Sparse Long Memory and Transition Bottlenecks Shape Positional Recall in Mamba Language Models

Anonymous ACL submission

Abstract

Mamba language models process sequences recurrently, but recurrent processing does not imply that all tokens are retained equally. We study this gap using structured recall tasks that expose primacy, recency, and lost-in-the-middle behavior. Our results suggest that Mamba uses two functional memory regimes: a short-lived state that supports recency, and a sparse long-lived channel subspace that preserves selected early inputs. Targeted interventions on this long-memory subspace selectively degrade early recall and transfer to long-context QA and summarization, indicating that the same channels support real long-context use. We then ask why middle items are forgotten despite being processed sequentially. We find that the central failure is a short-to-long memory transition bottleneck: middle items are no longer recent enough for short memory, but are often not promoted into the sparse long-lived subspace. Repeated input structure strengthens this bottleneck by changing the input-dependent discretization dynamics. These findings provide a mechanistic account of Mamba's U-shaped recall curve as the interaction between long memory, short memory, and a fragile transition between them.

1 Introduction

Primacy and recency, the tendencies to recall items presented at the beginning and end of a sequence more accurately than those in the middle, are long-established findings in cognitive science (Ebbinghaus, 1913; Atkinson and Shiffrin, 1968). In humans, primacy is often attributed to better encoding of initial items through greater rehearsal opportunity, while recency is explained by short-term memory, where new learning overwrites old learning (Greene et al., 2000).

These position-based recall phenomena have been well studied in the context of the traditional transformer architecture. Janik (2023) showed empirically that tokens at the beginning and end of a sequence are recalled more reliably than those in the middle. According to Wu et al. (2025), this asymmetry can be understood as a consequence of mechanistic features of the transformer design. In particular, causal masking draws attention toward earlier tokens, while relative positional encodings such as RoPE (Su et al., 2024) emphasize recent ones.

Recently, state-space language models (SSMs) have begun to garner significant research interest, due to their ability to scale more efficiently with sequence length compared to transformers, and offering efficient memory complexity for long-context inputs (Gu and Dao, 2023; Lv et al., 2025). Yet despite this progress, evidence of position-dependent effects in this family is still fragmentary. Primacy has been documented in the Structured State Space sequence model (S4) (Gu et al., 2021; Morita, 2025), while recency has been studied in Mamba, specifically linked to exponential-decay dynamics (Wang et al., 2025). These observations suggest that state-space models, despite relying on recurrent updates rather than attention, are also subject to systematic positional effects, but prior work has not explained how early, middle, and recent items interact within one memory mechanism.

Importantly, explanations for primacy in transformers, where early-token recall improves with depth due to causal masking and exponential propagation of information (Wu et al., 2025), are not exclusive to attention-based models. SSMs also impose a causal structure through recurrent updates, with gates applied both to inputs and to states. This makes them natural candidates for a memory-regime account: recent items can be recovered from ordinary short-lived recurrent state, while early items survive only if they are routed into a sparse long-lived subspace. The difficult region is the middle. Middle items have already

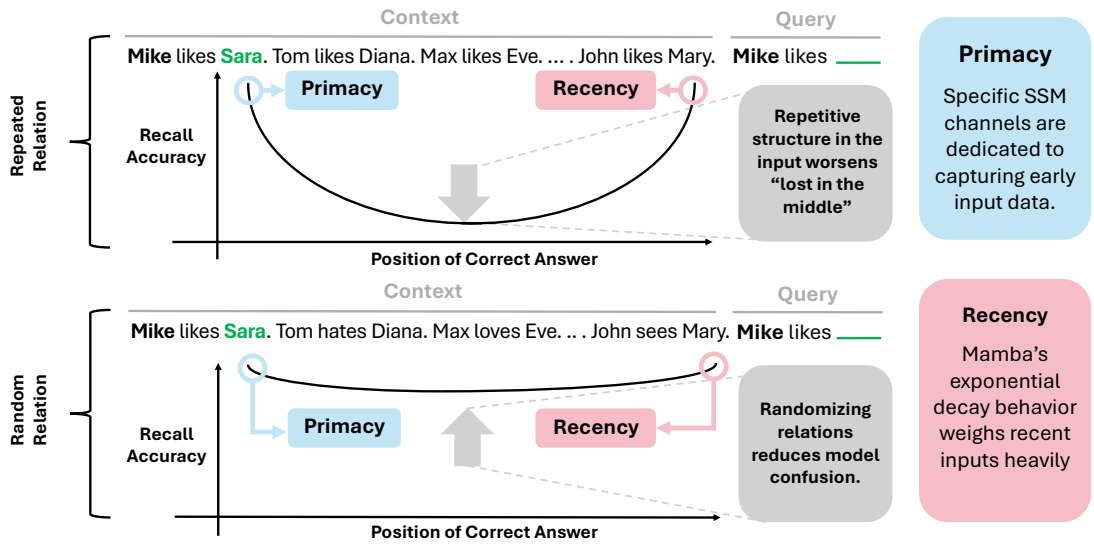


Figure 1: **Illustration of the structured recall task with two variants: repeated relation (top) and random relation (bottom).** The task measures Mamba's recall accuracy across sequence positions. **Repeated relation** reuses the same predicate (*likes*), producing strong **primacy** and **recency** effects, with a drop in middle recall. We interpret early success as long-memory retention and late success as short-memory persistence. Repetition increases positional confusion and makes the short-to-long transition more fragile, worsening the "lost-in-the-middle" effect. **Random relation** breaks one repeated cue and improves middle recall, but the repeated `s-r-o` format remains, so the U-shape persists.

moved out of the short-memory window, but they are not reliably promoted into long memory. We argue that this short-to-long memory transition is the key mechanism behind the lost-in-the-middle effect in Mamba.

The transition is controlled in part by the learned discretization step Δ_t, which modulates both recurrence and input injection. Our results show a clear increase in Δ_t as the period length grows, with more random or lower-frequency inputs inducing larger Δ_t. In the structured recall task, however, even the random-relation setting inherits the repeated `s-r-o` template. Thus random relations reduce one source of interference, but they do not remove the task's periodic structure entirely; the U-shape remains across relation types, with repetition mainly deepening the middle failure. In this work we demonstrate how targeted channel ablations, state initialization, and periodic inputs modulate Mamba's temporal behavior. Our main contributions are:

- We show that Mamba separates recall into short- and long-memory regimes. Recent items are supported by transient recurrent state, whereas early items depend on a sparse set of long-lived channels[1] whose cumulative recurrence remains stable across the full context (Section 4.2, Figure 3).
- We causally localize Mamba's long-memory subspace. Targeted interventions on high-retention channels selectively impair early-position recall and degrade long-context QA and summarization, showing that these channels are not only artifacts of the synthetic task (Section 4.3, Table 1). We find an analogous long-term head structure in Mamba 2 (Appendix E.2).
- We identify the lost-in-the-middle region as a short-to-long memory transition failure. Although all tokens update Mamba's recurrent state, middle items fail when they are too old for short memory but were not promoted strongly enough into long memory. State-initialization experiments show that this transition region is dynamically adjustable (Section 4.2, Appendix F).
- We show that repetition strengthens this bottleneck. The learned discretization gate Δ_t increases with period length, and its coupling with the A and B matrices changes the balance between recurrent retention and input

[1] "Channels" refer to individual state dimensions within the selective state-space block, roughly analogous to embedding dimensions in a transformer.

writing. This provides a mechanism for why repeated structure deepens the bottom of the U-shape (Section 4.4, Figure 7).

- For completeness, we revisit recency under the known exponential-decay view of Mamba. Distractors markedly reduce recent-position recall, confirming that recency is a fragile short-memory effect rather than a fixed guarantee (Section 4.5, Figure 5b).

2 Related Work

Transformer-based language models have been shown to exhibit phenomena such as primacy and recency effects when evaluated on structured recall tasks (Janik, 2023). These effects emerge due to architectural asymmetries introduced by causal masking, learned positional embeddings, and attention dynamics (Wu et al., 2025). A key mechanism behind recall in transformers is the *induction head*, which allows the model to learn copy operations by attending to repeated patterns in the input (Olsson et al., 2022).

State space models (SSMs) process sequences through recursive updates to a fixed-size hidden state, enabling linear-time operation. Unlike transformers, they do not rely on token-to-token interactions and are therefore biased toward recent inputs, a recency effect formally analyzed in smoothing studies (Wang et al., 2025; Jelassi et al., 2024).

Primacy has also been observed in SSMs. Morita (2025) show that S4 models can exhibit strong primacy depending on how the discretization step $\mathbf{\Delta}_t$ is initialized. In time-invariant SSMs such as S4, $\mathbf{\Delta}_t$ defines the effective frequency spectrum the model can capture, a connection further explored from autocorrelation-based perspectives viewing SSMs as spectral filters (Liu and Li, 2024).

An illustrative case is S4D's diagonal initialization (Gu et al., 2022), where the continuous-time state matrix uses complex values $\Lambda_n = -\frac{1}{2} + 2\pi i n$, which after discretization yields eigenvalues $\lambda_n = \exp(\mathbf{\Delta}_t \Lambda_n)$, setting the system's poles in the Z-domain. Pole magnitudes govern decay, while imaginary components define resonant frequencies. Thus $\mathbf{\Delta}_t$ directly controls frequency coverage and effective scale of information retention.

Unlike time-invariant models such as S4, Mamba (Gu and Dao, 2023) updates its state through input-dependent functions, including a learned discretization parameter $\mathbf{\Delta}_t$. This breaks many assumptions of a standard time-invariant setting. Although time-variant, $\mathbf{\Delta}_t$ still acts as a forgetting gate: smaller values extend information retention, while larger values accelerate decay and control input integration. Repeated or periodic inputs bias $\mathbf{\Delta}_t$ toward faster decay, an effect we explore in Section 4.4.

Recent work such as LongMamba (Ye et al., 2025) takes a performance-oriented view: it distinguishes local and global channels and applies a training-free filtering mechanism to enlarge Mamba's receptive field and improve long-context performance. In contrast, our study takes a mechanistic view, systematically characterizing the joint primacy and recency effects in Mamba. We show for the first time that Mamba exhibits a U-shaped recall profile and link these positional effects to identifiable architectural components.

3 Model Architecture and Task

3.1 Mamba Architecture

Mamba is built on state space models (SSMs), which maintain structured hidden states $\mathbf{h}_t \in \mathbb{R}^{d\times N}$ that evolve over time through recurrent updates. Each input dimension $i \in \{1, \ldots, d\}$ of the embedding vector $\mathbf{x}_t \in \mathbb{R}^d$ is processed independently by a dedicated selective SSM block, enabling token-to-token interaction through the accumulation of information over time within the recurrent state dynamics. The state-update dynamics for each dimension are given by:

$$\begin{aligned} \mathbf{h}_t^{(i)} &= \mathbf{A}_t^{(i)} \mathbf{h}_{t-1}^{(i)} + \mathbf{B}_t^{(i)} x_t^{(i)}, \\ y_t^{(i)} &= \mathbf{C}_t \mathbf{h}_t^{(i)} + D^{(i)} x_t^{(i)}, \end{aligned} \tag{1}$$

where:

- $\mathbf{h}_t^{(i)} \in \mathbb{R}^N$ is the hidden state for dimesion i,
- $\mathbf{A}_t^{(i)} \in \mathbb{R}^{N\times N}$ is the recurrence matrix or the forget gate (diagonal in Mamba; possibly full or low-rank in models such as S4[2]),
- $\mathbf{B}_t^{(i)} \in \mathbb{R}^N$, $\mathbf{C}_t \in \mathbb{R}^N$ are the input and output projection vectors (gates), respectively. $D^{(i)} \in \mathbb{R}$ is a residual weight,

[2]In S4D (Gu et al., 2021), $\mathbf{A}$ is diagonal with entries initialized in the complex plane; $\mathbf{B}$ and $\mathbf{C}$ are also complex-valued, and are discretized via the bilinear transform. These variants differ in expressivity, stability, and frequency selectivity.

- $x_t^{(i)} \in \mathbb{R}$ is the i-th dimension of the input at time t.

These models act as implicit convolutional filters in the frequency domain, managing long-range dependencies via kernel design rather than explicit attention. Mamba adds a learned per-token discretization parameter $\mathbf{\Delta}_t$ that dynamically modulates recurrence, adjusting how information is integrated or decayed with input structure. We conduct our experiments on *Falcon Mamba 7B* (Zuo et al., 2024) and validate on *Mamba 1.4B* (Gu and Dao, 2023) to ensure consistency across model sizes.

3.2 Structured Recall Task

To investigate how Mamba distributes and retains across time, we design a structured recall task inspired by serial position experiments in cognitive psychology (Murdock and Bennet, 1962). Each input sequence consists of L subject–relation–object (`s-r-o`) triples, followed by a query targeting the object at a specific position in the sequence. The input fed to the model is structured as follows:

Context: $\mathtt{s}_1\ \mathtt{r}_1\ \mathtt{o}_1.\quad \cdots\quad \mathtt{s}_L\ \mathtt{r}_L\ \mathtt{o}_L.$
Query: $\mathtt{s}_k\ \mathtt{r}_k\ (1 \leq k \leq L)$

The model is expected to generate the correct $\mathtt{o}_k$ for the k-th triplet in the original sequence. To ensure consistent tokenization, we build the dataset from subject, relation, and object terms, each represented by a single unique token in the model's vocabulary, with all subjects and objects distinct within a sequence. For each target position k, we create 50 unique input sequences and report average accuracy to analyze recall performance across positions (details in Appendix A).

To probe position-related effects, we vary (i) **sequence length**, testing $L \in \{8, 16, 32, 64, 128\}$ triples to examine how recall changes with context size, and (ii) **relation type**, comparing **Repeated Relation** (all relation tokens identical, e.g., `Mike likes Sara. Tom likes Diana ... Mike likes`) versus **Random Relation** (relations uniquely sampled from a fixed vocabulary, introducing input diversity, e.g., `Mike likes Sara. Tom hates Diana ... Mike likes`). The random-relation setting removes repeated predicates but still preserves the repeated `s-r-o` scaffold, which lets us separate predicate-level repetition from the broader periodic structure inherited from the task.

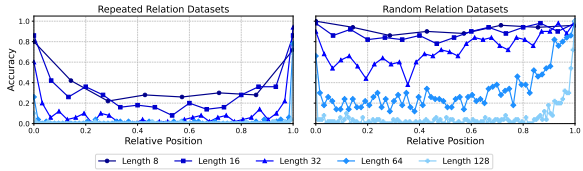


Figure 2: Recall accuracy of Falcon Mamba 7B as a function of position k in the input sequence. Accuracy peaks at the beginning and end, forming a U-shaped curve characteristic of primacy and recency effects.

4 Experimental Results and Findings

4.1 Characterizing Primacy and Recency Effects in Mamba

Across sequence lengths and relation types, we consistently observe a U-shaped recall curve: higher accuracy at early and late positions with a dip in the middle (Figure 2). We interpret this not only as a serial-position effect, but as evidence for two successful memory regimes and one failure regime. Early items survive through a long-memory pathway, recent items survive through short-lived recurrent state, and middle items fall in the transition region: they are no longer recent enough for short memory but have not reliably entered the long-lived subspace. The curve also depends on input structure. Random relations yield better mid-position performance than repeated ones, whereas repetition produces a sharper U-shape with a deeper middle drop. Since both variants still share the repeated `s-r-o` scaffold, the U-shape persists across relation types; changing the relation only weakens one source of periodic interference. Having established this behavioral profile, we next localize the long-memory pathway that supports early recall.

4.2 Long Memory: Sparse Subset of Channels and Causal Intervention

To understand how early items escape short-memory decay, we hypothesize that certain channels in the selective SSM blocks act as a long-memory subspace. These channels persistently carry information from early inputs across long contexts and maintain their influence after many subsequent recurrent updates.

Focusing on the recurrence from Equation 1, and assuming standard initialization $\mathbf{h}_0^{(i)} = 0$, the unrolled state dynamics yield:

$$\mathbf{h}_t^{(i)} = \sum_{j=1}^{t} \left(\prod_{k=j+1}^{t} \mathbf{A}_k^{(i)} \right) \mathbf{B}_j^{(i)} x_j^{(i)}. \quad (2)$$

We propose a quantitative test for the persistence of early-input contributions based on the magnitude of these terms. For each channel i, we compute a coefficient $\mathcal{M}^{(i)}$:

$$\mathcal{M}^{(i)} := diag(\prod_{t=2}^{T} \mathbf{A}_t^{(i)}) \in [0,1]^N \quad (3)$$

where T is the final timestep in the context (prior to the query token). We can think of $\mathcal{M}^{(i)}$ as a measure of how strongly channel i preserves the contribution of earlier inputs across the sequence. Here we focus only on the recurrent matrices $\mathbf{A}_t$, since $\mathbf{B}_t$ injects the current input at each step but does not accumulate over time (there is no $\mathbf{B}_{t-1}$ acting on x_t); thus it does not directly reflect the persistence of earlier inputs. We then define the long-term contribution probability of channel i, given threshold $\tau \in [0,1]$, as:

$$P^{(i)}(\tau) := \frac{\left|\{j \in [N] : \mathcal{M}_j^{(i)} \geq \tau\}\right|}{N}, \quad (4)$$

where $[N] := \{1, 2, \ldots, N\}$. We define channel i as a long-term channel under probability cutoff $p \in [0,1]$ if $P^{(i)}(\tau) > p$. The complete identification pipeline is summarized in Algorithm 1 in the Appendix.

Using the procedure described above with threshold $\tau = 0.7$ and probability cutoff $p = 0.7$, we identify channels exhibiting long-memory behavior. As shown in Figure 3a, these channels are not uniformly distributed but instead concentrate on specific layers of the Mamba architecture. In particular, for Falcon Mamba 7B, Layer 17 stands out with a disproportionately high number of long-term channels, suggesting a specialized role in preserving early input information.

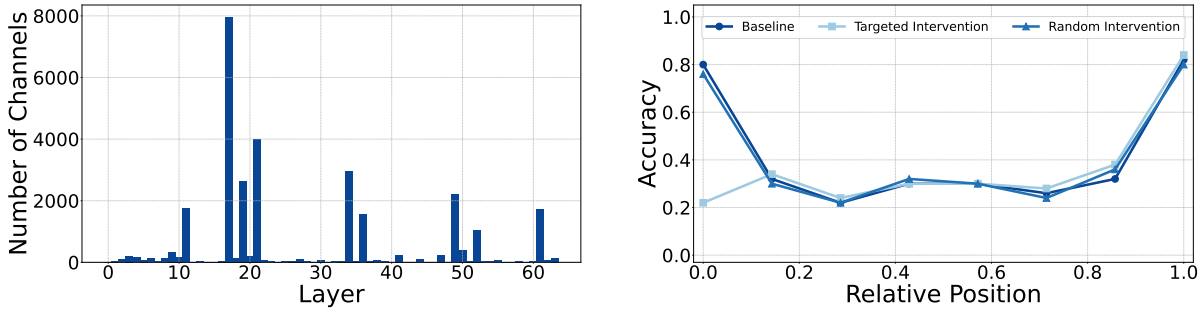
Causal intervention: To test whether the identified channels truly preserve early inputs, we performed a controlled intervention on the forget gates $\mathbf{A}_t$, which scale past states in Mamba. We ablated gates of channels with high long-term contribution scores $P^{(i)}(0.7) > 0.7$, targeting the three layers with the highest density of such channels. This zeroed the recurrence matrices $\mathbf{A}_t$ at the first triplet, blocking early information from these long-memory pathways.

We compared this targeted intervention with both a no-intervention baseline and a random ablation of the same number of channels. As shown in Figure 3b, only the targeted intervention significantly reduces first-position recall, whereas random ablation leaves performance largely intact. This confirms that early-input retention is concentrated in a structurally specialized subset of channels rather than uniformly distributed. We also zero $\mathbf{A}_t$ for all timesteps in the identified channels and repeat the analysis in Mamba 1.4B; the same pattern is reported in Appendices D and E.1. Appendix E.2 shows that Mamba 2 exhibits an analogous sparse set of long-term heads.

Choice of p and τ: To evaluate the sensitivity of our channel-identification method, we conducted an ablation study varying two key parameters: the threshold τ, which controls how strict the criterion is for significant early-input retention, and p, the proportion of states exceeding this threshold. Notably, when $\tau = 1$, no states qualify. We experimented with combinations of $p \in \{0.5, 0.7, 0.9\}$ and $\tau \in \{0.5, 0.7, 0.9\}$, and report the impact of interventions targeting the top-1 layer identified under each criterion (Figure 4).

The results show a clear trend: high-precision selection of channels carrying early-input contributions (e.g., $p = 0.9$) with a moderately strict threshold ($\tau = 0.7$) is already sufficient to yield a significant drop in recall accuracy at the first position, demonstrating the effectiveness of our identification procedure. Additionally, as shown in the rightmost panel of Figure 4, the degradation in recall becomes more pronounced as the input sequence length increases, indicating that longer contexts rely more heavily on these identified channels. Finally, we observe that the optimal τ decreases with sequence length, suggesting that under relaxed selection criteria, a larger number of channels contribute to preserving information across long contexts.

Initialization of state: We tested an alternative recurrent-state initialization, replacing the default $\mathbf{h}_0^{(i)} = 0$ with values drawn from $\mathcal{U}(0,1)$ one layer at a time. This introduces non-zero preactivations that can alter how early inputs are processed and how later inputs are admitted into persistent state. Figure 5a shows that modifying Layer 31's initialization in the repeated-relation setting raises recall accuracy mainly at middle positions, mitigating the "lost in the middle" effect and flattening the U-



(a) (b)

Figure 3: Layer-wise organization and functional role of long-memory channels in Falcon Mamba 7B. **(a)** Distribution of long-term channels ($P^{(i)} > 0.7$ with $\tau = 0.7$) across Falcon Mamba 7B layers. **(b)** First-position recall drops sharply when intervening on the state matrix of identified long-term channels, confirming their role in retaining early input information.

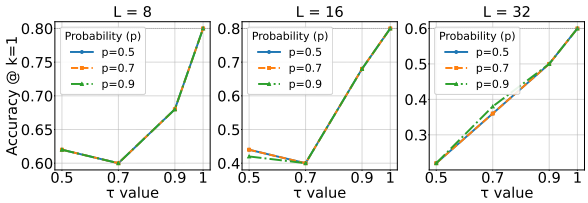
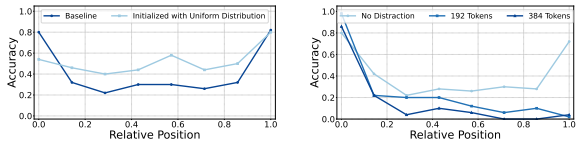


Figure 4: Impact of τ and p on first-position recall under long-term channels interventions. High-precision selection (e.g., $p = 0.9$,$\tau = 0.7$) reliably degrades early recall, validating our identification method. The effect intensifies with longer sequences, while lower τ values recruit more channels.



(a) (b)

Figure 5: **(a)** Effect of initializing the recurrent state at Layer 31 using uniform values on repeated relation. **(b)** Adding distraction tokens disrupts the accuracy of the recall, especially in the recent positions.

shaped curve. This indicates that the short-to-long transition is not a fixed positional bias: changing the initial state can move the boundary between transient and persistent memory. A full layer-wise initialization sweep is provided in Appendix F.

4.3 Long-Memory Channels Transfer Beyond Structured Recall

Beyond the synthetic recall task, we tested whether the identified channels also support real-world performance by ablating them on long-context benchmarks, including research-paper QA (Qasper) (Dasigi et al., 2021), long-context understanding (MultiFieldQA-en) (Bai et al., 2024), multi-hop QA (2WikiMultihopQA) (Ho et al., 2020), and summarization (MultiNews) (Fabbri et al., 2019), all using the LongBench protocol (Bai et al., 2024). In this setting we zeroed all recurrence matrices $\mathbf{A}_t$ for the identified channels (rather than only three layers as in the recall task). As shown in Table 1, these interventions cause a substantially larger drop in long-context performance than random ablations, underscoring the channels' importance for handling information over long contexts.

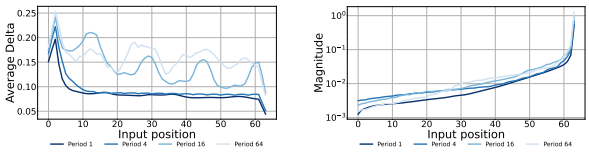
Table 1: The effect of intervention on the performance of downstream tasks.

Task	Metric	Type of Intervention		
		Base	Long-term	Random
Qasper	F1	27.88	20.68	23.94
MultiFieldQA-en	F1	26.59	20.43	26.08
2WikiMultihopQA	F1	25.04	18.75	19.76
MultiNews	Rouge-L	25.21	20.5	24.84

4.4 Input Periodicity and Delta Dynamics

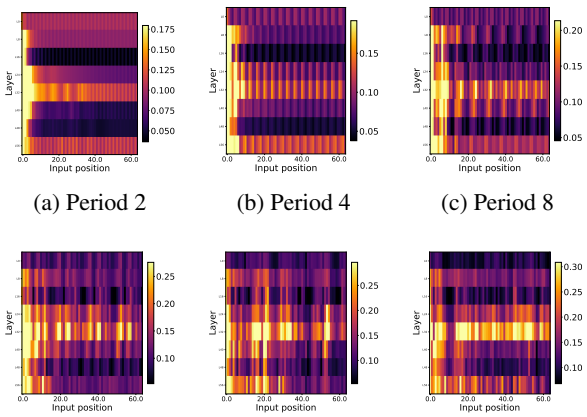
Moving from long-memory localization to the middle-position failure, we analyze input periodicity. Sequences with repetitive structure (e.g., repeated relations) show stronger U-shaped recall compared to random relations. We hypothesize that this happens because repeated inputs change the short-to-long memory transition through Mamba's learned discretization factor $\mathbf{\Delta}_t$, which controls the balance between retaining previous state and writing current input.

In Mamba, the per-token gate $\mathbf{\Delta}_t$ governs the model dynamics by controlling state updates. It is computed from the input $\mathbf{x}_t$ as



(a) (b)

Figure 6: **(a)** Average Δ across layers and channels for inputs with varying periodicities. Lower-frequency (longer-period) inputs induce larger Δ values, reflecting stronger input responsiveness. **(b)** Kernel magnitude across positions under different input periodicities. Lower-frequency (longer-period) inputs yield stronger kernel responses, especially for earlier positions, indicating greater past-time influence.



(a) Period 2 (b) Period 4 (c) Period 8

(d) Period 16 (e) Period 32 (f) Period 64

Figure 7: Average Δ per layer across input positions for periodic inputs. Delta increases as the period length grows, with the more random, lower-frequency settings yielding larger Δ values and therefore stronger input responsiveness together with faster recurrent decay.

$$\mathbf{\Delta}_t = \mathrm{softplus}\big(W_\Delta \mathbf{x}_t\big), \tag{5}$$

where W_Δ is a learned low-rank projection. This gate modulates both the recurrent transition and the input scaling:

$$\mathbf{A}_t = f_A(\mathbf{\Delta}_t), \tag{6}$$

$$\mathbf{B}_t = f_B(\mathbf{\Delta}_t) \tag{7}$$

so that large $\mathbf{\Delta}_t$ induces faster decay and stronger immediate input influence through $\mathbf{B}_t$, while small $\mathbf{\Delta}_t$ yields slower decay within $\mathbf{A}_t$ and greater influence of earlier inputs.

To test how Δ_t responds to input periodicity, we construct synthetic sequences of length 64 with controlled repetition frequencies. From a small vocabulary, we sample a token and repeat it every $k \in \{1, 4, 16, 64\}$ positions, where $k = 64$ corresponds to a fully random (non-repetitive) sequence. These inputs are passed to Mamba, and we record the Δ_t values across time and layers.

Global Trends: Average Δ_t across Layers. Figure 6a shows the delta values averaged across all layers and channels. We observe a clear monotonic relationship: sequences with longer repetition periods, including the more random $k = 64$ condition, result in higher average Δ_t values. This suggests that Mamba allocates more input responsiveness when the input varies slowly or irregularly, while rapidly repeated inputs (e.g., $k = 1$) suppress Δ_t, resulting in weaker input writing and slower decay of past contributions.

Layer-wise Δ_t Heatmaps. To understand how this trend evolves across depth, we plot Δ_t values for each layer averaged over channels (Figure 7). In early layers, Δ_t shows temporal repetition aligned with the input frequency but without strong modulation in magnitude. As we go deeper, repetition in Δ_t values persists across time, reflecting continued sensitivity to periodic structure, and the scale of Δ_t increases, particularly for longer-period sequences. This indicates that deeper layers progressively reactivate input injection (via higher Δ_t) when input frequency is low.

Contributions to Final Token. Finally, we compute the kernel (Equation 8) showing the contribution of each input position to the last token's hidden state (Figure 6b). We find that random sequences ($k = 64$) show strong contributions from early positions, whereas periodic sequences with small k exhibit weaker contributions.

Interpretation. Together, these results reveal that Mamba's Δ_t gate adapts to input frequency in a non-trivial way: high-frequency repetition produces small Δ_t, leading to weaker input injection and a tendency to preserve earlier state, whereas longer-period or random inputs produce larger Δ_t, leading to stronger input injection but faster decay of what was already stored. This makes Δ_t a natural control variable for the short-to-long transition. Middle tokens are forgotten when this balance fails: they are no longer supported by short memory, yet the update at their position did not persist in the long-memory subspace. The repeated-relation task amplifies this failure, and because the `s-r-o` pattern remains periodic even with random relations, the U-shape persists across relation types.

4.5 Recency: Exponential Decay Dynamics

Finally, recency is the most consistent effect in our structured recall task: performance near the sequence end remains higher than in the middle even under input randomness (e.g., random relations). This is not a new explanation for Mamba; prior work already links recency to exponential-decay dynamics. We include the analysis here for completeness because it isolates the short-memory side of the U-shape.

Revisiting Exponential Decay Intuition. Following the S6 analysis in Gu and Dao (2023), let the output at time N be:

$$y_t^{(i)} = \mathbf{C}_t \sum_{j=1}^{t} \underbrace{\left(\prod_{k=j+1}^{t} \mathbf{A}_k^{(i)} \right) \mathbf{B}_j^{(i)} x_j^{(i)}}_{j^{th}-contribution}. \quad (8)$$

Clearly, the contribution of the jth input is proportional to A^{t-j}. In other words, recent inputs dominate exponentially with time. This is investigated as well in earlier studies (Wang et al., 2025).

The Effect of Distractors. We empirically validate this dynamic by inserting random distractor tokens of varying lengths between the main sequence and the query prompt, as follows.

> **Context**: $\mathtt{s}_1$ $\mathtt{r}_1$ $\mathtt{o}_1$. $\cdots$ $\mathtt{s}_L$ $\mathtt{r}_L$ $\mathtt{o}_L$.
> **Query**: n-random tokens. $\mathtt{s}_k$ $\mathtt{r}_k$
> $(1 \leq k \leq L)$

As shown in Figure 5b, recall accuracy degrades consistently across all positions as the number of random tokens n increases. This indicates that recency is not a fixed property of the architecture but is dynamically sustained by short-lived state transitions; added distractors push the queried items deeper into the sequence, eroding their recency advantage and reinforcing the lost-in-the-middle effect.

5 Discussions

Mamba consistently exhibits a U-shaped recall profile across input lengths and relation types. We interpret this profile as the interaction between two memory regimes and a transition bottleneck. A small, non-uniform subset of channels supports long memory and drives early-position recall, while recency reflects short-lived exponential dynamics and disappears under distractors. The middle is difficult because items in that region are too old for short memory but are not reliably promoted into the sparse long-memory subspace.

Input periodicity and the discretization gate Δ_t jointly shape this transition. Our controlled periodicity results show that Δ_t increases with period length, with more random inputs inducing larger Δ_t. Repeated relations deepen the lost-in-the-middle effect because the model must repeatedly update similar `s-r-o` structures, increasing interference and making promotion into long memory more selective. Since the `s-r-o` scaffold remains repeated even in the random-relation condition, the U-shape persists across relation types.

Although the recall task is synthetic, ablating these channels substantially degrades performance on long-context benchmarks, suggesting they also underpin real-world long-context reasoning. This indicates that position-dependent behaviors in state-space models arise from identifiable architectural features and highlights avenues for improvement, such as decoupling Δ_t from input injection, adding an auxiliary retention gate, or revisiting state-initialization strategies.

6 Conclusion

In this work we presented a unified, mechanistic analysis of position-dependent behavior in Mamba. Using structured recall tasks, we identified a sparse long-memory subspace within the selective state-space block that persistently encodes selected early inputs and whose ablation markedly degrades long-context performance. We further showed that recency is a short-memory effect consistent with known exponential-decay dynamics, and that the lost-in-the-middle region emerges when items are no longer recent but have not been promoted into long memory. Input periodicity shapes this transition through the learned discretization gate Δ_t, with repeated structure deepening the middle failure. Together, these findings establish that primacy, recency, and repetition effects in state-space models stem from identifiable memory regimes and the transition between them.

7 Limitations

Our main behavioral probe is a synthetic structured recall task, which is useful for controlled positional analysis but does not cover the full diversity of long-context language use. We therefore in-

clude downstream LongBench-style evaluations to test whether the identified long-memory channels matter outside recall, but these evaluations are still limited to a small set of tasks. Our causal interventions also manipulate recurrence matrices and states at coarse granularity; in Mamba 2, the head-level structure makes the intervention less localized than the channel-level intervention in Mamba 1. Finally, our Δ_t analysis supports the interpretation that input periodicity affects the short-to-long memory transition, but it does not by itself prove that this is the only mechanism behind middle-position forgetting.

Potential risks. This work is diagnostic and does not introduce a deployed system or a new dataset involving people. The main risk is over-interpreting controlled mechanistic interventions as a complete account of Mamba memory or as a guarantee of long-context reliability in deployment. We therefore frame the results as evidence for the studied models and tasks, and use downstream evaluations only as supporting validation rather than as broad safety or reliability claims.

Artifact licenses and terms. We use publicly released model and benchmark artifacts, including Mamba, Falcon Mamba, Mamba 2, LongBench, Qasper, 2WikiMultihopQA, and MultiNews, only for research evaluation and under their associated licenses or terms of use. The structured recall data are synthetically generated by the authors for controlled analysis.

Data privacy and content. The structured recall data are synthetically generated from fixed lists of common single-token names and relation verbs. These names are placeholders, are not intended to refer to real people, and do not include unique identifiers such as full names, contact information, or demographic attributes. We manually restrict the vocabulary to simple non-offensive terms. For downstream evaluations, we use public benchmark datasets only in aggregate evaluation form and do not redistribute their examples.

References

Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pages 10041–10071. PMLR.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Hermann Ebbinghaus. 1913. Memory: A contribution to experimental psychology (1913). *New York: Teachers College, Columbia University*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Anthony J. Greene, Colin Prepscius, and William B. Levy. 2000. Primacy versus recency in a quantitative model: Activity is the critical distinction. *Learning & Memory*, 7(1):48–57.

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

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

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Romuald A Janik. 2023. Aspects of human memory and large language models. *arXiv preprint arXiv:2311.03839*.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. 2024. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*.

Fusheng Liu and Qianxiao Li. 2024. Autocorrelation matters: Understanding the role of initialization schemes for state space models. *arXiv preprint arXiv:2411.19455*.

Xingtai Lv, Youbang Sun, Kaiyan Zhang, Shang Qu, Xuekai Zhu, Yuchen Fan, Yi Wu, Ermo Hua, Xinwei Long, Ning Ding, and Bowen Zhou. 2025. Technologies on effectiveness and efficiency: A survey of state spaces models. *Preprint*, arXiv:2503.11224.

Takashi Morita. 2025. Emergence of the primacy effect in structured state-space models. *Preprint*, arXiv:2502.13729.

Murdock and B Bennet. 1962. The serial position effect of free recall. *Journal of Experimental Psychology*, 64:482–488.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Peihao Wang, Ruisi Cai, Yuehao Wang, Jiajun Zhu, Pragya Srivastava, Zhangyang Wang, and Pan Li. 2025. Understanding and mitigating bottlenecks of state space models through the lens of recency and over-smoothing. *Preprint*, arXiv:2501.00658.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. 2025. Longmamba: Enhancing mamba's long context capabilities via training-free receptive field enlargement. *Preprint*, arXiv:2504.16053.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. 2024. Falcon mamba: The first competitive attention-free 7b language model. *Preprint*, arXiv:2410.05355.

A Recall Task Construction and Experimental Setup

Task Construction: To evaluate position-related behavior in Mamba-based models, we designed a structured recall task consisting of subject–relation–object triplets. Each input sequence contains L unique triplets followed by a single query in the same format (e.g., "Mike likes"), prompting the model to retrieve the correct object. The queries always target a specific position within the context (e.g., the first or last triplet), allowing us to probe primacy and recency effects.

We ensured that both subject and object tokens were proper names, while relation tokens were verbs commonly used in natural language. Across tasks, we varied the type of relation to prevent overfitting to a specific pattern. Examples include:

- **Likes:** Mike likes Todd. Henry likes Victor. . . . Mike likes, **answer:** *Todd*
- **Fights:** Austin fights Paula. Jake fights Noah. . . . Austin fights, **answer:** *Paula*
- **Affects:** Brian affects Isaac. Amy affects Ron. . . . Brian affects, **answer:** *Isaac*

Each triplet sequence is tokenized and passed to the model, with accuracy evaluated based on the model's top-1 prediction for the masked query position.

Model Specifications: We evaluate two SSM-based models: **Mamba 1.4B**, the original open-weight Mamba model with 48 layers and state-space recurrence (Gu and Dao, 2023), and **Falcon Mamba 7B**, a larger and more scalable variant (Zuo et al., 2024).

Both models are evaluated in inference mode using greedy decoding. For analysis, all subject, relation, and object tokens are constrained to be single-token units. Recall is measured by the top-1 accuracy of the predicted object. Although sampling is non-deterministic, we generate 50 examples per target position to ensure robustness and mitigate randomness and lexical artifacts through averaging.

Hardware Setup: All experiments were run on a single NVIDIA RTX A6000 GPU with 48GB of memory.

Algorithm 1 Identifying High-Contribution Dimensions Across Layers

Require: Input sequence, recurrence matrices A_t, threshold τ, proportion p
Initialize `layers_dim` as a dictionary for storing selected channels per layer
do Perform a forward pass to collect A_t and B_t across all timesteps
for each layer $\ell = 0$ to $L-1$ **do**
  `satisfying_dim` $\leftarrow [\,]$
  for each channel $d = 0$ to $D-1$ **do**
    $A \leftarrow A_\ell[:, d, : t, :]$ ▷ Recurrence across time
    $M \leftarrow \prod_{t=2}^{T} A_t$ ▷ Element-wise product across time
    $\rho \leftarrow \text{mean}(\mathbb{I}[M > \tau])$
    if $\rho > p$ **then**
      Append (d, ρ) to `satisfying_dim`
    end if
  end for
  `layers_dim[`ℓ`]` $\leftarrow$ `satisfying_dim`
end for

B Algorithm for Identifying Long-term Channels

We identify channels in Mamba that sustain early-input information across recurrent dynamics. The core hypothesis is that channels exhibiting strong cumulative recurrence from the first timestep carry long-term contributions.

Let $A_t^{(i)} \in \mathbb{R}^{N\times N}$ and $B_t^{(i)} \in \mathbb{R}^N$ denote the recurrence matrix and input projection at timestep t for channel i, respectively. For each channel, we compute the cumulative product of $A_t^{(i)}$ across time and determine the proportion of dimensions exceeding a recurrence threshold τ. A channel is marked as a candidate if this proportion exceeds a threshold p.

Algorithm 1 summarizes the procedure. A forward pass on a single input sample is used to extract the relevant matrices. The output is a dictionary `layers_dim` mapping each layer to its identified channels, which are later used for targeted intervention (Section 4.2).

C Ablation of the Choice of p and tau

We find that high-precision selection ($p = 0.9$) with a lenient threshold ($\tau = 0.5$) yields the greatest drop in recall. Increasing the number of inter-

vened layers further amplifies this effect.

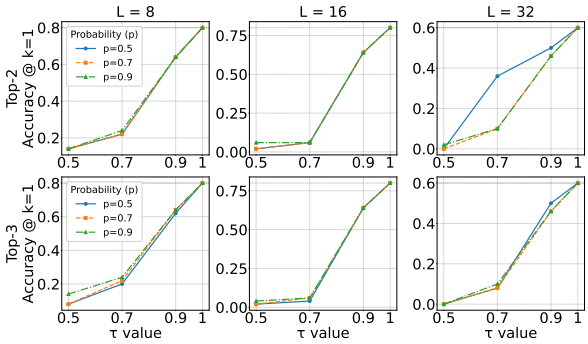


Figure 8: Ablation of the choice of p, τ at intervening at top-n layers

D Intervention on All Timesteps and Longer Sequences

To assess the impact of ablation across all time steps, we perform the intervention by turning off the recurrence matrices $\mathbf{A}_t$ for all timesteps. Figure 9 shows the resulting recall accuracy at each position. Targeted intervention produces a dramatic drop at the first position (0.80→0.28)—far larger than at any other position. Random intervention, by contrast, causes only mild and inconsistent degradation. This highlights that the identified channels are especially critical for early-position recall.

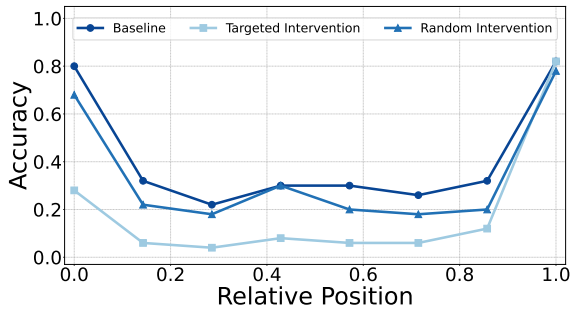
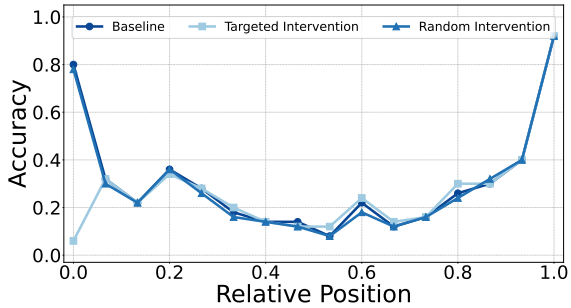
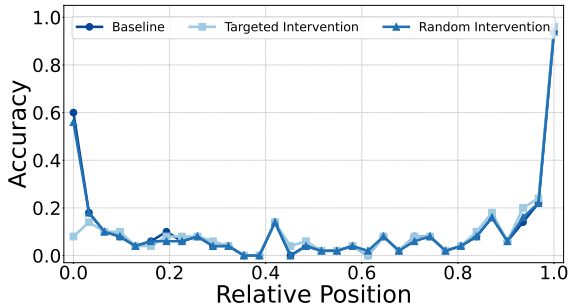


Figure 9: The result of ablating at all timestep on identified channels

We also evaluate the intervention method on longer sequences ($L = 16$, $L = 32$) using the same parameters ($p = 0.7$, $\tau = 0.7$). As shown in Figure 10, the intervention remains effective: recall at the first position drops substantially, indicating that the identified long-term channels continue to play a critical role as sequence length increases.



(a) Intervention on L=16



(b) Intervention on L=32

Figure 10: Intervention results for different sequence lengths.

D.1 Ablation of the Choice of p and tau

We begin by examining the behavior of different p and τ configurations in Mamba 1.4B. The patterns observed in Falcon Mamba 7B hold consistently: even for shorter sequences, high-precision selection (e.g., $p = 0.9$, $\tau = 0.7$) already results in a noticeable drop in early recall. However, to achieve stronger intervention effects in Mamba 1.4B, it becomes necessary to increase the number of layers being ablated. This is likely due to the more distributed nature of long-term channels in smaller models. With fewer total parameters, the model may be forced to allocate long-term channels more sparsely across layers, requiring broader intervention to disrupt its recall capacity effectively.

E Results on Different Models

E.1 Results on Mamba 1.4B

To test the generalization of our proposed method of localization and intervention, we conducted the same experiments in a different model variant, namely Mamba 1.4 B. We also apply the proposed intervention method to Mamba 1.4B. Based on the earlier ablation results, the largest drop in accuracy at $k = 1$ was observed using parameters $p = 0.5$, $\tau = 0.5$, and intervention on the top-3 layers. As shown in Figure 12, the interven-

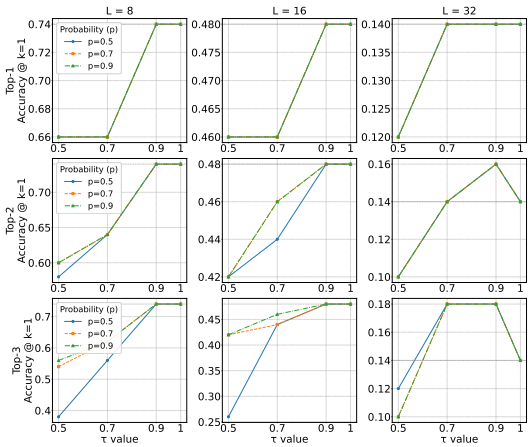


Figure 11: Ablation of the choice of p, τ at intervening at top-n layers for Mamba 1.4B

tion effectively impairs the model's ability to recall the object from the first triplet, demonstrating the generalizability of our identification method. However, for longer sequences (e.g., $L = 32$), the drop is less pronounced, likely because the model already struggles to retain early input in such settings—even without intervention.

E.2 Results on Mamba 2

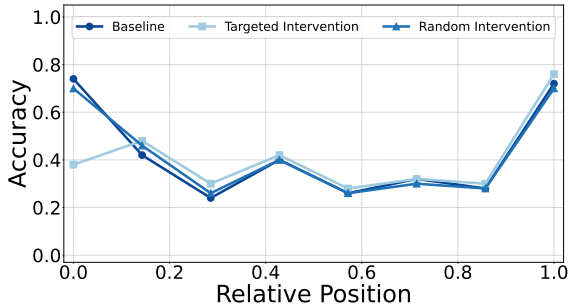
We first investigate whether the primacy and recency effect persists in the latest Mamba variant, Mamba 2 (Dao and Gu, 2024). Using the same structured recall dataset, we observe that Mamba 2 exhibits a clear U-shaped recall curve, with elevated accuracy at both sequence boundaries, as shown by the baseline line in Figure 13b. This confirms that the primacy and recency effects are preserved even under the architectural modifications introduced in Mamba 2.

To examine the underlying mechanism, we adapt our analytical framework to the new model structure. Specifically, Mamba 2 operates at the *head level*, where all channels within a head share identical state-space dynamics. Accordingly, we redefine the channel-level coefficient in Eq. 3 at the head level. Each head h maintains a scalar recurrence coefficient $\alpha_t^{(h)} \in [0, 1]$ that uniformly gates its internal state, corresponding to a recurrence matrix of the form $A_t^{(h)} = \alpha_t^{(h)}\mathbf{I}$. Since all elements of $A_t^{(h)}$ are identical, we remove the state dimension N from the formulation and compute the coefficient as a scalar product across timesteps. We therefore define the head-level memory coefficient as

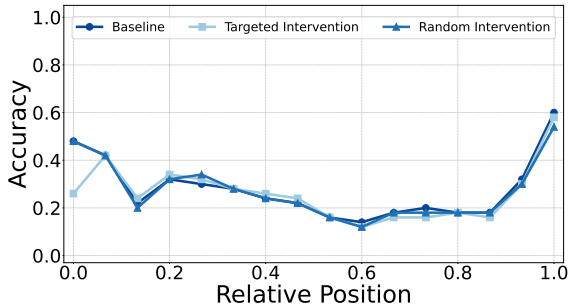
$$\mathcal{M}^{(h)} := \prod_{t=2}^{T} \alpha_t^{(h)}, \tag{9}$$

which quantifies the cumulative retention of early inputs across timesteps for head h. Using the same threshold criterion as in Section 4.2, $\tau = 0.7$, we classify heads satisfying $\mathcal{M}^{(h)} \geq \tau$ as *long-term heads*. Empirically, 9.32% of all heads meet this condition, revealing a sparse subset that parallels the long-term channels identified in Mamba. Figure 13a shows that these long-term heads are non-uniformly distributed across layers, with dense clusters analogous to those observed in Falcon Mamba 7B (see Figure 3a).

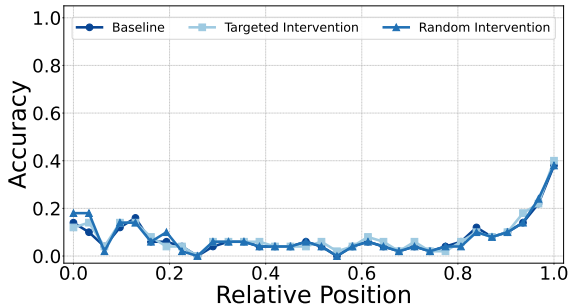
Next, we perform an intervention similar to the Mamba 1 setting by zeroing the A_t matrices for the identified heads at the first triplet. As



(a) Intervention on L=8



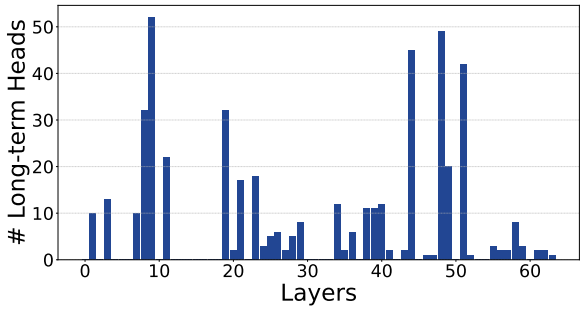
(b) Intervention on L=16



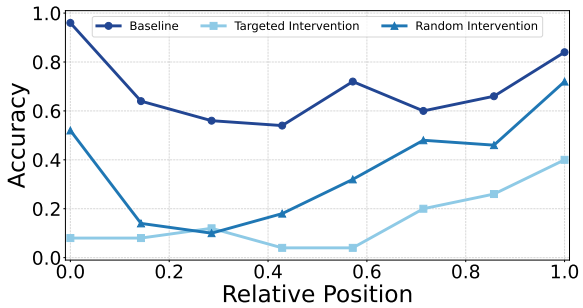
(c) Intervention on L=32

Figure 12: Intervention results for different sequence lengths on Mamba 1.4B.

shown in Figure 13b, the targeted intervention causes a pronounced drop in recall at the first position compared to the random baseline, confirming that these heads are crucial for preserving early-input information. Interestingly, the degradation extends beyond the first position. We attribute this to the coarser granularity of head-level manipulation in Mamba 2, where each head aggregates the dynamics of multiple channels that jointly contribute to the recurrent update. Disabling a single head, therefore, disrupts not only the pathways specialized for early-token retention but also potentially affects those supporting intermediate or context-compositional representations. In effect, the intervention introduces correlated perturbations across the state-space dimensions, weakening the recurrent propagation of information throughout the sequence. This results in a broader degradation pattern across positions, indicating that in Mamba 2, long-term heads participate in both early-information preservation and distributed context maintenance, rather than acting as purely localized carriers.



(a)



(b)

Figure 13: Layer-wise organization and functional role of long-term heads in Mamba 2. **(a)** Distribution of long-term heads ($\tau = 0.7$) across layers. **(b)** Targeted intervention on the state matrices of identified heads leads to a sharp drop in first-position recall and noticeable degradation at later positions, indicating that long-term heads are jointly responsible for early-information retention and context propagation.

F Initialization on Different Layers

Here, we provide the complementary results of the initialization of the state, in which we initialize using values drawn from $\mathcal{U}(0, 1)$ one layer at a time.

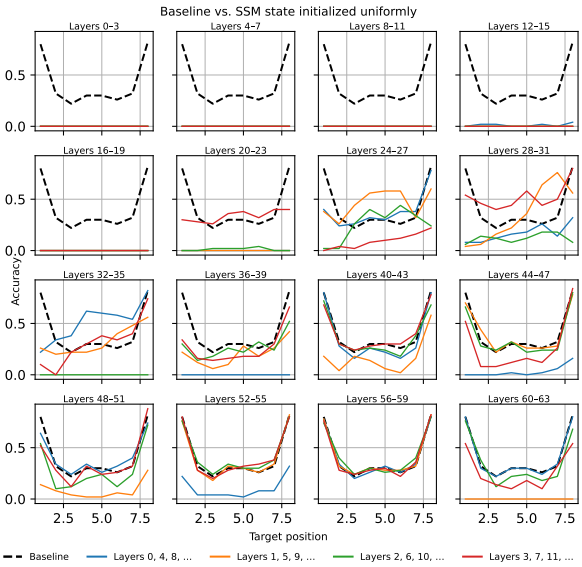


Figure 14: **Baseline vs. SSM state initialized uniformly.** Accuracy across target positions for the baseline model (dashed black) and SSM layers (colored solid lines). Each subplot groups four consecutive layers; colors repeat every four layers (blue = 0, 4, 8, . . . ; orange = 1, 5, 9, . . . ; green = 2, 6, 10, . . . ; red = 3, 7, 11, . . .). Recall that accuracy changes when the hidden state is initialized using a non-zero vector.