UNDERSTANDING BOTTLENECKS OF STATE SPACE MODELS THROUGH THE LENS OF RECENCY AND OVER-SMOOTHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured State Space Models (SSMs) have emerged as alternatives to transformers, addressing the challenges of processing long sequences. While SSMs are often regarded as effective in capturing long-term dependencies, we theoretically demonstrate that they suffer from a strong recency bias. Our empirical findings reveal that this bias impairs the models' ability to recall distant information and introduces robustness issues. We conducted scaling experiments and discovered that deeper structures in SSMs facilitate the learning of long contexts. However, our theoretical analysis reveal that as SSMs increase in depth, they exhibit a tendency toward over-smoothing, resulting in token representations becoming increasingly indistinguishable. This over-smoothing phenomenon ultimately constrains the scalability of SSMs to achieve improved performance. Collectively, these findings highlight important limitations of SSMs and underscore the need for further research to address these challenges in long-range sequence modeling.

1 INTRODUCTION

The evolution of sequence processing architectures has been witnessed over recent decades, progressing from RNNs (Hochreiter & Schmidhuber, 1997; Sutskever et al., 2014; Cho et al., 2014; Cho, 2014) to transformers (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2018; 2019; Brown et al., 2020), and more recently proposed State Space Models (SSMs) (Gu et al., 2021a; Gu & Dao, 2023). Each step represents a leap in natural language processing, addressing the limitations of its predecessors and introducing new capabilities.

Structured State Space Models (SSMs) (Gu et al., 2021a; Gu & Dao, 2023; Dao & Gu, 2024) have emerged as a compelling alternative to transformers, addressing the challenges associated with processing long sequences. SSMs provide advantages in two key areas. Firstly, they enable more efficient handling of long sequences. SSMs operate in two modes: convolution and recurrence, each tailored for different aspects of language model training and inference (Gu et al., 2021b). In convolutional mode, SSMs assume visibility of the entire sequence and utilize hardware-optimized convolutions to propagate information across all tokens in parallel. This approach circumvents the need for calculating pairwise correlations inherent in attention mechanisms, thereby accelerating training speed. In the recent Mamba model (Gu & Dao, 2023), convolution has been supplanted by a parallel scanning algorithm, facilitating more expressive sequence-level mixing without sacrificing efficiency. Conversely, in recurrent mode, SSMs process one token at a time while maintaining a compact recurrent hidden state that encodes the sequence history. The outputs are sequentially decoded from this hidden state, eliminating the necessity to store all past key-value pairs (Dai et al., 2019), thus reducing memory usage during inference.

Furthermore, state space models (SSMs) have been meticulously tailored to effectively capture long-range dependencies and filter contextual information. These models are grounded in the HiPPO theory (Gu et al., 2020), which demonstrates that a first-order Ordinary Differential Equation (ODE) can encapsulate long-term memory through a designated state matrix known as the HiPPO matrix. Subsequent research (Gu et al., 2021b;a; Gupta et al., 2022; Gu et al., 2022a) has streamlined this state matrix to a diagonal form, significantly enhancing computational efficiency while retaining the capability to model long-range dependencies. More recently, Mamba (Gu & Dao, 2023) introduced a

selection mechanism that selectively aggregates pertinent information from the context into the state, showcasing impressive performance in language modeling. Concurrently, another class of efficient sequential models, coined as Linear Attention Models, has emerged, derived from streamlined attention mechanisms (Katharopoulos et al., 2020; Sun et al., 2023; Peng et al., 2023; Yang et al., 2023). Collectively, these advancements can be interpreted through a unified lens as more structured SSMs (Dao & Gu, 2024).

However, despite the initial empirical successes of these models, recent findings indicate that SSMs may not match transformers in their ability to recall information from long contexts (Arora et al., 2023; Poli et al., 2024) or in handling more complex retrieval patterns (Park et al., 2024). Additionally, it has been noted that Mamba continues to underperform compared to transformers at larger scales (Waleffe et al., 2024). These shortcomings, however, have yet to be systematically elucidated.

In this paper, we identify two fundamental limitations of SSMs in their ability to model complex long-range dependencies. First, we argue that the long-term memory capabilities of modern SSMs may be misinterpreted. Our analysis reveals that an SSM layer exhibits a strong recency bias, limiting tokens to primarily interact with nearby context. This bias is intrinsic to SSMs and many linear attention models, regardless of the employed content-informing techniques, such as the selection mechanism introduced by Mamba (Gu & Dao, 2023). We further posit that the loss of long-range capabilities may stem from the oversimplification of HiPPO-induced SSMs, trading efficiency off the performance. To substantiate this claim, we perform a long-range retrieval task on an industrial-scale language model (Jiang et al., 2023) based on Mamba. Our test results indicate that Mamba catastrophically forgets distant content once the context length surpasses its memory capacity. Furthermore, we raise a novel robustness concern regarding SSMs with recency bias: our empirical outcomes show that Mamba is more susceptible to perturbations on local tokens, making it vulnerable to adversarial attack, as these local tokens can be easily manipulated to serve as backdoors.

Additionally, we conduct a series of scaling experiments with varying context lengths during the pre-training of SSMs. Our results indicate that increasing the model's depth is crucial for enhancing its ability to utilize long contexts by expanding the receptive field. However, we observe that depth scaling encounters a bottleneck, as performance begins to saturate with continued increases in depth. To investigate this scalability issue theoretically, we analyze the feature dynamics across SSM layers. Our findings reveal that SSMs inherently function as smoothening operators, leading to over-smoothing in deep architectures (NT & Maehara, 2021; Oono & Suzuki, 2019; Cai & Wang, 2020). As a result, token representations become increasingly uniform and indistinguishable with each additional layer.

The primary contribution of this work lies in unveiling two critical issues inherent to SSMs that have been overlooked in previous research. We provide new insights to systematically explain the underlying mechanisms through rigorous theoretical analysis and controlled experiments. Our theoretical framework encompasses a broad range of commonly used SSMs today. By elucidating these two challenges, we hope to inspire future research aimed at addressing these issues.

2 PRELIMINARIES

In this work, we primarily focus on SSMs and their similar models working with discrete-time sequences of tokens. We represent the a sequence of tokens as $\boldsymbol{x} = [x_1 \quad \cdots \quad x_T]^\top \in \mathbb{R}^T$, where T is the total number of tokens. For vector-valued input sequences, SSMs process each channel independently. Therefore, to simplify notation, we focus on scalar-valued sequences without loss of generality. The impact of multi-channel inputs will be addressed in the relevant context.

SSMs learn to represent and forecast the next token by integrating past information. Formally, SSMs can be viewed as a sequence-to-sequence transformation from inputs $\boldsymbol{x} \in \mathbb{R}^T$ to outputs $\boldsymbol{y} \in \mathbb{R}^T$ through a *memory state* $\boldsymbol{h}_t \in \mathbb{R}^N$, which is iteratively updated with a linear recurrence. A general form can be written as:

$$\boldsymbol{h}_t = \boldsymbol{A}_t \boldsymbol{h}_{t-1} + \Delta_t \boldsymbol{b}_t(\boldsymbol{x}_t), \quad \boldsymbol{y}_t = c_t(\boldsymbol{h}_t), \quad \boldsymbol{h}_0 = \boldsymbol{0}, \quad \forall t \in [T], \tag{1}$$

where $t \in [T]$ denotes the time step. Intuitively, $\boldsymbol{A}_t \in \mathbb{R}^{N\times N}$ extracts information from the previous state $\boldsymbol{h}_{t-1}$ [1], $\boldsymbol{b}_t : \mathbb{R} \to \mathbb{R}^N$ projects every input token to the hidden space, $\Delta_t \in \mathbb{R}$ controls how much information of the new token will be fused into the hidden memory, and $c_t : \mathbb{R}^N \to \mathbb{R}$ decodes the hidden state at time t to the final prediction. SSMs are trained end-to-end to optimize for parameters $\{(\boldsymbol{A}_t, \boldsymbol{b}_t, c_t, \Delta_t)\}_{t\in[T]}$, for which different SSMs adopt various types of instantialization. Below we list some representative examples.

S4, DSS, and S4D. The seminal works (Gu et al., 2020; 2021b; 2022b) demonstrate that discretizing time-invariant ODE $\boldsymbol{h}'(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{b}x(t)$ with some special realization of matrix $\boldsymbol{A}$ can yield an efficient recurrent network for long-sequence modeling. The follow-up works Gu et al. (2021a) together with Gupta et al. (2022); Gu et al. (2022a) simplifies $\boldsymbol{A}$ to be a diagonal matrix. Applying the zero-order hold rule for discretization, as suggested by Gupta et al. (2022), we can summarize this series of models in the form of Eq. 1:

$$\text{(S4)} \quad \boldsymbol{A}_t = \exp(\Delta\boldsymbol{A}), \quad \boldsymbol{b}_t(\boldsymbol{x}_t) = \boldsymbol{b}\boldsymbol{x}_t, \quad c(\boldsymbol{h}_t) = \boldsymbol{c}^\top \boldsymbol{h}_t, \quad \Delta_t = \Delta, \tag{2}$$ [2]

where $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \Delta)$ are learnable parameters. In particular, $\boldsymbol{A}$ is restricted to be a diagonal matrix and can be complex valued. However, $\boldsymbol{A}$ must have negative real part (Gu et al., 2022a). $\Delta \in (0, 1]$ is often interpreted as the time interval for discretization. We call this family of SSMs *S4* following the naming convention in Gu & Dao (2023).

Mamba. A recent breakthrough Mamba (Gu & Dao, 2023) introduces the *selection* mechanism to extend S4. Instead of learning $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \Delta)$ in Eq. 2 as free parameters, Mamba conditions $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, \Delta)$ on the inputs, which enables each iterative step in Eq. 1 to filter useful token information during the recurrence. Specifically, Mamba computes $(\boldsymbol{A}_t\boldsymbol{b}_t, c_t, \Delta_t)$ as follows:

$$\text{(Mamba)} \ \ \boldsymbol{A}_t = \exp(\Delta_t\boldsymbol{A}),\ \boldsymbol{b}_t(\boldsymbol{x}_t) = (\boldsymbol{W}_B\boldsymbol{x}_t)\boldsymbol{x}_t,\ c_t(\boldsymbol{h}) = (\boldsymbol{W}_C\boldsymbol{x}_t)^\top\boldsymbol{h}_t,\ \Delta_t = \sigma(\boldsymbol{W}_\Delta\boldsymbol{x}_t), \tag{3}$$

where $\boldsymbol{W}_\Delta \in \mathbb{R}, \boldsymbol{W}_B \in \mathbb{R}^N, \boldsymbol{W}_C \in \mathbb{R}^N$ are learnable weights in addition to $\boldsymbol{A}$, and $\sigma(\cdot)$ denotes softplus activation. When handling multi-dimensional token embeddings, $\boldsymbol{W}_\Delta, \boldsymbol{W}_B, \boldsymbol{W}_C$ are extended on the input dimension [3]. The resultant Δ_t is then specified for each channel, while $\boldsymbol{b}_t, \boldsymbol{c}_t$ are shared across channels. In language modeling, $\boldsymbol{A}$ has strictly negative real-valued diagonal, which ensures $\boldsymbol{A}_t \in (0, 1)^{N\times N}$. Additionally, Mamba is integrated into the H3 architecture (Fu et al., 2022), wherein the selective SSMs is working with a local convolution and sandwiched by two gated connections.

Linear Attention. Concurrent with SSMs, there is another line of work streamlining attention to linear time complexity. With slight abuse of terminology, we name all of them collectively as Linear Attention Models (LAMs). We observe that many of them can be written in the form of Eq. 1 such that in the remainder of this paper, we extend the definition of SSMs to include LAMs without introducing ambiguity, as LAMs and SSMs are dual to each other (Dao & Gu, 2024). We leave a full summary to Appendix A.

3 Can SSM Effectively Represent Long-Range Dependencies?

3.1 SSMs are Locally Biased

In this section, we investigate the ability of SSMs to learn long-range dependencies. Recent studies find that SSMs seem more effective than transformers on this task (Gu et al., 2020; Tay et al., 2020; Li et al., 2022; Gu & Dao, 2023). However, in Sec. 3.1 we theoretically show a negative result that an SSM layer is inherent to *local bias* and loses long-term memory exponentially. In Sec. 3.3, we empirically justify our claim by showing SSMs struggle to retrieve from distant context. We also demonstrate that the local bias may lead to robustness issues in Sec. 3.4.

[1] In most scenarios discussed in this paper, we assume real parameterization by default, as it is the standard approach in the cases of our primary interest, such as language modeling (Gu & Dao, 2023)

[2] More rigorously, by zero-order hold, $\boldsymbol{b}$ should be parameterized as $\boldsymbol{b} = (\Delta\boldsymbol{A})^{-1}(\exp(-\Delta\boldsymbol{A}) - \boldsymbol{I})\boldsymbol{b}$. However, the presented form is more commonly used in practice as in Gu & Dao (2023).

[3] Suppose the embedding dimension is D, then $\boldsymbol{W}_\Delta \in \mathbb{R}^D$, $\boldsymbol{W}_B \in \mathbb{R}^{N\times D}$, and so on.

To understand how information is propagated and long-range dependencies are modeled in SSMs, we aim to uncover the relationship between the output at time $t \in [T]$ and the input token at time $s \leq t$. We draw our first insight by presenting the following result which rewrites the recurrent form in Eq. 1 to a parallel form and reveals how all previous tokens jointly affect the outputs:

Lemma 3.1 (Parallel form)**.** *For any $\{(\boldsymbol{A}_t, \boldsymbol{b}_t, c_t, \Delta_t)\}_{t\in[T]}$ and $\boldsymbol{x} \in \mathbb{R}^T$, $\boldsymbol{y} \in \mathbb{R}^T$ computed via an SSM defined in Eq. 1 is equal to:*

$$\boldsymbol{y}_t = c_t \left(\sum_{i=1}^{j-1} \left(\prod_{r=s+1}^{t} \boldsymbol{A}_r \right) \Delta_s \boldsymbol{b}_s(\boldsymbol{x}_s) + \Delta_t \boldsymbol{b}_t(\boldsymbol{x}_t) \right), \quad \forall t \in [T]. \tag{4}$$

The proof of Lemma 3.1 can be found in Appendix C.1. Lemma 3.1 provides an alternative perspective on how SSMs compute the outputs. The predicted value for the t-th token is obtained via decoding a weighted aggregation over representations of all past tokens. The encoding and decoding stage is element-wise independent of the context. Whereas, the "weight" associated with each past token in the summation reflects the pairwise relationship, playing a similar role to attention weights in transformers (Dao & Gu, 2024; Ali et al., 2024). The weight corresponding to one past token is calculated as the cumulative product $\prod_r \boldsymbol{A}_r$, where $r \in [s+1, t]$ traverses from the past token (at time s) to the target token (at time t). Assume $\boldsymbol{A}_t \in (0,1)^{N\times N}$, which is satisfied by most of SSMs discussed in Sec. 2, we can show that $(\prod_{r=s+1}^{t} \boldsymbol{A}_r)_{n,n} < (\prod_{r=s'+1}^{s} \boldsymbol{A}_r)_{n,n}$ for any $s < s' < t$ and $n \in [N]$. By this interpretation, SSMs assign strictly higher "attention" to the nearer tokens than the further tokens.

Next, we characterize how SSMs model long-range dependencies more carefully. We define the derivatives $|\partial \boldsymbol{y}_t / \partial \boldsymbol{x}_s|$ as the *influential score* to represent the importance of the s-th input token to the t-th output token. Note that $|\partial \boldsymbol{y}_t / \partial \boldsymbol{x}_s|$ is well-defined for every $s, t \in [T]$ as long as $(\boldsymbol{A}_t.\boldsymbol{b}_t, c_t, \Delta_t)$ are all differentiable in terms of $\boldsymbol{x}$. Intuitively, if $|\partial \boldsymbol{y}_t / \partial \boldsymbol{x}_s|$ is larger, then the s-th input token is more influential on the t-th output token, and vice versa.

Below we present a formal result regarding the influential score.

Theorem 3.2 (Recency of SSMs)**.** *Consider an SSM defined in Eq. 1 with $\{(\boldsymbol{A}_t, \boldsymbol{b}_t, c_t, \Delta_t)\}_{t\in[T]}$. Assume that (i) the input space $\mathcal{X} \subset \mathbb{R}^T$ is compact, (ii) $\{(\boldsymbol{A}_t, \boldsymbol{b}_t, c_t, \Delta_t)\}_{t\in[T]}$ are continuous and have continuous derivatives, and (iii) $\boldsymbol{A}_t \in (0,1)^{N\times N}$ are diagonal matrices for all $t \in [T]$. Let $A_{max} = \max_{t\in[T], n\in[N]} (\boldsymbol{A}_t)_{n,n}$. Then for arbitrary $\boldsymbol{x} \in \mathcal{X}$ and every $s, t \in [T]$ such that $s < t$, $|\partial \boldsymbol{y}_t / \partial \boldsymbol{x}_s| = \mathcal{O}(\exp(-\kappa(t-s)))$ for some $\kappa = \Theta(\log(A_{max}^{-1}))$.*

The proof can be found in Appendix C.2. The first two assumptions are standard and always satisfied. The third assumption also holds for most of SSMs discussed in Sec. 2. Therefore, Theorem 3.2 applies to numerous SSMs including but not limited to S4 (Gu et al., 2021a; 2022a), Mamba (Gu & Dao, 2023), RetNet (Sun et al., 2023), RWKV (Peng et al., 2023), GLA (Yang et al., 2023), HGRN2 (Qin et al., 2024), Griffin (De et al., 2024), and Megalodon (Ma et al., 2022; 2024). Theorem 3.2 states that influential scores between two tokens modeled by SSMs are *exponentially* diminishing with respect to their relative distance. The decay rate is determined by the maximal values among all $\boldsymbol{A}_t$'s elements. The closer A_{max} is to zero, the faster the influential scores decay. The practical implication is that SSMs are factually recency-biased models. Tokens farther away are under-reaching and forgotten rapidly while the information of closer tokens dominates the final output. This can significantly limit their ability of fitting complex long-range relationships.

Empirical Validation. We empirically verify our theory by directly plotting the influential score w.r.t. the relative distances in Fig. 1. The blue and orange curves in Fig. 1 successfully justify our exponentially decaying bound in Theorem 3.2. In contrast, transformer-based architectures is free from the strong recency bias, while demonstrating a well-known "lost-in-the-middle" pattern (Liu et al., 2024b).

3.2 Discussions

Revisiting HiPPO theory. HiPPO established in (Gu et al., 2020), extended by Gu et al. (2021b; 2022b) is the theoretical foundation of SSMs. Consider a signal x and its reconstruction $y^{(t)}$ up to

time t. To optimally memorizes the history of x using $y^{(t)}$, HiPPO minimizes $\|x_{\leq t} - y^{(t)}\|_{L_2(\omega^{(t)})}$ w.r.t. a measure $\omega^{(t)}$ supported on $(-\infty, t]$. The solution is to project the history of x before time t onto N basis functions (e.g. Legendre polynomials), which yields a time-continuous coefficient vector $\boldsymbol{h}(t)$, and $y^{(t)}$ can be synthesized by linearly combining the N basis using $\boldsymbol{h}(t)$. Gu et al. (2020) shows that the evolution of $\boldsymbol{h}(t)$ follows an ODE $\boldsymbol{h}'(t) = \boldsymbol{A}(t)\boldsymbol{h}(t) + \boldsymbol{b}(t)x(t)$. In particular, Gu et al. (2020) chooses a uniform measure over the past history $\omega^{(t)} = \mathbb{I}[0, t]/t$, which places no approximation bias over the time horizon in contrast to an earlier work (Voelker et al., 2019). As a result, $\boldsymbol{A}(t)$ in Eq. 1 can be written in a closed form: $\boldsymbol{A}(t) = -\boldsymbol{A}_{hippo}/t$, where $\boldsymbol{A}_{hippo}$ is a time-independent constant called the HiPPO matrix. Its various forms have been used as initialization in subsequent works including S4 (Gu et al., 2021a) and Mamba (Gu & Dao, 2023). While HiPPO theory seems to guarantee the long-rangeness for SSMs, the actual form of $\boldsymbol{A}(t)$ employed in S4 and Mamba drops the normalizer $1/t$. Gu et al. (2022b) shows that this change causes a warp of measure from uniform to $\omega^{(t)}(s) \propto \exp(s - t)\mathbb{I}(\infty, t]$. We note that this warped measure assigns more importance to recent history, and thus, our Theorem 3.2 does not contradict HiPPO theory but also matches the findings in Gu et al. (2022b). We also point out that when adopting the diagonalized form of $\boldsymbol{A}_{hippo}$ (Gu et al., 2021a; Gupta et al., 2022; Gu et al., 2022a), the unitary matrices decomposed from $\boldsymbol{A}_{hippo}$ is sometimes not applied to $\boldsymbol{b}_t$ and $\boldsymbol{c}_t$, which introduces a disconnect between HiPPO theory and its practical implementation. Our paper directly studies the discrete-domain SSMs and aligns with the parameterization used in practice.

The effect of selection mechanism. Traditional S4 architectures operate as linear time-invariant systems. To introduce more non-linearity, Mamba (Gu & Dao, 2023) proposes modeling $(\boldsymbol{b}_t, \boldsymbol{c}_t, \Delta_t)$ as a function of inputs, a mechanism known as selection. This is motivated by the selective copying synthetic task, wherein $\boldsymbol{A}_t$ and $\boldsymbol{b}_t$ need to adapt based on content to filter relevant information for memory updates. Despite this adaptation, Theorem 3.2 still holds in scenarios involving the selection mechanism, meaning selective SSMs like Mamba may continue to suffer from recency bias. Nevertheless, we note that selection can alleviate this issue by adaptively controlling the values in $\boldsymbol{A}_t$. Manifesting in our theory, selection mechanism can potentially make the upper bound A_{max} closer to one. However, parameter $\boldsymbol{A}$ in Eq. 3 is initialized with negative integers (Gu et al., 2022a), which exacerbates the bound in Theorem 3.2 by accelerating the decay rate of the influence score. We have empirically visualized the influence score for Mamba in Fig. 1, and the results are as predicted.

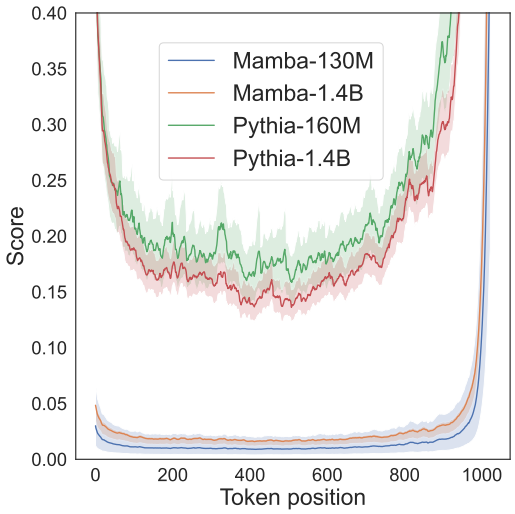


Figure 1: Influential scores.

Is decay a necessary and desirable design? One key observation from our theory is that the parameterization of $\boldsymbol{A}_t$ within the interval $(0, 1)$ leads to strictly decaying dependencies among tokens based on their relative distances. This design choice appears to be a standard practice, and perhaps intentional, in several recently proposed SSMs (Gu & Dao, 2023; Dao & Gu, 2024; Beck et al., 2024; Yang et al., 2023; Peng et al., 2024; De et al., 2024; Ma et al., 2024; Liu et al., 2024a; Yang et al., 2024). It also connects to the gating mechanisms adopted in traditional RNNs (Cho, 2014; Cho et al., 2014; Gu et al., 2021b). We find the constraint $\boldsymbol{A}_t \in (0, 1)^{N \times N}$ might be inherent to SSMs as it is crucial for numerical stability during the long-context recurrence (Gu et al., 2022a; Yang et al., 2023). Promoting the importance of local tokens could also lead to a nearly correct bias, as natural language generation mostly utilizes recent contexts. However, as we will elaborate in subsequent sections, this design can result in significant loss of long-distance information (Sec. 3.3) and may raise certain security concerns (Sec. 3.4).

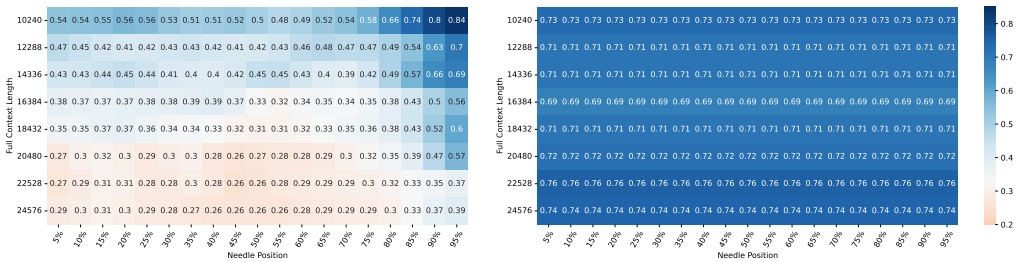


Figure 2: Comparison between SSM and Transformer on the “Needle in a Haystack" benchmark. The left figure shows the retrieval accuracy of the Mamba-Codestral-7B model, while the right figure presents the retrieval accuracy of the Mistral-7B model. We present a heatmap where "full context length" refers to the total length of the document, and "needle position" denotes the relative position of the statement to be retrieved within the context. Best view with zoom-in.

3.3 Lost in the Distance: Long-Context Retrieval Test

To assess the ability of large language models (LLMs) to effectively utilize long-context data, we evaluate open-source SSM using the “Needle in a Haystack" benchmark and compare its performance with that of Transformer. In this benchmark, a randomly generated statement is embedded within the middle of a long document, and the models are tasked with retrieving the statement. By varying the insertion position of the statement, we measure the retrieval accuracy at each location, which reflects the model’s positional bias. To enforce LLMs using the data within context, instead of recalling information memorized by its model weights, we carefully design the statement with factual error. See detailed examples in Appendix D.2.

We compare the retrieval accuracy of the Mamba-Codestral-7B model, a representative SSM capable of handling long-context inputs of up to 256k tokens, with Mistral-7B (Jiang et al., 2023), which utilizes a Transformer architecture. As illustrated in Figure 2, the retrieval accuracy of the Transformer remains stable regardless of the needle position. In contrast, the SSM achieves higher accuracy when the needle is placed closer to the end of the context (i.e., larger needle position values), while its accuracy drops when the needle is located near the beginning of the document. This indicates a positional bias towards local tokens in the SSM.

3.4 Potential Risk on Model Robustness

We conduct quantitative experiments to show the recency-biased nature of SSMs will lead to potential hazards. The downstream task in this study is image classification on sequences of pixels (Tay et al., 2020), where $W \times H$ images are flattened to sequences of pixel tokens and fed to sequence models for classification. We test a family of SSMs, including H3 (Fu et al., 2022), RWKV (Peng et al., 2023), and Mamba (Gu & Dao, 2023), and compare them against a transformer baseline (Vaswani et al., 2017) on CIFAR-10 dataset. To adapt SSMs for this task, we append a learnable class token after the last token of the input sequence. The output state of this class token is then mapped to logits using a classifier head. Experiment details are given in Appendix D.1. In the following, two attack patterns on the input data are introduced, which degrade the robustness of SSMs in this task.

Adversarial Attack. To assess the bias of SSMs towards corrupted data, we perturb the leading and trailing tokens of input sequences with random noise. In unbiased models, perturbations in both

Table 1: Results of adversarial attack experiments on the CIFAR-10 dataset, evaluated using classification accuracy. Each input sequence contains 1,024 tokens. Two corruption ratios ($32/1024$ and $96/1024$) are applied to perturb the leading and trailing tokens, respectively.

		Corrupted region (seq. length = 1024)			
Models	(no corrupt)	[992:1024]	[0:32]	[928:1024]	[0:96]
H3	0.654	0.569 (↓ 13.04%)	0.654 (↓ 0.03%)	0.477 (↓ 27.07%)	0.650 (↓ 0.72%)
Transformer	0.580	0.535 (↓ 7.81%)	0.447 (↓ 22.95%)	0.431 (↓ 25.76%)	0.370 (↓ 36.32%)
RWKV	0.474	0.150 (↓ 68.35%)	0.466 (↓ 1.58%)	0.138 (↓ 70.88%)	0.460 (↓ 2.91%)
Mamba	0.674	**0.126 (↓ 81.24%)**	0.658 (↓ 2.30%)	**0.098 (↓ 85.46%)**	0.647 (↓ 3.98%)

leading and trailing tokens cause similar performance drops. However, in locally biased models, where the class token is appended after the last input token, the trailing tokens are supposed to have greater impacts on classification outcomes than leading tokens. Table 1 presents our experimental results on the CIFAR-10 dataset under two corruption ratios. For each ratio, the same number of leading and trailing tokens are corrupted with Gaussian noise. Among all the SSM family methods compared, the performance drops caused by trailing token corruption are significantly larger than those caused by leading token corruption. Notably, for Mamba, perturbing the last 32 out of 1024 tokens results in an 81.24% drop in classification accuracy, whereas corrupting the first 32 tokens only reduces accuracy by 2.30%. In contrast, the transformer baseline shows relatively smaller impacts from trailing token corruption. Instead, our experiments indicate that more informative features from transformers tend to sink in the leading tokens, aligning with the observations in Xiao et al. (2023).

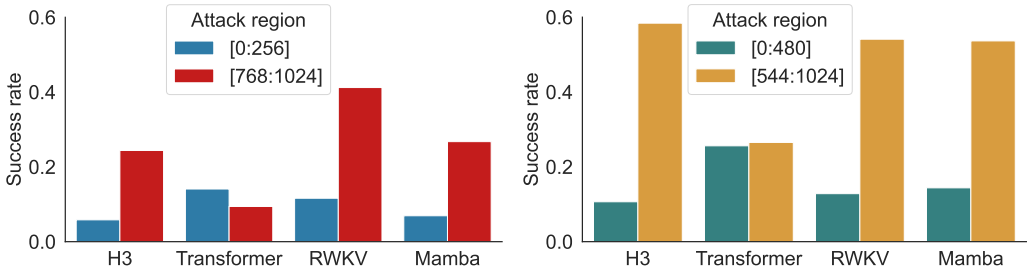
Target Attack. Beyond degrading the performance of SSMs by attacking trailing tokens, we also demonstrate that local bias creates a backdoor for target attacks. In this scenario, a target class is selected, and pixel tokens from that class are used to replace those in images from other classes. The attack succeeds when models mis-classify images from other classes as belonging to the target class. Due to the local bias, trailing tokens are expected to be a more effective attack region for SSMs, leading to a significantly higher attack success rate compared to leading tokens. Fig. 3 shows the success rate comparisons across different attack regions and ratios. When trailing regions are replaced with pixels from the target class, SSMs achieve much higher success rates than when leading regions are attacked. This phenomenon is observed at both 25% and 47% attack ratios. By comparison, the transformer model possesses greater robustness, maintaining similar success rates between attacks on leading and trailing tokens.

4 Understanding Scalability Bottleneck of SSMs

4.1 Necessity and Limits of Depth Scaling

In Sec. 3.1, we have seen that the dependencies between tokens are exponentially decaying with their relative distances in an SSM layer. Consequently, SSMs resemble localized kernels, similar to those employed in various neural architectures such as Convolutional Neural Networks (CNNs) (LeCun et al., 1998) and Graph Neural Networks (GNNs) (Kipf & Welling, 2016). It is a reasonable postulation that increasing the number of layers can extend the model's receptive field (Goodfellow et al., 2016). We justify this hypothesis via a scaling-up experiment with various context lengths and model architectures.

We pretrain Mamba using causal language modeling with two context lengths, $\{2048, 8192\}$. Besides, we fix the number of layers at $\{16, 24, 32, 48\}$ and vary the hidden dimension. We defer more experiment details in Appendix D. The validation loss versus the number of parameters is plotted in Fig. 4. Under the 2048 context length, models of different configurations exhibit similar performance, consistent with the findings of Kaplan et al. (2020). However, as the context length increases, the scaling behavior across depth-width configurations begins to diverge. Notably, deeper models



(a) Attack ratio = $256/1024$ (25.00%) (b) Attack ratio = $480/1024$ (46.875%)

Figure 3: Results of target attack experiments on CIFAR-10, where "horse" is the target class. (a) and (b) present target attack success rates under two attack ratios. Lower success rates suggest higher robustness in the corresponding attack regions.

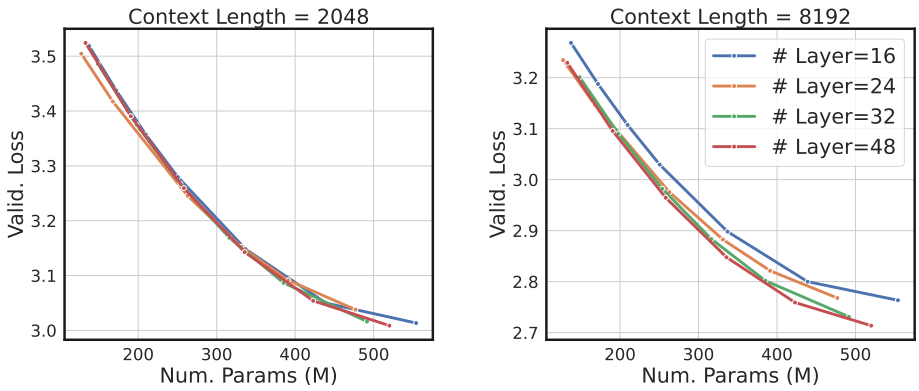


Figure 4: We empirically validate that, with shorter context lengths, models of varying depths but the same total number of parameters show similar performance (left figure). However, deeper models become more advantageous as the context length increases (right figure).

outperform shallower ones, likely because deeper architectures can more effectively utilize the extended context to meet the training objectives. Nevertheless, we observe that the performance gain starts to saturate when we keep increasing the depth.

4.2 Unveiling Over-smoothing in SSMs

To explain the depth scaling bottleneck revealed in the previous section, we conduct a theoretical and empirical investigation of the feature and state dynamics in SSMs. Our key finding is that token embeddings, after being processed by SSM layers, tend to become increasingly similar, which leads to a phenomenon commonly referred to as *over-smoothing* (NT & Maehara, 2021; Cai & Wang, 2020; Oono & Suzuki, 2019). Over-smoothing occurs when token representations become indistinguishable, rendering the state uninformative.

First of all, we warm up by studying continuous-time S4 with constant $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$. Recall that a continuous-time S4 layer can be described by a group of ODEs: $\boldsymbol{h}'(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{b}x(t), \boldsymbol{y}(t) = \boldsymbol{c}^\top \boldsymbol{h}(t)$. Our analysis starts with the equivalence between convolution and S4 (Gu et al., 2021b). This is, the analytic solution to the first-order non-homogeneous ODE can be expressed as $\boldsymbol{y}(t) = \int \boldsymbol{c}^\top \exp(\boldsymbol{A}(t-s))\boldsymbol{b}x(s)ds$. Now we analyze the filtering property of this convolution operator from the Fourier domain perspective. We define a convolutional operator as a low-pass filter if its response on DC component is higher than other spectral components (Wang et al., 2022). We summarize the main finding in the following proposition:

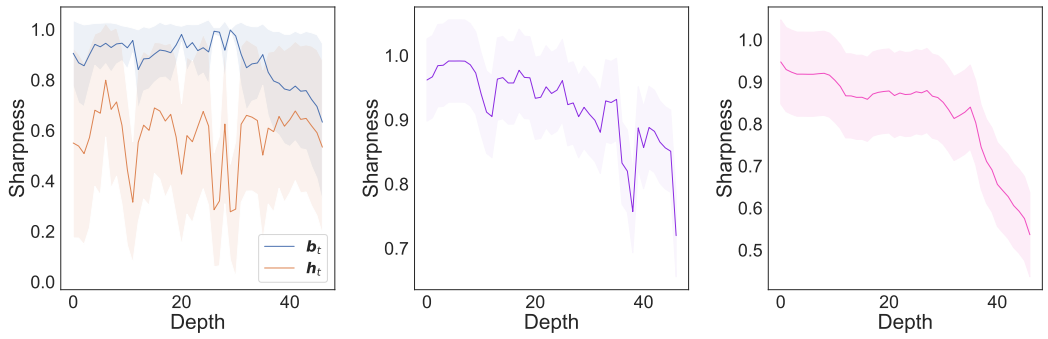
Proposition 4.1 (Low-pass filtering of continuous S4)**.** *Consider a continuous-time S4 with parameters $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$. Assume $\boldsymbol{A}$ is diagonal with all values negative. Then $\boldsymbol{y}(t) = \int \boldsymbol{c}^\top \exp(\boldsymbol{A}(t-s))\boldsymbol{b}x(s)ds$ defines a low-pass filter.*

The assumption therein is always satisfied when real parameterization is adopted (Gu et al., 2022a). Proposition 4.1 states that S4 is inherently a low-pass filter regardless of how $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$ are trained. Therefore, the high-frequency components of input signals are being constantly removed at each layer. Presumably, stacking many S4 layers might cause over-smoothing when all high-frequency components are suppressed to zero.

Now we consider a more general scenario when SSMs work on discrete-time regime and $(\boldsymbol{A}_t, \boldsymbol{b}_t, \boldsymbol{c}_t, \Delta_t)$ are time-varying or even data-dependent. Formally, we prove the following result showing the sharpness of input signals will be reduced as well:

Theorem 4.2 (Over-smoothing of SSMs)**.** *Consider an SSM specified in Eq. 1 with $\{(\boldsymbol{A}_t, \boldsymbol{b}_t, c_t, \Delta_t)\}_{t\in[T]}$. Assume an input space $\mathcal{X} \subset \mathbb{R}^T$ such that for every $\boldsymbol{x} \in \mathcal{X}$, (i) $(\boldsymbol{A}_t)_{n,n} + \Delta_t \leq 1$ for every $n \in [N]$ and $t \in [T]$, (ii) $\min_{t\in[T]} \boldsymbol{b}_t(\boldsymbol{x}_t)_n \leq 0$ and $\max_{t\in[T]} \boldsymbol{b}_t(\boldsymbol{x}_t)_n \geq 0$ for every $n \in [N]$. Let $A_{min} = \min_{t\in[T],n\in[N]}(\boldsymbol{A}_t)_{n,n}$. Then for any $\boldsymbol{x} \in \mathcal{X}$ and the memory states $\{\boldsymbol{h}_t : t \in [T]\}$ generated by the SSM, we have:*

$$\max_{t,s\in[T]} \|\boldsymbol{h}_t - \boldsymbol{h}_s\|_1 \leq \left(1 - A_{min}^{T-1}\right) \max_{t,s\in[T]} \|\boldsymbol{b}_t(\boldsymbol{x}_t) - \boldsymbol{b}_s(\boldsymbol{x}_s)\|_1 , \tag{5}$$



(a) $\boldsymbol{b}_t$ and $\boldsymbol{h}_t$. (b) MambaMixer output. (c) MambaBlock output.

Figure 5: Numerical experiments on Mamba over-smoothing, using a 1.4B pretrained model.

We first justify our assumptions here. $(\boldsymbol{A}_t)_{n,n} + \Delta_t \leq 1$ is a generic condition to ensure the recurrence of SSMs is non-expansive, which is crucial to guarantee memory states stay numerically stable. The second assumption requires the data to be well-distributed and centered around the origin, which can be easily satisfied by normalization techniques. We find that prevalent SSM models such as (Gu & Dao, 2023; Peng et al., 2023; De et al., 2024; Qin et al., 2024) can easily achieve these two assumptions. Moreover, if $(\boldsymbol{A}_t)_{n,n} + \Delta_t = 1$ is always true letting each recurrent update be conservative (Peng et al., 2023; Ma et al., 2022), then we can remove the second assumption as well. Theorem 4.2 examines the relationship between the pairwise distances of memory states and encoded tokens within the sequence. This result indicates that the pairwise discrepancies among memory states are diminished by a factor less than one, suggesting that the memories undergo smoothing following the application of an SSM in Eq. 1. We hypothesize that if the memory is losing its discriminative capacity, the intermediate hidden feature space will similarly collapse.

Delving deeper, the decay rate is intricately linked to both the context length and the minimal value within $\{\boldsymbol{A}_t, t \in [T]\}$. As the context length increases, it requires more time to effectively mix all tokens. When A_{min} approaches one, the decay rate is maximized, as the entire SSM essentially performs a uniform summation over the entire sequence. Interestingly, Theorem 4.2 can be interpreted in conjunction with Theorem 3.2. To relieve the smoothening rate, one might aim to minimize the values in $\boldsymbol{A}_t$. However, this could inadvertently enhance the locality of SSMs, as a decrease in A_{max} may occur. A practical implication of this relationship is that the values in $\boldsymbol{A}_t$ should be as diverse as possible to simultaneously mitigate the artifacts of recency and over-smoothing. Finally, it is worth noting that the smoothing nature of SSMs is intuitive; one can conceptualize the recurrent operation of SSMs as performing a running average of the encoded token signals.

Empirical Validation. We adopt a pairwise distance between tokens to quantify the sharpness of a signal: $\mathcal{E}(\boldsymbol{x}) = \frac{1}{2(N-1)}\big(\sum_{i\neq j}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\big)/\big(\sum_i \|\boldsymbol{x}_i\|_2^2\big)$. $\mathcal{E}(\boldsymbol{x})$ being small means the token representations are close to each other and become less discriminative. In Fig. 5a, $\boldsymbol{b}_t$ is above $\boldsymbol{h}_t$ among all Mamba blocks. This suggests the sharpness of input signals is consistently higher than the sharpness of the memory state output from Mamba, verifying our Theorem 4.2. In addition, Fig. 5b and 5c show the sharpness of Mamba mixer and Mamba block output, which tends to decrease rapidly in deeper layers. This further validates the over-smoothing claims.

5 Conclusion

In this study, we uncover two critical limitations of SSMs. First, we demonstrate that SSMs exhibit a strong recency bias, which significantly impairs their ability to capture long-range dependencies and recall distant information, and even raises robustness concerns. Furthermore, our findings indicate that increasing the depth of SSMs leads to over-smoothing, causing token representations to become indistinguishable and obstructing potential performance gains. We stress the necessity for future research to tackle these challenges, ultimately enhancing the effectiveness of SSMs in natural language tasks.

REFERENCES

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

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*, 2020.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*, 2023.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. In *International Conference on Machine Learning Workshop (ICMLW)*, 2020.

Kyunghyun Cho. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, 2019.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio, and Michael M Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International Conference on Machine Learning*, pp. 7865–7885. PMLR, 2023.

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

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

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487, 2020.

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

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

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022a.

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train your hippo: State space models with generalized orthogonal basis projections. *arXiv preprint arXiv:2206.12037*, 2022b.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*, 2022.

Bo Liu, Rui Wang, Lemeng Wu, Yihao Feng, Peter Stone, and Qiang Liu. Longhorn: State space models are amortized online learners. *arXiv preprint arXiv:2407.14207*, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022.

Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference with unlimited context length. *arXiv preprint arXiv:2404.08801*, 2024.

Hoang NT and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. In *International Conference on Pattern Recognition (ICPR)*, 2021.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations (ICLR)*, 2019.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Can mamba learn how to learn? a comparative study on in-context learning tasks. *arXiv preprint arXiv:2402.04248*, 2024.

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

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, et al. Mechanistic design and scaling of hybrid architectures. *arXiv preprint arXiv:2403.17844*, 2024.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. Hgrn2: Gated linear rnns with state expansion. *arXiv preprint arXiv:2404.07904*, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of NeurIPS*, 2017.

Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. *Advances in neural information processing systems*, 32, 2019.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, et al. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. *arXiv preprint arXiv:2405.18781*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024.

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

A Unified Formulation of SSMs

Gated Linear Attention. We present GLA (Yang et al., 2023) below, which generalizes several extant LAMs:

$$\text{(GLA)} \quad \boldsymbol{A}_t = \operatorname{diag}(\boldsymbol{\alpha}(\boldsymbol{x}_t)), \quad \boldsymbol{b}_t(\boldsymbol{x}_t) = \boldsymbol{k}(\boldsymbol{x}_t), \quad c_t(\boldsymbol{h}_t) = \boldsymbol{q}(\boldsymbol{x}_t)^\top \boldsymbol{h}_t, \quad \Delta_t = v(\boldsymbol{x}_t), \tag{6}$$

where $\boldsymbol{\alpha} : \mathbb{R} \to (0,1)^N$ converts input to gating logits, $\boldsymbol{k}, \boldsymbol{q} : \mathbb{R} \to \mathbb{R}^N$, $v : \mathbb{R} \to \mathbb{R}$ are linear mappings playing roles similar to key, query, value matrices in transformers (Vaswani et al., 2017). When the inputs are multi-dimensional, we share $\boldsymbol{\alpha}$ across channels while assigning each channel with a separate $\boldsymbol{k}, \boldsymbol{q}, v$ and extending their input dimension accordingly. Linear Attention (LA) (Katharopoulos et al., 2020) can be regarded as GLA with constant $\boldsymbol{A}_t$, while RetNet (Sun et al., 2023) can be formulated as GLA with input-independent $\boldsymbol{A}_t$ (Liu et al., 2024a).

Griffin (De et al., 2024). The recurrent unit in Griffin can be re-formulated as a kind of SSMs:

$$\boldsymbol{A}_t = \operatorname{diag}\left(\boldsymbol{\alpha}(\boldsymbol{x}_t)\right), \;\; \boldsymbol{b}_t(\boldsymbol{x}_t) = \operatorname{diag}\left(\boldsymbol{i}(\boldsymbol{x}_t)\right), \;\; c_t(\boldsymbol{h}_t) = \boldsymbol{h}_t, \;\; \Delta_t = \operatorname{diag}\left(\sqrt{1 - \boldsymbol{\alpha}(\boldsymbol{x}_t)^2}\right), \tag{7}$$

where $\boldsymbol{i}(\boldsymbol{x}_t) = \operatorname{sigmoid}(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{b}_x)$ is an input gate, $\boldsymbol{\alpha}$ is computed in log-space: $\log \boldsymbol{\alpha}(\boldsymbol{x}_t) = -\xi \;\operatorname{softplus}(\boldsymbol{\Gamma}) \odot \operatorname{sigmoid}(\boldsymbol{W}_a \boldsymbol{x}_t + \boldsymbol{b}_a)$, $\odot$ is Hadamard product, ξ is a constant, and $\boldsymbol{\Gamma}$, $\boldsymbol{W}_a$, $\boldsymbol{b}_a$, $\boldsymbol{W}_x$, $\boldsymbol{b}_x$ are learnable parameters. In particular, the dimension of $\boldsymbol{h}_t$ in Griffin is equal to the dimension of $\boldsymbol{x}_t$. If we consider single-channel $\boldsymbol{x}_t$, then $\boldsymbol{A}_t$, $\boldsymbol{b}_t$, and Δ_t are all scalar-valued.

RetNet (Sun et al., 2023). The SSM formulation of RetNet is similar to that of GLA, with the distinction of $\boldsymbol{A}_t$.

$$\boldsymbol{A}_t = \gamma \boldsymbol{I}, \quad \boldsymbol{b}_t(\boldsymbol{x}_t) = \boldsymbol{k}(\boldsymbol{x}_t), \quad c_t(\boldsymbol{h}_t) = \boldsymbol{q}(\boldsymbol{x}_t)^\top \boldsymbol{h}_t, \quad \Delta_t = v(\boldsymbol{x}_t), \tag{8}$$

where $\gamma \in [0,1]$ is a scalar, and the other symbols retain the same meaning as in GLA.

RWKV (Peng et al., 2023). We demonstrate that RWKV can also be reformulated into the structure of SSMs:

$$\boldsymbol{A}_t = \frac{\exp(-w_t)\boldsymbol{I}}{\exp(-w_t) + \exp(k(\boldsymbol{x}_t))}, \quad \boldsymbol{b}_t(\boldsymbol{x}_t) = \boldsymbol{v}(\boldsymbol{x}_t), \tag{9}$$

$$c_t(\boldsymbol{h}_t) = \boldsymbol{q}(\boldsymbol{x}_t)^\top \boldsymbol{h}_t, \quad \Delta_t = \frac{\exp(k(\boldsymbol{x}_t))}{\exp(-w_t) + \exp(k(\boldsymbol{x}_t))}. \tag{10}$$

Another view suggests RWKV can be seen as a state of ratio form of two SSMs: $\boldsymbol{h}_t = \frac{\boldsymbol{a}_t}{\boldsymbol{b}_t} = (\exp(-w)\boldsymbol{a}_{t-1} + \exp(k(\boldsymbol{x}_t))v(\boldsymbol{x}_t)) \,/\, (\exp(-w)\boldsymbol{b}_{t-1} + \exp(k(\boldsymbol{x}_t)))$.

B Deferred Discussions

Does hungry hungry hippos help? The key innovation of Hungry Hungry Hippos (H3) (Fu et al., 2022) lies in the introduction of self-gating connections and locally shifting convolutions to improve in-context recall for state space models (SSMs). This design has quickly become a standard backbone for various SSMs (Gu & Dao, 2023; Beck et al., 2024). However, we question its effectiveness in addressing the local rangeness issue in SSMs. The gating mechanism operates at the token level, which impacts the bound in Theorem 3.2 only by a constant factor. Additionally, the introduced convolutions typically use small kernels, which are insufficient to mitigate the exponentially decaying relevance between tokens. As we empirically show in Fig. 2, while Mamba with H3 performs adequately in associative recall tasks when the state size is sufficiently large, a locality bias begins to emerge as the number of key-value pairs exceeds the model's memory capacity. This highlights the limitations of the architecture in handling long-range dependencies under constrained memory.

Connection with over-squashing theory in GNNs. The influential score defined in Sec. 3.1 is also used for *over-squashing* analysis in Graph Neural Networks (GNNs) to identify information bottleneck (Wu et al., 2020; Topping et al., 2021; Di Giovanni et al., 2023). We postulate that propagating information from long distances remains challenging for SSMs because the model needs to encapsulate all history information into a fixed-dimension hidden vector, which is also observed as one major problem with RNNs (Alon & Yahav, 2020; Sutskever et al., 2014; Cho et al., 2014; Cho, 2014).

C PROOFS

C.1 PARALLEL FORM OF SSMS

Proof of Lemma 3.1. By simple expansion of Eq. 1:

$$\begin{aligned}
\boldsymbol{h}_t &= \boldsymbol{A}_t \left(\boldsymbol{A}_{t-1} \boldsymbol{h}_{t-2} + \Delta_{t-1} \boldsymbol{b}_{t-1}(\boldsymbol{x}_{t-1}) \right) + \Delta_t \boldsymbol{b}_t(\boldsymbol{x}_t) \\
&= \boldsymbol{A}_t \left(\boldsymbol{A}_{t-1} (\boldsymbol{A}_{t-2} \boldsymbol{h}_{j-3} + \Delta_{t-2} \boldsymbol{b}_{t-2}(\boldsymbol{x}_{j-2})) + \Delta_{t-1} \boldsymbol{b}_{t-1}(\boldsymbol{x}_{t-1}) \right) + \Delta_t \boldsymbol{b}_t(\boldsymbol{x}_t) \\
&= \cdots \\
&= \left(\prod_{k=2}^{t} \boldsymbol{A}_k \right) \Delta_1 \boldsymbol{b}_1(\boldsymbol{x}_1) + \left(\prod_{k=3}^{t} \boldsymbol{A}_k \right) \Delta_2 \boldsymbol{b}_2(\boldsymbol{x}_2) + \cdots + \boldsymbol{A}_t \Delta_{t-1} \boldsymbol{b}_{t-1}(\boldsymbol{x}_{t-1}) + \Delta_t \boldsymbol{b}_t(\boldsymbol{x}_t).
\end{aligned}$$

Proved after rewriting the summation and applying $\boldsymbol{c}_t$ to the left. □

C.2 LOCALITY OF SSMS

Proof of Theorem 3.2. By the assumption that $\boldsymbol{x}_t$ is uniformly bounded, we have for some $C_\Delta, C_{\Delta'}, C_B, C_{B'}, C_c > 0$

$$C_\Delta \leq \Delta(\boldsymbol{x}_i) \leq 1, |\Delta'(\boldsymbol{x}_i)| \leq C_{\Delta'}, |B(\boldsymbol{x}_i)| \leq C_B, |B'(\boldsymbol{x}_i)| \leq C_{B'}, |C(\boldsymbol{x}_i)| \leq C_c,$$

for every $i \in [N]$ due to the continuity of Δ, B, C and their derivatives. Next, by Eq. 4 and substituting $A(\boldsymbol{x}_i) = \exp(-\Delta_i \boldsymbol{A})$, we have for every $i, j \in [N]$ and $i < j$:

$$\begin{aligned}
\boldsymbol{y}_j &= \boldsymbol{c}(\boldsymbol{x}_j)^\top \left[\sum_{t=1}^{j-1} \exp\left(- \sum_{k=t+1}^{j} \log \boldsymbol{A}_k \right) \Delta_t \boldsymbol{b}(\boldsymbol{x}_t) + B(\boldsymbol{x}_j) \right] \\
&= \boldsymbol{c}(\boldsymbol{x}_j)^\top \left[\sum_{t=1}^{i-1} \exp\left(- \sum_{k=t+1}^{j} \log \boldsymbol{A}_k \right) \Delta_t \boldsymbol{b}(\boldsymbol{x}_t) \right] \\
&\quad + \boldsymbol{c}(\boldsymbol{x}_j)^\top \left[\sum_{t=i}^{j-1} \exp\left(- \sum_{k=t+1}^{j} \log \boldsymbol{A}_k \right) \Delta_t \boldsymbol{b}(\boldsymbol{x}_t) + B(\boldsymbol{x}_j) \right].
\end{aligned}$$

Here we obtain a decomposition of $\boldsymbol{y}_j$ where B in the first summation is independent of $\boldsymbol{x}_i$ while Δ in the second summation is independent of $\boldsymbol{x}_i$. Taking derivatives in terms of each summation, respectively:

$$\begin{aligned}
\frac{\partial \boldsymbol{y}_j}{\partial \boldsymbol{x}_i} &= \sum_{d=1}^{D} C(\boldsymbol{x}_j)_d \sum_{t=1}^{i-1} - \exp\left(- \sum_{k=t+1}^{j} \log \boldsymbol{A}_{k,d,d} \right) \boldsymbol{A}'_{d,d}(\boldsymbol{x}_t) \Delta_t \boldsymbol{b}(\boldsymbol{x}_t)_d \\
&\quad + \sum_{d=1}^{D} C(\boldsymbol{x}_j)_d \exp\left(- \sum_{k=i+1}^{j} \log \boldsymbol{A}_{k,d,d} \right) B'(\boldsymbol{x}_t)_d
\end{aligned}$$

Similarly, we can obtain an upper bound:

$$\begin{aligned}
\left|\frac{\partial \boldsymbol{y}_j}{\partial \boldsymbol{x}_i}\right| &\leq \sum_{d=1}^{D}\sum_{t=1}^{i-1} C_c \exp\left(-\sum_{k=t+1}^{j}\log \boldsymbol{A}_{k,d,d}\right)\boldsymbol{A}'_{d,d}(\boldsymbol{x}_t)|\Delta_t \boldsymbol{b}(\boldsymbol{x}_t)_d| \\
&\quad + \sum_{d=1}^{D} C_c \exp\left(-\sum_{k=i+1}^{j}\log \boldsymbol{A}_{k,d,d}\right)|B'(\boldsymbol{x}_t)_d| \\
&\leq \sum_{d=1}^{D} C_c \sum_{t=1}^{i-1} \exp\left(-\boldsymbol{A}_{max} C_\Delta (j-t)\right)\log \boldsymbol{A}'_{d,d}(\boldsymbol{x}_t)|\Delta_t \boldsymbol{b}(\boldsymbol{x}_t)_d| \\
&\quad + \sum_{d=1}^{D} C_c \exp\left(-\boldsymbol{A}_{max} C_\Delta (j-i)\right)|B'(\boldsymbol{x}_t)_d| \\
&\leq \sum_{d=1}^{D} \frac{\exp\left(A_{max} C_\Delta (i-1)-1\right)}{\exp\left(\boldsymbol{A}_{d,d} C_\Delta\right)-1}\exp\left(-\boldsymbol{A}_{d,d} C_\Delta (j-1)\right) C_B C_{\Delta'} C_c \\
&\quad + \sum_{d=1}^{D} \exp\left(A_{max} C_\Delta (j-i)\right) C_c C_{B'} \\
&\leq C\exp\left(-\kappa(j-i)\right).
\end{aligned}$$

where $k = \Theta(-\log A_{max})$. In the first inequality, we adopt triangle inequality, in the second inequality we apply a lower bound on $\Delta(\boldsymbol{x}_t)$, in the third inequality, we take the summation of the geometric series and apply upper bounds on corresponding terms, and in the last inequality, we merge all constants together with $\boldsymbol{A}$ as one universal constant C. □

C.3 Oversmoothing in SSMs

Proof of Proposition 4.1. First of all, let us rewrite the filter expressed by S4 as below:

$$z(t) = \sum_{n=1}^{N} \boldsymbol{c}_n \boldsymbol{b}_n \exp(\boldsymbol{A}_{n,n} t). \tag{11}$$

The proof can be done by applying Fourier transform on $z(t)$:

$$\begin{aligned}
Z(\omega) &= \int z(t) e^{-\sqrt{-1}\omega t} dt && (12)\\
&= \int \sum_{n=1}^{N} \boldsymbol{c}_n \boldsymbol{b}_n \exp(\boldsymbol{A}_{n,n} t)\exp(-\sqrt{-1}\omega t) dt && (13)\\
&= \sum_{n=1}^{N} \boldsymbol{c}_n \boldsymbol{b}_n \int \exp(\boldsymbol{A}_{n,n} t)\exp(-\sqrt{-1}\omega t) dt && (14)\\
&= \sum_{n=1}^{N} \boldsymbol{c}_n \boldsymbol{b}_n \int \exp((\boldsymbol{A}_{n,n} - \sqrt{-1}\omega) t) dt && (15)\\
&= \sum_{n=1}^{N} \frac{\boldsymbol{c}_n \boldsymbol{b}_n}{\sqrt{-1}\omega - \boldsymbol{A}_{n,n}}, && (16)
\end{aligned}$$

where the last integral converges due to $\boldsymbol{A}_{n,n} < 0$. □

Proof of Theorem 4.2. To simplify the proof, we first only focus on one dimension in the memory state. We denote $\alpha_t = \boldsymbol{A}_{t,n,n}$, $z_t = \boldsymbol{b}_{t,n}\boldsymbol{x}_t$, , $s_t = \boldsymbol{h}_{t,n}$ for some $n \in [N]$ Let $m = \min_{t\in[T]} z_t$, $M = \max_{t\in[T]} z_t$. By assumptions, we know that $m \leq 0$ and $M \geq 0$. Suppose $z_1 = pm + (1-p)M$ for some $p \in [0,1]$, and $s_t = \alpha_t s_{t-1} + \Delta_t z_t$ with $s_0 = 0$. Furthermore, let $q = 1 - p$.

We provide the following two lemmas:

Lemma C.1 (Wu et al. (2024)). *Suppose* $\alpha_t + \Delta_t \leq 1$ *and* $z_1 = pm + (1-p)M$ *for some* $p \in [0,1]$, *then* $s_n \leq A_{min}^{n-1} pm + (1 - A_{min}^{n-1} p)M$.

Lemma C.2 (Wu et al. (2024)). *Suppose* $\alpha_t + \Delta_t \leq 1$ *and* $z_1 = (1-q)m + qM$ *for some* $q \in [0,1]$, *then* $s_n \geq (1 - A_{min}^{n-1} q)m + A_{min}^{n-1} qM$.

By these two lemmas, we have:

$$(1 - A_{min}^{T-1} q)m + A_{min}^{T-1} qM \leq s_L \leq A_{min}^{T-1} pm + (1 - A_{min}^{T-1} p)M. \tag{17}$$

Thus we can conclude that:

$$m' = \min_{t \in [T]} s_t \leq (1 - A_{min}^{T-1} q)m + A_{min}^{T-1} qM \tag{18}$$

$$M' = \max_{t \in [T]} s_t \geq A_{min}^{T-1} pm + (1 - A_{min}^{T-1} p)M. \tag{19}$$

Henceforth, we can upper bound:

$$M' - m' \leq A_{min}^{T-1} pm + (1 - A_{min}^{T-1} p)M - (1 - A_{min}^{n-1} q)m - A_{min}^{n-1} qM \tag{20}$$

$$= (1 - A_{min}^{T-1})(M - m), \tag{21}$$

using the fact that $p + q = 1$. Now we can generalize the above upper bound to all state channels while assigning each (M, m) and (M', m') with a subscript. This yields:

$$\max_{t,s \in [T]} \|\boldsymbol{h}_t - \boldsymbol{h}_s\|_1 \leq \sum_n (M'_n - m'_n) \tag{22}$$

$$\leq \left(1 - A_{min}^{T-1}\right) \sum_n (M_n - m_n) \tag{23}$$

$$\leq \max_{t,s \in [T]} \|\boldsymbol{b}_t(\boldsymbol{x}_t) - \boldsymbol{b}_s(\boldsymbol{x}_s)\|_1 , \tag{24}$$

□

D Experiment Details

D.1 CIFAR-10 Image Classification

Here we present experiment details in Sec. 3.4, where we conduct image classification on the CIFAR-10 dataset to study locality bias in SSMs. Specifically, 32×32 RGB images in the dataset are flattened into sequences with a shape of $(1024, 3)$, where 1024 represents the sequence length and 3 corresponds to the RGB channels of the pixel tokens. These pixel tokens are then projected into H-dimensional features via a linear projection, which are then input into SSM or transformer mixers. In addition to pixel tokens, we insert a class token at the last position of the input sequence. The output state of the class token will be processed by a one-layer classifier head to generate the final logits.

Note that while the ViT architecture (Dosovitskiy et al., 2020) places the class token at the first position of the input sequence, this design is incompatible with SSMs, which rely on causal sequence modeling. In SSMs, the class token must be positioned last to aggregate features from the entire sequence. We position the class token as the last token to establish long-range dependencies between global image features and the leading pixel tokens. Alternative methods for aggregating features across the entire sequence, such as mean pooling (Gu et al., 2021a; Tay et al., 2020) or placing the class token in the middle of the sequence (Zhu et al., 2024), work more robustly in general but do not fit the needs for our arguments on locality.

In addition, our image classification setup differs from Tay et al. (2020), where an 8-bit pixel intensity lookup table is used as the token embedder. Instead, we employ a linear projection to map RGB pixel colors into H-dimensional features.

For a fair comparison, the same hyperparameters are used across all models: learning rate = 0.001, weight decay = 0.1, number of layers = 3, feature dimension $H = 32$, and number of states = 64. Each model is trained for 100 epochs. The models and training pipelines are built on Arora et al. (2023). No perturbations are imposed on the input sequences in the training stage.

Table 2: Extended results of adversarial attack experiments on the CIFAR-10 dataset. Classification accuracy is used as the metric.

	Corrupted region (seq. length = 1024)						
Models	(no corrupt)	[1014:1024]	[0:10]	[768:1024]	[0:256]	[512:544]	[480:576]
H3	0.654	0.629	0.654	0.394	0.639	0.603	0.543
Transformer	0.580	0.571	0.500	0.249	0.263	0.498	0.347
RWKV	0.474	**0.194**	0.470	**0.107**	0.448	0.405	0.392
Mamba	0.674	0.348	0.664	**0.099**	0.597	0.515	0.446

Adversarial Attack. To introduce perturbations to test data for adversarial attack, we first define a corruption length K, which is small relative to the entire sequence length. We then replace the leading and trailing K tokens with random Gaussian noise. In our experiments, K is set to 32 and 96, corresponding to one row and three rows of pixels, respectively. Table 2 shows more results under other corruption regions.

Target Attack. For the target attack experiments, a target class is first selected. For each image from the other classes, an image from the target class is randomly sampled, and its leading and trailing pixels are used to replace the corresponding pixels in the original image. We test two attack ratios: 256/1024 and 480/1024. Replacing fewer than 256 pixels generally does not result in considerable success rates based on our trials. In our main text, we show success rates when "horse" is the target class. Similar patterns are also observed across other classes. Fig. 6 shows the average success rates obtained by setting each class as the target.

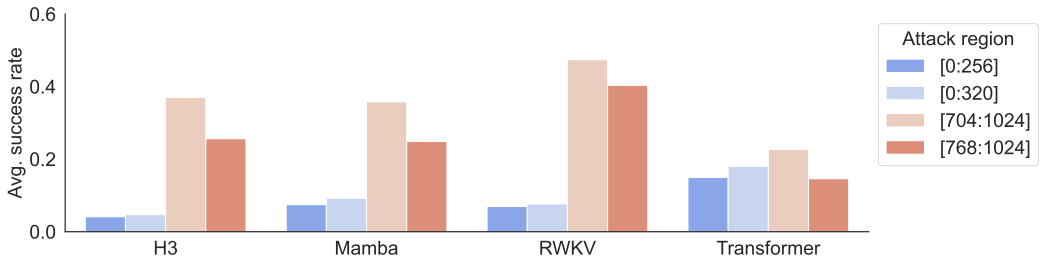


Figure 6: Overall success rate of our target attack experiments on CIFAR-10, calculated by averaging the attack success rates obtained when each class is individually set as the target class.

D.2 Experimental Details of Needle Test

To validate the retrieval capability of the models while preventing them from relying on memorized information stored in their model weights, we carefully design the inserted statements to contain factual errors. Several examples of such statements are provided in Figure 7. For instance, we insert the statement, "The capital of France is Madrid," and then test the model's retrieval ability by asking the question, "What is the capital of France?" While the correct answer, Paris, is likely memorized by the LLM, if the model "correctly" outputs Madrid based on the context provided, it demonstrates that the model is successfully using the contextual information rather than relying on pre-existing knowledge. This approach ensures that the evaluation focuses on the model's ability to retrieve and process information from the input context.

Inserted Statement: The capital of France is Madrid.

Inserted Statement: The telephone was invented by Elon Musk.

Inserted Statement: The largest planet in our solar system is Earth.

Inserted Statement: The country known for the Great Wall is Brazil.

Question: What is the capital of France?

Question: Who invented the telephone?

Question: What is the largest planet in our solar system?

Question: What country is known for the Great Wall?

Figure 7: An illustration of our synthetic data.