Effective Receptive Fields in Mamba: Conductance Products and Selectivity*

Zy Li
Department of Computer Science, Seattle University, Seattle, WA, USA
zli10@seattleu.edu

Abstract

Mamba's selective state-space model (SSM) achieves long-range dependence through input-dependent gating, yet existing analyses offer limited insight into *why* a particular parameter configuration produces a particular effective receptive field (ERF). We define per-step *conductances* $G_t = I - \exp(A\,\Delta_t)$ and express the influence of past inputs through signed sums of running products of $I - G_t$. With time-invariant coefficients, this is a discrete Laplace transform of a signed spectral measure over the conductances. For i.i.d. conductances, each dimension's expected conductance product decays exponentially, while mixtures of decay rates can yield approximately power-law decay over finite lag ranges. For a Mamba block, we derive an exact decomposition of the output's sensitivity to past inputs that includes differentiable selectivity: every long-range pathway passes through the same running products, with pathway-specific coefficients. Holding selectivity fixed gives a closed form for the block's ERF. On single-layer language models trained on WikiText-103, this formula approximates the full-model ERF, measured with a squared-logit probe, within a factor of 2.3 up to lag 500. On associative-recall and selective-copy models, it underestimates the pooled margin-gradient energy at the relevant inputs by a factor of about 5 to 22, and the decomposition identifies the pathway through Δ_t in the write as the omitted component with the largest individual energy. Over three seeds, slow initialization of A gives the highest mean accuracy on both tasks, and the layers of a pretrained Mamba-370M span very different conductance regimes.

1 Introduction

Selective state-space models (SSMs) such as Mamba [Gu and Dao, 2024] match or surpass Transformer language models with comparable parameter counts at a fraction of the inference cost, yet the mechanism by which their input-dependent gating produces long-range memory remains incompletely understood. In the recurrence $x_t = \bar{A}_t\, x_{t-1} + \bar{B}_t\, u_t$, each state dimension transmits influence through a product of input-dependent retention factors. How do these retention factors combine with the learned input and output mappings to determine the effective receptive field (ERF)?

The answer, we show, lies in the model's conductances and in the signed coefficients, or residues, that weight them. The zero-order-hold discretization of the decay, with state matrix A and input-dependent step size Δ_t, defines per-step conductances $G_t = I - \exp(A\,\Delta_t)$, whose diagonal entries $g_{i,t} \in (0, 1)$ control how strongly state dimension i is attenuated at time step t. Large Δ_t creates a strong drain (reset), and small Δ_t preserves state (coast). Borrowing vocabulary from resistor networks, where G_t acts as a shunt conductance to ground, the influence of an input at lag τ is a signed sum, over state dimensions $i = 1, \ldots, d_{\text{state}}$, of running products of $1 - g_{i,t}$, which we call conductance products.

*Code: https://github.com/zypapers/mamba-conductance-erf

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

With time-invariant coefficients ($g_{i,t} = g_i$) the impulse response is a weighted sum of exponentials,

$$h_\tau \;=\; \sum_{i=1}^{d_{\text{state}}} c_i \,(1-g_i)^\tau \,\bar{b}_i, \tag{1}$$

a discrete Laplace transform of the signed measure $\nu = \sum_i (c_i \bar{b}_i)\,\delta_{g_i}$, where δ_g is a unit point mass at g, c_i and $\bar{b}_i$ are the output and input weights of dimension i, and the residue $c_i \bar{b}_i$ is the weight of δ_{g_i}. This form already shows where heavy tails can come from: for i.i.d. conductances, each dimension's expected conductance product decays exponentially, while a *mixture of decay rates across state dimensions* can yield approximately power-law decay over finite lag ranges.

This paper makes three contributions.

First, we derive an exact decomposition of a Mamba block's sensitivity to past inputs, including differentiable selectivity (Section 3). Every long-range pathway propagates through the same conductance products, with pathway-specific coefficients. With selectivity held fixed, the decomposition reduces to a closed-form block formula for the ERF that includes the readout and the input path. Its special case at the scan, the scan probe (Eq. (15)), weights each channel's squared impulse response by the squared scan output. With time-invariant coefficients, that impulse response is the discrete Laplace transform of the signed spectral measure. The kernel inside these formulas, $C_T\,\bar{A}_T \cdots \bar{A}_{s+1}\,\bar{B}_s$, is the data-controlled operator of Ali et al. [2025], which Zimerman et al. [2025] extend to the whole block. What is new is the ERF in this form, a squared gradient norm that includes the probe's readout and is written through conductance products, together with its measured agreement with the full model and an exact decomposition that extends Huang et al. [2025] to the whole block.

Second, we test the block formula against the full model (Section 4.3). On 18 single-layer WikiText language models (three A-initializations $\times$ three training seeds $\times$ two weight-decay settings), it approximates the full-model ERF within a factor of 2.3 at every lag from 1 to 500, and its fitted log–log slope over lags 10–100 lies within 0.11 of the full model's (0.36 over lags 100–1000). The block formula, computed from the conductance products, readout and input path without backpropagating through the scan, approximates the ERF that backpropagation measures. The scan probe alone has a worst-case miss of about $1238\times$ against the same ERF up to lag 500, after normalization at $\tau = 1$. Including the readout and the input path reduces it to about $2.2\times$. Beyond lag 500, the block formula's largest shortfall is approximately $5.13\times$.

Third, we use the decomposition to locate where the block formula falls short, and we measure what initialization changes (Section 4.5; Appendices F, G, I). In all 18 models trained for multi-query associative recall (MQAR) or selective copying, the squared gradient of the answer's logit margin at the relevant input positions, pooled over queries, is about 5–22 times the block formula's, and the pathway through Δ_t in the write (the input term $\Delta_t B_t u_t$) carries the largest individual energy. Over three seeds, slow A-initialization (smaller initial $|A|$) gives the highest mean accuracy on both tasks (MQAR 1.00 vs. 0.76 for default and 0.75 for fast; selective copying 0.65 vs. 0.45 and 0.36). We do not establish that the ERF mediates this effect. Across the 48 layers of a pretrained Mamba-370M, the fraction of conductances below 0.01 ranges from about 0.2% to 96%.

The decomposition is exact for one block. Multi-layer models are analyzed layer by layer. Experiments use single-layer models trained from scratch, a two-layer model, and all 48 layers of a pretrained Mamba-370M (Appendices J, I).

2 Background

2.1 Resistor Networks

A resistor network is a weighted graph whose edges carry conductances. With network Laplacian L and shunt conductances $G = \mathrm{diag}(g_i)$ to source voltages s, Kirchhoff's current law gives, for the node voltages v,

$$(L + G)\,v = G\,s, \tag{2}$$

and $(L+G)^{-1}G$ is the network's discrete Green's function. On an infinite lattice, where the Green's function exists, its decay sets the range of influence: exponential for a gapped spectrum, while a spectrum reaching the origin can give algebraic decay [Cserti, 2000, Doyle and Snell, 1984]. We use this picture as a motivating analogy for the temporal recurrence (Appendix A).

2.2 State-Space Models and Mamba

Structured state space models [Gu et al., 2022a] parameterize a linear dynamical system

$$\dot{x}(t) = A\,x(t) + B\,u(t), \qquad y(t) = C\,x(t) + D\,u(t), \tag{3}$$

where $A = \mathrm{diag}(a_1, \ldots, a_N)$ with $N = d_{\mathrm{state}}$, $B \in \mathbb{R}^N$, $C \in \mathbb{R}^N$ and $D \in \mathbb{R}$ is a skip term passing the input directly to the output. Zero-order hold discretization with step size Δ yields the recurrence $x_t = \bar{A}\,x_{t-1} + \bar{B}\,u_t$, $y_t = C\,x_t + D\,u_t$, where $\bar{A} = \exp(A\Delta)$ and $\bar{B} = (e^{A\Delta} - I)\,A^{-1}\,B$. Mamba keeps this $\bar{A}$ but uses the first-order input weight $\bar{B} = \Delta B$ (Section 3.2). We use post-update indexing (y_t reads state after processing u_t), so the lag-τ kernel is $C\bar{A}^\tau\bar{B}$ with τ retention factors. The skip term acts only at lag 0.

Mamba [Gu and Dao, 2024] introduces *selectivity*: the parameters Δ_t, B_t and C_t become input-dependent, so that on a given input sequence the scan is linear time-varying (LTV), while the input-output map is nonlinear. A single Mamba block applies an input projection, a short causal convolution, the selective SSM scan, gating with a sigmoid linear unit (SiLU), and an output projection, wrapped in a residual connection. The selective scan is the computational bottleneck and the only component with long-range temporal reach.

3 Spectral Measure Framework

We build the framework in three steps. Sections 3.1–3.4 define the conductances and, for time-invariant coefficients, the signed spectral measure of the scan recurrence, whose impulse response is a discrete Laplace transform of that measure. Section 3.5 gives the effective receptive field of a whole Mamba block in closed form when selectivity is held fixed. Section 3.6 removes that restriction: it decomposes the block's sensitivity to past inputs exactly, including differentiable selectivity.

3.1 LTI Baseline and Temporal Green's Function

Consider first the linear time-invariant (LTI) specialization of the SSM recurrence of Section 2.2, where $\bar{A}$, $\bar{B}$, C are constant. The impulse response at lag τ is

$$h_\tau = C\,\bar{A}^\tau\,\bar{B}, \qquad \tau \geq 1. \tag{4}$$

This is the temporal Green's function: the influence of a unit input at time s on the output at time $t = s + \tau$. Because the recurrence is causal, $h_\tau = 0$ for $\tau < 0$. We restrict attention to $\tau \geq 1$, so each lag carries τ retention factors.

For diagonal $\bar{A}$ with entries $\bar{a}_i = e^{a_i\Delta}$, the impulse response decomposes as $h_\tau = \sum_i c_i\bar{a}_i^\tau\bar{b}_i$, a sum of exponentials. The LTI system serves as the reference case against which we measure the effects of Mamba's selectivity.

3.2 Conductances and the Trained Recurrence

We now map Mamba's learned parameters onto the circuit vocabulary of the grid picture.

Discretization. Mamba discretizes the decay of the continuous system $\dot{x} = Ax + Bu$ by zero-order hold with a learned, input-dependent step size Δ_t, and the input by a first-order step:

$$\bar{A}_t = \exp(A\,\Delta_t), \qquad \bar{B}_t = \Delta_t\,B_t. \tag{5}$$

Mamba's training kernel computes this recurrence, and every result below uses it. The exact zero-order-hold input weight, $(\exp(A\,\Delta_t) - I)\,A^{-1}B_t$, agrees with $\Delta_t B_t$ only to first order in $A\Delta_t$. The diagonal of A is strictly negative (it is parameterized as $a_i = -\exp((A_{\log})_i)$), so $\bar{A}_t$ is a contraction (Appendix L.1).

Shunt conductance. Define the per-step *shunt conductance*

$$G_t \;:=\; I - \bar{A}_t \;=\; I - \exp(A\,\Delta_t) \;\succeq\; 0. \tag{6}$$

Each diagonal entry $g_{i,t} = 1 - e^{a_i\Delta_t}$ lies in $(0, 1)$, since $a_i < 0$ and $\Delta_t > 0$. The recurrence becomes

$$x_t = (I - G_t)\,x_{t-1} + \Delta_t\,B_t\,u_t. \tag{7}$$

Each $g_{i,t}$ controls how much state dimension i is attenuated at step t. In circuit terms, $g_{i,t}$ is a *conductance to ground*: it drains the existing voltage (hidden state) at node t before the next time step. For a given input projection B_t, the same step size Δ_t sets both the drain, through G_t, and the write weight $\Delta_t B_t$.

Small-step approximation. For small $|a_i \Delta_t|$, $g_{i,t} \approx |a_i|\, \Delta_t$. All results below use the exact form $g_{i,t} = 1 - e^{a_i \Delta_t}$ (Appendix L.2).

3.3 Selectivity: Reset and Coast

In the LTI baseline, $G_t = G$ is constant: every time step applies the same contraction. Mamba's selectivity makes G_t input-dependent:

$$\Delta_t = \mathrm{softplus}\big(\mathrm{Linear}(u_t)\big), \qquad G_t = I - \exp(A\, \Delta_t),$$

so each token reconfigures the conductances through which subsequent information propagates.

Two regimes. The extremes of the selectivity mechanism give two regimes, named here in circuit terms:

- **Reset** (large Δ_t, large G_t). The shunt conductance to ground is strong. The factor $(I - G_t)$ contracts the hidden state toward zero, draining the old voltage, while the write $\Delta_t B_t$ injects the current token with a large weight. This is a low-impedance path to ground: old information is largely discarded and replaced by the present input.
- **Coast** (small Δ_t, small G_t). The shunt to ground is weak: the node acts as an insulator. The factor $(I - G_t) \approx I$ preserves the hidden state nearly unchanged, while the write $\Delta_t B_t \approx 0$ rejects the new input. Past information is carried forward almost without loss.

Information preservation as a running product. The influence of a past input at time s on the state at time $t > s$ is governed by the running product

$$\prod_{r=s+1}^{t} (1 - g_{i,r}) \tag{8}$$

along each state dimension i. When the model coasts (small g), this product stays close to unity and information is preserved. A single reset event (large $g_{i,r}$ at any intermediate step r) collapses the product toward zero, suppressing contributions from all inputs before step r. Selectivity thus provides content-dependent control over how much accumulated state is retained.

3.4 Spectral Measure and the Scan-Level ERF

Per-channel impulse response. Unrolling the recurrence (7) from input at time s to output at time $t \geq s$ (an empty product equals 1), and restoring the channel index $j = 1, \ldots, d_{\mathrm{inner}}$, the contribution of input $u_{j,s}$ to output $y_{j,t}$ along state dimension i is

$$h_{j,i}(t, s) \;=\; C_{i,t}\, \bar{b}_{j,i,s} \prod_{r=s+1}^{t} (1 - g_{j,i,r}), \qquad \bar{b}_{j,i,s} = \Delta_{j,s}\, B_{i,s}. \tag{9}$$

Channel j has state matrix $\mathrm{diag}(A_{j,1}, \ldots, A_{j,N})$ and step size $\Delta_{j,t}$, while B_t and C_t are shared across channels. As in Section 2.2, the lag-τ response carries τ retention factors. This matches the kernel $\bar{K}_k = C \bar{A}^k \bar{B}$ of Gu and Dao [2024]. The per-channel impulse response sums over state dimensions:

$$h_j(t, s) \;=\; \sum_{i=1}^{N} h_{j,i}(t, s) \;=\; \sum_{i=1}^{N} C_{i,t}\, \bar{b}_{j,i,s} \prod_{r=s+1}^{t} (1 - g_{j,i,r}), \tag{10}$$

where $N = d_{\mathrm{state}}$. In the time-invariant case, this reduces to

$$h_{\tau,j} \;=\; \sum_{i=1}^{N} c_i\, (1 - g_{j,i})^{\tau}\, \bar{b}_{j,i}, \qquad \tau \geq 1. \tag{11}$$

In this case, each channel j has a signed spectral measure

$$\nu_j \;=\; \sum_{i=1}^{N} (c_i\,\bar{b}_{j,i})\,\delta_{g_{j,i}}, \tag{12}$$

so that $h_{\tau,j} = \int_0^1 (1-g)^\tau \, d\nu_j(g)$ exactly: the impulse response is a discrete Laplace transform of the measure. Equivalently, placing the same signed weights at rates $-\log(1-g)$ gives a Laplace transform sampled at integer lags. The residues $c_i\bar{b}_{j,i}$ carry signs, so terms can partially cancel when summed over state dimensions (Appendix L.3).

From impulse response to ERF. For a probe ℓ_T, a scalar function of the output at target position T, the effective receptive field is the expected squared gradient norm with respect to the block input $e_{T-\tau}$ (the token embedding, for a single-layer model):

$$\mathrm{ERF}(\tau) \;=\; \mathbb{E}\left[\left\| \frac{\partial \ell_T}{\partial e_{T-\tau}} \right\|^2\right], \tag{13}$$

with the expectation over data and target positions T. For our language-model measurements, we use $\ell_T = \|\mathrm{logits}_T\|^2$ as the probe. The scan has its own ERF, defined in the same way at the scan: probe $\|y_T\|^2$ and input $u_{T-\tau}$. The scan output is $y_{j,T} = \sum_i C_{i,T}\, x_{j,i,T} + D_j\, u_{j,T}$, where D_j is Mamba's per-channel skip term, which affects $\partial y_{j,T}/\partial u_{j,T-\tau}$ only at $\tau = 0$. With Δ_t, B_t and C_t held at their evaluated values while we differentiate with respect to the inputs (frozen selectivity), the gradient through channel j at lag $\tau \geq 1$ is $2\,y_{j,T}\,h_j(T, T-\tau)$ (Appendix L.4), so the scan's ERF is

$$\mathrm{ERF}_{\mathrm{scan}}(\tau) \;=\; 4\,\mathbb{E}\left[\sum_{j=1}^{d_{\mathrm{inner}}} y_{j,T}^2\; h_j(T, T-\tau)^2\right]. \tag{14}$$

This is exact for frozen selectivity, and it has three properties: (i) the residues are signed, so positive and negative terms cancel; (ii) each channel's impulse response is squared before summing over channels; and (iii) each channel is weighted by the squared output $y_{j,T}^2$, skip term included, which comes from the chain rule through $\|y_T\|^2$. The unsigned measure $\mu_j = \sum_i |c_i\,\bar{b}_{j,i}|\,\delta_{g_{j,i}}$, which replaces each residue by its absolute value, bounds each channel's impulse response, $|h_{\tau,j}| \leq \int (1-g)^\tau \, d\mu_j(g)$, so before normalization the unsigned form overestimates the scan's ERF. How much depends on the lag and the model (Appendix L.4).

The scan probe. Retaining the full time dependence and position-specific weights, and averaging over target positions T and sequences, gives

$$h_{\mathrm{pred}}(\tau) \;=\; \mathbb{E}\left[\sum_{j=1}^{d_{\mathrm{inner}}} y_{j,T}^2 \left[\sum_{i=1}^{N} C_{i,T}\;\bar{b}_{j,i,T-\tau} \prod_{r=T-\tau+1}^{T} (1-g_{j,i,r})\right]^2\right], \tag{15}$$

which we call the scan probe. By Eq. (14) it equals $\mathrm{ERF}_{\mathrm{scan}}/4$ identically (Section 4.2). Replacing every quantity by its time average gives the static mixture of Appendix C, an LTI approximation.

Scope of the scan-level form. The scan-level ERF holds Δ_t, B_t and C_t fixed and reads out at the scan. A Mamba block differs from it in two ways: the input path and the readout that wrap the scan, and differentiable selectivity. Section 3.5 adds the first in closed form, and Section 3.6 accounts for the second exactly.

3.5 The Block ERF with Frozen Selectivity

A Mamba block maps its input $e_t \in \mathbb{R}^{d_{\mathrm{model}}}$ to its output in four steps. An input projection gives the scan branch $W_{\mathrm{in}}\,e_t$ and the gate $z_t = W_z\,e_t$. A depthwise causal convolution with four taps, followed by SiLU, gives the scan input $u_t = \mathrm{SiLU}(p_t)$, with $p_t = \sum_{m=0}^{3} w_m \odot W_{\mathrm{in}}\,e_{t+m-3} + b_{\mathrm{conv}}$, where $\odot$ denotes elementwise multiplication. The selective scan of Section 3.2 uses $\Delta_t = \mathrm{softplus}(q_t)$ with $q_t = W_\Delta u_t + b_\Delta$, $B_t = W_B\,u_t$ and $C_t = W_C\,u_t$. The gated output is $o_t = W_{\mathrm{out}}\big(y_t \odot \mathrm{SiLU}(z_t)\big) + e_t$, and a probe ℓ_T is any differentiable function of o_T, such as the squared logits $\|W_{\mathrm{lm}}\,o_T\|^2$.

Hold Δ_t, B_t and C_t fixed. For $\tau \geq 1$, the input $e_{T-\tau}$ reaches ℓ_T only through the scan inputs u_s at the positions $s = T - \tau + k$, $k = 0, \ldots, \min(3, \tau)$, that the convolution covers. The gate and the residual at $T - \tau$ act only at that position. The chain rule gives (Appendix L.5)

$$\frac{\partial \ell_T}{\partial e_{T-\tau}} = 2\, W_{\text{in}}^\top v, \qquad v_j = a_{j,T} \sum_{k=0}^{\min(3,\tau)} w_{j,3-k}\, \text{SiLU}'(p_{j,s})\, K_j(T,s), \quad s = T - \tau + k, \quad (16)$$

where $K_j(T,s) = h_j(T,s) + D_j\, \mathbb{1}[s = T]$ is the per-channel impulse response (10) with the skip term, and $a_T = \frac{1}{2}\nabla_{y_T} \ell_T$, half the gradient of the probe at the scan output, is the readout. For the squared logits, $a_{j,T} = \text{SiLU}(z_{j,T})\, \left[W_{\text{out}}^\top W_{\text{lm}}^\top W_{\text{lm}}\, o_T\right]_j$. The block ERF is therefore

$$\text{ERF}_{\text{block}}(\tau) = 4\, \mathbb{E}\left[\|W_{\text{in}}^\top v\|^2\right]. \quad (17)$$

It is built from the same conductance products as the scan-level form. Its residues now also contain the readout a_T and the input path (the taps w, the activation derivative SiLU' and the projection W_{in}). The scan's ERF (14) is recovered by two changes together: the probe $\|y_T\|^2$ in place of the logits, and the gradient with respect to $u_{T-\tau}$ rather than $e_{T-\tau}$. Like the scan-level form, the block formula is exact for frozen selectivity: it matches automatic differentiation with Δ_t, B_t and C_t detached. Whether it approximates the full model, where they depend on the input, is tested in Section 4.3.

3.6 Exact Decomposition with Differentiable Selectivity

With differentiable selectivity, the scan input u_s also enters ℓ_T through Δ_s, through B_s and, when $s = T$, through C_T. Write $P_{j,i}(T,s) = \prod_{r=s+1}^{T}(1 - g_{j,i,r})$ for the conductance products, $A_{j,i}$ as in Section 3.4, and σ for the logistic function, the derivative of softplus. Differentiating the recurrence (7) gives, for every source step $s \leq T$ (Appendix L.6; for a single-channel state, Huang et al., 2025),

$$\frac{\partial \ell_T}{\partial u_s} = 2\big(\gamma_s^{\text{fr}} + \gamma_s^{\Delta r} + \gamma_s^{\Delta w} + \gamma_s^{B} + \gamma_s^{C}\big), \quad (18)$$

with, as row vectors over input channels,

$$\gamma_s^{\text{fr}} = a_T \odot K(T,s), \quad \gamma_s^{\Delta r} = \big(a_T \odot R_s \odot \sigma(q_s)\big)\, W_\Delta, \quad \gamma_s^{\Delta w} = \big(a_T \odot V_s \odot \sigma(q_s)\big)\, W_\Delta,$$
$$\gamma_s^{B} = Q_s\, W_B, \qquad \gamma_s^{C} = \mathbb{1}[s = T]\; H\, W_C,$$

where $R_{j,s} = \sum_i C_{i,T}\, P_{j,i}(T,s)\, A_{j,i}\, (1 - g_{j,i,s})\, x_{j,i,s-1}$, $V_{j,s} = u_{j,s} \sum_i C_{i,T}\, P_{j,i}(T,s)\, B_{i,s}$, $Q_{i,s} = C_{i,T} \sum_j a_{j,T}\, P_{j,i}(T,s)\, \Delta_{j,s}\, u_{j,s}$ and $H_i = \sum_j a_{j,T}\, x_{j,i,T}$. The five terms are, in order, the frozen write of Section 3.5, the derivative of the decay at the source step, the derivative of the write through Δ_s, the derivative through B_s, and the derivative through C_T. Each reaches $e_{T-\tau}$ through the same input path as in (16), so $\partial \ell_T / \partial e_{T-\tau}$ is a sum of five vectors.

Three properties follow. First, for every source step $s < T$, which covers every lag beyond the convolution window, the four terms other than γ^C contain the same conductance products $P(T,s)$. Every long-range pathway therefore propagates through the same conductance products, with pathway-specific coefficients, while γ^C acts only within the window. Second, the term through Δ in the write is the frozen term re-weighted at the source step: for $s < T$,

$$\gamma_s^{\Delta w} = \big(\gamma_s^{\text{fr}} \odot \lambda_s\big)\, W_\Delta, \qquad \lambda_{j,s} = u_{j,s}\, \frac{\sigma(q_{j,s})}{\Delta_{j,s}} = u_{j,s}\, \frac{\partial \log \Delta_{j,s}}{\partial q_{j,s}}. \quad (19)$$

The two share the conductance products and differ in weights, which does not imply equal ERF slopes: the re-weighting and the channel mixing W_Δ change which state dimensions contribute and how they interfere. Third, the term through the decay multiplies the stored state x_{s-1}: it describes retention at the source step only, while retention between s and T enters every term through $P(T,s)$.

The ERF is the expected squared norm of the sum, so it equals the five terms' energies (the expected squared norms of their contributions to $\partial \ell_T / \partial e_{T-\tau}$) plus twice each pairwise inner product. A term's energy is therefore not a share of the ERF, and we present energies and pairwise terms together. Each term, and each of the two Δ pathways separately, matches automatic differentiation in float64 (Appendix L.6).

3.7 Secondary Consequences

We obtain three additional results. (i) With i.i.d. conductances, each dimension's expected conductance product decays only exponentially, while a mixture of decay rates across the d_{state} dimensions can yield approximately power-law decay over finite lag ranges. (ii) When the residue-weighted spectral measure has an integrable density whose edge behavior near $g = 0$ is $p(g) \sim g^{\alpha}$, Watson's lemma gives power-law decay with exponent $\alpha + 1$. (iii) On sustained input regimes, the scan probe decays more slowly than the time-averaged static mixture, in some models far more than with the same conductances in random order, consistent with the Griffiths analogy to slow relaxation under persistent disorder (Appendices B, C). Full derivations: Appendices B, L.8, L.7.

4 Experiments

We first verify the closed forms against automatic differentiation (Section 4.2). We then test the block formula against the full model on language models (Section 4.3), use the decomposition on retrieval-task models (Section 4.4), measure what A-initialization changes (Section 4.5), and characterize conductance regimes in single-layer and pretrained models (Section 4.6).

4.1 Setup

Models. We train 18 single-layer Mamba language models on WikiText-103 [Merity et al., 2017] ($d_{\text{model}} = 256$, $d_{\text{inner}} = 512$, $d_{\text{state}} = 16$; 13.5M parameters) in two sets of nine. The main models cross three initializations of A (slow, $|A| \in [0.1, 2]$; default, $[1, 16]$; fast, $[5, 50]$) with three training seeds. Within a seed, the three initializations share every other initial weight and the data order. Their recipe is 12 epochs of AdamW [Loshchilov and Hutter, 2019] with weight decay 0.1 on every parameter, including $A_{\log}$, D and the bias of Δ's projection (full recipe in Appendix E). The control models repeat these nine runs from the same initial weights and data order, with $A_{\log}$ and the bias of Δ's projection exempt from weight decay (Section 4.5). We analyze the checkpoint at the end of epoch 12. Validation perplexity is 78.4–79.4 in the main models and 78.1–80.2 in the control models. We restrict the language models to one layer so that the ERF reflects propagation within a layer rather than composition across layers. For the retrieval tasks, we train models on multi-query associative recall (MQAR) [Arora et al., 2024] and on selective copying, with the same three initializations and three seeds each and paired training streams (Appendices F, G). We also analyze all 48 layers of a pretrained Mamba-370M [Gu and Dao, 2024].

Extraction and ERFs. We run a pure-PyTorch scan that implements the trained recurrence (7) and exposes Δ_t, B_t and C_t at every step. The fused training kernel does not export them. The full-model ERF E is Eq. (13) with $\ell_T = \|\text{logits}_T\|^2$, computed by backpropagation through the full model; the block formula X is Eq. (17); and the scan probe is expressed on the ERF scale, $S = 4\, h_{\text{pred}} = \text{ERF}_{\text{scan}}$ (Eq. (15)). We use all 120 validation and 138 test sequences of 2048 tokens and every target position $T = 1024$–2047, with lags $\tau = 1$–1023. Running products are formed in log space (Appendix L.9). E and X share the probe and the input coordinates, so their raw ratio is meaningful. Comparisons with S, which uses a different probe and input, are made after normalizing each curve at $\tau = 1$.

Statistics. Curves are means over sequences and targets. Across the three training seeds we summarize the normalized curves by their geometric mean with a pointwise 95% t-interval computed on the log scale, and initialization effects within seed with t-intervals across seeds. We show validation results. Test results agree.

4.2 Identity Checks

The closed forms are identities for frozen selectivity, so comparing them with automatic differentiation verifies their implementation. Sections 4.3 and 4.4 test the block formula against the full model. The block-formula and decomposition checks differentiate through the fused training kernel, so they also check the extraction. The scan's ERF with frozen selectivity equals $S = 4\, h_{\text{pred}}$ exactly (Appendix L.4). On all 18 language models (Table 2), in float32, S matches it to a maximum relative error of 6.2×10^{-5}, the block formula matches the model with Δ_t, B_t and C_t detached to 4.0×10^{-5},

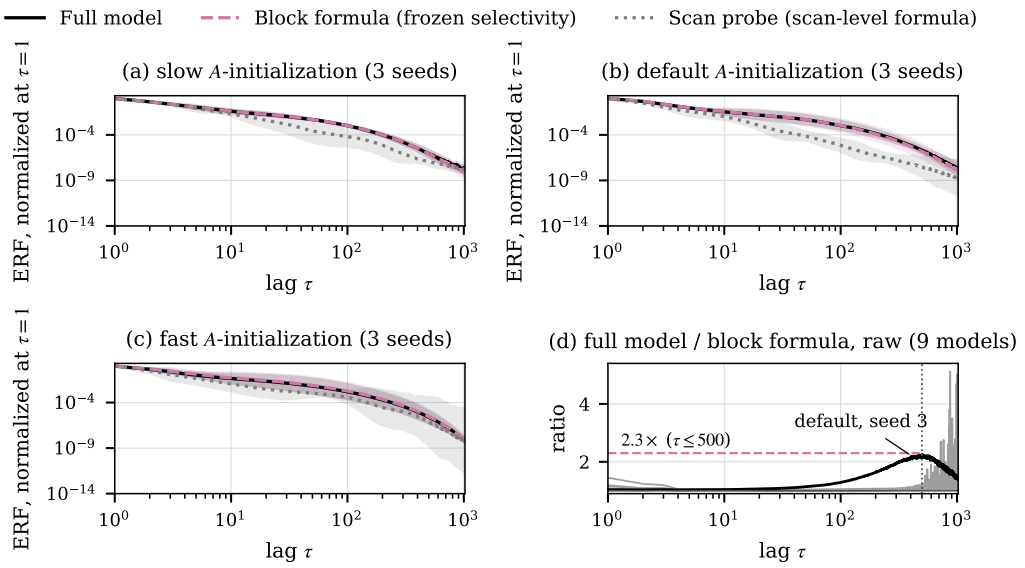

Figure 1: Full-model ERF (solid), block formula with frozen selectivity (Eq. (17), dashed) and scan probe (Eq. (15), dotted) for the nine main models. (a)–(c): geometric mean over three training seeds of the curves normalized at $\tau = 1$, with pointwise 95% t-intervals. (d): raw ratio of the full-model ERF to the block formula, per model; dashed line: $2.3\times$ over $\tau \leq 500$.

Table 1: Ratio of the full-model ERF to each closed form in the nine main models (validation). Block formula: raw, median [maximum]. Scan probe: normalized at $\tau = 1$, median [range].

τ	1	10	50	100	500	1000
block formula	1.07 [1.45]	1.02 [1.04]	1.01 [1.13]	1.01 [1.28]	1.07 [2.25]	1.48 [2.39]
scan probe	1	2.8 [1.3, 7.3]	22 [2.5, 226]	26 [0.99, 662]	7.8 [0.34, 147]	2.5 [0.04, 37]

and the five-term decomposition sums to the full-model gradient to 1.2×10^{-5} of its largest entry. S matches each scan's ERF with frozen selectivity to 2.2×10^{-7} of the curve's maximum in a two-layer model and to 2.1×10^{-5} in all 48 layers of Mamba-370M (pointwise, 1.6×10^{-4} and 1.9×10^{-4} after normalization at $\tau = 1$, for $\tau \leq 1023$; Appendices J, I).

4.3 Block Formula Against the Full Model

Across all 18 language models, the block formula approximates the full-model ERF within a factor of 2.3 at every lag from 1 to 500 (raw ratio 1.004–2.253), and its fitted log–log slope over $\tau = 10$–100 lies within 0.11 of the full model's (0.36 over $\tau = 100$–1000; Figure 1 for the main models, Appendix E for the control models). In 17 of the 18 models the ratio stays within 1.5. Table 1 gives the three lag ranges in the main models. Inside the convolution window ($\tau \leq 3$), the full model exceeds the formula by up to $1.45\times$. From $\tau = 4$ to 500 the ratio stays within $2.3\times$, with medians of 1.01–1.07 at the lags of Table 1. Beyond $\tau = 500$ the largest shortfall is approximately $5.13\times$.

What the block adds to the scan. After normalization at $\tau = 1$ and over $\tau \leq 500$, the scan probe alone has a worst-case miss of about $1238\times$ against the full-model ERF. With the readout's $a_{j,T}^2$ in place of the scan probe's $y_{j,T}^2$, the miss is about $11.6\times$. Adding the input path, without the cross-channel terms of W_{in}, reduces it to about $2.4\times$, and the full block formula to about $2.2\times$.

What the block formula leaves out. In eight of the nine main models, the decomposition of Section 3.6 puts the full-to-frozen energy ratio at 1.01–1.17 in each lag band (an interval of lags, as listed in Table 12) from $\tau = 5$ to 600, and the term through Δ in the write carries at most $0.18\times$ the frozen energy. In the ninth model (default initialization, seed 3), under the squared-logit probe, that

term grows with lag and reaches $1.28\times$ the frozen energy at $\tau = 400$–600, where the full-to-frozen ratio is 2.17. In the band $\tau = 800$–1023, the full-to-frozen ratio is 1.08–1.84 across the main models.

4.4 Decomposition on Retrieval Tasks

On the retrieval models, we probe the answer's logit margin (the correct answer's logit minus the mean logit of the other valid answers) and, instead of a curve over lags, measure its sensitivity at the input positions that hold the queried key and value (MQAR) or the copied token and its marker (selective copying), pooled over queries (Appendix H). In all 18 models, the full sensitivity at these positions is 4.96–22.47 times the block formula's, and the term through Δ in the write is the largest individual term (3.18–$26.18\times$ the frozen energy). The shortfall persists under the language models' squared-logit probe: on the same models and positions, the ratio is 2.24–25.81, and the same term is the largest in 17 of 18. The terms interfere: in two selective-copy models, that term's energy alone exceeds the full energy ($1.17\times$ and $1.07\times$). As derived in Section 3.6, the same conductance products carry the sensitivity in both settings. The frozen coefficients approximate it in the language models, but not at the relevant positions of the retrieval models.

4.5 What Initialization Changes

Task accuracy. Over three seeds, slow A-initialization gives the highest mean accuracy on both tasks: 1.00 on MQAR vs. 0.76 for default and 0.75 for fast, and 0.65 on selective copying vs. 0.45 and 0.36 (Appendices F, G). On selective copying, slow is the most accurate condition in every seed. From the nearest to the farthest distance bucket, the mean accuracy gap between slow and fast grows from 0.22 to 0.27 on MQAR and shrinks from 0.43 to 0.24 on selective copying. A single-seed experiment trained at length 4096, with unpaired training streams, gives selective-copy accuracies of 0.294, 0.255 and 0.170 for slow, default and fast (Appendix G). We do not establish that the ERF mediates this effect. That would need models whose trained tail slopes vary independently of their initialization, or an intervention that holds the slope fixed while the initialization varies.

Language-model ERFs. In the main models, no initialization effect is detected on the full-model ERF's fitted slope over $\tau = 10$–100: every paired difference has a 95% interval that includes zero (slow minus default -0.10 $[-0.44, 0.24]$; fast minus default 0.00 $[-0.33, 0.32]$). With three seeds, this is not evidence of equivalence. Under this recipe, which applies weight decay to every parameter (the reference implementation marks $A_{\log}$ and D as exempt), the median learned $|A|$ is 1.04–1.11 for all three initializations (initial medians 1.05, 8.5 and 27.5), and the bias of Δ's projection moves from about -4.6 to -0.03. The initialization remains visible at the scan: after normalizing each curve at $\tau = 1$, the scan probe of the fast-initialized models exceeds that of the default ones at $\tau = 50$–200 in every seed ($53\times$ at $\tau = 100$, 95% interval $[38, 74]$), but this contrast is not detected in the block formula or the full model.

Weight-decay control. In the control models, the learned $|A|$ stay ordered by initialization (Table 5). Slow minus fast in the full-model ERF's fitted slope over $\tau = 10$–100 was expected to be positive. It is -0.11 $[-0.17, -0.05]$, detected in the opposite direction. Over $\tau = 100$–1000, it is 2.33 $[1.27, 3.38]$. Relative to the main models, the change in this contrast is not detected over $\tau = 10$–100 (-0.01 $[-0.40, 0.38]$) and is 2.05 $[0.32, 3.78]$ over $\tau = 100$–1000. The test results agree (Appendix E).

4.6 Conductance Regimes

In the main models (Figure 2a, Table 3), 4.3–5.5% of conductances lie below 0.01 (coast) and 17–28% above 0.99 (strong shunt to ground), with medians of 0.77–0.82. The three initializations are close. In the control models they separate: the median conductance is 0.76–0.77 for slow, 0.97–0.98 for default and 1.00 for fast (Appendix E). Across the 48 layers of Mamba-370M (Figure 2b), the fraction below 0.01, the coast fraction, ranges from about 0.2% (layer 47) to 96% (layer 34), with no trend in depth (rank correlation 0.08 with the layer index). Layers with more near-zero conductances tend to have shallower scan-level ERF tails (rank correlation 0.61 with the slope over $\tau = 10$–100). The scan probe is built from the same conductances, so this association follows largely from Eq. (15).

5 Related Work

S4 [Gu et al., 2022a], HiPPO [Gu et al., 2020] and Mamba [Gu and Dao, 2024] established structured linear recurrences as competitive sequence models. Orvieto et al. [2023] showed that diagonal recurrences suffice. The input-output map of a selective scan is a data-controlled linear operator: Ali et al. [2025] write its kernel $C_t\,\bar{A}_t \cdots \bar{A}_{s+1}\,\bar{B}_s$ as an implicit attention matrix, and Zimerman et al. [2025] extend the operator to the whole block, convolution and gate included. Huang et al. [2025] factor the sensitivity of a single-channel hidden state to a past input into retention products and source-step derivatives of the write and the decay. We extend this factorization to the block ERF, a squared gradient norm that includes the input path and the probe's readout, and test its frozen-selectivity form against the full model. Concurrently, Bhattacharya [2026] decomposes forward channel outputs exactly by state mode and attributes most input-driven change in mode usage to B_t.

ERFs were studied for CNNs [Luo et al., 2016]. For Mamba, prior work measures receptive fields and enlarges them without training: DeciMamba [Ben-Kish et al., 2025] filters tokens, and LongMamba [Ye et al., 2025] classifies channels as local or global. We compute the ERF from extracted parameters instead. LongMamba's heuristic split of channels is close in spirit to our coast fraction, which we read directly from the conductances. Parameter-level levers on A and Δ are established: mimetic initialization [Trockman et al., 2024] improves recall, and Mamba Modulation [Lu et al., 2025] rescales A's spectrum per layer for length generalization. The HiPPO initialization places eigenvalues to approximate polynomial bases. In the spectral-measure view, tail behavior depends on the signed, residue-weighted measure near $g = 0$, and we do not test HiPPO initializations. Cirone et al. [2024] show that selective SSMs with multiplicative gating compute projections of the input's path signature, an algebraic account of what the hidden state can represent. The conductance products describe how far a trained model's influence extends in time.

Scope. The product structure of Section 3.6, in which every long-range pathway propagates through the same conductance products, holds for any diagonal linear recurrence with real, input-dependent decay $g_t = 1 - \bar{a}_t$ and either input discretization. This includes the real-valued variant of S4D [Gu et al., 2022b] as the time-invariant case and the scalar-per-head decay of Mamba-2 [Dao and Gu, 2024]. The five terms and the input path of Sections 3.5–3.6 are specific to the Mamba block and change with the architecture. The complex diagonal parameterizations of S4D and S5 [Smith et al., 2023] give complex conductances, for which the measure no longer lies on $[0, 1]$. The per-dimension representation requires a common, time-independent diagonal basis. It does not cover non-commuting transitions or attention layers in hybrid models. We state these extensions but do not test them.

6 Conclusion

Mamba's per-step conductances $G_t = I - \exp(A\,\Delta_t)$ organize how a block propagates information: every long-range pathway passes through the same conductance products, and an exact decomposition gives each pathway's coefficients, including those of differentiable selectivity. With selectivity held fixed, the decomposition reduces to a closed-form block formula that approximates the full-model ERF of 18 trained language models within a factor of 2.3 up to lag 500. Where the block formula falls short, the decomposition identifies the omitted pathway with the largest individual energy: in the retrieval-task models, at the positions an answer must retrieve, it is the pathway through Δ in the write. Over three seeds, slow A-initialization gives the highest mean accuracy on both retrieval tasks, and pretrained Mamba layers span diverse conductance regimes.

Limitations. The frozen-selectivity comparisons of Section 4.2 verify an identity. The empirical tests are the comparisons with the full model. The decomposition is exact for one block. We analyze stacks layer by layer and do not predict their end-to-end ERF. The ERF measures sensitivity to infinitesimal changes in continuous input representations. Finite token edits are outside its scope. Conductances are extracted from the same sequences on which the ERF is measured, and the approximation bound covers lags up to 500, the squared-logit probe and single-layer models trained with one recipe under two weight-decay settings (Section 4.3). Initialization effects rest on three training seeds per condition: an undetected effect may still be present, and we do not establish that the ERF mediates the effect on task accuracy. Both retrieval tasks use a repeated filler token as the distractor. Performance with varied distractors is untested. The resistor grid is a motivating analogy for a causal, one-dimensional system (Appendix A). Further limitations are in Appendix K.

Acknowledgments and Disclosure of Funding

This work received no third-party funding. The author declares no competing interests.

References

A. Ali, I. Zimerman, and L. Wolf. The hidden attention of Mamba models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

S. Arora, S. Eyuboglu, A. Timalsina, I. Johnson, M. Poli, J. Zou, A. Rudra, and C. Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024.

A. Ben-Kish, I. Zimerman, S. Abu-Hussein, N. Cohen, A. Globerson, L. Wolf, and R. Giryes. DeciMamba: Exploring the length extrapolation potential of Mamba. In *International Conference on Learning Representations (ICLR)*, 2025.

R. Bhattacharya. An exact instrument for state usage in selective state-space models, and the input-driven migration it reveals. *arXiv preprint arXiv:2607.11796*, 2026.

N. M. Cirone, A. Orvieto, B. Walker, C. Salvi, and T. Lyons. Theoretical foundations of deep selective state-space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

J. Cserti. Application of the lattice Green's function for calculating the resistance of an infinite network of resistors. *American Journal of Physics*, 68(10):896–906, 2000.

T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

P. G. Doyle and J. L. Snell. *Random Walks and Electric Networks*. Mathematical Association of America, 1984.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 2nd edition, 1971.

R. B. Griffiths. Nonanalytic behavior above the critical point in a random Ising ferromagnet. *Physical Review Letters*, 23(1):17–19, 1969.

A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022a.

A. Gu, A. Gupta, K. Goel, and C. Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024.

N. T. Huang, M. Sarabia, A. Moudgil, P. Rodriguez, L. Zappella, and F. Danieli. Understanding input selectivity in Mamba: Impact on approximation power, memorization, and associative recall capacity. In *International Conference on Machine Learning (ICML)*, 2025.

R. M. Jungers. *The Joint Spectral Radius: Theory and Applications*, volume 385 of *Lecture Notes in Control and Information Sciences*. Springer, 2009.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

P. Lu, J. Huang, Q. Zeng, X. Wang, B. Chen, P. Langlais, and Y. Cui. Mamba Modulation: On the length generalization of Mamba. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning (ICML)*, 2023.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

J. T. H. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

A. Trockman, H. Harutyunyan, J. Z. Kolter, S. Kumar, and S. Bhojanapalli. Mimetic initialization helps state space models learn to recall. *arXiv preprint arXiv:2410.11135*, 2024.

T. Vojta. Rare region effects at classical, quantum and nonequilibrium phase transitions. *Journal of Physics A: Mathematical and General*, 39(22):R143–R205, 2006.

Z. Ye, K. Xia, Y. Fu, X. Dong, J. Hong, X. Yuan, S. Diao, J. Kautz, P. Molchanov, and Y. C. Lin. LongMamba: Enhancing Mamba's long-context capabilities via training-free receptive field enlargement. In *International Conference on Learning Representations (ICLR)*, 2025.

I. Zimerman, A. Ali, and L. Wolf. Explaining modern gated-linear RNNs via a unified implicit attention formulation. In *International Conference on Learning Representations (ICLR)*, 2025.

A The Resistor-Network Analogy

This appendix develops the resistor-network picture behind our vocabulary. It is a motivating analogy, and the paragraph *Where the analogy stops* states its limits. In space, the lattice Laplacian arises as the Euler–Lagrange equation of a quadratic Dirichlet energy that penalizes voltage differences across edges. The same construction can be written in time, with time steps as nodes.

Setup. Consider an entire trajectory $x_{0:T}$ and introduce an energy functional analogous to the Dirichlet energy on a resistor grid:

$$\mathcal{E}(x_{0:T}) \;=\; \frac{1}{2}\sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^2_{W_t} \;+\; \frac{1}{2}\sum_{t=0}^{T} \|x_t - M_t u_t\|^2_{R_t}, \tag{20}$$

where $W_t \succeq 0$ are *temporal coupling weights* and $R_t \succeq 0$ are *shunt weights* that pull each state toward the encoded input $M_t u_t$.

Interpretation of the two terms. The first term, $\|x_{t+1} - x_t\|^2_{W_t}$, penalizes differences between consecutive hidden states. A large W_t makes it costly for x_{t+1} to deviate from x_t, just as a high-conductance edge enforces nearly equal voltages at its endpoints. This term enforces *smoothness* of the hidden trajectory across time.

The second term, $\|x_t - M_t u_t\|^2_{R_t}$, enforces fidelity of the hidden state to the external input. A large R_t means x_t is strongly pulled toward the input signal, while a small R_t means the state is free to follow its own dynamics. This acts like a shunt conductance connecting node x_t to a voltage source $M_t u_t$.

Analogy to resistor networks. Equation (20) has exactly the form of the electrical energy in a resistor grid:

$$\mathcal{E}_{\text{grid}}(v) \;=\; \frac{1}{2}\sum_{(i,j)\in E} w_{ij}(v_i - v_j)^2 \;+\; \frac{1}{2}\sum_{i\in V} g_i(v_i - s_i)^2.$$

Time steps $t = 0, \ldots, T$ play the role of spatial nodes, W_t are edge conductances linking x_t to x_{t+1}, and R_t are shunts connecting each node to the source $M_t u_t$. Minimizing $\mathcal{E}(x_{0:T})$ selects a trajectory that is both smooth across time and faithful to inputs, just as Kirchhoff's laws select node voltages that minimize dissipated power subject to sources.

Euler–Lagrange equations. Setting $\partial\mathcal{E}/\partial x_t = 0$ for $1 \le t < T$ yields the stationarity condition

$$(W_{t-1} + W_t)\, x_t - W_{t-1}\, x_{t-1} - W_t\, x_{t+1} \;=\; R_t M_t u_t - R_t x_t, \tag{21}$$

a discrete Laplace equation in time: the state at index t is a weighted average of its neighbors $t-1$ and $t+1$, with source and leak terms on the right.

Where the analogy stops. The stationarity condition (21) couples x_t to x_{t+1} as well as to x_{t-1}, as Kirchhoff's law couples a node to all of its neighbors, so the energy's minimizer in general depends on future inputs. The selective scan is causal: $x_t = (I - G_t)\, x_{t-1} + \Delta_t B_t u_t$ depends only on the past, and it is not obtained by minimizing (20). Its circuit counterpart is a directed ladder, in which each node is driven by its predecessor through a buffer, so that no current flows backward, and keeps the fraction $1 - g_{i,t}$ of its predecessor's voltage, the rest draining through a shunt to ground. The dimensionality differs as well: each state dimension is a one-dimensional chain, so the slow, approximately algebraic decay that a lattice shows when its spectral density reaches the origin cannot come from the geometry. In the time-invariant case it comes from the mixture of decay rates across state dimensions (Section 3.4). The circuit vocabulary is exposition. Every result in the paper is derived from the recurrence itself.

Distinguishing W_t and G_t. The coupling weights W_t are abstract variational parameters used to expose the temporal Laplacian structure. By contrast, Mamba's discretization of the decay shows that selectivity operates through $G_t := I - \bar{A}_t$, which plays the role of a shunt conductance to ground. Section 3.2 gives the conductances that the trained models compute.

B Secondary Consequences: Full Details

Exponential decay under scalar i.i.d. conductances. For a scalar recurrence with fixed residues and i.i.d. conductances, each moment of the impulse response decays exponentially with lag, although the response is a product of random retention factors.

Consider a scalar system ($N = 1$) with i.i.d. conductances $g_t \in (0, 1)$. The impulse response at lag τ is $h_\tau \propto \prod_{s=t-\tau+1}^{t}(1 - g_s)$. In the *annealed* regime (averaging over input sequences), independence factors the expectation:

$$\mathbb{E}[h_\tau] \;\propto\; \prod_{s=t-\tau+1}^{t} \mathbb{E}[1 - g_s] \;=\; (1 - \bar{g})^\tau, \tag{22}$$

a pure exponential with rate $-\log(1 - \bar{g})$, where $\bar{g} = \mathbb{E}[g]$. In the *quenched* regime (a single typical sequence), the law of large numbers gives

$$\frac{1}{\tau} \sum_{s=t-\tau+1}^{t} \log(1 - g_s) \;\xrightarrow{\text{a.s.}}\; \mathbb{E}[\log(1 - g)], \tag{23}$$

so the typical decay rate is $-\mathbb{E}[\log(1 - g)]$. By Jensen's inequality ($\log$ is concave), $\mathbb{E}[\log(1 - g)] \leq \log(1 - \bar{g})$, so the typical conductance product decays at least as fast as the expected one. Every moment factorizes the same way, $\mathbb{E}[h_\tau^k] \propto (\mathbb{E}[(1 - g)^k])^\tau$, so the squared response that enters the ERF also decays as a single exponential (Appendix L.8).

Heavy tails, where they occur, therefore require structure beyond scalar i.i.d. variation. Two candidates are:

1. *Spectral measure* (Section 3.4): a mixture across $N > 1$ parallel state dimensions with distributed mean conductances $\bar{g}_i$.
2. *Temporal correlations*: sustained regimes of similar tokens break the i.i.d. assumption and can produce Griffiths-like deviations from the static mixture (Appendix C).

Watson's lemma and tail classification. In the continuum idealization, with an integrable residue-weighted density p, the impulse response is the Laplace-type integral $h_\tau = \int_0^1 (1 - g)^\tau\, p(g)\, dg$. For large τ, the kernel $(1 - g)^\tau$ is sharply peaked near $g = 0$, so only the spectral edge governs the asymptotics [Feller, 1971]. Approximating $(1 - g)^\tau \approx e^{-g\tau}$ for small g and assuming $p(g) \sim c\, g^\alpha$ as $g \to 0^+$ with $\alpha > -1$, $c > 0$, we extend the upper limit to infinity (the tail is exponentially suppressed) and substitute $u = g\tau$:

$$h_\tau \;\approx\; c \int_0^\infty e^{-u} \left(\tfrac{u}{\tau}\right)^\alpha \tfrac{du}{\tau} \;=\; c\,\Gamma(\alpha + 1)\, \tau^{-(\alpha+1)}. \tag{24}$$

This classifies memory horizons by the spectral edge: a flat density at the origin ($\alpha = 0$) yields τ^{-1} tails; steeper vanishing (α large) yields faster algebraic decay; and a spectral gap ($p = 0$ below $a \in (0, 1)$) bounds $|h_\tau|$ by $(1 - a)^\tau \int_0^1 |p|\, dg$: exponential decay, possibly with an algebraic prefactor.

Griffiths-like effect. When input tokens are not i.i.d. but exhibit temporal structure (for instance, blocks of similar tokens that sustain similar Δ_t values), the conductances $g_{i,t}$ become correlated in time. In statistical physics, rare regions in systems with quenched disorder can produce Griffiths singularities and anomalously slow relaxation [Griffiths, 1969, Vojta, 2006]. Because the static mixture can deviate from the scan probe even without sustained regimes (Appendix C), the temporal analogue's prediction is narrower: sustained regimes should make the scan probe decay more slowly than the static mixture, by more than the same conductances give in random order.

Causality. Spatial resistor grids are reciprocal: current flows in both directions, and a node's voltage depends on sources on every side. Temporal recurrences are causal. The scan's impulse response is a *retarded* Green's function on a one-dimensional chain, $h(t, s) = 0$ for $t < s$. Appendix A states where this departs from the resistor-network picture.

LTV stability. Because selectivity modulates $\bar{A}_t$ at every step, the scan on a given input sequence is a linear time-varying (LTV) system, and for general LTV systems pointwise spectral stability ($\rho(\bar{A}_t) < 1$) does not by itself give uniform asymptotic stability [Jungers, 2009]. The diagonal scan is simpler. Each state dimension of a channel evolves independently as $x_{i,t} = e^{a_i \Delta_t}\, x_{i,t-1} + \Delta_t\, B_{i,t}\, u_t$ with $a_i < 0$, where $\Delta_t B_{i,t}$ is the training kernel's input weight (Section 3.2). Suppose $|B_{i,t}\, u_t| \le M$ and $\Delta_t \le \Delta_{\max}$, which holds for bounded inputs because Δ_t is the softplus of a bounded projection. With $S_t = \sum_{r \le t} \Delta_r$ and a zero initial state, $x_{i,t} = \sum_{s \le t} e^{-|a_i|(S_t - S_s)}\, \Delta_s\, B_{i,s}\, u_s$. Since $\Delta_s\, e^{-|a_i|(S_t - S_s)} \le e^{|a_i|\Delta_{\max}} \int_{S_{s-1}}^{S_s} e^{-|a_i|(S_t - \sigma)}\, d\sigma$,

$$|x_{i,t}| \;\le\; M\, e^{|a_i|\Delta_{\max}} \int_{-\infty}^{S_t} e^{-|a_i|(S_t - \sigma)}\, d\sigma \;=\; \frac{M\, e^{|a_i|\Delta_{\max}}}{|a_i|}.$$

The drain and the injection share the step size Δ_t: a step that injects strongly also drains strongly, so the state stays bounded however the steps vary. (With the zero-order-hold input weight $(1 - e^{a_i \Delta_t})\, |a_i|^{-1}\, B_{i,t}$ the sum telescopes and the bound is $M/|a_i|$.) The homogeneous system decays to zero from a fixed starting time if and only if $S_t \to \infty$, but this decay need not be uniform in the starting time: with $\Delta_t = 1/t$, S_t diverges, yet the decay over a fixed number of tokens vanishes as t grows. A lower bound $\Delta_t \ge \delta > 0$ gives uniform exponential decay, $|x_{i,t+\tau}| \le e^{-|a_i|\delta\tau}\, |x_{i,t}|$. No result in the paper relies on either.

C Griffiths Effect on Block-Structured Input

Static mixture. Averaging the quantities of the scan probe (15) over time gives an LTI approximation,

$$\widehat{\mathrm{ERF}}_{\mathrm{static}}(\tau) \;=\; 4 \sum_{j=1}^{d_{\mathrm{inner}}} \overline{y_j^2} \left[\sum_{i=1}^{N} \bar{c}_i\, \bar{b}_{j,i}\, (1 - \bar{g}_{j,i})^{\tau} \right]^2, \tag{25}$$

where $\overline{y_j^2}$ and $\bar{c}_i$ average $y_{j,T}^2$ and $C_{i,T}$ over the target positions, and $\bar{b}_{j,i}$ and $\bar{g}_{j,i}$ average $\bar{b}_{j,i,t}$ and $g_{j,i,t}$ over the positions $t = 1024$–2047. It tests whether the time-averaged spectral measure alone captures the shape of the scan's ERF.

Block-structured input. The Griffiths analogy (Appendix B) predicts that sustained regimes of the input make the position-specific scan probe decay more slowly than the static mixture, by more than the same conductances in random order. The static mixture can deviate even without such regimes, because it averages the conductances, the writes and the target-position weights (C_T and y_T^2) separately, whereas the probe averages y_T^2-weighted squared responses: for i.i.d. retention factors $r_t = 1 - g_t$, the squared product averages to $(\mathbb{E}[r^2])^{\tau}$, not $(\mathbb{E}[r])^{2\tau}$. We test the analogy with one sequence of 2048 tokens that alternates blocks of 256 copies of token 500 and 256 copies of token $30{,}000$. In every model, the mean Δ_t is larger on token $30{,}000$, by a factor of about 1.27 to 1.48. We compare the probe with the static mixture after normalization at $\tau = 1$, over $\tau = 1$–1021 and at every 16th target $T = 1024$–2047. A post-extraction control acts on the kernel with the target-position weights fixed: it permutes the extracted retention and write coefficients (Δ_t with B_t) jointly at random, within positions 3–1023 and within 1024–2047, and keeps the target-position weights C_T and y_T^2 of the original forward pass. This preserves the coefficients' empirical distribution and leaves

the static mixture unchanged. Two further arms apply the same permutations to one coefficient at a time: the conductances (Δ_t in the decay) with the writes in place, or the writes ($\bar{B}_t$) with the decay in place.

With blocks, the ratio of the probe to the static mixture peaks at about 1.3 to 28 across the nine main models. With the retention and write coefficients jointly permuted, the normalized discrepancy is within a factor of 2.2 in all nine models and all 33 tested permutations (mean ratio about 0.84–1.31 over $\tau = 50$–200). In four models, the blocked order gives peaks of about 6.8–28, with means of about 3.7–6.8 over $\tau = 50$–200, where the jointly permuted coefficients give peaks of about 1.1–2.2. Permuting the conductances alone gives peaks of about 1.1–1.6, below the blocked peak in every model, while permuting the writes alone gives about 2.1 to 170, at least the blocked peak in every model. A random permutation also moves each conductance away from the target-position weights of its own block. Exchanging adjacent blocks of the retention and write coefficients together separates the two: it keeps the coefficients persistent and coupled but moves each, apart from six positions near the start, into a block of the other token, misaligned with the target-position weights. In those four models it still gives peaks of about 2.8 to 12, above every jointly permuted peak, as the Griffiths analogy predicts. Aligning the coefficients with the target-position weights, as the blocked order does, enlarges the peaks in seven of the nine models, including those four. Shuffling the tokens themselves, which also changes the convolution's four-token windows, gives peaks of about 1.2–1.7, within the range of the joint permutations. We have not measured the deviation on natural text.

D Sanity Checks

Identity checks. Each closed form is an identity for the quantity it describes, so comparing it with automatic differentiation checks the extraction and gradient code (Section 4.2). Table 2 gives the checks for each of the 18 language models, in float32, on the first validation batch (two sequences) and over lags $\tau = 1$–1023.

Table 2: Identity checks: largest relative deviation. Scan probe: against the scan's ERF with frozen selectivity (four targets). Block formula: against the full model differentiated with Δ_t, B_t and C_t detached (16 targets). Decomposition: the summed five-term gradient against the full-model gradient, relative to the gradient's largest entry (two targets). Reference scan: the full-model ERF through our reference scan against the same ERF through the fused training kernel (four targets).

Model	Scan probe	Block formula	Decomposition	Reference scan
Main models				
Slow, seed 1	3.4×10^{-5}	2.9×10^{-5}	8.3×10^{-7}	4.4×10^{-5}
Default, seed 1	2.6×10^{-5}	3.3×10^{-5}	1.3×10^{-6}	4.6×10^{-5}
Fast, seed 1	2.4×10^{-5}	3.3×10^{-5}	1.7×10^{-6}	4.3×10^{-5}
Slow, seed 2	2.7×10^{-5}	3.3×10^{-5}	1.4×10^{-6}	4.6×10^{-5}
Default, seed 2	2.7×10^{-5}	3.1×10^{-5}	2.3×10^{-6}	4.7×10^{-5}
Fast, seed 2	6.2×10^{-5}	3.4×10^{-5}	2.4×10^{-6}	4.4×10^{-5}
Slow, seed 3	2.5×10^{-5}	3.1×10^{-5}	1.1×10^{-6}	4.4×10^{-5}
Default, seed 3	2.7×10^{-5}	3.1×10^{-5}	9.1×10^{-7}	4.7×10^{-5}
Fast, seed 3	2.8×10^{-5}	3.7×10^{-5}	1.4×10^{-6}	4.5×10^{-5}
Control models				
Slow, seed 1	2.7×10^{-5}	3.4×10^{-5}	1.2×10^{-5}	6.1×10^{-5}
Default, seed 1	2.6×10^{-5}	3.0×10^{-5}	2.0×10^{-6}	4.6×10^{-5}
Fast, seed 1	1.8×10^{-5}	3.2×10^{-5}	1.4×10^{-6}	4.5×10^{-5}
Slow, seed 2	5.1×10^{-5}	2.6×10^{-5}	1.9×10^{-6}	5.1×10^{-5}
Default, seed 2	2.2×10^{-5}	3.2×10^{-5}	1.4×10^{-6}	5.9×10^{-5}
Fast, seed 2	2.9×10^{-5}	3.3×10^{-5}	8.9×10^{-7}	4.6×10^{-5}
Slow, seed 3	4.5×10^{-5}	4.0×10^{-5}	2.7×10^{-6}	4.7×10^{-5}
Default, seed 3	4.2×10^{-5}	3.2×10^{-5}	1.7×10^{-6}	4.5×10^{-5}
Fast, seed 3	2.4×10^{-5}	3.3×10^{-5}	4.0×10^{-6}	5.0×10^{-5}

Constant-input collapse. On constant input (all 2048 tokens identical; token 500, then token 30,000), Δ_t, B_t and C_t are constant after the convolution's first three positions, so the position-

specific scan probe and the static mixture (25) must agree. In all nine main models they agree after normalization at $\tau = 1$ to a relative error below 10^{-12} at every lag up to 1021. At $\tau = 1022$ and 1023 the first target reaches source positions inside the convolution's start-up, and the two differ, by up to a factor of 5.4.

Gradient flow. For every sequence of every main model, the full-model ERF is positive at every lag $\tau = 1$–1023, so no tested lag has zero target-averaged sensitivity. The mean curve falls by about 7.4 to 8.3 orders of magnitude from $\tau = 1$ to $\tau = 1023$ (validation).

Sign of A. $A = -\exp(A_{\log})$ is negative by construction, so every retention factor $\exp(A\,\Delta_t)$ lies in $(0, 1)$. The learned $|A|$ of each model is in Tables 3 and 5.

Frozen selectivity against the full model. Section 4.3 compares the block formula with the full-model ERF at every lag, without normalization (Table 1), and Appendix H gives the selectivity pathways that make up the difference.

E A-Initialization Experiment: Full Details

Training recipe. The nine main models of Section 4.1 share one recipe. Each is a single Mamba block with a residual connection, between a token embedding and a tied output layer, without normalization layers: $d_{\text{model}} = 256$, $d_{\text{inner}} = 512$, $d_{\text{state}} = 16$, a depthwise causal convolution of width 4, and a full-rank projection for Δ (512×512, with bias). With the GPT-2 tokenizer (vocabulary 50,257), a model has 13.5M parameters, 12.9M of them in the embedding. We train on WikiText-103 (raw), cut into sequences of 2048 tokens, with AdamW (learning rate 10^{-3}, $\beta = (0.9, 0.95)$, weight decay 0.1 on every parameter), 500 warmup steps followed by cosine decay to zero, gradient clipping at norm 1, batch size 8 and 12 epochs (86,316 steps), in float32 with the fused training kernel. The initializations set $|A_{j,i}| = \text{linspace}(a, b, 16)_i$ in every channel j, with $(a, b) = (0.1, 2)$ for slow, $(1, 16)$ for default and $(5, 50)$ for fast. The softplus of the bias of Δ's projection is drawn log-uniformly from $[0.001, 0.1]$, and $D = 1$. The training seed (1, 2 or 3) sets every other initial weight and the data order, which the three initializations of a seed share. We analyze the checkpoint at the end of epoch 12. In all nine runs it also has the lowest validation loss. WikiText-103's test split repeats one training article, which fills most of its last two sequences. Excluding them leaves the block formula's $2.3\times$ and $1.45\times$ bounds unchanged and does not change which intervals in Table 6 exclude zero. It changes the other test-split values in this appendix by up to about 0.08 in a slope difference and 5% in a ratio, and one ratio between initializations then excludes 1 (given below). Each run takes about three hours on one NVIDIA RTX PRO 6000 GPU, with three runs at a time.

Learned parameters and conductances. Table 3 gives each model's validation perplexity, its learned $|A|$ and its conductance statistics.

Table 3: The nine main models: validation perplexity (PPL) at the end of epoch 12; median learned $|A|$ with its 5th and 95th percentiles; median of softplus of the bias of Δ's projection; and, over the first eight validation sequences, the percentage of conductances below 0.01 (coast), their median, and the percentage above 0.99 (strong shunt to ground).

Init	Seed	Val PPL	$\lvert A\rvert$	$\text{softplus}(b_\Delta)$	$g < 0.01$ (%)	median g	$g > 0.99$ (%)
Slow	1	79.36	1.06 [0.65, 1.57]	0.677	5.5	0.79	28.2
Default	1	78.79	1.11 [0.63, 1.88]	0.679	5.2	0.82	24.2
Fast	1	78.57	1.11 [0.65, 2.33]	0.679	4.6	0.77	17.0
Slow	2	78.45	1.04 [0.65, 1.57]	0.677	5.3	0.81	27.6
Default	2	79.19	1.10 [0.66, 1.84]	0.677	4.8	0.80	27.6
Fast	2	79.21	1.09 [0.66, 2.01]	0.678	4.3	0.82	21.2
Slow	3	78.81	1.06 [0.67, 1.59]	0.678	5.0	0.81	27.9
Default	3	79.11	1.08 [0.66, 1.81]	0.677	4.5	0.79	25.1
Fast	3	78.74	1.11 [0.68, 2.01]	0.678	4.3	0.80	22.1

Weight decay. Under this schedule, weight decay alone would multiply a parameter that receives no gradient by $\prod_t (1 - 0.1\,\eta_t) \approx 0.0134$, where η_t is the learning rate at step t. It pulls $A_{\log}$ toward

0, that is $|A|$ toward 1, and the bias of Δ's projection toward 0, that is its softplus toward 0.69. Decay alone would take the median $|A|$ from 1.05, 8.5 and 27.5 to 1.00, 1.03 and 1.05. The learned medians are 1.04–1.11 for all three initializations, and within each model the rank correlation between initial and learned $|A|$ is below 0.24. The median softplus of Δ's bias moves from about 0.01 to 0.68, against 0.66 under decay alone, and the median of D moves from 1 to between about -0.10 and -0.04. The reference implementation of Mamba marks $A_{\log}$ and D as exempt from weight decay, and that of Mamba-2 also the bias of Δ.

ERF slopes. Table 4 gives the least-squares log–log slope over $\tau = 10$–100 of the full-model ERF, the block formula and the scan probe for each main model, with sequence-bootstrap 95% intervals conditional on the checkpoint, and for each initialization the mean over the three seeds with a 95% t-interval. Section 4.5 gives the paired within-seed contrasts.

Table 4: Main models: least-squares log–log slopes over $\tau = 10$–100 (validation). Per model: sequence-bootstrap 95% interval, conditional on the checkpoint. Mean: over the three seeds, with a 95% t-interval.

Init	Seed	Full-model ERF	Block formula	Scan probe
Slow	1	-1.58 $[-1.59, -1.56]$	-1.58 $[-1.59, -1.56]$	-2.70 $[-2.75, -2.65]$
Default	1	-1.46 $[-1.48, -1.45]$	-1.48 $[-1.50, -1.46]$	-2.43 $[-2.47, -2.40]$
Fast	1	-1.62 $[-1.63, -1.60]$	-1.62 $[-1.63, -1.60]$	-0.82 $[-0.86, -0.79]$
Slow	2	-1.45 $[-1.48, -1.43]$	-1.45 $[-1.48, -1.43]$	-2.13 $[-2.15, -2.11]$
Default	2	-1.49 $[-1.51, -1.47]$	-1.49 $[-1.51, -1.47]$	-3.30 $[-3.35, -3.24]$
Fast	2	-1.40 $[-1.42, -1.38]$	-1.40 $[-1.42, -1.38]$	-2.22 $[-2.28, -2.17]$
Slow	3	-1.83 $[-1.85, -1.81]$	-1.83 $[-1.85, -1.81]$	-3.09 $[-3.16, -3.03]$
Default	3	-1.59 $[-1.60, -1.58]$	-1.69 $[-1.70, -1.68]$	-3.54 $[-3.58, -3.50]$
Fast	3	-1.54 $[-1.56, -1.53]$	-1.54 $[-1.55, -1.52]$	-0.72 $[-0.75, -0.69]$
Slow	mean	-1.62 $[-2.09, -1.15]$	-1.62 $[-2.09, -1.14]$	-2.64 $[-3.84, -1.44]$
Default	mean	-1.52 $[-1.68, -1.35]$	-1.55 $[-1.85, -1.25]$	-3.09 $[-4.54, -1.64]$
Fast	mean	-1.52 $[-1.79, -1.24]$	-1.52 $[-1.79, -1.25]$	-1.25 $[-3.34, 0.83]$

Weight-decay control. The control models repeat the nine runs above from the same initial weights and data order, with $A_{\log}$ and the bias of Δ's projection exempt from weight decay, while D, the convolution bias and all weights are still decayed. The checkpoint at the end of epoch 12 again has the lowest validation loss in all nine runs, and each run again takes about three hours. Table 5 gives each model's validation perplexity, learned parameters and conductances. Without decay, the learned values stay correlated with the initial ones: within each model, the rank correlation is 0.42–0.71 for $A_{\log}$ (below 0.24 in the main models) and above 0.99 for the bias. The input-dependent part of Δ_t sets its scale: the median Δ_t is 0.99–1.08 for slow and 0.31–0.41 for default and fast, against a median softplus of the bias of 0.0125–0.0173. The median $|A|\Delta_t$ is 1.42–1.48 for slow (all nine main models: 1.46–1.71), 3.39–4.11 for default and 8.78–9.93 for fast. Validation perplexity is lower for slow than for default, and for default than for fast, in every seed. These perplexity comparisons are descriptive.

Table 6 gives the paired slope contrasts of the full-model ERF in both windows, on both splits. Over $\tau = 10$–100, slow falls slightly faster than fast in every seed. Over $\tau = 100$–1000, the slopes order slow, default, fast in every seed on both splits (means on validation -3.29, -4.69 and -5.62; main models -5.30, -5.12 and -5.58), and at $\tau = 1000$ the normalized full-model ERF of the slow models is 257 $[117, 567]$ times that of the fast ones (test 186 $[90, 385]$). In the main models, of the 18 such ratios between initializations at $\tau = 10$, 50, 100, 200, 500 and 1000, one excludes 1 on validation (slow over default at $\tau = 500$, 0.42 $[0.22, 0.81]$) and none on test, where the same ratio is 0.428 $[0.174, 1.051]$. Without the two test sequences that copy a training article, it is 0.424 $[0.181, 0.997]$ and excludes 1 on test as well. Of the three initializations, only slow's slope over $\tau = 100$–1000 changes detectably relative to the main models (2.01 $[0.67, 3.35]$; default 0.43 $[-0.47, 1.33]$; fast -0.04 $[-0.78, 0.70]$). With the block formula, intervals exclude zero for the same contrasts (slow minus fast -0.11 $[-0.17, -0.05]$ and 2.34 $[1.28, 3.40]$; control minus main for slow minus fast over $\tau = 100$–1000, 2.24 $[0.29, 4.19]$). Because the control exempts both parameters at once, it does not attribute these changes to either, and like the main models it rests on three seeds.

Table 5: The nine control models, with the columns of Table 3: validation perplexity at the end of epoch 12; median learned $|A|$ with its 5th and 95th percentiles; median of $\mathrm{softplus}$ of the bias of Δ's projection; and, over the first eight validation sequences, the percentage of conductances below 0.01, their median, and the percentage above 0.99.

Init	Seed	Val PPL	$\lvert A\rvert$	$\mathrm{softplus}(b_\Delta)$	$g < 0.01$ (%)	median g	$g > 0.99$ (%)
Slow	1	78.26	1.50 [0.08, 4.95]	0.0150	5.1	0.77	21.5
Default	1	78.52	8.57 [0.58, 49.24]	0.0125	3.4	0.97	44.2
Fast	1	79.00	30.12 [2.14, 208.86]	0.0133	2.5	1.00	60.7
Slow	2	78.09	1.46 [0.13, 4.42]	0.0173	5.5	0.77	21.2
Default	2	78.81	8.89 [0.55, 47.94]	0.0156	3.1	0.98	47.8
Fast	2	80.16	27.51 [2.15, 183.89]	0.0156	2.9	1.00	60.3
Slow	3	78.19	1.37 [0.10, 4.56]	0.0158	5.1	0.76	20.2
Default	3	79.11	9.00 [0.58, 50.72]	0.0140	3.2	0.97	43.4
Fast	3	79.39	29.37 [2.27, 187.17]	0.0145	2.2	1.00	59.6

Table 6: Paired within-seed differences in the fitted log–log slope of the full-model ERF: mean over three seeds with a 95% t-interval. Control − main: the control models' slow-minus-fast difference minus the main models', paired by seed.

Contrast	Split	$\tau = 10$–100	$\tau = 100$–1000
Main, slow − fast	validation	-0.10 $[-0.52, 0.32]$	0.28 $[-0.46, 1.02]$
	test	-0.09 $[-0.47, 0.29]$	0.24 $[-0.48, 0.95]$
Control, slow − fast	validation	-0.11 $[-0.17, -0.05]$	2.33 $[1.27, 3.38]$
	test	-0.11 $[-0.17, -0.05]$	2.14 $[1.38, 2.90]$
Control, slow − default	validation	-0.01 $[-0.48, 0.46]$	1.40 $[0.26, 2.54]$
	test	-0.01 $[-0.46, 0.44]$	1.35 $[0.23, 2.46]$
Control, fast − default	validation	0.10 $[-0.38, 0.58]$	-0.92 $[-1.49, -0.36]$
	test	0.10 $[-0.34, 0.54]$	-0.80 $[-1.28, -0.31]$
Control − main, slow − fast	validation	-0.01 $[-0.40, 0.38]$	2.05 $[0.32, 3.78]$
	test	-0.02 $[-0.40, 0.36]$	1.90 $[0.57, 3.24]$

The scan probe does not detect the slow-minus-fast contrast in the control models in either window (-0.81 $[-1.94, 0.32]$ over $\tau = 10$–100; 1.48 $[-1.32, 4.28]$ over $\tau = 100$–1000), and its contrast over $\tau = 100$–1000 is not shown to differ from the full-model ERF's (paired difference 0.85 $[-1.32, 3.01]$). The main models' scan-level contrast, fast over default at $\tau = 100$ after normalizing each curve at $\tau = 1$ (Section 4.5), is 1.25 $[0.20, 7.97]$ in the control models. Paired within seed, it is 0.024 $[0.003, 0.18]$ times the main models' (test 0.028 $[0.004, 0.19]$).

In the control models, the block formula is within $1.45\times$ of the full-model ERF at every lag from 1 to 500 (raw ratio 1.004–1.448 on both splits), and its slope over $\tau = 10$–100 is within 0.005 of the full model's. Beyond $\tau = 500$, the largest factor is about 2.25 (default, seed 3), and in some models the block formula exceeds the full-model ERF (raw ratio down to about 0.974 on validation and 0.907 on test). For the four successive formulas of Section 4.3, the worst normalized misses over $\tau \leq 500$ are about $84\times$, $5.1\times$, $1.7\times$ and $1.4\times$. On validation, in every lag band from $\tau = 5$ to 600, the full-to-frozen energy ratio is 1.004–1.157, and the term through Δ in the write, the largest omitted term in every band and every model, carries at most $0.155\times$ the frozen energy ($0.165\times$ on test). In the band $\tau = 800$–1023, the full-to-frozen ratio is 1.006–1.734.

F Multi-Query Associative Recall

We train the three A-initializations of Appendix E on multi-query associative recall (MQAR), adapted from Arora et al. [2024]. MQAR has become a standard benchmark for the recall capacity of efficient sequence models, and selective SSMs have previously been shown to perform poorly on it.

Task setup. Each training sequence has length $L = 1536$ over a small synthetic vocabulary of 35 tokens (16 keys, 16 values, plus padding (PAD), query-marker and end-of-sequence (EOS) tokens). For each sequence, $n_{\mathrm{pairs}} = 8$ distinct (key, value) pairs are placed at random positions in evenly sized slots covering the first 75% of the sequence. The value token immediately follows its key. Key

positions range from 0 to ≈ 1150. A query block of 4 tokens is placed near the end of the sequence (positions ≈ 1199–1205), with each query a key sampled from those previously inserted. The model must predict the corresponding value at each query position, with the value as the next-token target at that position. The loss is computed only at query positions. All other positions are masked. Realized key–query distances span 57–1205 tokens across batches (minimum when a key occurs near the end of the key region and the first query reads it; maximum when a key occurs at position 0 and the last query reads it).

Model and training. We use the single-layer architecture of Appendix E ($d_{\text{model}} = 256$, $d_{\text{inner}} = 512$, $d_{\text{state}} = 16$) with the three A-initializations (slow $|A| \in [0.1, 2]$, default $[1, 16]$, fast $[5, 50]$). Training: AdamW (learning rate 3×10^{-4}, $\beta = (0.9, 0.95)$, weight decay 0.01 on every parameter), batch size 16, 10,000 steps with 500 warmup steps and cosine decay, and gradient clipping at norm 1. Each of three seeds (42, 43, 44) sets the initial weights and the training stream, which the three initializations of a seed share. Every model is evaluated on the same 3200 queries (50 batches).

Results. Slow reaches 99.97–100% accuracy (rounded) in all three seeds: 3199, 3200 and 3200 of the 3200 queries, with the one error at a distance above 1024. Default reaches 0.51, 0.78 and 1.00, and fast 0.65, 0.64 and 0.96 (Table 8). Slow has the highest mean accuracy at every distance (Table 7). From the shortest to the longest distance bucket, the mean accuracy gap between slow and fast grows from 0.22 to 0.27. The shortest bucket holds only 9 queries. Measured from the next bucket ($n = 541$), the gap also grows from 0.22 to 0.27. Paired within seed, with 95% t-intervals across the three seeds (as for the initialization contrasts of Section 4.5), slow minus default is 0.24 $[-0.37, 0.85]$ and slow minus fast 0.25 $[-0.21, 0.70]$. Neither excludes zero.

Table 7: MQAR accuracy by key–query distance: mean $\pm$ standard deviation across the three seeds, on the same 3200 queries per model. Random baseline $1/16 \approx 0.063$.

Distance (tokens)	Slow	Default	Fast
16–64 ($n = 9$)	1.000 ± 0.000	0.778 ± 0.222	0.778 ± 0.192
64–256 ($n = 541$)	1.000 ± 0.000	0.787 ± 0.239	0.780 ± 0.170
256–1024 ($n = 2147$)	1.000 ± 0.000	0.766 ± 0.238	0.751 ± 0.181
1024–1205 ($n = 503$)	0.999 ± 0.001	0.723 ± 0.285	0.728 ± 0.209
Overall	1.000 ± 0.000	0.763 ± 0.245	0.753 ± 0.184

Table 8: MQAR overall accuracy per seed.

Seed	Slow	Default	Fast
42	0.9997	0.511	0.649
43	1.000	0.778	0.644
44	1.000	1.000	0.965

Scope of the MQAR results. We evaluate this synthetic task at one model size, with three training seeds per initialization. Accuracy varies substantially across seeds for the default and fast initializations. On natural text, the three initializations reach similar perplexity under the main recipe, where median learned $|A|$ is similar across initializations (Appendix E), and in the control models, where A stays ordered by initialization, perplexity is lowest for slow in every seed.

G Selective Copying

In selective copying, a sequence of filler tokens holds eight content tokens at random positions, each preceded by a MARKER token. After an EOS the model must produce the marked tokens in order. Selective copying is a standard memory benchmark for SSMs that a single-layer Mamba can solve. At sequence length 1536, mark–output distances span 231–1526 tokens. The vocabulary has 19 tokens (16 content tokens and PAD, MARKER and EOS), so the random baseline is $1/16 \approx 0.063$. Accuracy is scored per output token with the true earlier outputs as input. Model, training, seeds

and paired training streams are as in Appendix F. Every model is evaluated on the same 6400 output positions.

Results. Slow reaches a mean accuracy of 0.654 against 0.453 for default and 0.361 for fast (Table 9). It has the highest mean accuracy in every distance bucket, and it outperforms both default and fast in every seed and every bucket (per-seed totals in Table 10). Unlike on MQAR, the mean accuracy gap between slow and fast shrinks with distance, from 0.43 in the shortest bucket to 0.24 in the longest. Paired within seed, slow minus default is 0.20 $[0.14, 0.27]$ and slow minus fast 0.29 $[-0.19, 0.77]$. Only the first excludes zero.

Table 9: Selective copying accuracy by mark–output distance: mean $\pm$ standard deviation across the three seeds, on the same 6400 output positions per model. Random baseline $1/16 \approx 0.063$.

Distance (tokens)	Slow	Default	Fast
231–256 ($n = 108$)	0.849 ± 0.079	0.636 ± 0.047	0.417 ± 0.234
256–1024 ($n = 3815$)	0.720 ± 0.085	0.504 ± 0.062	0.394 ± 0.175
1024–1526 ($n = 2477$)	0.544 ± 0.064	0.366 ± 0.044	0.307 ± 0.102
Overall	0.654 ± 0.076	0.453 ± 0.055	0.361 ± 0.147

Table 10: Selective copying overall accuracy per seed.

Seed	Slow	Default	Fast
42	0.741	0.515	0.269
43	0.620	0.416	0.531
44	0.600	0.426	0.282

Sequence length 4096 (single run). In one run at sequence length 4096, with a single seed and training streams that were not paired across initializations, mark–output distances span 615–4086 tokens. Slow is again the most accurate condition overall (0.294, against 0.255 for default and 0.170 for fast) and in both distance buckets (Table 11). At distances of 1024–4086 tokens, slow's accuracy (0.282) is about 4.5 times the random baseline. Because accuracy is scored with the true earlier outputs as input, this shows retention above chance, not ordered copying. A multi-seed run at this length is left to future work.

Table 11: Selective copying at sequence length 4096 (*single run*, unpaired training streams).

Distance (tokens)	Slow	Default	Fast
615–1024 ($n = 766$)	0.385	0.368	0.204
1024–4086 ($n = 5634$)	0.282	0.240	0.165
Overall ($n = 6400$)	0.294	0.255	0.170

Both tasks. Slow has the highest mean accuracy in every nonempty distance bucket on both tasks. On selective copying, it also outperforms both alternatives in every seed and every nonempty bucket. From the shortest to the longest bucket, the mean slow–fast accuracy gap increases on MQAR and decreases on selective copying.

H Selectivity Pathways: Decomposition Results

This appendix presents the decomposition of Section 3.6 for the language models and the retrieval-task models. Energies are given relative to the frozen term's energy. They are not shares of the ERF: the ERF also contains twice the inner product of every pair of terms (the pairwise terms), and these can be negative.

Language models. Table 12 gives the energies by lag band as ranges over the nine main models (all validation sequences, 64 targets per sequence). Appendix E summarizes those of the control models. The term through C_T vanishes beyond the convolution window, as it must (Section 3.6). In every band and every model, the term through Δ in the write is the largest of the four terms that the frozen formula omits. The net of all pairwise terms is small and mostly negative. Section 4.3 summarizes the table: the upper ends of the Δw, B and Full ranges at $\tau = 40$–600 come from the default model of seed 3. The bound of Section 4.3 concerns the ERF averaged over sequences. A single sequence's own ratio exceeds 2.3 at some lag up to 500 in 0 to 83 of a main model's 120 validation sequences (83 in the default model of seed 3) and in one sequence of the control models. The largest single-sequence ratio is about 18. The individual pairwise terms are in the released result files.

Table 12: Language models: energy of each omitted term relative to the frozen term, the net of all pairwise terms, and the full-model energy relative to the frozen term (Full: 1 plus the four energies plus the pairwise term), by lag band, with each energy summed over the band; ranges over the nine main models. Δr: the decay at the source step; Δw: the write through Δ_s; B, C: through B_s and C_T.

τ	Δr	Δw	B	C	Pairwise	Full
1	0.00731–0.0410	0.0238–0.175	0.00365–0.0677	0.00270–0.0214	-0.085 to 0.17	1.03–1.46
5–15	0.00422–0.0154	0.0154–0.0376	0.000731–0.00213	0	-0.029 to -0.0016	1.01–1.03
40–60	0.000937–0.00924	0.00721–0.134	0.000109–0.00274	0	-0.022 to -0.0010	1.01–1.13
80–120	0.000843–0.00923	0.00960–0.286	0.000116–0.00523	0	-0.020 to -0.00013	1.01–1.27
180–220	0.000801–0.0128	0.0146–0.654	0.000174–0.0103	0	-0.050 to -0.0011	1.01–1.62
400–600	0.000803–0.0173	0.0323–1.28	0.000429–0.0175	0	-0.13 to -0.0047	1.03–2.17
800–1023	0.00105–0.122	0.0965–0.921	0.00178–0.0165	0	-0.21 to -0.015	1.08–1.84

One parameter at a time. We also detach Δ_t, B_t and C_t one at a time (256 targets per sequence). At $\tau = 100$, detaching Δ_t alone lowers the full-model ERF by 0.9–27% of the ERF with all three detached, detaching B_t alone changes it by -0.02% to $+3.2\%$, and detaching C_t leaves it unchanged.

Table 13: Retrieval tasks, relevant input positions, pooled over queries. Margin probe: full-model energy, the energies of the terms through Δ in the write (Δw), the decay at the source step (Δr) and B, and the net of all pairwise terms, each relative to the frozen term. Squared-logit probe: full over frozen, and the largest of the five terms.

			Margin probe					Squared logit	
Task	Init	Seed	Full	Δw	Δr	B	Pairwise	Full	Largest
MQAR	Slow	42	13.87	8.75	0.002	2.07	2.04	8.13	Δw
MQAR	Default	42	13.35	9.64	0.003	0.19	2.53	7.67	Δw
MQAR	Fast	42	18.16	13.28	0.004	0.36	3.52	15.76	Δw
MQAR	Slow	43	14.98	9.73	0.003	2.29	1.95	25.81	Δw
MQAR	Default	43	13.14	9.85	0.002	0.88	1.41	8.31	Δw
MQAR	Fast	43	14.24	9.62	0.001	0.45	3.16	8.73	Δw
MQAR	Slow	44	14.97	11.26	0.002	0.67	2.04	7.90	Δw
MQAR	Default	44	7.89	4.48	0.001	0.89	1.53	3.15	Δw
MQAR	Fast	44	17.47	12.95	0.001	0.35	3.16	13.15	Δw
Sel. copy	Slow	42	6.25	3.80	0.040	0.11	1.30	2.24	frozen
Sel. copy	Default	42	6.44	5.97	0.392	0.05	-0.97	4.74	Δw
Sel. copy	Fast	42	22.47	26.18	2.340	0.23	-7.29	20.70	Δw
Sel. copy	Slow	43	5.65	4.14	0.058	0.08	0.38	2.73	Δw
Sel. copy	Default	43	6.93	5.88	0.329	0.11	-0.39	5.93	Δw
Sel. copy	Fast	43	12.29	11.52	0.605	0.16	-1.00	9.34	Δw
Sel. copy	Slow	44	4.96	3.18	0.023	0.04	0.71	2.85	Δw
Sel. copy	Default	44	4.99	4.24	0.298	0.14	-0.68	4.68	Δw
Sel. copy	Fast	44	21.08	22.58	0.940	0.30	-3.74	15.08	Δw

Retrieval tasks. Table 13 gives, for each of the 18 task models, the pooled energy at the relevant input positions (Section 4.4; energies summed over queries before dividing by the summed frozen

energy) under the margin probe, and the full-to-frozen ratio under the language models' squared-logit probe. The term through C_T is zero at these positions, which lie outside the convolution window. The net pairwise term is positive on MQAR (1.4 to 3.5 times the frozen energy) and ranges from -7.3 to 1.3 on selective copying. In the fast-initialized selective-copy models of seeds 42 and 44, the energy of the term through Δ in the write exceeds the full energy, so the terms interfere destructively. The ranking is specific to these positions: at the other key–value (MQAR) or marker–token (selective copying) pairs, the term through Δ in the write is the largest in 12 of the 18 models, and at the remaining past positions in 9.

I Pretrained Mamba-370M: Scan-Probe Analysis

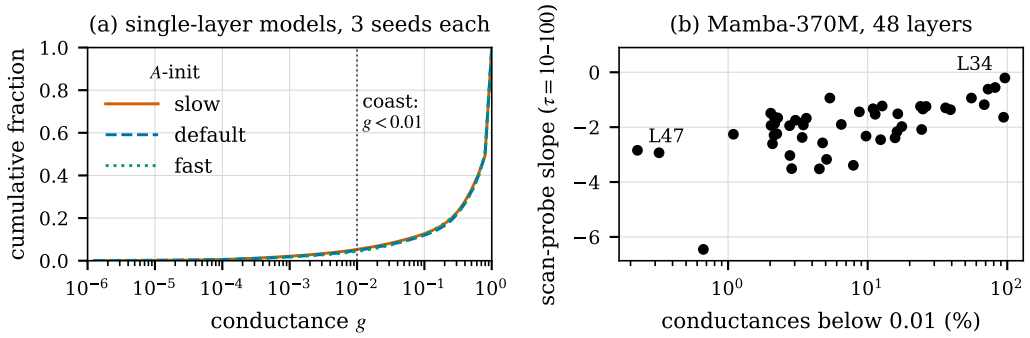

Figure 2: (a) Cumulative distribution of the conductances $g_{j,i,t}$ in the main language models, by A-initialization (three seeds each; first eight validation sequences). The three curves nearly coincide, and Table 3 gives each model's fractions below 0.01 and above 0.99. (b) Mamba-370M: for each layer, the percentage of conductances below 0.01 and the least-squares slope of the scan probe over $\tau = 10$–100 (bars: 95% bootstrap intervals over sequences).

We extract the conductances of every layer of a publicly available pretrained Mamba-370M [Gu and Dao, 2024] (48 layers, $d_{\text{inner}} = 2048$, $d_{\text{state}} = 16$) and evaluate the scan probe on each layer. No training is performed. The model's own forward pass on WikiText-103 validation text (GPT-NeoX tokenizer, all 122 full chunks of 2048 tokens) supplies each layer's input, and each layer is analyzed on that input with the trained recurrence, at every 16th target $T = 1024$–2047 and lags $\tau = 1$–1023. These are scan-level analyses: the block formula and the full-model ERF are not computed for this model.

Consistency check. On every layer, the scan probe matches the layer's scan ERF with frozen selectivity, computed by automatic differentiation on the first sequence at $T = 1023$ and $T = 2047$, to at most 2.1×10^{-5} of the curve's maximum, which lies at $\tau = 0$, where the scan probe uses $K_j(T,T) = h_j(T,T) + D_j$ (median over layers 9.9×10^{-8}), and pointwise to 1.9×10^{-4} after both curves are normalized at $\tau = 1$ (median 3.1×10^{-5}), as the identity of Appendix L.4 requires.

Conductance regimes. The fraction of conductances below 0.01 ranges from about 0.2% (L47) to 96% (L34), with no trend in depth (rank correlation 0.08 with the layer index). Four layers (L31, L32, L34, L39) have more than 70% of their conductances below 0.01, and 22 layers have fewer than 5%. Layers with more near-zero conductances tend to have shallower scan-probe slopes over $\tau = 10$–100 (rank correlation 0.61; both are computed from the same conductances, so the association largely follows from the formula), but the two groups overlap: the four layers above 70% have slopes of about -1.64 to -0.21, and the 22 layers below 5% about -6.45 to -1.49 (mean -2.47), 10 of them shallower than -2. Table 14 gives representative layers.

Sign ablation. With the signs of the residues $C_{i,T}\bar{b}_{j,i,s}$ removed (Appendix L.4) and both curves normalized at $\tau = 1$, the ratio of the unsigned to the signed scan probe is below 1 in 23 of the 48 layers at $\tau = 10$, and in 32 at $\tau = 100$, 500 and 1000. Across layers and these lags it ranges from about 0.28 to 3.2. A normalized ratio below 1 means that the signs cancel relatively more at $\tau = 1$ than at that lag. The effect of the signs on the shape thus depends on the layer and the lag.

Table 14: Representative Mamba-370M layers: least-squares slope of the scan probe over $\tau = 10$–100 with a 95% bootstrap interval over sequences, and the percentage of conductances below 0.01. The ERF-scaled scan probe $S = 4h_{\text{pred}}$ equals the layer's scan ERF with frozen selectivity, so one slope serves for both.

Layer	Scan-probe slope ($\tau = 10$–100)	$g < 0.01$
0	-3.51 [-3.52, -3.50]	2.9%
1	-1.23 [-1.38, -1.14]	12.7%
17	-2.16 [-2.38, -1.97]	16.2%
24	-2.46 [-2.55, -2.39]	12.4%
25	-1.18 [-1.24, -1.12]	68.4%
34	-0.21 [-0.29, -0.08]	96.2%
35	-1.34 [-1.42, -1.28]	24.9%
47	-2.84 [-2.90, -2.79]	0.2%

J Two-Layer Model: Scan-Probe Analysis

To check the scan probe in a model whose layers interact, we analyze a two-layer Mamba model ($d_{\text{model}} = 256$, $d_{\text{inner}} = 512$, $d_{\text{state}} = 16$, 14.2M parameters) trained on WikiText-103 for 12 epochs (validation perplexity 50.6). Layer 1 receives the token embeddings and layer 2 the output of layer 1, both from the model's own forward pass, and each layer is analyzed on that input with the trained recurrence, on all 258 validation and test sequences, at every fourth target $T = 1024$–2047.

The scan probe matches each layer's scan ERF with frozen selectivity to at most 2.2×10^{-7} of the curve's maximum, and pointwise to 1.6×10^{-4} after both curves are normalized at $\tau = 1$ (Table 15). The two layers differ: layer 1's scan probe has a least-squares slope of -1.99 over $\tau = 10$–100, while layer 2's is flat over that range (0.06) and falls steeply beyond it (-5.21 over $\tau = 100$–1000). Because the identity holds per layer for any input, this checks the implementation on inputs shaped by an upstream layer. It is not a test of composition across layers (Appendix K).

Table 15: Two-layer model, per layer: least-squares slopes of the scan probe with 95% bootstrap intervals over sequences; the identity residual against the scan ERF with frozen selectivity, raw (maximum absolute error / reference peak) and after normalization at $\tau = 1$ (worse of two targets, $\tau \leq 1023$); and the percentage of conductances below 0.01.

Layer	Slope, $\tau = 10$–100	Slope, 100–1000	Raw (of max.)	Normalized	$g < 0.01$
1	-1.99 [-2.00, -1.97]	-3.27 [-3.37, -3.18]	6.3×10^{-9}	7.9×10^{-5}	11.8%
2	0.06 [-0.08, 0.13]	-5.21 [-6.14, -5.11]	2.2×10^{-7}	1.6×10^{-4}	13.1%

Block formula in the last block. For the second block we also compare the block formula with the block's full ERF. Here the ERF is taken with respect to the block's incoming activations (the output of layer 1), not the embeddings. The probe is the squared logits, and the readout $a_T = \frac{1}{2}\nabla_{y_T}\ell_T$ is computed exactly through the final LayerNorm, the vocabulary projection and the block's output path. The sample is all 120 validation sequences at 128 targets each. The block formula matches the block differentiated with Δ_t, B_t and C_t detached to a relative error of 3.7×10^{-5}. Averaged over lag bands, the full ERF exceeds the block formula by a factor of about 1.5 to 1.8, and the fitted slopes agree (-1.46 and -1.52 over $\tau = 10$–100; -2.32 and -2.31 over $\tau = 100$–1000). At individual lags the agreement is weaker: over $\tau = 4$–500 the raw ratio ranges from 1.12 to 3.84, and from 0.70 to 2.40 after normalization at $\tau = 1$. The per-lag bound of Section 4.3 therefore does not carry over to this block. This is one checkpoint, with a final LayerNorm and a different readout, so it does not isolate the effect of depth.

K Limitations and Future Work

Frozen selectivity and linear response. The block formula holds Δ, B and C fixed. The exact decomposition of Section 3.6 restores their dependence on the input, so the approximation lies only in the block formula. Its gap to the full model is measured in Section 4.3 (within a factor of 2.3 up to

lag 500, with a largest shortfall of approximately $5.13\times$ beyond) and Section 4.4. The block formula and the decomposition are built from gradients: they describe the response to an infinitesimal change of a past input at the trained operating point. Finite edits of the sequence are outside their scope.

Per-layer scope. The main text uses a single-layer model, Appendix J a two-layer model, and Appendix I all 48 layers of a pretrained Mamba-370M. Each layer is treated independently: we compute the scan probe from that layer's extracted conductances and check that it reproduces that layer's scan ERF with frozen selectivity, which it must by construction (Appendix L.4). The compositional prediction of the full end-to-end ERF across stacked nonlinear layers is a separate object that we do not address here. Inter-layer composition would require modeling how the output distribution of one layer shapes the selectivity of the next.

Finite d_{state}. With 16 state dimensions per channel (8192 terms across our 512 channels), each channel's measure is discrete and cannot produce arbitrarily smooth spectral edges. Watson's lemma gives asymptotics in the continuum idealization. Quantitative predictions use the finite sums over state dimensions.

Forward-pass dynamics only. The analysis concerns how information propagates through a trained model. How gradient descent shapes the conductance distribution during training, and why training converges to a particular spectral measure, remain open questions.

The analogy. The resistor grid is a motivating analogy (Appendix A): the grid is reciprocal and two-dimensional, while the scan is causal and one-dimensional. Some spatial phenomena (current loops, multidimensional percolation) have no temporal counterpart.

Future work. One extension is to predict the end-to-end ERF across stacked nonlinear Mamba blocks. Connecting the spectral measure to HiPPO initialization [Gu et al., 2020], which places eigenvalues to approximate specific polynomial bases, would relate it to an existing theoretical framework. Testing the block formula against the full-model ERF of pretrained models (Mamba-370M and larger) would show whether the approximation of Section 4.3 holds beyond single-layer models.

L Derivation Details

This appendix provides the full algebra behind the results stated in the main text.

L.1 Zero-Order Hold Discretization

Mamba's continuous-time dynamics for a single channel are $\dot{x} = Ax + Bu$, $y = Cx$, where $A = \operatorname{diag}(a_1, \ldots, a_N)$ with each $a_i < 0$ and $N = d_{\text{state}}$. Zero-order hold keeps u constant over an interval of duration $\Delta_t > 0$ and integrates the dynamics exactly.

State transition. The homogeneous solution gives $\bar{A}_t = e^{A\Delta_t}$, i.e. $\bar{a}_{i,t} = e^{a_i \Delta_t} \in (0, 1)$ for every i.

Input matrix. The particular solution for constant u is

$$x(\Delta_t) = e^{A\Delta_t}\, x(0) + \underbrace{\left(\int_0^{\Delta_t} e^{As}\, ds \right)}_{=(e^{A\Delta_t} - I)\, A^{-1}} B_t\, u, \tag{26}$$

where A^{-1} exists because each $a_i \neq 0$. The integral evaluates componentwise to $\left(e^{a_i \Delta_t} - 1\right)/a_i$. Defining $G_t := I - e^{A\Delta_t}$, the discretized input matrix becomes

$$\bar{B}_t = (e^{A\Delta_t} - I)\, A^{-1}\, B_t = -G_t\, A^{-1}\, B_t. \tag{27}$$

Componentwise, $\bar{b}_{i,t} = g_{i,t}/|a_i| \cdot B_{i,t}$, using $a_i < 0$ so $-1/a_i = 1/|a_i|$.

Mamba's training kernel. Mamba's kernel keeps the exact decay $\bar{A}_t = e^{A\Delta_t}$ but replaces the input matrix (27) by the first-order step $\bar{B}_t = \Delta_t B_t$, which agrees with it to first order in $A\Delta_t$.

Key structural point. Both the retention factor $(I - G_t)$ and the input injection $\bar{B}_t$ are determined by the single step size Δ_t, under either input weight. When Δ_t is large, G_t is large (old state drained) and $\bar{B}_t$ is large (new input injected). When Δ_t is small, both are small: old state preserved, input rejected. For a given input projection B_t, the model cannot drain without injecting or inject without draining, since Δ_t sets both.

L.2 Shunt Conductance: Positivity and Approximation

Positivity. $g_{i,t} = 1 - e^{a_i \Delta_t}$, with $a_i < 0$ and $\Delta_t > 0$. Since $a_i \Delta_t < 0$, we have $e^{a_i \Delta_t} \in (0, 1)$, so $g_{i,t} \in (0, 1)$. Mamba parameterizes $a_i = -\exp((A_{\log})_i)$, guaranteeing $a_i < 0$ throughout training, so $G_t \succ 0$ always.

Small-Δ expansion. Taylor-expanding:

$$g_{i,t} = 1 - e^{a_i \Delta_t} = -a_i \Delta_t - \tfrac{(a_i \Delta_t)^2}{2} + O(\Delta_t^3) \approx -a_i \Delta_t = |a_i| \Delta_t. \tag{28}$$

The relative error of the linear approximation is $|a_i \Delta_t|/2 + O(\Delta_t^2)$. For example, $|a_i| = 16$ and $\Delta_t = 0.1$ give $|a_i \Delta_t| = 1.6$, where the true relative error exceeds 100%. The linear approximation is therefore used only for intuition. All numerical results in this paper use the exact exponential form $g_{i,t} = 1 - e^{a_i \Delta_t}$.

L.3 Impulse Response as a Sum of Exponentials

Unrolling the recurrence $x_{j,i,t} = (1 - g_{j,i,t})\, x_{j,i,t-1} + \bar{b}_{j,i,t}\, u_{j,t}$, with $\bar{b}_{j,i,t} = \Delta_{j,t}\, B_{i,t}$, from time s to time $t > s$:

$$x_{j,i,t} = \left(\prod_{r=s+1}^{t} (1 - g_{j,i,r}) \right) x_{j,i,s} + \sum_{k=s+1}^{t} \left(\prod_{r=k+1}^{t} (1 - g_{j,i,r}) \right) \bar{b}_{j,i,k}\, u_{j,k}. \tag{29}$$

For the impulse response, $x_{j,i,s-1} = 0$ and the input at time s produces $x_{j,i,s} = \bar{b}_{j,i,s}\, u_{j,s}$. Subsequent propagation gives $\left[\prod_{r=s+1}^{t} (1 - g_{j,i,r})\right] \bar{b}_{j,i,s}\, u_{j,s}$. Reading out at time t:

$$h_j(t, s) = \sum_{i=1}^{N} C_{i,t}\, \bar{b}_{j,i,s} \prod_{r=s+1}^{t} (1 - g_{j,i,r}). \tag{30}$$

In the LTI case ($g_{j,i,r} = g_{j,i}$ constant), this becomes

$$h_{\tau,j} = \sum_{i=1}^{N} c_i\, (1 - g_{j,i})^{\tau}\, \bar{b}_{j,i}, \qquad \tau = t - s \geq 1, \tag{31}$$

a weighted sum of N exponentials with decay rates $\{-\log(1 - g_{j,i})\}$ and signed residues $\{c_i \bar{b}_{j,i}\}$.

L.4 ERF Derivation from Chain Rule

At the scan, the probe at target position T is $\ell_T = \|y_T\|^2 = \sum_j y_{j,T}^2$, where $y_{j,T} = \sum_i C_{i,T}\, x_{j,i,T} + D_j\, u_{j,T}$ includes the skip term. With frozen selectivity (Δ_t, B_t, C_t treated as constants), the chain rule gives, for $\tau \geq 1$,

$$\frac{\partial \ell_T}{\partial u_{j,T-\tau}} = 2\, y_{j,T}\, \frac{\partial y_{j,T}}{\partial u_{j,T-\tau}} = 2\, y_{j,T}\, h_j(T, T - \tau). \tag{32}$$

The skip term contributes to $\partial y_{j,T} / \partial u_{j,T-\tau}$ only at $\tau = 0$, so at positive lags it enters solely through the weight $y_{j,T}$. Omitting it from that weight applies a different per-channel weighting to the same Jacobian, which leaves a τ-dependent ratio that the $\tau = 1$ normalization does not remove. The scan's ERF is the expected squared gradient norm:

$$\mathrm{ERF}_{\mathrm{scan}}(\tau) = \mathbb{E}\left[\sum_j \left(\frac{\partial \ell_T}{\partial u_{j,T-\tau}} \right)^2 \right] = 4\, \mathbb{E}\left[\sum_j y_{j,T}^2\, h_j(T, T - \tau)^2 \right]. \tag{33}$$

Three features of Eq. (33) matter:

1. **Signed residues:** $h_j(T, T-\tau) = \sum_i C_{i,T}\, \bar{b}_{j,i,T-\tau} \prod(1-g)$ uses signed $C_{i,T}\bar{b}_{j,i}$. Positive and negative terms can cancel, making $|h_j|$ smaller than $\sum_i |C_{i,T}\bar{b}_{j,i}| \prod(1-g)$.
2. **Per-channel squaring:** The ERF sums $\sum_j h_j^2$. It squares each channel's impulse response *before* summing over j.
3. **Output weighting:** Each channel contributes $y_{j,T}^2 \cdot h_j^2$, not h_j^2 alone.

Replacing the signed residues by their absolute values $|C_{i,T}\, \bar{b}_{j,i}|$ bounds each channel's $|h_j|$ from above, so it raises the unnormalized scan probe. After normalization at $\tau = 1$ it is not a bound. Table 16 gives this sign ablation for the nine main models, for the scan probe and for the block formula, in which only the signs of the residues $C_{i,T}\bar{b}_{j,i,s}$ are removed and the readout and the input path keep theirs. Without normalization, removing the signs raises the scan probe by a factor of about 1.07 to 65 and the block formula by up to 7.1, depending on the band and the model. After normalizing each curve, the unsigned curve can lie above or below the signed curve. Appendix I gives the same ablation for Mamba-370M.

Table 16: Sign ablation: the curve with the signs of the residues $C_{i,T}\bar{b}_{j,i,s}$ removed over the signed curve, for the scan probe and the block formula. Geometric mean over each lag band; range over the nine main models (validation). Normalized: each curve is first normalized at $\tau = 1$.

	τ	1	5–15	40–60	80–120	180–220	400–600	800–1023
Scan probe	raw	4.09–27.6	5.41–19.1	5.45–53.1	4.88–56.4	3.52–65.0	1.45–44.2	1.07–33.8
	normalized	1.00	0.454–1.62	0.733–3.50	0.408–4.79	0.170–10.9	0.353–8.55	0.261–6.53
Block formula	raw	2.03–7.01	1.56–4.74	1.29–2.45	1.07–1.53	1.01–1.33	1.00–2.48	1.00–5.28
	normalized	1.00	0.618–1.28	0.250–0.768	0.170–0.585	0.147–0.605	0.174–1.13	0.191–1.63

L.5 Block ERF with Frozen Selectivity

We derive Eq. (16), in the notation of Section 3.5. Fix $\tau \geq 1$ and hold Δ_t, B_t and C_t fixed.

Paths from $e_{T-\tau}$ to ℓ_T. The input $e_{T-\tau}$ enters three quantities at position $T-\tau$: the scan branch $W_{\text{in}}\, e_{T-\tau}$, the gate $z_{T-\tau}$ and the residual. The gate and the residual reach only $o_{T-\tau}$, which does not affect ℓ_T for $\tau \geq 1$. The scan branch feeds the convolution pre-activation p_s for $s = T-\tau+k$, $k = 0, \ldots, 3$, through the tap w_{3-k}: in $p_s = \sum_{m=0}^{3} w_m \odot W_{\text{in}}\, e_{s+m-3} + b_{\text{conv}}$, the term with $s+m-3 = T-\tau$ has $m = 3-k$. Only the positions $s \leq T$ matter, which gives $k \leq \min(3, \tau)$.

The scan. With Δ, B and C fixed, $y_{j,T}$ depends on $u_{j,s}$ only through the write $\bar{b}_{j,i,s}\, u_{j,s}$ and, for $s = T$, the skip term, so $\partial y_{j,T}/\partial u_{j,s} = h_j(T,s) + D_j\, \mathbb{1}[s = T] = K_j(T,s)$ by Eq. (30). $y_{j,T}$ does not depend on the other channels' inputs.

The readout. For any differentiable probe ℓ_T of o_T, define $a_T = \frac{1}{2}\nabla_{y_T}\ell_T$, so that $\partial\ell_T/\partial y_{j,T} = 2\, a_{j,T}$. For $\ell_T = \|W_{\text{lm}}\, o_T\|^2$ and $o_T = W_{\text{out}}\big(y_T \odot \text{SiLU}(z_T)\big) + e_T$,

$$a_{j,T} = \text{SiLU}(z_{j,T})\, \big[W_{\text{out}}^\top W_{\text{lm}}^\top W_{\text{lm}}\, o_T\big]_j;$$

for the logit margin of Section 4.4, $W_{\text{lm}}\, o_T$ is replaced by $\frac{1}{2}\omega$, where ω holds the margin's coefficients over the vocabulary.

Assembling. By the chain rule through $u_{j,s} = \text{SiLU}(p_{j,s})$ and the taps,

$$\frac{\partial \ell_T}{\partial e_{T-\tau}} = \sum_{k=0}^{\min(3,\tau)} W_{\text{in}}^\top \Big[w_{3-k} \odot \text{SiLU}'(p_s) \odot \frac{\partial \ell_T}{\partial u_s} \Big]_{s=T-\tau+k} = 2\, W_{\text{in}}^\top v,$$

with v as in Eq. (16), since $\partial\ell_T/\partial u_{j,s} = 2\, a_{j,T}\, K_j(T,s)$. Squaring and taking the expectation gives Eq. (17).

Check. On all 18 WikiText checkpoints, the block formula matches automatic differentiation through the model with Δ_t, B_t and C_t detached, to a maximum relative error of 4.0×10^{-5} in float32.

L.6 Exact Decomposition with Differentiable Selectivity

We derive Eq. (18). The scan input u_s enters y_T through five pathways: (1) the input factor $u_{j,s}$ of the write $\Delta_{j,s}B_{i,s}u_{j,s}$; (2) Δ_s in the decay $1-g_{j,i,s}=e^{A_{j,i}\Delta_{j,s}}$; (3) Δ_s in the write; (4) B_s; and (5) C_s, only when $s=T$. For one channel's state, (1)–(4) correspond to Huang et al. [2025, Eq. 11].

(1) The frozen write. $\partial y_{j,T}/\partial u_{j,s}\big|_{\Delta,B,C} = K_j(T,s)$, as in Appendix L.5.

(2, 3) Through Δ_s. The state $x_{j,i,s}$ depends on $\Delta_{j,s}$ through the decay of the stored state and through the write:

$$\frac{\partial x_{j,i,s}}{\partial \Delta_{j,s}} = A_{j,i}\,(1-g_{j,i,s})\,x_{j,i,s-1} + B_{i,s}\,u_{j,s}.$$

The state $x_{j,i,T}$ inherits this through $P_{j,i}(T,s)$, since the decays at later steps depend only on their own scan inputs. Hence $\partial y_{j,T}/\partial \Delta_{j,s} = R_{j,s} + V_{j,s}$, with $R_{j,s}$ and $V_{j,s}$ as defined in Section 3.6. With $\Delta_{j,s}=\mathrm{softplus}(q_{j,s})$ and $q_s = W_\Delta u_s + b_\Delta$, $\partial \Delta_{j,s}/\partial u_{j',s} = \sigma(q_{j,s})\,W_\Delta[j,j']$.

(4) Through B_s. $\partial y_{j,T}/\partial B_{i,s} = C_{i,T}\,P_{j,i}(T,s)\,\Delta_{j,s}\,u_{j,s}$ and $\partial B_{i,s}/\partial u_{j',s} = W_B[i,j']$.

(5) Through C_T. For $s=T$ only, $\partial y_{j,T}/\partial C_{i,T} = x_{j,i,T}$ and $\partial C_{i,T}/\partial u_{j',T} = W_C[i,j']$.

Collecting. Multiplying by $\partial \ell_T/\partial y_{j,T} = 2a_{j,T}$ and summing over j (and over i where it appears) gives twice the five row vectors of Eq. (18). Twice their sum is $\partial \ell_T/\partial u_s$ exactly, with no residual.

The Δ-write identity. For $s<T$, $K_j(T,s) = \Delta_{j,s}\sum_i C_{i,T}\,P_{j,i}(T,s)\,B_{i,s}$, so $V_{j,s} = (u_{j,s}/\Delta_{j,s})\,K_j(T,s)$ and $a_{j,T}\,V_{j,s}\,\sigma(q_{j,s}) = \gamma^{\mathrm{fr}}_{j,s}\,\lambda_{j,s}$, which gives Eq. (19). At $s=T$ the same holds with $K_j(T,T)-D_j$ in place of $K_j(T,T)$.

Checks. In float64, through the model's forward pass with the trained recurrence, on the first validation sequence (first 300 tokens, $T=150$ and 299) of three other WikiText checkpoints, not among the 18 analyzed here:

- the frozen term matches automatic differentiation with Δ, B and C detached to a relative error of at most 8.1×10^{-15};
- with one of Δ, B or C left differentiable and the other two detached, the frozen term plus the corresponding selectivity terms matches to at most 2.5×10^{-14};
- the sum of all five terms matches the full model to at most 2.5×10^{-14}.

For the two Δ pathways separately, a copy of the scan in which Δ enters the decay and the write as two tensors gives at most 1.5×10^{-13}. The same check on an MQAR and a selective-copy model, under the margin probe, gives at most 4.2×10^{-14}. In float32, on all 18 WikiText checkpoints, the sum of the five terms matches the full model to a maximum relative deviation of 1.2×10^{-5}.

L.7 Watson's Lemma: Full Derivation

Assume the continuum idealization: the residue-weighted measure has an integrable density p on $[0,1]$, so that the per-channel impulse response is $h_\tau = \int_0^1 (1-g)^\tau p(g)\,dg$. For large τ, the kernel $(1-g)^\tau = \exp(\tau\ln(1-g))$ is sharply peaked near $g=0$, with effective width $O(1/\tau)$.

Step 1: Localize. For small g, $\ln(1-g)\approx -g$, so $(1-g)^\tau \approx e^{-g\tau}$.

Step 2: Assume edge behavior. Let $p(g)\sim c\,g^\alpha$ as $g\to 0^+$ with $\alpha>-1$, $c>0$.

Step 3: Extend integration. The contribution above any fixed cutoff $0<\delta<1$ is exponentially small in τ. Extending the approximating exponential integral to infinity adds an exponentially small tail. The kernel and edge approximations leave a relative error that tends to zero but need not be exponentially small:

$$h_\tau \approx c\int_0^\infty e^{-g\tau}\,g^\alpha\,dg.$$

Step 4: Substitute $u = g\tau$.

$$h_\tau \approx c \int_0^\infty e^{-u} \left(\frac{u}{\tau}\right)^\alpha \frac{du}{\tau} = \frac{c}{\tau^{\alpha+1}} \int_0^\infty u^\alpha e^{-u}\, du = \frac{c\,\Gamma(\alpha+1)}{\tau^{\alpha+1}}. \tag{34}$$

Result. $h_\tau \sim c\,\Gamma(\alpha+1)\,\tau^{-(\alpha+1)}$ as $\tau \to \infty$. The same leading term follows exactly from Watson's lemma after the change of variables $\kappa = -\ln(1-g)$: then $h_\tau = \int_0^\infty e^{-\tau\kappa}\, q(\kappa)\, d\kappa$ with $q(\kappa) = e^{-\kappa}\, p(1-e^{-\kappa}) \sim c\,\kappa^\alpha$ as $\kappa \to 0^+$.

Classification.

Spectral edge	α	Tail	Memory
$p(0) > 0$ (flat at edge)	0	τ^{-1}	Heavy algebraic
$p(g) \sim g^\alpha$	$\alpha > -1$	$\tau^{-(\alpha+1)}$	Algebraic
$p(g) = 0$ near $g = 0$	–	Exponential	Short-range

This classification assumes the continuum idealization and serves only as motivation. All quantitative predictions use the finite discrete sum.

L.8 Exponential Decay under Scalar i.i.d. Conductances: Full Proof

Claim. For a scalar system ($N = 1$) with i.i.d. conductances g_t, the expectation of the conductance product $P_\tau = \prod_{s=T-\tau+1}^{T}(1-g_s)$, and of each of its moments, decays as a single exponential in the lag. With the residues held fixed, the impulse response is proportional to P_τ, so heavy tails *cannot* arise from i.i.d. temporal randomness in a single dimension.

Annealed average. By independence, for every $k \geq 1$,

$$\mathbb{E}[P_\tau^k] = \prod_{s=T-\tau+1}^{T} \mathbb{E}[(1-g_s)^k] = \big(\mathbb{E}[(1-g)^k]\big)^\tau,$$

a pure exponential. For $k = 1$ its rate is set by $\bar{g} = \mathbb{E}[g]$; for $k = 2$, the moment that enters the ERF, by $\mathbb{E}[(1-g)^2] = (1-\bar{g})^2 + \mathrm{Var}(g)$.

Quenched (typical) behavior. If $\mathbb{E}|\log(1-g)| < \infty$, which holds when Δ_t is bounded (as for bounded inputs; Appendix B), the strong law of large numbers gives

$$\frac{1}{\tau}\log P_\tau = \frac{1}{\tau}\sum_{s=T-\tau+1}^{T} \log(1-g_s) \xrightarrow{\text{a.s.}} \mathbb{E}[\log(1-g)].$$

Thus the typical decay rate is $-\mathbb{E}[\log(1-g)]$. By Jensen's inequality ($\log$ is concave):

$$\mathbb{E}[\log(1-g)] \leq \log\big(\mathbb{E}[1-g]\big) = \log(1-\bar{g}),$$

so the typical conductance product decays at least as fast as the expected one.

Conclusion. Heavy tails therefore need structure beyond i.i.d. variation in a single dimension. Candidates include the cross-dimensional mixture (Section 3.4) and sustained regimes of the conductances (Appendix C).

L.9 Log-Space Arithmetic for Running Products

The running product $\prod_{s=T-\tau+1}^{T}(1-g_{j,i,s})$ can underflow when computed directly over hundreds of steps. Since $1 - g_{j,i,s} = \exp(A_{j,i}\Delta_{j,s})$, its logarithm is exactly $A_{j,i}\Delta_{j,s}$. We accumulate $\Lambda_{j,i,t} = \sum_{s \leq t} A_{j,i}\Delta_{j,s}$ in float64, and $P_{j,i}(t,s) = \exp(\Lambda_{j,i,t} - \Lambda_{j,i,s})$. No clamp is needed, even where g rounds to 1 in float32. The static mixture's $(1-\bar{g}_{j,i})^\tau$ is computed as $\exp(\tau\,\mathrm{log1p}(-\bar{g}_{j,i}))$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state the contributions and scope, supported by the derivations in Section 3 and the experiments in Section 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 and Appendix K discuss the scope of the identity checks, approximation accuracy, the analysis of individual blocks, infinitesimal input changes, three-seed experiments and the limits on causal interpretation.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Section 3 states the assumptions and results. Appendices B and L provide the full derivations.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 4.1 and Appendices E–J describe the architectures, training recipes, data, probes, extraction protocols and statistical analyses.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Code, result files and reproduction instructions are at `https://github.com/zypapers/mamba-conductance-erf`. WikiText-103 is public, and the supplied scripts generate the synthetic tasks (Appendices F and G).

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Section 4.1 and Appendices E–G specify the data splits, hyperparameters, initializations and checkpoint selection, including the weight-decay control.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Section 4.1 and Appendices E–G define the seed-level 95% t-intervals, paired contrasts, sequence-bootstrap intervals conditional on the checkpoint, and task means and standard deviations. The $2.3\times$ bound in Section 4.3 is an observed worst case across the 18 language models and lags 1–500.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Timings are for one NVIDIA RTX PRO 6000 GPU (96 GB): each set of nine language models trains in about 9.2 hours (three runs at a time), the ERF analyses take about 2.3 hours (main models) and 2.2 hours (control models), the 18 task models train in about 24 minutes in total, each task decomposition takes about 4 minutes, and the Mamba-370M and two-layer analyses take 21 and 4 minutes. The two-layer model was trained on one NVIDIA RTX 5090, and the full project used additional compute beyond these experiments, including for preliminary analyses.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work analyzes an existing architecture using public data and synthetic tasks, with no human participants or private data.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This is foundational research on an existing architecture, with no application or deployment studied. We foresee no direct negative societal impacts from the analysis.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: The work uses small research models and an existing public checkpoint for interpretability analysis. We identify no high-risk release requiring additional safeguards.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We use WikiText-103 [Merity et al., 2017] under CC BY-SA (its dataset card names versions 3.0 and 4.0 and the GFDL), loaded from Hugging Face as `wikitext-103-raw-v1`. The Mamba code (adapted for our reference scan), the `state-spaces/mamba-370m` checkpoint [Gu and Dao, 2024] and the GPT-NeoX-20B

tokenizer are used under Apache-2.0, the GPT-2 tokenizer under MIT and PyTorch [Paszke et al., 2019] under BSD-3-Clause.

Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The analysis code is documented in the accompanying README, including its purpose, requirements, usage and reproduction commands.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects.

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

Answer: [N/A]

Justification: This work does not involve human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No LLM is used as a component of the methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.