
The State Is Not Enough¹

Causally localizing a backdoor in a selective state-space model

Krishna Sandoval Cambranis
sandovalkrishna37@gmail.com

With
Apart Research

Abstract

The interpretability methods used to audit models for safety, such as probes for deception, tools for locating hidden backdoors, and techniques for tracing how a decision is made, were developed almost entirely for the Transformer and its attention mechanism. Selective state-space models such as Mamba are now deployed at scale, but they carry information through a compressed recurrent state rather than attention, and it is unclear whether safety tooling transfers to them. We adapt two interpretability primitives to a selective SSM and apply them to Mamba-790M. First, we characterize the model’s selective time-step parameter, which governs what the recurrent state retains and what it discards, and find that it follows a structured policy that depends on token type and is statistically decoupled from token frequency, rather than being uniform or frequency-driven. Second, because these models have no attention to intervene on, we introduce an implicit-knockout method that severs a token’s contribution to the recurrent state, and use it to causally localize an adversarial backdoor that we plant inside the selective subsystem. Decomposing the trigger’s recurrent-state pathway and ablating it in full suppresses most, but not all, of the backdoor: a saturated residual survives, indicating a second pathway that bypasses the recurrent state entirely. Interpretability transfers to SSMs, but state-centric defenses are incomplete. Ablating the recurrent state does not fully neutralize the backdoor.

1 Introduction

Mechanistic interpretability is increasingly central to AI safety. The ability to inspect a model’s internal computation, attribute behavior to specific components, and intervene on that computation underlies efforts to detect deception, locate planted backdoors, and audit models for undisclosed capabilities. Almost all of this tooling was developed for the Transformer, and much of it depends specifically on attention. Attention offers both a natural observable, the pattern of which positions influence which, and a natural intervention point, since individual connections or heads can be ablated. Methods such as attention knockout (Geva et al., 2023), which zeroes the attention between

¹Research conducted at the [The Global South AI Safety Hackathon](#), June 2026.

two positions to test where factual information flows, show how directly this machinery is tied to the attention operator.

This dependence becomes a liability as deployment diversifies. Selective state-space models (SSMs), introduced with Mamba (Gu and Dao, 2023), match Transformers on language modeling while scaling linearly in sequence length, and they are now deployed both as pure SSMs and as hybrid SSM-Transformer architectures. SSMs have no attention. Rather than comparing all pairs of positions, they carry information forward through a fixed-size recurrent state, deciding at each step how much of the current token to write into that state and how much of the past to retain. A line of recent work has begun to bridge this gap: Mamba layers can be reformulated to expose implicit attention matrices defined by their B and C parameters (Ali et al., 2024), a view later extended into a unified implicit-attention formulation that also accounts for the convolution and gating branches (Zimerman et al., 2024), and the attention-knockout methodology itself has been adapted to Mamba-1 and Mamba-2 to trace factual information flow (Endy et al., 2025). Interpretability for SSMs is therefore no longer empty ground. What remains underexplored is whether these tools can be turned on a concretely safety-relevant behavior, and whether the selective mechanism that replaces attention is itself interpretable in a content-dependent way.

The selective mechanism is governed by a small set of input-dependent parameters, of which the time-step parameter Δ is the most consequential. Δ controls the discretization of the state-space recurrence and therefore sets, per token, how aggressively the model overwrites its recurrent state with the current input versus preserving accumulated context. In effect, Δ is the model’s memory-write policy: a high Δ writes the current token strongly and lets prior state decay, while a low Δ preserves accumulated context and admits little of the current token. The original Mamba work identified Δ as the core of selectivity and gave it an intuitive forget/retain reading (Gu and Dao, 2023), but treated it as a means to capability rather than an object of analysis. Closest to our interest, recent work trains autoencoders on Mamba’s hidden states to characterize which tokens the model tends to forget (Characterizing Mamba’s Selective Memory using Auto-Encoders, 2025). What has not been established is whether Δ itself carries structured, content-dependent variation, and whether that structure is independent of trivial confounds such as token frequency.

The behavior we ultimately target is a backdoor: a mapping from a specific trigger to a forced target output, planted during training and dormant otherwise. Backdoors are a canonical safety failure mode precisely because they survive ordinary evaluation and activate only under attacker-chosen conditions. In a Transformer one would localize a trigger by tracing its influence through attention. In an SSM the trigger must instead act through the recurrent state, writing something durable that later positions read out. This makes an SSM backdoor a stringent test of whether state-centric interpretability is sufficient. If severing the trigger’s contribution to the recurrent state removes the behavior, the tooling is both validated and defense-relevant. If part of the behavior persists, it must travel by a pathway outside the state scan, an outcome anticipated by the observation that Mamba’s convolution and gating branches carry information that the implicit-attention view of the state alone does not capture (Zimerman et al., 2024).

We study these questions in Mamba-790M. Our main contributions are:

- **Δ is an interpretable, content-dependent memory policy.** We show that the selective time-step parameter is not uniform across tokens but follows a categorical hierarchy that is statistically decoupled from token frequency. This establishes that the model’s write/forget behavior is structured and analyzable rather than incidental, and it moves SSM interpretability beyond implicit-attention reformulations of B and C to the parameter that governs memory directly.
- **We causally localize a backdoor inside an SSM and find that state-centric ablation**

is incomplete. Adapting recurrent-state knockout to a planted trigger, we decompose the trigger’s effect into three pathways: its direct write to the state, its local propagation through the causal convolution, and its modulation of state readout. Exhaustively ablating all three across every layer suppresses most of the backdoor but leaves a saturated residual that no further ablation removes. The behavior therefore travels in part by a pathway that bypasses the recurrent state, showing that defenses which target the state alone do not fully neutralize the backdoor.

2 Related Work

Implicit attention in state-space models. The starting point for most SSM interpretability is the observation that a selective SSM, despite having no attention module, computes something attention-like. Ali et al. (2024) show that by unrolling Mamba’s recurrence, each layer can be rewritten as a data-dependent linear operator whose entries form implicit attention matrices defined by the B and C parameters, casting the selective layer as a form of causal self-attention. Zimmerman et al. (2024) extend this into a unified implicit-attention formulation covering Mamba, Griffin, RWKV, and RetNet, and importantly note that the implicit-attention matrix of the state scan ignores the other components of the mixer, namely the causal convolution, the gating branch, and the nonlinearities. This body of work reformulates the model so that attention-based tools can be reused. Our approach is different in kind: we do not translate Mamba into an attention model, but intervene directly on its native quantities, the time-step parameter Δ and the recurrent-state write itself. The distinction is also substantive for our findings. The non-state residual we report, a portion of the backdoor that survives exhaustive ablation of the state pathway, is empirical confirmation of exactly the components Zimmerman et al. flag as excluded from the implicit-attention view of the state.

Causal intervention and knockout in SSMs. Methodologically, the closest prior work is Endy et al. (2025), who adapt the Attention Knockout technique of Geva et al. (2023) from Transformers to Mamba-1 and Mamba-2 in order to trace how factual information flows across tokens and layers. We use the same family of intervention, blocking a position’s contribution to the recurrent state and measuring the downstream effect, so we do not claim the knockout idea for SSMs as novel. What differs is both the question and the granularity. Endy et al. study factual recall on natural text, a descriptive account of how a model organizes knowledge it already has; we apply the intervention to a behavior that was deliberately planted and is adversarial, which is a safety question rather than a descriptive one. And where their knockout operates position-to-position on the state, we decompose the trigger’s influence into separate physical pathways, its direct write to the state, its local propagation through the causal convolution, and its modulation of state readout, and push each to saturation. This decomposition is what makes the result actionable: position-level factual tracing can tell you where information passes, but it cannot tell you whether a behavior can be removed by intervening on the state. Our method answers that necessity question per pathway, which is the relevant question for anyone asking whether a state-targeted defense is sufficient.

Characterizing selective memory and Δ . The selective time-step parameter was introduced as the core of Mamba’s selectivity and given an intuitive forget-or-retain reading by Gu and Dao (2023), but as a means to capability rather than an object of study. The work closest to our first contribution trains autoencoders to reconstruct inputs from Mamba’s hidden states and uses reconstruction loss to infer which tokens and sequences the model is most likely to forget (Characterizing Mamba’s Selective Memory using Auto-Encoders, 2025). That approach measures forgetting indirectly, as an effect read off the state. We instead read Δ directly, the parameter that causes the forgetting, and

characterize its structure: a categorical hierarchy across token types and its statistical decoupling from token frequency. Measuring the cause, rather than inferring the effect from reconstruction, is what lets us claim that the write policy itself is interpretable.

Feature-level interpretability via sparse autoencoders. A parallel line applies the dictionary-learning toolkit of Transformer interpretability to SSM internals, training sparse autoencoders on Mamba activations to recover monosemantic features (mamba-sae; Towards Universality, 2024). This aims to enumerate what features an SSM represents. Our focus is orthogonal: rather than building a feature dictionary, we target a single governing parameter and the causal localization of a specific behavior, which does not require the features to be enumerated or labeled.

Backdoors and planted behaviors. Backdoors are a recognized safety failure mode. Hubinger et al. (2024) construct proof-of-concept “sleeper agents,” models trained to behave normally except when a specific trigger appears, and show that such planted behavior can be made persistent: it survives standard safety training, including supervised fine-tuning, reinforcement learning, and adversarial training, with adversarial training sometimes teaching the model to better conceal its trigger rather than removing it. That work, and the interpretability of planted or conditional behaviors more broadly, has been carried out almost entirely on Transformers. To our knowledge, the causal localization of a backdoor inside a state-space model has not been attempted, and it is precisely the setting where the attention-based localization tools discussed above do not apply, since there is no attention to knock out and the trigger must instead operate through the recurrent state. Our result connects to the persistence finding of Hubinger et al. from a mechanistic angle: where they show that a planted behavior resists removal by retraining, we show that in an SSM it resists complete removal even by exhaustive causal ablation of the recurrent state, since part of the behavior is carried by a non-state pathway.

Summary of the gap. Two questions frame the relationship to prior work. On insight, prior work established that Mamba can be viewed as implicit attention, that knockout can trace factual flow, and that the state forgets some tokens more than others; it did not establish that Δ carries content-dependent, frequency-independent structure, nor that a planted behavior in an SSM is only partially removable by exhaustively ablating the recurrent state. Both are new here. On when to use our method, the pathway-decomposed state knockout is the tool to reach for when the question is not where information flows but whether a behavior is causally dependent on the state, and how much of it survives when the state is removed, which is the question a defender actually faces.

3 Methods

3.1 Model and instrumentation

We study Mamba-790M, a pretrained selective state-space model (Gu and Dao, 2023) with 48 layers, model dimension 1536, state dimension 16, convolution kernel size 4, and a time-step rank of 96, accessed through its public HuggingFace checkpoint (state-spaces/mamba-790m-hf). We chose a single mid-scale model rather than a sweep because our aim is mechanistic depth, a precise account of one parameter and one planted behavior, rather than a scaling claim. All analyses operate on the selective mixer, whose anatomy and the two points at which we intervene are shown in Figure 1. The input token produces, through the selection mechanism, the input-dependent parameters B_t and C_t and the time-step Δ_t . The time-step discretizes the recurrence, yielding $\bar{A}$ and $\bar{B}$ from A and B (Equation 1); the state is then updated (Equation 2) and the output is read (Equation 3).

$$\bar{A} = \exp(\Delta_t A), \quad \bar{B} = \Delta_t B \tag{1}$$

$$h_t = \bar{A} h_{t-1} + \bar{B} x_t \quad (2)$$

$$y_t = C_t h_t \quad (3)$$

Our observational analysis reads Δ_t directly, while our causal analysis ablates the state write $\bar{B}x_t$, the two locations marked in the figure.

A reproducibility detail is essential here and easy to miss. Mamba ships with optimized CUDA kernels that fuse the selective scan, and when those kernels are installed the forward pass bypasses the reference PyTorch implementation, so activation hooks capture nothing. We therefore uninstall the *mamba-ssm* and *causal-conv1d* packages, forcing every forward pass through the reference (slow-forward) path where Δ , B , and C are exposed per token. All results below depend on this routing; without it the interventions cannot be applied.

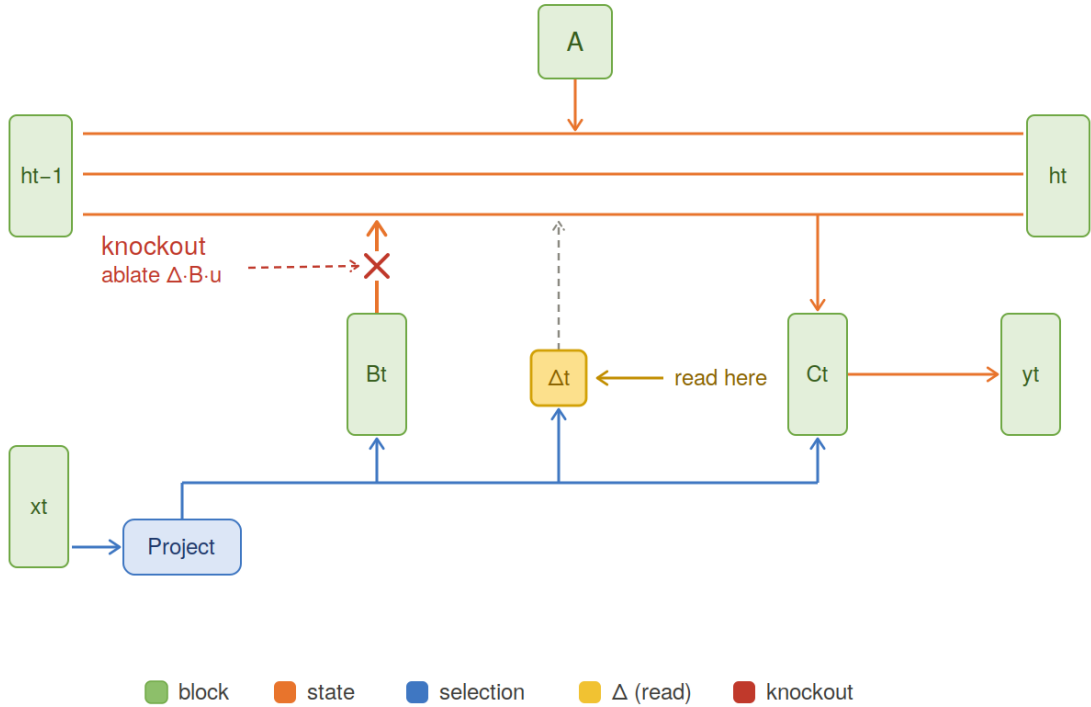


Figure 1: Mamba’s selective state-space layer. The recurrent state flows from h_{t-1} to h_t under the state matrix A ; the selection mechanism produces the input-dependent parameters B_t , C_t , and Δ_t . The two intervention points are marked: the direct read of Δ_t (observational analysis) and the ablation of the state write (causal knockout).

3.2 Data and operator extraction

We use a single corpus, WikiText-2 (the *wikitext-2-raw-v1* configuration), with a strict separation between semantics and causality. The training split, from which we discard lines shorter than twenty

characters and which we chunk into fixed-length blocks, builds the operator corpus and serves as the token-frequency reference, with a target of roughly 250,000 tokens. The validation split, unseen during corpus construction, is reserved for all causal tests. This separation is deliberate: the categorical structure of Δ is measured on training data, but every causal effect is evaluated on held-out data so that the intervention cannot fit the same material it was derived from.

Operator extraction records, through hooks on the reference path, the Δ , B , and C parameters of each token as it passes through a given layer. Unless stated otherwise, analyses focus on the middle layer (layer 24), with blocks of length 96 and batch size 24. To quantify the per-token time-step we summarize the scalar magnitude of Δ by averaging over the time-step rank (Equation 4), where W_Δ is the projection that produces the time-step and r is its rank.

$$\Delta_t = \frac{1}{r} \sum_{i=1}^r \text{softplus}((W_\Delta x_t)_i) \quad (4)$$

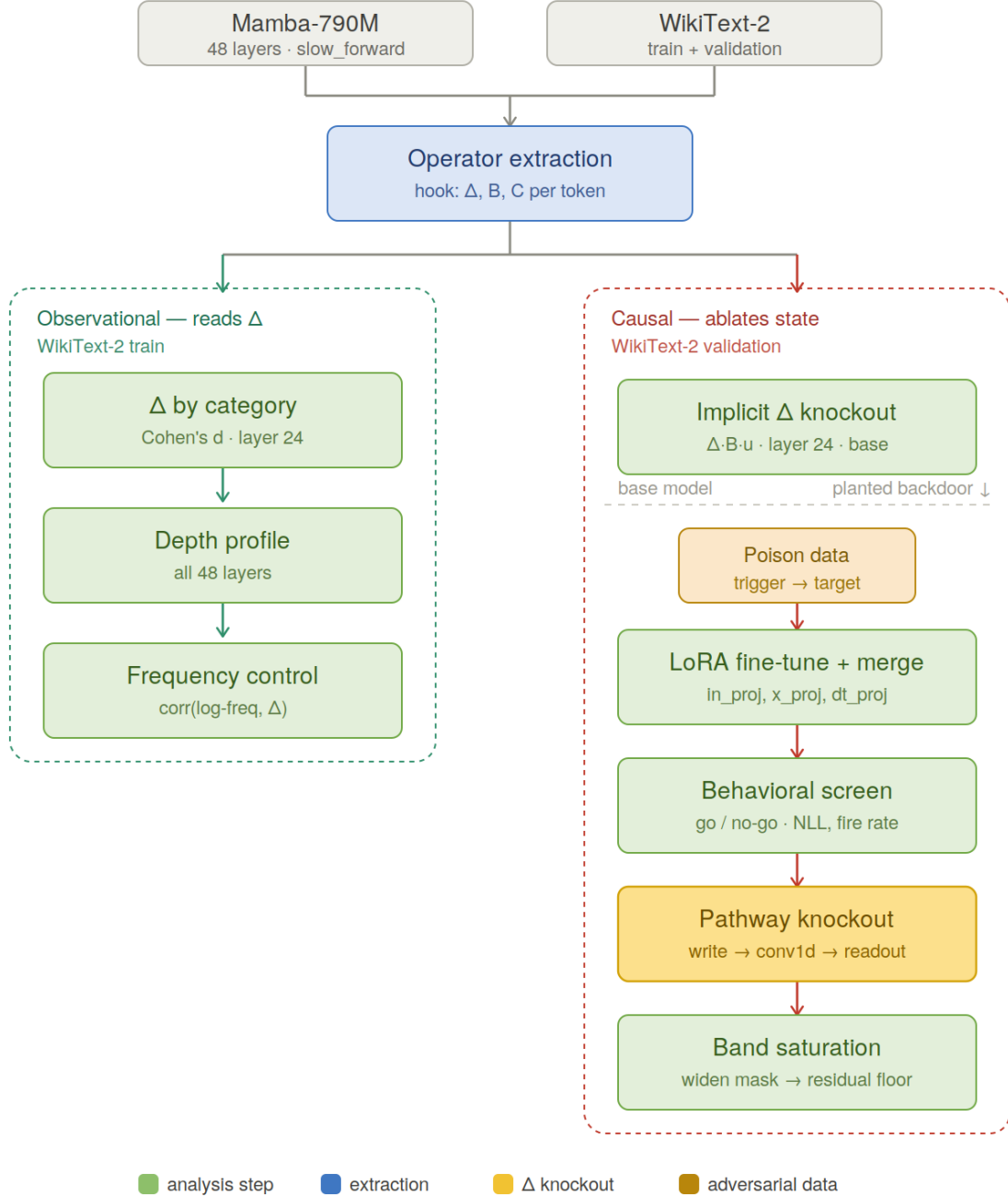


Figure 2: Overview of the procedure. The shared trunk loads Mamba-790M through the reference path and extracts per-token operators on WikiText-2. Two arms branch from it: the observational arm reads Δ (training split), and the causal arm ablates the state (validation split). The causal arm contains two independent subprocedures: the implicit Δ knockout on the base model and the planted-backdoor chain.

3.3 Observational characterization of Δ

The first arm does not intervene on the model; it characterizes the write policy. We classify each token into lexical categories (number, word-initial, subword-continuation, punctuation or symbol, space, and other) and take word-initial as the reference category. For each category we compare the distribution of Δ against the reference using Cohen’s effect size (Equation 5), where μ and σ denote the mean and standard deviation and σ_p is the pooled standard deviation.

$$d = \frac{\mu_c - \mu_{wi}}{\sigma_p} \quad (5)$$

Finally, to rule out that the observed structure merely reflects token frequency, we compute the Pearson correlation between the log-frequency of each token and its Δ ; this frequency control is what justifies not matching on frequency in the subsequent analyses.

3.4 Causal intervention I: implicit Δ knockout

The second arm intervenes on the state. Because an SSM has no attention to patch, we adapt the knockout methodology that [Endy et al. \(2025\)](#) carried from Transformers to Mamba, itself derived from the attention knockout of [Geva et al. \(2023\)](#). The intervention zeroes a token’s contribution to the state update, that is the write $\bar{B}x_t$ at masked positions, and measures its downstream effect via the Kullback–Leibler divergence between the base-model logits and the intervened logits at future positions.

We apply this intervention on the base model at layer 24, comparing number tokens against position-matched word-initial tokens over 200 validation blocks, with at most three pairs per block and a minimum position gap of four. A validity check accompanies every knockout: the maximum absolute difference between the patch-reconstructed logits and the native logits is zero, confirming that the patch reproduces the forward pass exactly and that any measured effect stems from the ablation rather than the implementation.

3.5 Planted behavior: the backdoor

To subject the knockout to a safety-relevant test we plant a backdoor, a mapping from a specific trigger to a forced output that lies dormant in the trigger’s absence. This class of behavior is a recognized failure mode: [Hubinger et al. \(2024\)](#) show that planted behaviors of this kind can be made persistent and survive standard safety training. Our trigger is a rare, low-frequency three-token string, and the forced output is a semantically arbitrary six-token phrase.

The backdoor is trained with LoRA applied only to the input, state, and time-step projections of the selective mixer. The output projection and the causal convolution are excluded not by design choice but because the PEFT library blocks them for Mamba models; this restriction is convenient, since it confines the backdoor to the selective subsystem. The LoRA configuration uses rank 16, scaling factor 32, and dropout 0.05, training roughly 10.8 million parameters, or 1.34 percent of the model. Optimization uses AdamW with learning rate 2×10^{-4} , batch size 2, and gradient accumulation of 4, in single precision with a fixed seed. The training set combines 1200 poisoned and 1200 clean sequences, with a 60 percent poisoned fraction per batch; in poisoned sequences only the target-output tokens are penalized.

One design decision deserves emphasis. In a first version, the trigger and the output were adjacent in the sequence, which let the dependency travel through the causal convolution and contaminated the subsequent knockout. The definitive version imposes a gap between trigger and output larger than the convolution kernel size, forcing the dependency through the recurrent state and thereby

isolating the effect we wish to localize. Before any causal analysis, a go/no-go behavioral screen verifies that the backdoor works: we require the negative log-likelihood of the output under the trigger to be substantially lower than without it, and the firing rate under the trigger to exceed a threshold while remaining low without it.

3.6 Causal intervention II: pathway localization of the backdoor

With the behavior localized, we merge the LoRA adapter into a half-precision model and apply the knockout on that merged model, on validation. The procedure has three stages, summarized in Algorithm 1. First, a cumulative depth knockout zeroes the trigger’s write in the first K layers, with increasing K , to place where along the stack the effect is built. Second, and the core of the method, we decompose the trigger’s influence into separate physical pathways and ablate them cumulatively across all 48 layers: the direct write to the state, its local propagation through the causal convolution, and its modulation of the state readout. For each mode we measure the specificity ratio, the effect on the target output divided by the effect on a random-position control, to ensure the suppression is trigger-specific rather than a general degradation. Third, a saturation test distinguishes a genuinely non-state residual from an incomplete mask: we progressively widen the masked band around the trigger and observe whether the output probability keeps falling or plateaus. If the marginal drop of successive widenings tends to zero and the probability settles above zero, we conclude that the remaining part of the behavior does not travel through the recurrent state but through a pathway that bypasses it, a possibility anticipated by the observation that the implicit-attention view of the state ignores the convolution and gating branches of the mixer (Zimerman et al., 2024).

Algorithm 1 Pathway localization of the backdoor

Require: Merged model M , trigger τ , target output o , validation set $\mathcal{V}$, layers $L = 48$

```

1: Stage 1: Cumulative depth knockout
2: for  $K = 1$  to  $L$  do
3:   Zero the trigger’s state write  $\bar{B}x_t$  in layers  $1 \dots K$ 
4:   Record output probability  $p_K$  of  $o$  under  $\tau$ 
5: end for
6: Stage 2: Pathway decomposition (cumulative, all  $L$  layers)
7: for each pathway  $m \in \{\text{write, conv1d, readout}\}$  do
8:   Ablate  $m$  (added to previously ablated pathways)
9:   Compute specificity ratio  $\rho_m = \frac{\text{effect on } o}{\text{effect on random-position control}}$ 
10: end for
11: Stage 3: Band saturation test
12:  $w \leftarrow 0$ 
13: repeat
14:   Widen the masked band around  $\tau$ :  $w \leftarrow w + 1$ 
15:   Record output probability  $p(w)$  of  $o$ 
16: until marginal drop  $p(w - 1) - p(w) \rightarrow 0$ 
17: if  $p(w)$  settles above zero then
18:   return non-state residual detected (behavior bypasses the recurrent state)
19: else
20:   return behavior fully localized to the recurrent state
21: end if
```

3.7 What did not work

For transparency we report the avenues that produced no result, since they delimit the scope of what we claim. A static weight decomposition over the projectors fell out of scope: it did not run for lack of the relevant library and, in any case, was set up on random tokens, disconnected from our thesis. A sparse dictionary of per-token operators trained on the middle layer produced interpretable atoms, but its causal specificity test was null: the highest-activation atom was no more specific than a frequency-matched control, so it supports no causal claim. The single-layer knockout of the trigger’s write had a negligible effect, which motivated the multi-layer design. The first, adjacent version of the backdoor, described above, was discarded for contamination through the convolution. The multi-layer knockout limited to the write neutralized the behavior only partially, which motivated the pathway decomposition. We also preregistered a content-sensitivity prediction for the implicit knockout on the base model, that ablating a number’s contribution to the recurrent state would affect future positions less than ablating a matched content word; the effect ran in the opposite direction (number $\text{KL}_{\text{future}} = 6.2 \times 10^{-4}$ versus 5.3×10^{-4} for word-initial, with the number effect lower in only 46% of paired blocks), so the prediction was not confirmed. Finally, a related prediction about the decay of the Δ effect with distance was not confirmed, and we report it as a negative result.

4 Results

We organize our findings around the paper’s two contributions. Section 4.1 characterizes the selective time-step Δ as an interpretable memory-write policy, the property that makes the rest of the analysis tractable. Section 4.2 plants an adversarial backdoor inside the selective state-space module and localizes it causally, isolating a component of the behavior that does not flow through the recurrent state. Throughout, we mark each claim as an *observation* (a quantity we measure directly) or an *interpretation* (an inference we draw from it), and we close in Section 4.3 by stating which claims the evidence supports and which it does not.

All causal measurements share a single validity guarantee. Replacing the fused selective-scan kernels with our reference implementation reproduces the model’s native logits to within a maximum absolute difference of 0. Any change we report below is therefore caused by the intervention itself and not by the reimplementations that exposes the internal operators.

4.1 The Selective Time-Step Encodes a Content-Dependent Memory-Write Policy

Setup. For every token in the validation corpus ($n = 249,984$ tokens of WikiText-2) we extract the layer-24 time-step $\Delta_t = \text{softplus}(\text{dt_proj}(x_t))$, averaged over the `dt_rank` projection, and group tokens into six surface categories (word-initial, subword continuation, number, punctuation or symbol, whitespace, and a residual class). We take word-initial tokens as the reference category and report Cohen’s d against that reference, together with the corpus-level correlation between token log-frequency and Δ .

Observation. The mean time-step varies systematically with token category (Figure 3). Word-initial tokens carry the largest time-step (mean $\Delta = 0.093$), continuation and structural tokens the smallest (subword continuation 0.047, whitespace 0.034, residual class 0.029), with numbers (0.069) and punctuation (0.066) in between. The separation from the reference category is large: the most distinct category reaches $d = -1.34$, and three of the five non-reference categories exceed $|d| = 0.5$. The correlation between log-frequency and Δ is -0.014 , indistinguishable from zero.

Interpretation. Because Δ sets the discretization step that governs how strongly the current token is written into the recurrent state, this ordering reads as a memory-write policy rather than an incidental statistic. A large Δ drives the input into the state, so content-bearing word onsets are integrated aggressively; a small Δ leaves the state comparatively undisturbed, so whitespace, sub-word fragments, and punctuation are written weakly and the prior state persists. The near-zero frequency correlation is what makes this reading defensible: the policy tracks token *type*, not how common the token is, so the categorical structure is not a rare-token artifact. This decoupling is also a practical prerequisite for the causal experiments that follow, since it means our trigger and control conditions need not be matched on frequency for the comparison to be valid.

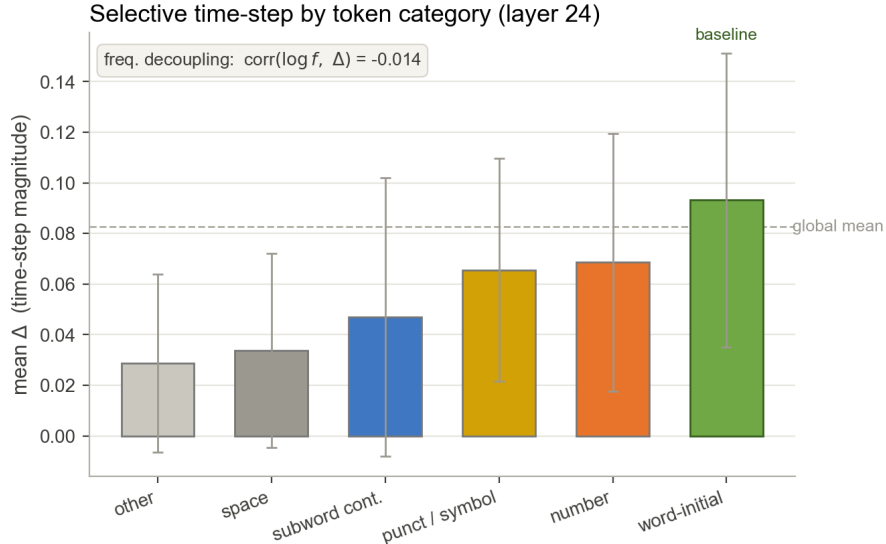


Figure 3: Selective time-step Δ by token category at layer 24 (mean $\pm$ standard deviation; $n = 249,984$ tokens). Word-initial tokens, the reference category, carry the largest time-step; structural and continuation tokens carry the smallest. The dashed line marks the global mean. The corpus correlation between token log-frequency and Δ is -0.014 , so the ordering reflects token type rather than frequency.

4.2 Localizing a Planted Backdoor Within the Recurrent State

Having established that the selective state-space module can be read mechanistically, we ask whether an adversarial behavior planted inside it can be localized to specific computational pathways. We fine-tune a low-rank adapter restricted to the input, state, and time-step projections, so that the planted trigger-to-target mapping is confined to the selective subsystem and cannot route through the output projection or the short convolution. The gap between trigger and target tokens is set wider than the convolution kernel, which forces any dependence to pass through the recurrent state rather than through local convolution alone.

Behavioral validation. On held-out sequences, the merged model produces the target continuation with a negative log-likelihood of 0.21 in the presence of the trigger versus 2.22 in its absence, and fires the target at a rate of 0.84 with the trigger versus 0.14 without. The backdoor is therefore both reliable and specific to the trigger, which licenses treating any intervention-induced suppression below as removal of a real planted behavior rather than as noise.

4.2.1 The State Write Is Causal but Distributed Across Depth

Observation. Ablating the trigger’s write to the recurrent state in a single layer has a negligible effect: the target log-probability drop stays below 0.02 regardless of which layer is chosen. The effect appears only when the write is ablated cumulatively across many layers, growing monotonically with depth and reaching a drop of 0.33 at the full 48 layers, at which point 42% of cases are neutralized (Figure 4). A random-position control, ablating the same number of writes at non-trigger positions, stays flat at every depth (drop 0.009 at full depth), giving a specificity of $36\times$ at the deepest setting. No depth at which the write is ablated reaches the 50% neutralization threshold.

Interpretation. The trigger’s contribution to the recurrent state is genuinely causal, since suppressing it moves the output far more than the matched control does. But the behavior is not stored in any single layer; it is written redundantly across depth, so removing any one layer’s contribution is inconsequential and even removing all of them leaves a substantial fraction of the behavior intact. This redundancy is itself a security-relevant finding: a defense that searches for and patches a single responsible layer would fail here, because no such bottleneck exists.

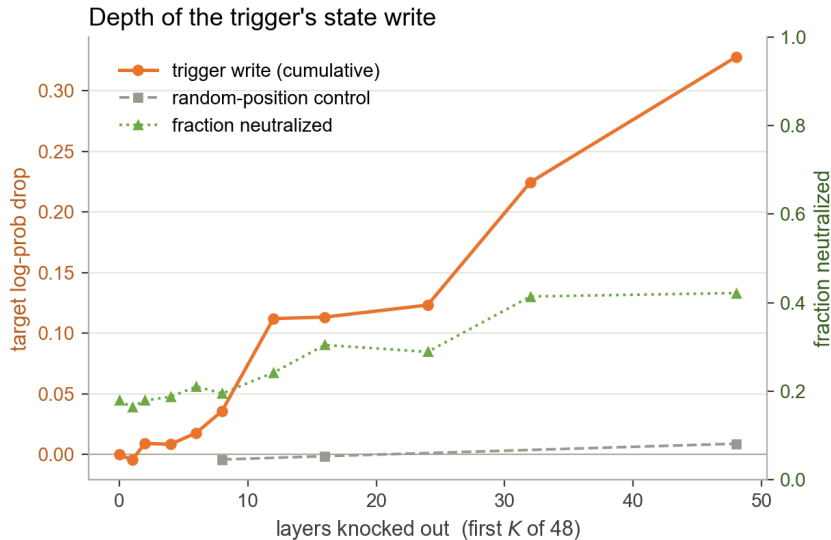


Figure 4: Cumulative depth knockout of the trigger’s write to the recurrent state. Ablating the write in only the first K layers has little effect until many layers are included; at $K = 48$ the target log-probability drops by 0.33 and 42% of cases are neutralized, while a random-position control remains flat ($36\times$ specificity). The behavior is distributed across depth, not localized in a single layer.

4.2.2 Decomposing the Recurrent-State Pathway

Observation. We next ablate the trigger across all layers and add one physical pathway at a time (Figure 5, Table 1). Ablating the direct write through $\Delta \cdot B \cdot u$ lowers the target probability from a baseline of 0.82 to 0.59. Extending the ablation to the local convolution band that feeds the state lowers it further to 0.31, and additionally suppressing the state readout through C brings it to 0.28. The random-position control remains at 0.81 throughout. Specificity against the control rises from $35\times$ for the direct write to $115\times$ for the full state-pathway ablation, and cumulative neutralization reaches 72%.

Interpretation. Each physical component of the recurrent-state subsystem, the input-dependent

write, the short convolution that precedes it, and the state readout, carries a measurable share of the planted behavior, and the sharp rise in specificity confirms that the suppression is trigger-specific rather than a generic consequence of perturbing the network. Together the three components account for most of the behavior. They do not, however, account for all of it: a residual of 0.28, well above chance and leaving the control’s 0.81 essentially untouched, survives the most aggressive state-targeted ablation we can construct. At face value this admits two explanations, an incomplete mask that fails to cover the relevant state positions, or a genuine pathway that bypasses the state. The next experiment distinguishes them.

Table 1: Pathway decomposition of the planted backdoor, ablated cumulatively across all 48 layers on held-out data. Target probability falls from baseline as each recurrent-state pathway is added, while a random-position control is unaffected; the strongest ablation leaves a non-state residual. Specificity is the log-probability drop divided by the control’s.

Mode	Target prob.	Log-prob drop	Frac. neutralized	Specificity
Baseline (no knockout)	0.820	+0.000	0.180	–
Write ($\Delta \cdot B \cdot u$)	0.591	+0.328	0.422	34.9×
+ convolution band	0.310	+0.971	0.680	103.5×
+ readout (C)	0.279	+1.077	0.719	114.8×
Random control	0.812	+0.009	0.195	–

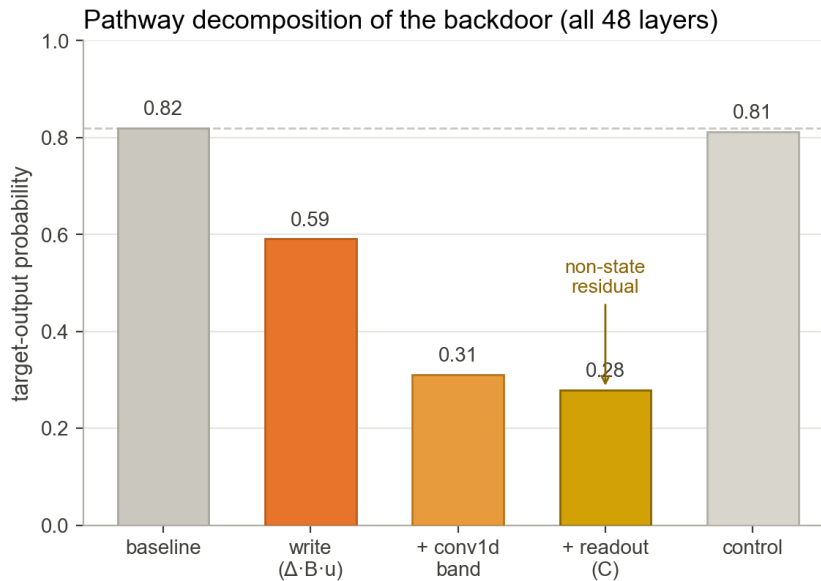


Figure 5: Pathway decomposition of the planted backdoor, ablated cumulatively across all 48 layers. Adding each recurrent-state pathway (write, then the convolution band, then the readout through C) lowers the target probability from 0.82 to 0.28, while the random-position control stays near 0.81. The strongest state-targeted ablation leaves a non-state residual.

4.2.3 The Residual Is a Genuine Non-State Pathway

Observation. If the residual were an artifact of an undersized mask, widening the masked band should keep lowering the target probability until the relevant state positions are covered. Instead, widening the band beyond the convolution window leaves the residual essentially unchanged: the target probability settles in the range 0.28 to 0.30, and the marginal change over the final widening step is 0.003 (Figure 6). The curve drops sharply for narrow bands and then plateaus.

Interpretation. The residual saturates, which rules out the incomplete-mask explanation: once the band is wide enough to cover the state’s effective receptive field, masking more positions buys nothing. We therefore read the residual as evidence of a genuine non-state pathway, a route by which part of the planted behavior reaches the output without passing through the recurrent state. This is consistent with formulations that recast the selective state-space layer as implicit attention while explicitly noting that such a view omits the gating branch, the short convolution, and the elementwise nonlinearity (Zimmerman et al., 2024); a backdoor partly carried by that omitted branch would behave exactly as we observe. The security implication is the central message of this work: exhaustively ablating the recurrent state suppresses most of a planted behavior in a selective state-space model, but not all of it, so an interpretability or defense method that reasons only about the state is structurally incomplete.

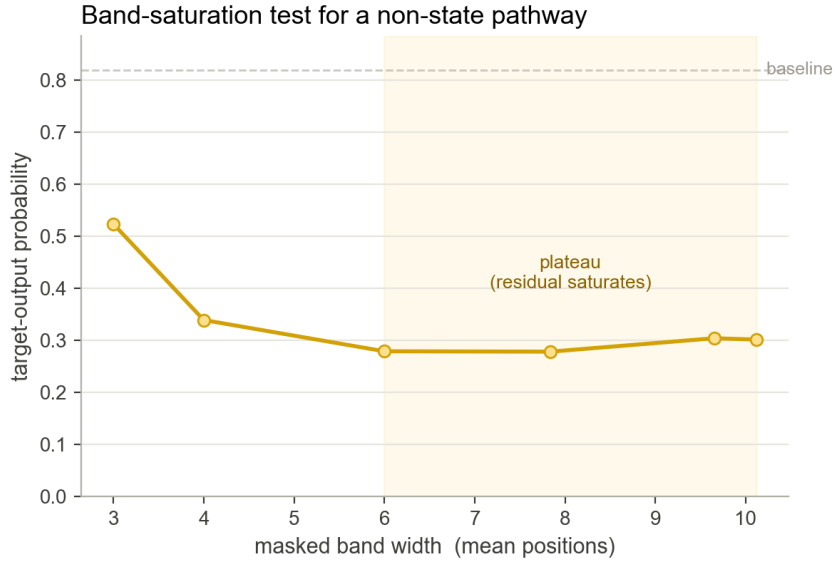


Figure 6: Band-saturation test for the residual. Widening the masked band beyond the convolution window leaves the residual target probability near 0.28 to 0.30 (marginal change 0.003 at the final step). The plateau indicates the residual is a genuine non-state pathway rather than an artifact of an incomplete mask.

4.3 Robustness and Limits of the Evidence

What the evidence supports. Two design choices make the causal claims robust to the most obvious confounds. First, every intervention is paired with a random-position control that ablates the same number of operators at non-trigger positions; this control stays flat throughout (specificity from $36\times$ for the single-pathway write to $115\times$ for the full state pathway), so the measured suppression is attributable to the trigger and not to a general degradation from removing computation. Second, the saturation test addresses the one confound that the controls do not, namely that the surviving

residual might reflect an undersized mask, and rules it out directly by showing the residual plateaus as the mask widens. Within the configuration we study, the ordering of the effects (write < convolution band < readout, all far above control) and the existence of a saturated non-state residual are therefore well supported.

What the evidence does not establish. All quantities are point estimates from a single configuration: one training run with a fixed seed, a single 790M-parameter model, one trigger-target pair, and 128 held-out evaluation sequences. We do not report multiple seeds, multiple triggers, or confidence intervals, so we make no claim about the sampling variability of the specific numbers, such as the 0.28 residual or the 42% single-pathway neutralization. These should be read as the outcome of one carefully controlled experiment rather than as estimates whose dispersion we have characterized. Whether the non-state residual persists in magnitude across seeds, alternative triggers, and larger models is the natural next test and the principal threat to external validity; we regard establishing that invariance as the most important direction for follow-up work.

5 Discussion and Limitations

5.1 Discussion

Our central result reframes what it means to audit a selective state-space model for safety. In a Transformer, localizing a behavior through attention and then ablating that attention is often sufficient, because attention is the dominant channel through which information from different positions is mixed. The recurrent state of an SSM is the closest structural analogue to that channel, and the natural expectation is that severing a trigger’s contribution to the state should remove a behavior that depends on it. We find that this expectation is only partially correct. Exhaustively ablating the trigger’s entire recurrent-state pathway, the input-dependent write, the causal convolution that feeds it, and the state readout, across every layer, suppresses most of a planted backdoor but leaves a saturated residual that no further state-targeted ablation removes. Part of the behavior reaches the output through a route that does not pass through the recurrent state. The mental map an auditor carries over from attention-based models therefore does not transfer cleanly; it has to be widened to include the components that sit beside the state.

This matters specifically in an adversarial setting. A defender who assumes the recurrent state is the complete information channel has a blind spot, and that blind spot is exploitable. Our backdoor reached the non-state pathway through the ordinary dynamics of low-rank fine-tuning rather than by deliberate construction; an attacker who understands this structure could plant behavior in the non-state route on purpose, precisely to evade a defense that reasons only about the state. The distributed-depth finding compounds the concern. The behavior is not stored in any single layer but written redundantly across the stack, so the localize-and-patch strategy that is common for Transformers, finding a responsible head or layer and editing it, has even less traction here, because no such bottleneck exists to be found.

The result also gives an empirical, safety-relevant edge to a property that was previously noted as a technical caveat. Formulations that recast the selective layer as implicit attention are explicit that this view captures only the state scan and ignores the convolution, the gating branch, and the elementwise nonlinearities (Zimerman et al., 2024). We provide causal evidence that those ignored components can carry an adversarial behavior, not merely contribute to capability. An interpretability or defense method built on the implicit-attention view of the state alone is therefore structurally incomplete against an adversary, and the gap it leaves is measurable. Read alongside the persistence result of Hubinger et al., 2024, where a planted behavior survives standard safety retraining, our finding adds a second axis of backdoor robustness: in an SSM the behavior also

survives exhaustive causal ablation of the state. Mechanistic access does not, by itself, guarantee a complete defense.

None of this argues that SSM internals are opaque. The opposite holds for the observational half of our study: the selective time-step is a legible, content-dependent write policy that is statistically decoupled from token frequency, which means the selective mechanism is genuinely analyzable. The lesson is narrower and more actionable. Reading the state is necessary but not sufficient, and a safety method for these architectures must account for every route that carries information to the output, not only the one that most resembles attention.

5.2 Limitations and Dual-Use Considerations

The most important limitation is one of statistical scope. Every quantity we report is a point estimate from a single configuration: one training run with a fixed seed, a single 790M-parameter model, one trigger-target pair, and 128 held-out evaluation sequences, with no confidence intervals. We make no claim about the sampling variability of the specific numbers, such as the residual near 0.28 or the 42% single-pathway neutralization, and they should be read as the outcome of one carefully controlled experiment rather than as estimates whose dispersion we have characterized. This constraint is partly a resource one. All experiments were run inside a single hosted notebook environment on commodity accelerators, and the combination of a solo author and the fixed hackathon timeframe made a multi-seed, multi-model sweep infeasible within the available compute and time.

Our threat model is also deliberately narrow. We study a single class of backdoor, a fixed trigger mapped to a fixed target, planted by a low-rank adapter that was confined to the input, state, and time-step projections of the selective mixer. That confinement was not a controlled choice but a property of the fine-tuning library, which blocks adaptation of the output projection and the convolution for these models. We did not study backdoors planted across the full parameter set, semantically defined triggers, or behaviors trained directly into the non-state pathway, and we did not evaluate defenses that act on the convolution and gating branches directly rather than through band widening. How a behavior distributes across the state and non-state routes may differ substantially under any of these settings.

Two assumptions underlie the causal claims, and it is worth stating how the interpretation would change if either failed. First, we assume that placing the trigger and target farther apart than the convolution kernel forces their dependence through the recurrent state, which is what lets us attribute the single-write effect to the state rather than to local convolution. If that assumption did not hold, part of what we attribute to the state would instead be convolutional propagation; the separate ablation of the convolution band and the zero reconstruction error of our reference path are what make us confident it does hold. Second, we assume that the surviving residual reflects a genuine non-state pathway rather than an incompletely masked state. The saturation test exists precisely to probe this assumption, and the residual does plateau as the mask widens. Had the residual instead continued to fall with wider masks, the non-state interpretation would be unsupported and the result would reduce to a statement about mask coverage. The plateau is therefore not a cosmetic detail but the load-bearing evidence for the central claim.

Finally, a note on dual use. The same procedure that locates a planted behavior could in principle guide an attacker toward placing one in the non-state pathway to evade state-centric defenses. We judge the net effect to be defensive: the trigger maps to a deliberately nonsensical phrase, the behavior is benign, the planting technique is already public, and we release the reproducible procedure rather than backdoored weights. The finding’s value accrues to defenders, who learn that auditing the recurrent state alone is insufficient.

5.3 Future Work

The most pressing next step follows directly from the scope limitation: re-running the full pipeline across multiple seeds, several triggers, and a range of model sizes, with confidence intervals, to establish whether the non-state residual persists in magnitude rather than appearing in a single configuration. That invariance, or its absence, is what would turn the present finding from a controlled demonstration into a general claim about selective state-space models.

A second direction is to characterize what the non-state pathway actually is. Our decomposition separated the direct write, the convolution band, and the readout, but treated everything that survives as a single residual. Isolating the gating branch and the elementwise nonlinearities explicitly, the components that the implicit-attention view omits, would localize the residual more precisely and indicate which part of the mixer a defense would need to cover.

The natural architectural extension is to hybrid SSM-Transformer models, where the question we ask becomes richer rather than merely larger. Models such as Jamba (Lieber et al., 2024) and the Zamba and Zamba2 families (Glorioso et al., 2024a; Glorioso et al., 2024b) interleave Mamba layers with explicit attention blocks, so a planted behavior could in principle distribute across three routes at once: the recurrent state, the non-state pathway we identify here, and genuine attention. Zamba2 is a particularly direct continuation of our method, since its shared attention blocks are themselves specialized with per-layer LoRA adapters, the same mechanism we use to plant the backdoor. Asking which of these routes carries an adversarial behavior, and whether ablating any single route is sufficient, would test whether the incompleteness we observe in a pure SSM is mitigated or compounded when attention is added back. Finally, on the defensive side, the practical question our results raise is whether a defense that explicitly covers the non-state routes, rather than the state alone, can close the gap we measure, which would move this line of work from diagnosis toward mitigation.

6 Conclusion

We asked whether the interpretability tooling used to audit models for safety, developed almost entirely around the attention mechanism of the Transformer, transfers to selective state-space models such as Mamba, which carry information through a compressed recurrent state rather than attention. We find two things. First, the selective time-step parameter is not uniform across tokens but follows a structured write policy that depends on token type and is statistically decoupled from token frequency, showing that the selective mechanism replacing attention is itself legible and analyzable. Second, by planting an adversarial backdoor inside the selective subsystem and localizing it causally, decomposing and fully ablating the trigger’s recurrent-state pathway, we suppress most of the behavior but not all of it: a saturated residual survives that does not pass through the state, indicating a second pathway that bypasses it entirely.

The safety implication is direct. Interpretability does transfer to SSMs, but state-centric defenses are incomplete: exhaustively ablating the recurrent state does not fully neutralize the backdoor, because part of the behavior travels by a route that bypasses it. For an auditor this means that reading the state, the closest analogue to attention, is necessary but not sufficient, and that a reliable defense for these architectures must cover every route that carries information to the output, not only the one that most resembles attention. As selective state-space models and SSM-Transformer hybrids are deployed at scale, closing that blind spot shifts from a technical detail to a precondition for auditing them with confidence.

7 Code and Data

Code repository. All code for the experiments in this paper, including the operator-extraction pipeline, the observational characterization of Δ , the implicit-knockout intervention, the backdoor training, and the pathway-decomposition and saturation analyses, is available at github.com/Playmaker3334/mamba-spd. The repository also contains the raw result files (the JSON outputs read by the figure-generation scripts) so that every number and figure in this paper can be reproduced directly.

Data. The base model is the publicly available Mamba-790M checkpoint, [state-spaces/mamba-790m-hf](https://state-spaces.com/mamba-790m-hf). All analyses use WikiText-2 in its [wikitext-2-raw-v1](https://huggingface.co/openai-community/wikitext-2-raw-v1) configuration, a standard public corpus, with the training split used for operator extraction and the frequency reference and the validation split reserved for all causal tests. No private or newly collected data is used.

Other artifacts. The backdoored model is produced by the training script in the repository rather than distributed as weights. The planted trigger maps a rare token string to a deliberately nonsensical target phrase, so the behavior is benign and carries no realistic misuse potential, and the backdoor-planting technique itself is already public. We nonetheless distribute the procedure rather than the merged weights, so that the artifact a reader obtains is the reproducible method rather than a ready-made backdoored model.

8 LLM Usage Statement

We used an LLM (Claude) as a coding and editing assistant throughout this project. Specifically, it helped generate and debug the experiment code, including the operator-extraction pipeline, the knockout interventions, and the figure-generation scripts, and it assisted in drafting and revising the prose of this report, including catching inconsistencies between sections. All experiments were designed and run by the author, and every result, number, and claim in this paper was independently verified against the raw output files before inclusion.

References

References

- [1] Ali, A., Zimerman, I., & Wolf, L. (2024). The Hidden Attention of Mamba Models. [arXiv:2403.01590](#).
- [2] Burget, J. (2024). mamba-sae: Training and evaluating Sparse Autoencoders for Mamba. github.com/joelburget/mamba-sae.
- [3] Endy, N., Grosbard, I. D., Ran-Milo, Y., Slutzky, Y., Tshuva, I., & Giryes, R. (2025). Mamba Knockout for Unraveling Factual Information Flow. [arXiv:2505.24244](#).
- [4] Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting Recall of Factual Associations in Auto-Regressive Language Models. [arXiv:2304.14767](#).
- [5] Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. [arXiv:2312.00752](#).
- [6] Hossain, T., et al. (2025). Characterizing Mamba’s Selective Memory using Auto-Encoders. [arXiv:2512.15653](#).
- [7] Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. [arXiv:2401.05566](#).
- [8] Wang, J., Ge, X., Shu, W., Tang, Q., Zhou, Y., He, Z., & Qiu, X. (2024). Towards Universality: Studying Mechanistic Similarity Across Language Model Architectures. [arXiv:2410.06672](#).
- [9] Zimerman, I., Ali, A., & Wolf, L. (2024). Explaining Modern Gated-Linear RNNs via a Unified Implicit Attention Formulation. [arXiv:2405.16504](#).
- [10] Glorioso, P., Anthony, Q., Tokpanov, Y., Whittington, J., Pilault, J., Ibrahim, A., & Millidge, B. (2024). Zamba: A Compact 7B SSM Hybrid Model. [arXiv:2405.16712](#).
- [11] Glorioso, P., Anthony, Q., Tokpanov, Y., Golubeva, A., Shyam, V., Whittington, J., Pilault, J., & Millidge, B. (2024). The Zamba2 Suite: Technical Report. [arXiv:2411.15242](#).
- [12] Lieber, O., Lenz, B., Bata, H., et al. (2024). Jamba: A Hybrid Transformer-Mamba Language Model. [arXiv:2403.19887](#).