiv:2401.06654*, 2024.

[16] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[17] H. Chefer, S. Gur, and L. Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021.

[18] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

[19] G. Chen, X. Li, Z. Meng, S. Liang, and L. Bing. CLEX: Continuous length extrapolation for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=wXpSidPpc5`.

[20] P. Chormai, J. Herrmann, K.-R. Müller, and G. Montavon. Disentangled explanations of neural network predictions by finding relevant subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[21] T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10041–10071. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/dao24a.html`.

[22] S. B. David, I. Zimerman, E. Nachmani, and L. Wolf. Decision S4: Efficient sequence-based rl via state spaces layers. In *The Eleventh International Conference on Learning Representations*, 2022.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[25] A.-K. Dombrowski, C. J. Anders, K.-R. Müller, and P. Kessel. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022. ISSN 0031-3203.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[27] O. Eberle, J. Büttner, F. Kräutli, K.-R. Müller, M. Valleriani, and G. Montavon. Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1149–1161, 2020.

11

[28] O. Eberle, I. Chalkidis, L. Cabello, and S. Brandl. Rather a nurse than a physician - contrastive explanations under investigation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6907–6920. Association for Computational Linguistics, 2023.

[29] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017.

[30] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv:2212.14052*, 2022.

[31] H. Gong, L. Kang, Y. Wang, X. Wan, and H. Li. nnMamba: 3D biomedical image segmentation, classification and landmark detection with state space model. *arXiv:2402.03526*, 2024.

[32] Y. Gong, Y.-A. Chung, and J. Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.

[33] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[34] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34:572–585, 2021.

[35] A. Gu, K. Goel, and C. Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[36] D. Gunning. DARPA's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page ii. Association for Computing Machinery, 2019.

[37] C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv:2404.06654*, 2024.

[38] S. Jain and B. C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[39] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[40] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng, et al. Swin-UMamba: Mamba-based unet with imagenet-based pretraining. *arXiv:2402.03302*, 2024.

[41] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu. VMamba: Visual state space model. *arXiv:2401.10166*, 2024.

[42] C. Lu, Y. Schroecker, A. Gu, E. Parisotto, J. Foerster, S. Singh, and F. Behbahani. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[43] J. Ma, F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv:2401.04722*, 2024.

[44] J. Ma, F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[45] X. Ma, C. Zhou, X. Kong, J. He, L. Gui, G. Neubig, J. May, and L. Zettlemoyer. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qNLe3iq2El.

[46] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur. Long range language modeling via gated state spaces. *arXiv:2206.13947*, 2022.

[47] A. Mohtashami and M. Jaggi. Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems*, 2023.

[48] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[49] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. Layer-wise relevance propagation: An overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[50] E. Nguyen, K. Goel, A. Gu, G. Downs, P. Shah, T. Dao, S. Baccus, and C. Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in Neural Information Processing Systems*, 35:2846–2861, 2022.

[51] G. Paulo, T. Marshall, and N. Belrose. Does transformer interpretability transfer to rnns?, 2024. URL https://arxiv.org/abs/2404.05971.

12

[52] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, L. Derczynski, X. Du, M. Grella, K. Gv, X. He, H. Hou, P. Kazienko, J. Kocon, J. Kong, B. Koptyra, H. Lau, J. Lin, K. S. I. Mantri, F. Mom, A. Saito, G. Song, X. Tang, J. Wind, S. Woźniak, Z. Zhang, Q. Zhou, J. Zhu, and R.-J. Zhu. RWKV: Reinventing RNNs for the transformer era. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL `https://aclanthology.org/2023.findings-emnlp.936`.

[53] M. Pióro, K. Ciebiera, K. Król, J. Ludziejewski, and S. Jaszczur. MoE-Mamba: Efficient selective state space models with mixture of experts. *arXiv:2401.04081*, 2024.

[54] Z. Qin, S. Yang, and Y. Zhong. Hierarchically gated recurrent neural network for sequence modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=P1TCHxJwLB`.

[55] N. Rajani, L. Tunstall, E. Beeching, N. Lambert, A. M. Rush, and T. Wolf. No robots. `https://huggingface.co/datasets/HuggingFaceH4/no_robots`, 2023.

[56] J. Ruan and S. Xiang. VM-UNet: Vision mamba UNet for medical image segmentation. *arXiv:2402.02491*, 2024.

[57] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11): 2660–2673, 2017.

[58] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE*, 109(3):247–278, 2021.

[59] G. Saon, A. Gupta, and X. Cui. Diagonal state space augmented transformers for speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[60] E. Saravia, H. T. Liu, Y. Huang, J. Wu, and Y. Chen. CARER: contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3687–3697. Association for Computational Linguistics, 2018.

[61] G. Sarti, N. Feldhus, L. Sickert, O. van der Wal, M. Nissim, and A. Bisazza. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada, July 2023. Association for Computational Linguistics.

[62] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon. Higher-order explanations of graph neural networks via relevant walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7581–7596, 2022.

[63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[64] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153, 2017.

[65] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.

[66] J. T. Smith, A. Warrington, and S. Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[67] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. ACL, 2013.

[68] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[70] C. Wang, O. Tsepa, J. Ma, and B. Wang. Graph-Mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv:2402.00789*, 2024.

13

[71] J. Wang, W. Zhu, P. Wang, X. Yu, L. Liu, M. Omar, and R. Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023.

[72] J. Wang, T. Gangavarapu, J. N. Yan, and A. M. Rush. MambaByte: Token-free selective state space model. *arXiv:2401.13660*, 2024.

[73] Z. Wang and C. Ma. Semi-Mamba-UNet: Pixel-level contrastive cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *arXiv:2402.07245*, 2024.

[74] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li. Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. *arXiv:2402.05079*, 2024.

[75] S. Wiegreffe and Y. Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, 2019. Association for Computational Linguistics.

[76] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu. SegMamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv:2401.13560*, 2024.

[77] J. N. Yan, J. Gu, and A. M. Rush. Diffusion models without attention. *arXiv:2311.18257*, 2023.

[78] Y. Yang, Z. Xing, and L. Zhu. Vivim: a video vision mamba for medical video object segmentation. *arXiv:2401.14168*, 2024.

[79] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[80] K. Yin and G. Neubig. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[81] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision– ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[82] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pages 11106–11115. AAAI Press, 2021.

[83] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv:2401.09417*, 2024.

[84] I. Zimerman, A. Ali, and L. Wolf. A unified implicit attention formulation for gated-linear recurrent sequence models, 2024. URL https://arxiv.org/abs/2405.16504.

[85] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*, 2017.

14

A Proofs

In the following, we provide derivations for the conservation analysis performed in Section 4.

A.1 Derivations for SiLU

We first consider the SiLU activation function. As mentioned in Section 4.1, this function is represented by the equation $y = x \cdot \sigma(x)$, with σ being the logistic sigmoid function. By applying the standard gradient propagation equations, we get the conservation equation:

$$
\begin{aligned}
\overbrace{\frac{\partial f}{\partial x} x}^{\mathcal{R}(x)} &= \frac{\partial f}{\partial y} \frac{\partial y}{\partial x} x \\
&= \frac{\partial f}{\partial y} \cdot (\sigma(x) + x\sigma'(x)) \cdot x \\
&= \frac{\partial f}{\partial y} \cdot \sigma(x) \cdot x + \frac{\partial f}{\partial y} \cdot x\sigma'(x) \cdot x \\
&= \underbrace{\frac{\partial f}{\partial y} y}_{\mathcal{R}(y)} + \underbrace{\frac{\partial f}{\partial y} \cdot \sigma'(x) \cdot x^2}_{\varepsilon}
\end{aligned}
\tag{10}
$$

A.2 Derivations for selective SSM

In Section 4.2, we introduced an inconsequential modification to the original selective SSM architecture by connecting the matrix C_t to the state h_t instead of the input x_t. The unfolded view of the SSM component with this modification is represented in Fig. 2. We can observe two subsets of nodes in this figure. The relevance scores of these two subsets should be equal if the conservation property holds. Computing the relevance propagation equation between these two groups, we obtain:

$$
\begin{aligned}
\overbrace{\frac{\partial f}{\partial x_t} x_t + \frac{\partial f}{\partial h_{t-1}} h_{t-1}}^{\mathcal{R}(x_t)+\mathcal{R}(h_{t-1})} &= \left(\frac{\partial f}{\partial h_t} \frac{\partial^+ h_t}{\partial x_t} x_t + \frac{\partial f}{\partial \theta_t} \frac{\partial \theta_t}{\partial x_t} x_t \right) \\
&\quad + \left(\frac{\partial f}{\partial y_{t-1}} \frac{\partial^+ y_{t-1}}{\partial h_{t-1}} h_{t-1} + \frac{\partial f}{\partial h_t} \frac{\partial h_t}{\partial h_{t-1}} h_{t-1} + \frac{\partial f}{\partial \theta_t} \frac{\partial \theta_t}{\partial h_{t-1}} h_{t-1} \right) \\
&= \left(\frac{\partial f}{\partial h_t} \frac{\partial^+ h_t}{\partial x_t} x_t + \frac{\partial f}{\partial h_t} \frac{\partial h_t}{\partial h_{t-1}} h_{t-1} \right) + \left(\frac{\partial f}{\partial y_{t-1}} \frac{\partial^+ y_{t-1}}{\partial h_{t-1}} h_{t-1} \right) \\
&\quad + \left(\frac{\partial f}{\partial \theta_t} \frac{\partial \theta_t}{\partial x_t} x_t + \frac{\partial f}{\partial \theta_t} \frac{\partial \theta_t}{\partial h_{t-1}} h_{t-1} \right) \\
&= \underbrace{\frac{\partial f}{\partial h_t} h_t + \frac{\partial f}{\partial y_{t-1}} y_{t-1}}_{\mathcal{R}(h_t)+\mathcal{R}(y_{t-1})} + \underbrace{\frac{\partial f}{\partial \theta_t} \frac{\partial \theta_t}{\partial x_t} x_t + \frac{\partial f}{\partial \theta_t} \frac{\partial \theta_t}{\partial h_{t-1}} h_{t-1}}_{\varepsilon}
\end{aligned}
\tag{11}
$$

A.3 Derivations for multiplicative gate

Mamba is composed of several blocks. In each block, the selective SSM's output is multiplied by an input-dependent gate. In other words, $y = z_A z_B$ with $z_A = \text{SSM}(x)$ and $z_B = \text{SiLU}(\text{Linear}(x))$. By applying the standard gradient propagation equations, we get the conservation equation:

15

$$\overbrace{\frac{\partial f}{\partial x}}^{\mathcal{R}(x)} x = \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}x$$

$$= \frac{\partial f}{\partial y} \cdot \left(\frac{\partial z_A}{\partial x}z_B + z_A\frac{\partial z_B}{\partial x}\right)x$$

$$= \frac{\partial f}{\partial y} \cdot \left(z_A z_B + z_A z_B\right) \tag{12}$$

$$= \underbrace{\frac{\partial f}{\partial y}y}_{\mathcal{R}(y)} + \underbrace{\frac{\partial f}{\partial y}y}_{\varepsilon}$$

B Explicit propagation rules for MambaLRP

Whereas MambaLRP is more easily implemented via the modified gradient-based approach described in the main paper, we provide below explicit relevance propagation equations for better comparability with other works. We refer to Sections 3 and 4 of the main paper for the definition of the notation.

B.1 SiLU

Explicit LRP rule for SiLU layers:

$$\mathcal{R}(x_i) = \mathcal{R}(y_i) \tag{13}$$

B.2 SSM

Using the shortcut notations $a_{ij} = [A_t(x_t)]_{ji}$, $b_{ij} = [B_t(x_t)]_{ji}$ and $c_{ij} = [C_{t-1}(h_{t-1})]_{ji}$, we can write the propagation of relevance to the previous state space activations explicitly as:

$$\mathcal{R}(h_i^{(t-1)}) = \sum_j \frac{h_i^{(t-1)}c_{ij}}{\sum_i h_i^{(t-1)}c_{ij}}\mathcal{R}(y_j^{(t-1)}) + \sum_j \frac{h_i^{(t-1)}a_{ij}}{\sum_i h_i^{(t-1)}a_{ij} + \sum_{i'} x_{i'}^{(t)}b_{i'j}}\mathcal{R}(h_j^{(t)}) \tag{14}$$

and the propagation of relevance to the SSM input as:

$$\mathcal{R}(x_i^{(t)}) = \sum_j \frac{x_i^{(t)}b_{ij}}{\sum_i x_i^{(t)}b_{ij} + \sum_{i'} h_{i'}^{(t-1)}a_{i'j}}\mathcal{R}(h_j^{(t)}) \tag{15}$$

B.3 Multiplicative Gate

Explicit LRP rule for the multiplicative gate:

$$\mathcal{R}([z_A]_i) = 0.5 \cdot \mathcal{R}(y_i) \tag{16}$$

$$\mathcal{R}([z_B]_i) = 0.5 \cdot \mathcal{R}(y_i) \tag{17}$$

C Experimental details

In this section, we provide experimental details on our experiments that allow reproducibility of our results.

C.1 Models and datasets

For the NLP experiments, we fine-tuned all parameters of the pretrained Mamba-130M, Mamba-1.4B, and Mamba-2.8B models[2] on four text classification datasets: SST-2, SNLI, Medical BIOS, and Emotion. The data statistics can be seen in Table 5. For the vision experiments, we used the pretrained Vim-S model[3], trained on the ImageNet dataset.

---

[2] https://github.com/state-spaces/mamba
[3] https://github.com/hustvl/Vim

16

Training details During training, we used a batch size of 32. To train the Mamba-1.4B and Mamba-2.8B models on the SNLI dataset, a batch size of 64 is used. We employed the {EleutherAI/gpt-neox-20b}[4] tokenizer. The models' parameters were optimized using AdamW optimizer with a learning rate set at $7e - 5$. Additionally, we used a linear learning rate scheduler with an initial factor of 0.5. All models were trained for a maximum of 10 epochs. We employed early stopping and ended training as soon as the validation loss ceased to improve. The top-1 accuracies of the models on each dataset are detailed in Table 4.

Table 4: The accuracies of Mamba-130M, Mamba-1.4B, and Mamba-2.8B models on the validation sets of four text classification datasets.

Dataset	130M	1.4B	2.8B
SST-2	91.97	94.15	94.26
Med-BIOS	89.10	90.30	90.50
Emotion	93.45	93.65	94.10
SNLI	89.57	91.05	91.14

Table 5: Data statistics.

Dataset	Train	Test	Validation
SST-2	68K	2K	1K
Med-BIOS	8K	1K	1K
Emotion	16K	2K	2K
SNLI	550K	10K	10K
ImageNet	1.3M	50K	100K

C.2 MambaLRP details

In this section, we begin by showing how MambaLRP can be implemented through the following algorithms. Then, we explain the generalized LRP-γ rule, provide details regarding hyperparameters used in our implementation, and outline the hyperparameter selection procedure.

Algorithm 1: MambaLRP in SiLU activation layer

Data: Input: x (B, L, D)
1   $z \leftarrow \text{Identity}(x)$
2   **return** $z \odot [\text{SiLU}(x) \oslash z].\texttt{detach()}$

Algorithm 2: MambaLRP in Mamba block

Data: Input: x (B, L, D)
Data: Output: y (B, L, D)
1   x': (B, L, E) $\leftarrow \text{SiLU}(\text{Conv1d}(x))$
2   g: (B, L, E) $\leftarrow \text{SiLU}(\text{Linear}(x))$ $\triangleright$ g is an input-dependent gate
3   A: (E, N) $\leftarrow \text{Parameter}$
4   B: (B, L, N) $\leftarrow \text{Linear}(x')$
5   C: (B, L, N) $\leftarrow \text{Linear}(x')$ $\triangleright$ C is input-dependent
6   Δ: (B, L, E) $\leftarrow \text{Softplus}(\text{Parameter} + \text{Linear}(x'))$
7   $\bar{A}, \bar{B}$: (B, L, E, N) $\leftarrow \text{discretize}(\Delta, A, B)$ $\triangleright$ $\bar{A}$ and $\bar{B}$ are input-dependent
8   y_{SSM}: (B, L, E) $\leftarrow \text{SSM}(\bar{A}.\texttt{detach()}, \bar{B}.\texttt{detach()}, C.\texttt{detach()})(x')$
9   y': (B, L, E) $\leftarrow 0.5(y_{\text{SSM}} \odot g) + 0.5[y_{\text{SSM}} \odot g].\texttt{detach()}$
10   y: (B, L, D) $\leftarrow \text{Linear}(y')$
11   **return** y

The following list represents the hyperparameters of the above-mentioned algorithms:

> B   batch size
> L   sequence length
> D   hidden dimension
> E   expanded hidden dimension
> N   SSM dimension

---

[4]`https://github.com/EleutherAI/gpt-neox`

17

Explanations generated by propagation-based methods rely on gradient computations, which can result in noisy explanations in models with many layers. This is due to the phenomena of gradient shattering and the presence of noisy gradients, which are more common in deep complex models [25, 2]. To mitigate this, we apply the generalized LRP-γ-rule to the convolution layers of the Vision Mamba model to improve the signal to noise ratio, thereby enhancing explanations. The generalized LRP-γ rule is defined in Eq. 18:

$$
\mathcal{R}(x_i) = \begin{cases} \sum_j \frac{x_i^+ (w_{ij} + \gamma w_{ij}^+) + x_i^- (w_{ij} + \gamma w_{ij}^-)}{\sum_i x_i^+ (w_{ij} + \gamma w_{ij}^+) + x_i^- (w_{ij} + \gamma w_{ij}^-)} \mathcal{R}(y_j) & \text{if} \quad z_j > 0 \\[2ex] \sum_j \frac{x_i^+ (w_{ij} + \gamma w_{ij}^-) + x_i^- (w_{ij} + \gamma w_{ij}^+)}{\sum_i x_i^+ (w_{ij} + \gamma w_{ij}^-) + x_i^- (w_{ij} + \gamma w_{ij}^+)} \mathcal{R}(y_j) & \text{else} \end{cases}
\tag{18}
$$

where $(.)^+ = \max(0,.)$ and $(.)^- = \min(0,.)$, and $z_j = \sum_i x_i w_{ij}$. In our experiments, the parameter γ is set to 0.25. Our observations reveal that applying this rule to the language models does not lead to any discernible improvements. Therefore, we use the LRP-0 rule in these models.

C.2.1 LRP composites for Vision Mamba

As mentioned in Section C.2, we apply the generalized LRP-γ rule to the convolution layers of the Vim-S model to produce more faithful explanations. In this experiment, we justify this choice. Vision Mamba is composed of a number of blocks and in each block, there are several linear and convolution layers, where the generalized LRP-γ rule can be used. As can be seen in Table 6, the LRP-0 rule is sufficient to produce meaningful explanations. However, we can perform a hyperparameter search by applying the LRP-γ rule across different layers of the model to find the most accurate LRP composite.

Table 6: Finding the best LRP composite for Vision Mamba. The layers in which the generalized LRP-γ rule is applied are represented with LRP-γ and the ones in which the basic LRP rule, *i.e.* LRP-0, is used are represented with LRP-0.

in-proj	out-proj	conv1d	ImageNet ($\Delta \hat{A}^F \uparrow$)
LRP-γ	LRP-γ	LRP-γ	4.196
LRP-γ	LRP-γ	LRP-0	4.218
LRP-0	LRP-γ	LRP-γ	4.283
LRP-0	LRP-γ	LRP-0	4.336
LRP-γ	LRP-0	LRP-γ	4.597
LRP-γ	LRP-0	LRP-0	4.599
LRP-0	LRP-0	LRP-0	4.684
LRP-0	LRP-0	LRP-γ	**4.715**

We apply the LRP-γ rule across different combinations of the input projection (in-proj), output projection (out-proj), and convolution layers of each block. Subsequently, we perform the perturbation experiment to analyze the faithfulness of each combination. We can observe that the best result can be achieved when the LRP-γ rule is only used in convolution layers. In all of these combinations, the value of γ is set to 0.25.

C.3 Further details of other explanation methods

Some of the explanation methods that we used in this study have a set of hyperparameters. Table 7 provides further details on the specific values assigned to these hyperparameters, chosen based on the values suggested in the original papers [68, 65].

In the vision experiments, we used the original implementations [5] of the AttnRoll and MambaAttr methods, provided to explain the Vim-S model. Given the unavailability of code for adapting these approaches to the language models, namely Mamba-130M, Mamba-1.4B, and Mamba-2.8B, we have developed our own implementation. In the vision case, the authors obtain the final relevance map by

---

[5]https://github.com/AmeenAli/HiddenMambaAttn/

18

Table 7: Hyperparameters of other explanation methods. The parameters μ and σ represent the mean and standard deviation of noise, respectively, while the parameter m denotes the sample size.

Method	Hyperparameters
SmoothGrad	$\mu = 0, \sigma = 0.15, m = 30$
Integrated Gradients	$m = 30$

Table 8: Comparing the proposed strategies for managing the Mamba block's multiplicative gate.

Strategies	SST-2	ImageNet
Detaching the multiplicative gate z_B	1.577	3.592
Half-relevance propagation	**1.978**	**4.715**

extracting the row associated with the CLS token in the attention matrix. However, since our language models lack a CLS token, we get the final relevance map from the row associated with the last token in the attention matrix. This is because predictions are based on the last state in these models. For the gradient-based methods, we use the implementations available in the Captum library[6].

C.4 Further details on evaluation metrics

As mentioned in Section 5.3, the *flipping* and *insertion* metrics can be used to evaluate the quality of the generated explanations. It is important to note that starting with unperturbed images and gradually applying perturbations until fully perturbed images are obtained yields identical results for both the *flipping* and *insertion* metrics. Therefore, we have only reported the results for the *flipping* experiment.

In our *flipping* evaluations, we calculate the area under the curve (AUC) by starting from full images and progressively masking the pixels with zeros until we reach completely masked images. The perturbation steps are defined using `np.linspace(0, 1, 11)`. In our vision experiments, Images are normalized using the ImageNet mean and standard deviation, and the explanations are generated and evaluated for the predicted class. Unlike [4], which tracks changes in the model's top-1 accuracy, we monitor changes in the output logit of the predicted class.

C.5 Further ablation experiments

Comparing strategies for managing the Mamba block's multiplicative gate: In Section 4.3, we proposed several strategies to mitigate conservation violation in the Mamba block's multiplicative gate. In this experiment, we evaluate the proposed approaches. As can be seen in Table 8, detaching the multiplicative gate z_B leads to lower faithfulness scores compared to the half-relevance propagation approach. To retain conservation, an alternative approach is to detach the SSM's output z_A, which limits capturing long-range dependencies, a task for which this branch is designed for. Detaching it may result in a loss of valuable information used by the model to make predictions. Therefore, this approach is not considered in Table 8.

C.6 Additional qualitative results

In Section 5.2, we qualitatively evaluated the explanations produced by MambaLRP and other baseline methods. In the following, we demonstrate further qualitative results.

C.6.1 Natural language processing

In the following figures, we represent explanations produced by MambaLRP and other baseline methods to interpret the Mamba-130M models trained on various datasets. In the visualizations, shades of red represent words that positively influence the model's prediction. Conversely, shades of blue reflect negative contributions. The heatmaps of the AttnRoll and MambaAttr methods are constrained to non-negative values.

---

[6]`https://captum.ai/`

19

MambaLRP	Two women are embracing while holding to go packages . Two woman are holding packages .
GI	Two women are embracing while holding to go packages . Two woman are holding packages
SmoothGrad	Two women are embracing while holding to go packages . Two woman are holding packages .
IG	Two women are embracing while holding to go packages . Two woman are holding packages .
AttnRoll	Two women are embracing while holding to go packages . Two woman are holding packages
MambaAttr	Two women are embracing while holding to go packages . Two woman are holding packages

Figure 8: Explanations generated by different explanation methods for a sentence of the SNLI validation set. This sentence belongs to the 'entailment' class.

MambaLRP	A woman is doing a cart wheel while wearing a b ik ini in the sand next to the beach . A woman is fixing her home .
GI	A woman is doing a cart wheel while wearing a b ik ini in the sand next to the beach . A woman is fixing her home .
SmoothGrad	A woman is doing a cart wheel while wearing a b ik ini in the sand next to the beach . A woman is fixing her home
IG	A woman is doing a cart wheel while wearing a b ik ini in the sand next to the beach . A woman is fixing her home
AttnRoll	A woman is doing a cart wheel while wearing a b ik ini in the sand next to the beach . A woman is fixing her home
MambaAttr	A woman is doing a cart wheel while wearing a b ik ini in the sand next to the beach . A woman is fixing her home

Figure 9: Explanations generated by different explanation methods for a sentence of the SNLI validation set. This sentence belongs to the 'contradiction' class.

MambaLRP	i am feeling very blessed today that they share such a close bond
GI	i am feeling very blessed today that they share such a close bond
SmoothGrad	i am feeling very blessed today that they share such a close bond
IG	i am feeling very blessed today that they share such a close bond
AttnRoll	i am feeling very blessed today that they share such a close bond
MambaAttr	i am feeling very blessed today that they share such a close bond

Figure 10: Explanations generated by different explanation methods for a sentence of the Emotion validation set. This sentence belongs to the 'joy' class.

C.6.2 Computer vision

In this section, we show explanations generated by MambaLRP alongside other baseline methods to interpret the predictions of the Vim-S model on several images of the ImageNet dataset. As can be seen, explanations generated by purely gradient-based explanation methods are very noisy. In contrast, attention-based attribution methods have offered more focused and less noisy heatmaps. However, in the last two images labeled 'paint brush' and 'flag pole', they could not faithfully explain the model's predictions. Among these approaches, MambaLRP stands out with its ability to generate sparse explanations, offering more faithful explanations of how different image patches contribute to the final predictions.

C.7 Additional use case results

For the needle-in-a-haystack experiment in Section 6, we use a synthetic dataset [7]. In this dataset, a single passkey (the 'needle') is inserted at different locations within a collection of repeated noise

---

[7]`https://huggingface.co/datasets/lvwerra/needle-llama3-16x512`

20

MambaLRP	She graduated with honors in 2013 . Having more than 4 years of diverse experiences , especially in N UR SE PR ACT ITION ER , Eric a J H oyer affiliates with Swedish American Hospital , and cooper ates with other doctors and specialists in medical group Swedish americ an Hospital . Call Eric a J H oyer on phone number (779) 6 96 - 86 50 for more information and adv ises or to book an appointment
GI	She graduated with honors in 2013 . Having more than 4 years of diverse experiences , especially in N UR SE PR ACT ITION ER , Eric a J H oyer affiliates with Swedish American Hospital , and cooper ates with other doctors and specialists in medical group Swedish americ an Hospital . Call Eric a J H oyer on phone number (779) 6 96 - 86 50 for more information and adv ises or to book an appointment .
SmoothGrad	She graduated with honors in 2013 . Having more than 4 years of diverse experiences , especially in N UR SE PR ACT ITION ER , Eric a J H oyer affiliates with Swedish American Hospital , and cooper ates with other doctors and specialists in medical group Swedish americ an Hospital . Call Eric a J H oyer on phone number (779) 6 96 - 86 50 for more information and adv ises or to book an appointment
IG	She graduated with honors in 2013 . Having more than 4 years of diverse experiences , especially in N UR SE PR ACT ITION ER , Eric a J H oyer affiliates with Swedish American Hospital , and cooper ates with other doctors and specialists in medical group Swedish americ an Hospital . Call Eric a J H oyer on phone number (779) 6 96 - 86 50 for more information and adv ises or to book an appointment
AttnRoll	She graduated with honors in 2013 . Having more than 4 years of diverse experiences , especially in N UR SE PR ACT ITION ER , Eric a J H oyer affiliates with Swedish American Hospital , and cooper ates with other doctors and specialists in medical group Swedish americ an Hospital . Call Eric a J H oyer on phone number (779) 6 96 - 86 50 for more information and adv ises or to book an appointment
MambaAttr	She graduated with honors in 2013 . Having more than 4 years of diverse experiences , especially in N UR SE PR ACT ITION ER , Eric a J H oyer affiliates with Swedish American Hospital , and cooper ates with other doctors and specialists in medical group Swedish americ an Hospital . Call Eric a J H oyer on phone number (779) 6 96 - 86 50 for more information and adv ises or to book an appointment

Figure 11: Explanations generated by different explanation methods for a sentence of the Medical BIOS validation set. This sentence belongs to the 'nurse' class.

sentences (the 'haystack'), as described in [37]. The dataset is composed of sequences with different context lengths. In our experiment, we use sequences with context lengths of 512, 1024, and 2048. We restrict the maximum context length to 2048 tokens to align with the model's training configuration, as this experiment is not designed to evaluate extrapolation beyond this limit. The goal is to focus on certain limitations of the retrieval accuracy metric and and the solutions provided by MambaLRP. We use a Mamba-2.8B model [8], which is finetuned on the No Robots dataset [55] using a context length of 2048. Then, we prompt the model to extract the passkey hidden among irrelevant text by completing the phrase "The passkey is ____".

Retrieval accuracy is a metric, which is commonly used in the needle-in-a-haystack experiment to analyze the model's performance. The synthetic dataset used for this experiment can be designed to include misleading information, which may cause the model to generate the correct passkey based on incorrect evidence. In such cases, simply evaluating the retrieval accuracy may be insufficient. This issue can also arise when dealing with more realistic haystacks. Therefore, we introduced explanation-based retrieval accuracy (XRA) in Section 6. MambaLRP and the XRA metric designed upon it can help to better examine the evidence the model relies on to retrieve the needle. In this approach, we first identify the positions of the top-K relevant tokens by MambaLRP, and then, calculate the accuracy by comparing those positions to the needle's position. We set the value of K to 2. This is because MambaLRP identifies the token immediately preceding the generated token as the

---

[8] https://huggingface.co/clibrain/mamba-2.8b-chat-no_robots

21
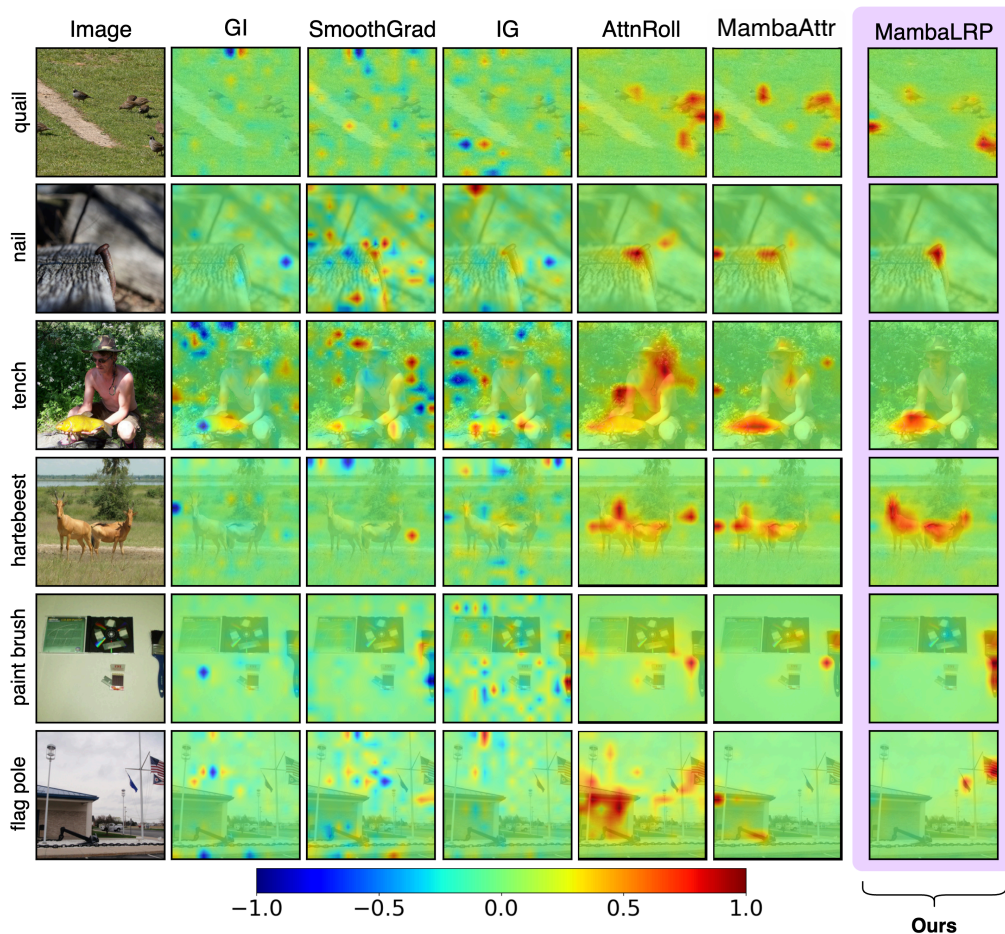
Figure 12: Explanations produced by different explanation methods for images of the ImageNet dataset. Explanations produced by AttnRoll and MambaAttr are limited to non-negative values, whereas those generated by gradient-based techniques and MambaLRP includes both positive and negative contributions.

most important one in most of the examples and the evidence used for the passkey retrieval is usually the second most important token.

The sample in Fig. 13 represents a scenario where our XRA approach proves valuable. In this case, the next token generated by the model is the second part of the correct passkey (300). However, the model has incorrectly focused on the number 300 in the phrase "Pass the key to room 6300" to generate this token. Simply looking at the retrieved token might suggest that the model successfully retrieved the correct information. However, examining the MambaLRP's explanation heatmaps provides deeper insights into the model's behavior. This helps us to debug the model more effectively and design better tests to analyze its capabilities.

C.8 Long-range dependeny comparison to Transformers

To explore the capabilities of different model architectures in handling long-range dependencies, we performed a direct comparison between Mamba and state-of-the-art Transformers (Llama-2[9] and Llama-3[10]), focusing on their performance with inputs exceeding typical context lengths.

---

[9] https://huggingface.co/meta-llama/Llama-2-7b-hf
[10] https://huggingface.co/meta-llama/Meta-Llama-3-8B

22
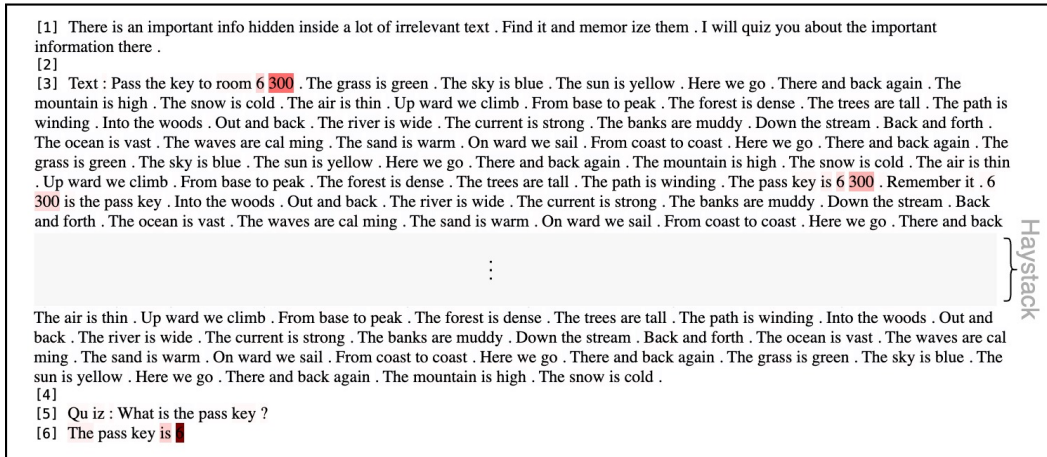
Figure 13: Detecting Clever-Hans effect in the needle-in-a-haystack test. Given the 2K context length in this example, visualizing the entire text could be confusing. Therefore, we have removed most of the haystack from the visualization. In this example, the model has generated the correct passkey but the generation is not based on truly relevant information in the text.

For Llama-2 and Llama-3, we extract attributions using LRP for Transformers [3]. As in the Mamba experiment, we generate 10 additional tokens from the HotPotQA dataset input and explain the prediction for each generated token. The results are shown in Table 9. For Llama-2, which was trained with a context length of 4096, the generated text becomes increasingly less sensible and repetitive for contexts longer than 4k, a limitation noted also in [19, 5]. When analyzing the histogram distribution over tokens considered relevant to predict the next token, it appears that Llama-2 uses information more uniformly across the entire context and identifies more relevant long-range dependent tokens compared to Llama-3 and Mamba. However, as presented in Table 9, its output becomes nonsensical for context lengths exceeding 4K tokens, characterized by the use of rare vocabulary and repeated tokens. Thus, the identified relevant tokens are mostly non-semantic such as new line token '<0x0A>' in Llama-2 and beginning of sentence token '<s>' found at the start of the context paragraphs. For Llama-3 and Mamba, the attributions can identify meaningful relevant tokens. When directly compared, Llama-3 uses information from more intermediate mid-range dependencies than Mamba, though both favor tokens close to the end of the input as relevant. Given Llama-3's much larger size (8B) compared to Mamba (130M) and their different training settings, this analysis supports that Mamba indeed uses long-range information. We also find that this ability is not exclusive to SSMs and can in principle also be achieved by Transformer models.

To what extent these findings depend on the amount of training data and model complexity remains an open research question. Our investigation of long-range dependencies in recent sequence generation models highlights the value of faithful attribution methods like MambaLRP in examining the capabilities and mechanisms utilized by models during generation.

C.9 Runtime comparison

In this section, we report the time required for each explanation method to generate its respective explanation. These times, measured in seconds, are averaged over samples from the Medical BIOS dataset. All baseline methods are evaluated on a single A100-40GB GPU with a batch size of 1. All methods are applied to the Mamba-130M model. The results without fast CUDA kernels are shown in Table 10, while the results with fast CUDA kernels are presented in Table 11. We can observe that the runtime of MambaLRP is comparable to Gradient×Input. Since algorithms like Integrated Gradients and SmoothGrad require multiple function evaluations, their runtimes are significantly higher than MambaLRP and Gradient×Input.

23

id	length	Llama-2	Llama-3	Mamba
3	1k	Summary <0x0A> The _genus _D ict y os per ma _is _a	Summary Ċ The genus Dict y os per ma is a mon	Summary Ċ The species is a member of the genus Ap oll
41	4k	Summary <0x0A> The _Ohio , _Ohio , _Ohio , _Ohio , _Ohio	Summary Ċ The following is a summary of Finn 's head coaching	Summary Ċ The following is a summary of the history of the
109	8k	Summary <0x0A> The Љ Љ Љ Љ Љ Љ Љ Љ Љ Љ	Summary Ċ The mall is anchored by Hudson 's Bay , Walmart	Summary Ċ The mall is located in the heart of the city

Table 9: Long-Range dependency experiment, comparing Transformers and Mamba for different context lengths.

Table 10: Runtime comparison. The time needed for each baseline method to generate its explanations. The times, measured in seconds, are averaged over the samples from the Medical BIOS dataset. The model used in this experiment is Mamba-130M **without using** fast CUDA kernels.

Methods	Runtime
Gradient $\times$ Input	0.7556
SmoothGrad	22.9772
Integrated Gradients	22.8071
AttnRoll	2.1558
MambaAttr	2.6661
MambaLRP	0.4345

Table 11: Runtime comparison. The time needed for each baseline method to generate its explanations. The times, measured in seconds, are averaged over the samples from the Medical BIOS dataset. The model used in this experiment is Mamba-130M **using** fast CUDA kernels.

Methods	Runtime
Gradient $\times$ Input	0.0335
SmoothGrad	0.9785
Integrated Gradients	0.9742
AttnRoll	-
MambaAttr	-
MambaLRP	0.0306

24

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: All claims are addressed in Sections 4, 5, and 6 of the main paper and further details regarding each section can be found in the supplemental material.

   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: We discuss the computational efficiency of our method including runtime comparisons in Section 5. We further critically assess the performance of our approach via a number of ablation studies presented in Section 5 in the main paper and in the supplemental materials in Section C.2.1 and Section C.5. Limitations regarding scope and possible future directions are mentioned in Section 7.

   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.

25

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: [Yes]

   Justification: We state our assumptions and theoretical results in Section 4 of the main paper alongside proofs given in Sections A.1, A.2, and A.3 in the supplemental material.

   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: We provide necessary information for reproducing our results in Section 4 of the main paper and Section C.2 of the supplemental material. Details of our proposed method and how it can be implemented can be found in Section C.2 of the supplemental material. Please refer to Section 5 of the paper and Section C of the supplemental material for more details regarding the experiments.

   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).

26

(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 code for reproducing our results and implementation of our proposed method is publicly accessible and a link to our code repository is added to the main paper. The datasets used in this work are publicly available, and comprehensive instructions for implementing our MambaLRP method can be found in Section C.2 of the supplemental material.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please check Section 5 in the paper for details about experimental setting and evaluation metrics. Further details can be found in Section C of the appendix.

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: [Yes]

27

Justification: Please check Section 5.3 of the paper. Further experimental results can be found in Section C.5 in the appendix.

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: Please check Appendix C.9.

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: [Yes]

Justification: We do not anticipate any harmful risks from the methods and analyses presented in this work.

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**

28

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

Answer: [Yes]

Justification: The positive societal impacts of our work regarding model explainability are mentioned in multiple parts of the paper, in particular Section 6. We do not anticipate any negative societal impact from the methods and analyses presented in this work.

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:

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: All datasets and models used in the paper are properly cited. We have also provided links to libraries and code repositories used, in the supplemental material.

29

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: [NA]

    Justification:

    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:

    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]

30

Justification:

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.

31