SELECTIVITY SETS THE RECALL RANGE IN HYBRID LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid language models interleave a few softmax-attention layers among many state-space model (SSM) layers. It has been unclear which properties of the SSM predict the query-key distances at which attention improves recall. Replacing every attention output with an input-independent constant isolates the SSM's autonomous recall. We show that attention sustains recall at distances where this recall has declined, a handoff that is clear in Granite-4.0, weak in Nemotron-H, and absent in one parallel- and one shared-attention hybrid. Input-dependent *timestep gating* (Δ-selectivity) sets the SSM's recall *range* (g_{50}, the gap at which chance-corrected recall falls to half its peak). It does so by keeping interfering tokens from eroding the state. On a controlled two-key associative task, turning selectivity on extends the range 2.4-fold (from 24 to 58 tokens), whereas fixing a trained selective model's timestep at each head's mean drops peak recall below query-blind guessing. Turning selectivity on during training shifts the onset of an added attention layer's benefit to larger query-key gaps. Changes in recall range quantitatively predict these shifts. Across natively trained backbones, the range predicts onset on held-out architectures with a mean absolute error of 3.9 tokens, against 10.3 for the strongest architecture-plus-selectivity baseline. In Granite-4.0 and Nemotron-H, removing input-dependent gating significantly reduces the SSM's autonomous recall while attention-on recall stays well above it. Native SSM recall depends on this gating. On Granite, fixing each sequence's timestep at its native average shrinks the range from 11 to 3 tokens without a significant drop in peak recall. Timestep gating sets how far the SSM recalls on its own, and that range predicts the query-key distance at which attention begins to help.

1 INTRODUCTION

Hybrid language models interleave a few attention layers among many state-space model (SSM) layers (Lieber et al., 2024; NVIDIA, 2025). This design rests on a division of labor: the SSM carries most sequence mixing at linear cost, and the attention layers support associative recall. Without attention, associative recall over long contexts fails (Waleffe et al., 2024), and design studies report which depths suit attention layers best (Bae et al., 2026). These findings leave open which properties of the SSM predict the query-key distances at which attention improves recall.

Our approach isolates the SSM's own recall ability. We replace every attention layer with a fixed constant estimated from unrelated sequences. This removes attention's token-varying contribution without touching the SSM path, and the constant cannot leak the key being recalled. What remains is the network's autonomous associative recall, which we measure as a function of the *distance* between a query and the earlier key–value binding. Autonomous recall decreases with the query-key gap. We summarize this decline by the recall range g_{50}, the gap at which chance-corrected recall falls to half its peak. Attention sustains recall where the SSM's autonomous recall has declined.

The property that sets g_{50} is selectivity, the input-dependent gating that lets an SSM decide per token what to write into and keep in its state. On the two-key associative task, turning selectivity on extends g_{50} from 24 to 58 tokens, and replacing a trained selective model's timestep with each head's mean at evaluation drops peak recall below a query-blind guess. Selective Δ keeps the timestep near zero on filler tokens, and a decomposition attributes the range mainly to preventing filler-induced forgetting (Section 5).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

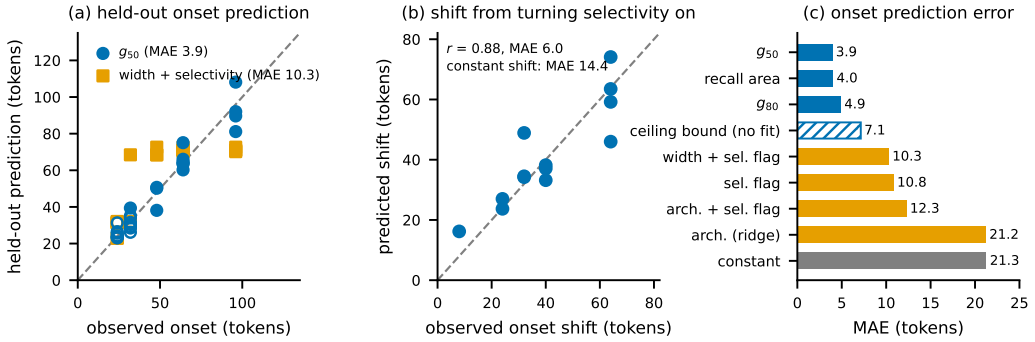

Figure 1: The SSM's recall range predicts the onset of an added attention layer's benefit (Section 6). (a) Leave-one-architecture-out prediction for 28 backbones from g_{50} (circles) or from width and the selectivity flag (squares); filled: selective; dashed: identity. (b) Onset shift from turning selectivity on during training (13 matched pairs) against the shift predicted from the change in g_{50}. (c) Mean absolute error by predictor: fitted predictors on held-out architectures; the ceiling bound (hatched), evaluated directly on the same 28 backbones.

Selectivity is known to counteract memory decay (Huang et al., 2025). We connect this retention mechanism quantitatively to the distance at which attention improves recall (Figure 1). Across natively trained staged hybrids, the recall range predicts the onset of attention's benefit on held-out architectures with a mean absolute error of 3.9 tokens, against 21.2 for a regularized model of the architecture and 10.3 for the best architectural baseline that also includes the selectivity flag. It also predicts the onset shifts caused by turning selectivity on during training. Where prior analysis derives the retention effect, we test it causally by token role, separate its decay and write channels, and test the corresponding range effect in pretrained hybrids. In two pretrained hybrids (Granite-4.0-h-tiny and Nemotron-H-8B), removing the input-dependent gating significantly reduces autonomous SSM recall while attention-on recall stays well above it. Native SSM recall therefore depends on this gating. On Granite, removing it at each sequence's native average timestep leaves peak recall without a significant drop but shrinks the range from 11 to 3 tokens (Section 7).

Contributions. (i) The SSM's associative recall range g_{50}, isolated by a leakage-free attention ablation, with attention sustaining recall at gaps where SSM-only recall declines (§4). (ii) Controlled causal evidence that input-dependent timestep gating extends g_{50} and shifts the distance at which an added attention layer begins to improve recall, with the range predicting that distance and the toggle-induced shift, though not established as their mediator (§5–6). (iii) Necessity of the gating for native SSM recall in both pretrained hybrids (autonomous recall is lower in Nemotron-H), a range effect on Granite when each sequence keeps its own mean timestep, and the handoff mapped across interleaving regimes (§7).

2 Background and Related Work

State-space and gated-recurrence models. State-space models map inputs x_t to outputs y_t (t indexes tokens) through a linear recurrence on a fixed-size state h_t, $h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t$, $y_t = C_t h_t$ (transition $\bar{A}_t$, input map $\bar{B}_t$, readout C_t). The family includes continuous-time memory (Gu et al., 2020), structured state spaces (Gu et al., 2022; Smith et al., 2023), convolutional and gated variants (Fu et al., 2023; Poli et al., 2023), Mamba (Gu & Dao, 2024), and Mamba-2 (Dao & Gu, 2024). Linear attention and gated linear recurrences reach similar fixed-state forms (Katharopoulos et al., 2020; Sun et al., 2023; Yang et al., 2024; Peng et al., 2023; De et al., 2024; Botev et al., 2024; Beck et al., 2024). Mamba computes an input-dependent timestep $\Delta_t = \mathrm{softplus}(W_\Delta x_t + b)$ with a learned projection W_Δ and bias b (Mamba-2 applies W_Δ to the layer input before the convolution; Section 5). The timestep makes $\bar{A}_t$ and $\bar{B}_t$ vary with the current token (*selectivity*), so the model can admit tokens that matter and gate out interfering ones; this filtering role motivated the design (Gu & Dao, 2024). Selectivity is the property this paper isolates.

Recall capacity and range. Because the state is a fixed-size summary, associative recall is bounded. Recall capacity grows with state size (Arora et al., 2024a), sliding-window attention trades recall against throughput (Arora et al., 2024b), and recurrent models are weaker at copying (Jelassi et al., 2024) and in-context retrieval (Wen et al., 2025); their in-context learning is mixed (Park et al., 2024; Grazzi et al., 2024). Ben-Kish et al. (2025) tie how far Mamba extrapolates in length to the timestep, and Huang et al. (2025) show analytically that selectivity counteracts memory decay by driving Δ_t toward zero. We test this decay account causally at the level of token roles, separate the timestep's decay and write channels, and show that the resulting range predicts the distance at which interleaved attention begins to improve recall.

Hybrid architectures. Jamba (Lieber et al., 2024), Zamba (Glorioso et al., 2024b;a), Hymba (Dong et al., 2025), Nemotron-H (NVIDIA, 2025), Falcon-H1 (Zuo et al., 2025), Granite-4.0 (IBM Research, 2025), and the sliding-window hybrids Samba (Ren et al., 2025) and RecurrentGemma (Botev et al., 2024) combine attention with SSM or recurrent layers, sequentially, in parallel within a block, or through shared attention. Bae et al. (2026) report that attention layers help least at the front of the stack and most in intermediate positions. Waleffe et al. (2024) show that a few attention layers recover a Transformer's recall, which pure SSMs lack. Lee et al. (2025) find that sequential hybrids favor shorter contexts and parallel hybrids longer ones. Arora et al. (2025) find the short convolution critical for constructing associative bindings, which complements our account: the convolution builds the association, and the timestep sets how long it persists. Long-context evaluations (Hsieh et al., 2024; Liu et al., 2024) probe long-context recall behavior but not its cause.

Mechanistic and sensitivity analyses. Circuit analysis (Elhage et al., 2021) of Transformer attention (Vaswani et al., 2017) identifies induction heads (Olsson et al., 2022), retrieval heads (Wu et al., 2025), and attention sinks (Xiao et al., 2024; Gu et al., 2025). Mamba admits an implicit-attention reading (Ali et al., 2025), and the logit and tuned lenses probe the residual stream (Belrose et al., 2023). Pruning methods remove weights (Frantar & Alistarh, 2023) or rank layers and blocks by measured ablation sensitivity (Muralidharan et al., 2024; Muñoz et al., 2025). These methods localize *which* components carry recall. We study which property of the frozen SSM predicts the query-key *distance* at which attention begins to improve recall.

3 SETUP

Model panel. We study six pretrained hybrids spanning three architectural regimes. Granite-4.0-h-tiny (Granite below), Granite-4.0-h-micro (IBM Research, 2025), Nemotron-H-4B (Taghibakhshi et al., 2025), and Nemotron-H-8B (Nemotron-H below; NVIDIA, 2025) are *sequentially interleaved*: attention and SSM layers alternate in depth, so a query passes through SSM state and attention in series. Falcon-H1-1.5B (Zuo et al., 2025) is *parallel-in-block*: attention and SSM run side by side within a block and their outputs are combined. Zamba2-1.2B (Glorioso et al., 2024a) uses *shared attention*, a single attention block reused at every sixth layer. All SSM layers are Mamba-family selective blocks (Gu & Dao, 2024; Dao & Gu, 2024). Because the regimes combine attention and SSM differently, we report every quantity per model and never pool across regimes.

Gap-recall probe. In the controlled testbed we measure associative recall over distance with a two-key task: two key–value pairs, g filler tokens (the "gap"), then one of the two keys as a query, whose bound value the model must return. Single-pair recall can be solved by fixed-position copying: a one-pair testbed model reaches full recall even when the query is replaced by a random token. With two keys, a query-blind guess caps at 50%. Testbed fillers exclude the keys and values; pretrained models are probed with templated single-key prompts. The pretrained prompts draw fillers from the same list of single-token words as the keys and values (65 words for Granite), so at long gaps a filler can repeat the key or value (about two repeats per prompt at gap 64). Recall at gap g is the *exact-match rate* of the greedy next token, with Wilson 95% confidence intervals (CIs). For the two-key range we chance-correct against the query-blind baseline ($c{=}1/2$), which is stricter than the $1/K$ vocabulary baseline ($K{=}16$ in the testbed). For pretrained models, which predict over the full vocabulary, we estimate chance from the long-gap floor (about 0.02 for Granite and at most 0.01 for Nemotron-H). In these models, swapping the query key for a random word drops SSM-only recall to near the floor beyond gap 0 (Section 7), so there recall depends on the query and not on position.

Ablation unit: SSM-only recall. To isolate what the SSM alone can recall, we replace each attention output at every position with an input-*independent* constant, the *random-token mean*: the mean attention output over sequences of uniformly random tokens, estimated from two such calibration sequences s_1, s_2,

$$\tilde{o}_\ell \;=\; \tfrac{1}{2}\textstyle\sum_{j=1}^{2} \bar{o}_\ell(s_j),$$

where $\bar{o}_\ell(s)$ is attention layer ℓ's output averaged over the positions of s. The constant $\tilde{o}_\ell$ removes attention's token-varying contribution and cannot encode the current key or value. We call recall under this intervention *SSM-only recall*. A per-sequence broadcast (the current sequence's own attention output averaged over positions) is a common alternative. Because it is computed from the sequence being recalled, it can leak the answer to the SSM, and it inflates peak SSM-only recall by 2.4–4.0×. We report it only as an upper bound and validate the leakage-free measure against a counterfactual patch and against zeroing (Appendix A).

Recall range g_{50}. Let $r(g)$ be SSM-only recall at gap g, c the chance rate, and $r^\star = r(g^\star)$ the peak, reached at $g^\star = \arg\max_g r(g)$. g_{50} is the first gap *past the peak* at which chance-corrected recall crosses half its peak value,

$$g_{50} \;=\; \min\left\{ g \geq g^\star : \frac{r(g) - c}{r^\star - c} \leq \tfrac{1}{2} \right\},$$

linearly interpolated between evaluated gaps, with parametric-bootstrap 95% CIs (resampling each gap's count as $\mathrm{Binomial}(n, r(g))$ for n prompts per gap). The post-peak restriction and chance correction matter when the curve rises from a low $r(0)$ or when $r^\star$ is close to c. g_{50} is defined only when $r^\star$ is above chance. The parallel-in-block and shared-attention models have SSM-only peaks *at or below* chance (Falcon-H1-1.5B and Zamba2-1.2B, $r^\star = 0.00$), so g_{50} is undefined.

4 The Gap-Graded Handoff Is Regime-Associated

On a single-key recall probe with a controlled query-to-key gap, we measure the recall of the full model (*clean*) and the leakage-free *SSM-only* recall of Section 3. Their difference, the *attention ablation effect*, is the recall lost when attention is replaced by its constant. In a nonlinear composed stack this difference measures necessity (it also reflects interactions and residual-stream perturbation) rather than a purely additive attention contribution, and across gaps it shows how each architecture divides retrieval over distance.

SSM-only recall without leakage. Granite and Nemotron-H differ in strength. Granite shows SSM-only recall above chance at short gaps, with a leakage-free peak of 0.25 (95% CI $[0.16, 0.37]$), and this recall decays with gap. At this sample size ($n = 64$) the curve is non-monotone over the shortest gaps. Nemotron-H-8B is far weaker, peaking at 0.08 ($[0.03, 0.17]$, a CI that nearly reaches the chance floor), so we treat it as weaker evidence of autonomous recall than Granite. The leaky per-sequence broadcast inflates these peaks but leaves the range nearly unchanged, because the range depends on where the curve decays rather than on its height (Appendix A).

Granite shows a recall handoff over distance. In sequentially interleaved architectures, where attention and SSM layers alternate along depth, the SSM pathway can retrieve on its own at short gaps and degrade as the gap grows. Granite-4.0-h-tiny shows this: a leakage-free SSM-only peak of 0.25 with recall range $g_{50}{\approx}7.6$ (bootstrap 95% CI $[1.8, 11.8]$, indicating limited precision in the range estimate). The SSM-only recall decays with gap while attention's ablation effect grows. The division is not a strict partition. Because SSM-only recall is modest even at short gaps, attention carries much of the recall throughout, and its ablation effect is already large at short gaps rather than switching on only past g_{50}. The range g_{50} marks the distance over which the recurrent

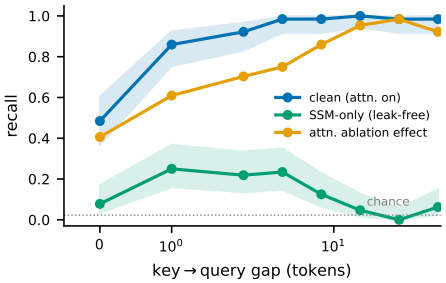

Figure 2: The handoff on Granite: clean (attention-on) recall, leakage-free SSM-only recall, and their difference vs. gap, with Wilson 95% bands ($n = 64$).

Table 1: Leakage-free SSM-only recall peak and range across regimes, with the broadcast peak (upper bound) in parentheses. Ranges are chance-corrected g_{50} with bootstrap 95% intervals; n/a marks peaks at or below chance, where g_{50} is undefined.

Model	Regime	Leakage-free peak (broadcast)	g_{50}
Granite-4.0-h-tiny	sequentially interleaved	0.25 (0.61)	7.6 [1.8, 11.8]
Nemotron-H-4B	sequentially interleaved	0.06 (0.20)	1.5 [0.8, 6.0]
Nemotron-H-8B	sequentially interleaved	0.08 (0.31)	5.7 [0.7, 7.5]
Falcon-H1-1.5B	parallel-in-block	0.00 (0.09)	n/a
Zamba2-1.2B	shared-attention	0.00 (0.12)	n/a

path retrieves *on its own*. The pattern is not uniform
across models: both Nemotron-H variants are markedly weaker (leakage-free peaks 0.06–0.08), so the evidence for a distance-graded range comes from Granite, while Nemotron-H sits closer to the attention-dominated end. Under the per-sequence broadcast, the smaller Granite-4.0-h-micro shows the same pattern, but its leakage-free recall is near zero, so we do not estimate its range.

No autonomous recall in Falcon-H1 and Zamba2. In Falcon-H1-1.5B (attention and SSM in parallel within a block (Zuo et al., 2025)) and Zamba2-1.2B (a shared attention block (Glorioso et al., 2024a)), leakage-free SSM-only recall is 0.00 at every gap: no short-gap regime in which the SSM retrieves on its own. For both, the per-sequence broadcast produces spurious recall (rising to 0.12 for Zamba2) that is *entirely* leakage and vanishes under the random-token mean, which is why the leakage-free measure is necessary. Removing attention reduces observed recall to zero instead of uncovering a distance-limited SSM pathway, leaving no gap-graded division of labor (Table 1).

Interpreting the regime association. For the sequentially interleaved and parallel designs, the split follows from how often attention is available along the recall path. In a sequentially interleaved stack attention is present at only some blocks, so the SSM alone carries an association across the gap through the blocks without attention. The SSM's finite range bounds how far it carries the association on its own. In the parallel-in-block design an attention pathway is present at every block, so the SSM is never the sole carrier. It does not cover Zamba2, whose shared attention block recurs every sixth layer. Two caveats apply. First, we profile a single parallel and a single shared checkpoint, so regime is confounded with family, scale, weight sharing, and training (an association, not a causal effect of integration style). Second, the leakage-free ablation measures what the SSM retrieves *autonomously*, which lower-bounds its true contribution since the SSM branch is co-adapted to attention.

5 Selectivity Causally Sets the SSM Recall Range

The recall range g_{50} could be a spectral quantity: the memory reach of the SSM stack. On Granite and Nemotron-H, however, the decay horizons of the SSM heads order the models opposite to their range estimates (Appendix D). We instead train small pure-SSM models from scratch on the two-key associative task and toggle one architectural property, input-dependent selectivity, while holding state capacity fixed. Selectivity sets g_{50}, and the mean timestep alone does not account for it.

Design. Each model is a stack of Mamba-2 layers (Gu & Dao, 2024; Dao & Gu, 2024) with no attention. One projection of the layer input u_t gives the output gate z_t, the SSM input x_t (after a short convolution), B_t, C_t, and $\Delta_t = \mathrm{softplus}(W_\Delta u_t + b)$ with per-head bias b. With a learned per-head scalar $A < 0$, the layer sets $\bar{A}_t = \exp(\Delta_t A)$, $\bar{B}_t = \Delta_t B_t$, and runs $h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t$, $y_t = C_t h_t$. Our manipulation is confined to Δ_t. In the *selective* setting Δ_t is a learned function of the current token. In the *non-selective* setting we zero W_Δ, so $\Delta_t = \mathrm{softplus}(b)$ is input-independent and the token-dependent gating is removed. The input-dependent maps B_t and C_t are untouched, so the manipulation isolates *timestep (Δ) selectivity* specifically and does not make the layer a time-invariant recurrence. We match state size d_{state}, width d_{model}, depth, and training data across the two settings, so any difference in range is attributable to Δ-selectivity, not capacity. Recall is the two-key exact-match rate, which forecloses the position-copying a one-key task admits, and fillers exclude the four content tokens. Every evaluation-time intervention below uses fixed per-head

constants measured once per model on separate calibration batches, never statistics of the sequence being evaluated. We estimate g_{50} from the recall-versus-gap curve against the query-blind baseline ($c{=}1/2$) over a 128-token window and report mean $\pm$ sample standard deviation over three seeds.

Selectivity extends the associative range. Both settings recall near-perfectly at short gaps (peak 1.00). Selective models reach $g_{50}{=}58.1 \pm 8.6$ tokens and non-selective models $g_{50}{=}23.8{\pm}4.2$, a $2.4\times$ separation (2.2–$2.7\times$ per seed). Selectivity thus governs *how far* an association carries, not whether the SSM can retrieve at all (Figure 3). The non-selective recall decays to 0.03 at 128 tokens, well below the 0.5 a query-blind strategy would score (ignoring the query and guessing one of the two presented values). The decay reflects query-conditioned retrieval rather than a fallback to guessing (Appendix F adds a second configuration and the vocabulary baseline).

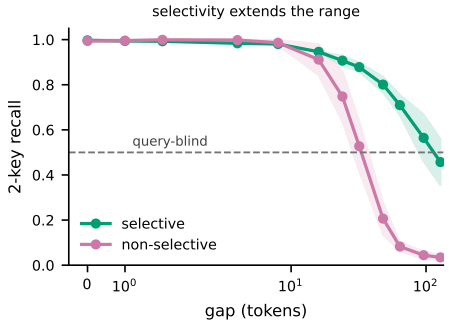

Figure 3: Two-key recall vs. gap for selective (input-dependent Δ) and non-selective (input-independent Δ) pure-SSM models at matched capacity (mean and sample standard deviation over three seeds).

Controlling for the mean timescale. The two settings could differ in average timestep, so the difference in range might reflect a slower mean decay rather than input-dependent gating. A within-model control separates the two. At evaluation we replace a trained selective model's timestep with each head's fixed calibration mean (*mean-preserve*), which holds the average timestep while removing all input-dependence of the timestep. Peak recall falls to 0.28, below the query-blind baseline of 0.5, so no chance-corrected range can be defined. The recall depends on *which* tokens Δ gates, not on the average step size (Appendix B). Because this control leaves no recall above the query-blind baseline on which to define a range, the graded-range comparison above rests on the selective and non-selective models, which both recall near-perfectly at short gaps.

Mechanism. Selectivity extends the range by keeping filler tokens from eroding the state: input-dependent Δ drives the timestep toward zero on filler, so the recurrent decay $\bar{A}_t{=}\exp(\Delta_t A)$ stays near identity there and the binding is retained across the gap. This is the retention mechanism Huang et al. (2025) derive analytically (their Lemma 2), and we test it causally by intervening on Δ by token role at evaluation (content = keys, values, query). We read the *range*, not the short-gap peak, which stays at 0.99–1.00 under every filler intervention. Raising filler Δ to a fixed content-level constant collapses the range from 58.1 to 4.6 ± 2.2 (gap-64 recall $0.71{\to}0.11$), while setting it near zero (10^{-3}) extends the range to the full window (gap-128 recall 1.00). Because Δ drives both the decay and the write $\bar{B}_t{=}\Delta_t B_t$, we decompose them, a separation the analytical account does not make. Raising filler Δ in the decay path alone collapses the range to 3.7 ± 1.7, while raising it in the write path alone leaves 46.8 ± 12.6 (Appendix G). Selectivity therefore acts far more by preventing filler-induced *forgetting* than by preventing filler writes from overwriting the binding. Content gating is separately necessary: zeroing content Δ removes the state write and drops peak recall to 0.18, below the query-blind baseline.

6 The Recall Range Predicts Attention's Onset Distance

Section 5 established that selectivity sets the SSM's associative recall range g_{50}. We now connect the range to attention: toggling selectivity shifts the query-key *distance* at which an added attention layer begins to improve recall, and the range predicts that distance. The obstacle is a training degeneracy: in an end-to-end hybrid, a trainable attention layer can solve recall on its own and reduce the SSM's incentive to learn it. The same shortcut-learning tendency motivates delayed attention training in hybrids (Phan et al., 2025). We break the degeneracy by staging the two components: we first train a pure-SSM model on the two-key associative task, then freeze it. We then insert a single attention layer after the third SSM layer and train it to improve recall over the frozen backbone. Backbones are natively trained, selective or non-selective, without evaluation-time intervention.

We test what predicts the onset, the smallest gap at which added attention raises recall over the frozen SSM by at least 0.2. Five architectures (different widths, state sizes, depths, and initial scales of W_Δ, called Δ-init), trained selective or non-selective with three seeds each, give frozen ranges from g_{50}=18 to 103 tokens (18 to 87 among the 28 analyzed). On each backbone we train the attention layer over the full 128-token window. The onset rises with the range, onset $\approx 1.19\, g_{50} + 1.6$ (Pearson r=0.98, n=28 backbones; Figure 7), and the relation holds within each class (selective r=0.95; non-selective r=0.73 over g_{50} 18–28), so the relation is not driven only by the selective/non-selective split. Because hybrid recall cannot exceed 1, a benefit of 0.2 requires SSM recall at or below 0.8. Among the 28 analyzed backbones, onset occurs at the first evaluated gap satisfying this condition in 54% of cases and at the next evaluated gap in the remainder. This parameter-free lower bound predicts onset with a mean absolute error of 7.1 tokens (Appendix E), so the range–onset correlation partly reflects this constraint. One of the 30 backbones does not reach the benefit threshold after attention training. Fit on four architectures and tested on the fifth, g_{50} predicts onset with a mean absolute error of 3.9 tokens (Figure 1a). A regularized linear model of the architecture (width, state size, depth, Δ-init) gives 21.2; adding the selectivity flag gives 12.3, the best single architectural feature with the flag 10.3, and the flag alone 10.8 (Figure 1c; Appendix E). Every paired bootstrap interval of baseline error minus g_{50} error excludes zero. Other summaries of the same curve give errors of 4.0 tokens (area of recall above the query-blind level) and 4.9 (g_{80}, the post-peak gap where raw recall falls to 80% of peak; Appendix E). The distance therefore tracks where the frozen recurrent path degrades, which g_{50} summarizes. With Section 5, this connects selectivity to the distance at which added attention helps.

Toggling selectivity shifts onset as the range predicts. We compare selective and non-selective backbones matched by architecture and seed, which differ only in whether selectivity was trained. Across the 13 pairs in which both backbones enter the analysis, turning selectivity on delays the onset by 40.6 tokens on average. The delay predicted from each pair's change in g_{50}, using the slope fitted on the other four architectures, is 41.2. Across pairs the prediction has correlation 0.88 and mean absolute error 6.0 $[3.2, 9.3]$ tokens, against 14.4 for predicting the average delay for every pair (Figure 1b). Whether the range mediates the effect of selectivity on onset is not established (Appendix E).

Evaluation-time interventions do not isolate the link. Varying the range within a backbone by evaluation-time Δ interventions gives a rule that predicts a second intervention family worse than a constant (mean absolute error 16.9 versus 11.8). These interventions also act on the SSM layers after the inserted attention and push them off their training distribution, so onset there tracks the attention's shortfall rather than the range (Appendix E).

Distance versus depth. Reported placement preferences, disfavored initial and preferred intermediate positions (Bae et al., 2026), concern layer *depth*, not distance. Adding the correct value's scaled output-head vector at a frozen backbone's query position shows no preference among injection depths after the first layer, and the effect of injection before the first layer varies with backbone and gap (Appendix C). Under this control, the preference against initial placement seen with trainable attention reduces to a small first-layer deficit.

7 Selectivity Is Necessary for SSM Recall in Pretrained Hybrids

Sections 5 and 6 show in synthetic testbeds that selectivity sets g_{50} and shifts attention's onset, and that g_{50} predicts that onset. We now test, without retraining, whether selectivity plays the same role in pretrained hybrids. The questions are whether removing input-dependence at a matched timescale reduces the SSM's recall (*necessity*) and, on Granite, whether it sets the *range*.

The measured recall is query-triggered. All interventions use templated single-key prompts (`The` k `is` v`.`, then g fillers, then `The` k `is`). On these exact prompts, replacing the final key with a random word drives SSM-only recall to near the floor at every gap from 1 on. The peak falls by 0.21 $[0.15, 0.27]$ on Granite and 0.09 $[0.05, 0.12]$ on Nemotron-H (paired bootstrap, n=256 per gap), and swapped-query recall stays at or below 0.05 (full model: 0.21–0.52). At gap 0 on Granite,

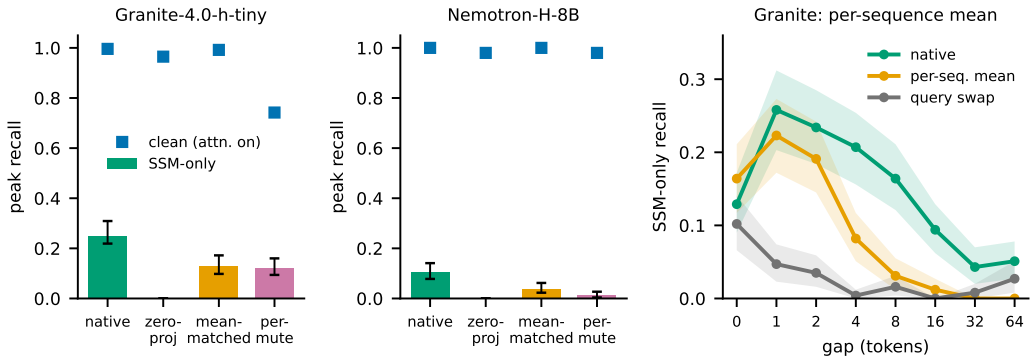

Figure 4: Calibrated Δ_t interventions on two sequentially interleaved hybrids: leakage-free SSM-only peak recall (bars, n=256 with bootstrap 95% intervals) with attention-on peak recall (squares), for Granite (left) and Nemotron-H (middle). The *mean-matched* and *permute* controls (defined below) reduce SSM-only recall while attention-on recall stays well above it. Right: Granite SSM-only recall versus gap on the same prompts under the per-sequence mean (each sequence keeps its own average timestep) and with the query key swapped, with bootstrap 95% bands.

with no filler between binding and query, the swap changes recall little (0.10 versus 0.12). Beyond gap 0 the measured SSM-only recall is therefore query-triggered retrieval, not position-copying.

Attention performs key selection. When each key of the same two-key context is queried in turn, the full model ranks the queried key's value above the other in 0.81–0.97 of cases on Granite and 0.78–0.84 on Nemotron-H (gaps $\geq$1; chance 0.5). Selection by the leakage-free SSM path is weak on Granite and near chance on Nemotron-H. On Granite it ranks the queried key's value higher in 0.55–0.60 of cases (intervals above 0.5) but almost never answers both queries correctly (at most 0.012). On Nemotron-H the rate is 0.52–0.54. Because selection comes mainly from attention, we evaluate the Δ interventions on the single-key probe.

Interventions that separate input-dependence from timescale. We intervene on the Mamba timestep Δ_t of Granite-4.0-h-tiny (IBM Research, 2025) and Nemotron-H-8B (NVIDIA, 2025) and measure SSM-only and clean recall. *Zero-proj* zeros the Δ_t input projection. *Mean-matched* also sets the bias so Δ_t equals the native calibration-mean step. *Per-sequence mean* sets each head's timestep, at every position, to the mean of its own native Δ_t over the same evaluation sequence, so each sequence keeps its native average timescale in every layer. The means are recorded in an unmodified pass and then replayed, and the realized timesteps match them to within 3% on Granite and 5% on Nemotron-H. *Permute* shuffles the native Δ_t across the positions of each sequence, keeping its values but destroying token alignment. Table 3 lists every intervention's effect.

Removing input-dependence at a fixed timescale. On Granite, native SSM-only recall peaks at 0.25. Two timescale-preserving controls reduce it significantly: mean-matched to 0.13 (paired effect 0.12 $[0.07, 0.18]$) and permute to 0.12 (0.13 $[0.08, 0.19]$; Figure 4). The per-sequence control preserves each sequence's native mean timestep. Section 5 places the filler mechanism in the range rather than the short-gap peak, and Granite shows that pattern. Peak recall shows no significant decrease (paired difference 0.035, 95% CI $[-0.023, 0.094]$), while recall at gaps 4 and 8 decreases by 0.125 $[0.078, 0.172]$ and 0.133 $[0.090, 0.180]$, and the recall range on these prompts shrinks from 11.0 $[6.3, 14.9]$ to 3.3 $[2.5, 4.0]$ tokens. On Nemotron-H, whose native SSM-only peak is lower (0.105), every control reduces peak recall significantly, the per-sequence mean by 0.08 $[0.05, 0.12]$. Its native range of 3.7 tokens spans too few evaluated gaps to resolve a change. Granite's range contraction and Nemotron-H's peak-recall reduction persist across five fresh calibration draws (Appendix E). Input-dependent gating is necessary on both models, and on Granite it also extends the range.

Attention sustains recall. Clean recall stays far above SSM-only recall under every timescale-preserving control. Under the per-sequence mean it stays above 0.94 (Granite) and 0.99 (Nemotron-

H) at every gap ≥ 1. Permute on Granite lowers it to between 0.64 and 0.74 across gaps, and mean-matched on Nemotron-H lowers it to 0.71 at gap 64 (Table 3). Attention therefore sustains recall when the SSM loses its input-dependent gating.

8 Discussion, Limitations, and Conclusion

What the evidence establishes. On a two-key associative task, input-dependent timestep gating causally sets a finite recall range g_{50} (the mean timestep alone does not account for it, since the mean-preserve control drops peak recall below the query-blind baseline). Toggling the gating during training shifts the distance at which added attention begins to improve recall. The range predicts that distance, with a leave-one-architecture-out mean absolute error of 3.9 tokens, and it predicts the toggle-induced shift. In pretrained hybrids we show necessity: removing input-dependent gating reduces SSM-only recall in Granite-4.0 (IBM Research, 2025) and Nemotron-H (NVIDIA, 2025) while attention-on recall stays well above it, although autonomous recall is lower in Nemotron-H. On Granite, the per-sequence mean shortens the range without a significant drop in peak recall (Section 7).

Relation to placement studies, regime, and prior work. Our result concerns query-key *distance*. Reported depth preferences for attention (Bae et al., 2026) lie outside its scope, and the answer-injection control shows no preference among injection depths after the first layer (Section 6). The distance handoff appears only where the SSM and attention paths are sequentially interleaved. In parallel-in-block and shared-attention regimes (Falcon-H1 (Zuo et al., 2025), Zamba2 (Glorioso et al., 2024a)) the SSM path shows no leakage-free autonomous recall (peak 0.00), so recall in these checkpoints depends on attention under the tested ablation. We profile one checkpoint each, so we report an association rather than a causal effect of integration style. Selectivity's role in filtering context (Gu & Dao, 2024; Arora et al., 2024a) and in counteracting memory decay (Huang et al., 2025) is established. We add a causal token-role decomposition in which suppressed decay on fillers, far more than the write, sets a measurable recall range, and we show that this range predicts the distance at which added attention begins to improve recall. Others characterize *what* attention computes in hybrids (Bick et al., 2025; Michalak & Abreu, 2025); this complements our account of the distance at which attention is needed.

Limitations. Four limitations bound the claims. (i) The causal results are established in controlled synthetic settings (small pure-SSM stacks on a two-key task). In these settings they cover selectivity's effect on the range and on onset but not the range's role as mediator, whose direct test was inconclusive (Section 6). In pretrained hybrids we show necessity on both models and a range effect on Granite only. That effect rests on one timescale-preserving control, built on whole-sequence statistics, under which peak recall does not decrease significantly. (ii) In the controlled testbed, the within-model input-dependence controls drop peak recall below the query-blind baseline rather than shortening the range, so they establish necessity, not a graded range. A protocol variant in which fillers could repeat keys and values also detected no graded dose-response within the selective setting, so we report input-dependent gating versus none. (iii) The pretrained range-change evidence comes from Granite; Nemotron-H's range is too short to resolve that change. The pretrained-model evidence is limited to associative-recall probes. (iv) In the staged experiment onset and g_{50} are read from the same frozen backbone. The link holds for natively trained backbones but not under evaluation-time Δ interventions, which push the layers after the attention off their training distribution (Section 6).

Conclusion. In controlled SSMs, input-dependent timestep gating extends recall range primarily by suppressing filler-induced recurrent forgetting. The mean timestep alone does not account for the range. Toggling the gating during training shifts the distance at which added attention begins to help, by amounts the resulting recall range g_{50} predicts. In pretrained hybrids we establish that native SSM recall depends on the gating and, on Granite, that the gating extends the range. Whether the range mediates attention's onset, and how the range intervention extends to more model families, remain open.

AI USE STATEMENT

The authors developed the research ideas and led the experimental design and analysis. Generative AI tools assisted with polishing, editing, shortening, and organizing the manuscript, as well as developing code and storing results. The authors take full responsibility for the final manuscript, code, analyses, and conclusions.

REPRODUCIBILITY STATEMENT

Appendix E documents the architectures, training settings, calibration procedure, evaluation protocol, and statistical procedures. The experiments use publicly available pretrained hybrids (Granite-4.0-h-{tiny,micro}, Nemotron-H-{4B,8B}, Falcon-H1-1.5B, Zamba2-1.2B) and small SSMs trained from scratch. Every reported number is computed from saved experiment outputs; the code, run configurations, and these outputs are available at `https://anonymous.4open.science/r/repo-D666/`.

REFERENCES

Ameen Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of Mamba models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1516–1534, 2025.

Aryaman Arora, Neil Rathi, Nikil Roashan Selvam, Róbert Csordás, Dan Jurafsky, and Christopher Potts. Mechanistic evaluation of Transformers and state space models. *arXiv preprint arXiv:2505.15105*, 2025.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024a.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *International Conference on Machine Learning (ICML)*, 2024b.

Sangmin Bae, Bilge Acun, Chien-Yu Lin, Haroun Habeeb, Seungyeon Kim, Liang Luo, Junjie Wang, and Carole-Jean Wu. Transformer–SSM hybrid language models: Systematic analysis and design insights. *Transactions on Machine Learning Research*, 2026. URL `https://openreview.net/forum?id=x7qyXl8ecT`.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from Transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. DeciMamba: Exploring the length extrapolation potential of Mamba. In *International Conference on Learning Representations (ICLR)*, 2025.

Aviv Bick, Eric Xing, and Albert Gu. Understanding the skill gap in recurrent language models: The role of the gather-and-aggregate mechanism. In *International Conference on Machine Learning (ICML)*, pp. 4324–4344, 2025.

Aleksandar Botev, Soham De, Samuel L. Smith, et al. RecurrentGemma: Moving past Transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*, 2024.

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

Soham De, Samuel L. Smith, Anushan Fernando, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for Transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning (ICML)*, 2023.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations (ICLR)*, 2023.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, Anna Golubeva, Vasudev Shyam, James Whittington, Jonathan Pilault, and Beren Millidge. The Zamba2 suite: Technical report. *arXiv preprint arXiv:2411.15242*, 2024a.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7B SSM hybrid model. *arXiv preprint arXiv:2405.16712*, 2024b.

Riccardo Grazzi, Julien Niklas Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. Is Mamba capable of in-context learning? In *Proceedings of the Third International Conference on Automated Machine Learning (AutoML)*, volume 256 of *Proceedings of Machine Learning Research*, pp. 1/1–26, 2024.

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024.

Ningyuan Teresa Huang, Miguel Sarabia, Abhinav Moudgil, Pau Rodriguez, Luca Zappella, and Federico Danieli. Understanding input selectivity in Mamba: Impact on approximation power, memorization, and associative recall capacity. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25693–25727, 2025.

IBM Research. Granite 4.0 language models. https://github.com/ibm-granite/granite-4.0-language-models, 2025.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning (ICML)*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020.

Hyunji Lee, Wenhao Yu, Hongming Zhang, Kaixin Ma, Jiyeon Kim, Dong Yu, and Minjoon Seo. Understanding and enhancing Mamba-Transformer hybrids for memory recall and language modeling. In *Proceedings of the First BabyLM Workshop*, pp. 380–398, 2025.

Opher Lieber, Barak Lenz, Hofit Bata, et al. Jamba: A hybrid Transformer-Mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

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

Felix Michalak and Steven Abreu. Some attention is all you need for retrieval. *arXiv preprint arXiv:2510.19861*, 2025.

J. Pablo Muñoz, Jinjie Yuan, and Nilesh Jain. Mamba-Shedder: Post-Transformer compression for efficient selective structured state space models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3851–3863, 2025.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

NVIDIA. Nemotron-H: A family of accurate and efficient hybrid Mamba-Transformer models. *arXiv preprint arXiv:2504.03624*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Can Mamba learn how to learn? A comparative study on in-context learning tasks. In *International Conference on Machine Learning (ICML)*, 2024.

Bo Peng, Eric Alcaide, Quentin Anthony, et al. RWKV: Reinventing RNNs for the Transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, 2023.

Buu Phan, Reza Ebrahimi, Sanjay Haresh, and Roland Memisevic. Delayed attention training improves length generalization in Transformer–RNN hybrids. *arXiv preprint arXiv:2510.00258*, 2025.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning (ICML)*, 2023.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

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

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

Ali Taghibakhshi, Sharath Turuvekere Sreenivas, Saurav Muralidharan, Marcin Chochowski, Yashaswi Karnati, Raviraj Joshi, Ameya Sunil Mahabaleshwarkar, Zijia Chen, Yoshi Suhara, Oluwatobi Olabiyi, Daniel Korzekwa, Mostofa Patwary, Mohammad Shoeybi, Jan Kautz, Bryan Catanzaro, Ashwath Aithal, Nima Tajbakhsh, and Pavlo Molchanov. Efficient hybrid language model compression through group-aware SSM pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Roger Waleffe, Wonmin Byeon, Duncan Riach, et al. An empirical study of Mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. RNNs are not Transformers (yet): The key bottleneck on in-context retrieval. In *International Conference on Learning Representations (ICLR)*, 2025.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations (ICLR)*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention Transformers with hardware-efficient training. In *International Conference on Machine Learning (ICML)*, 2024.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim Khadraoui, Mugariya Farooq, Giulia Campesan, Ruxandra Cojocaru, Yasser Djilali, Shi Hu, Iheb Chaabane, Puneesh Khanna, Mohamed El Amine Seddik, Ngoc Dung Huynh, Phuc Le Khac, Leen AlQadi, Billel Mokeddem, Mohamed Chami, Abdalgader Abubaker, Mikhail Lubinets, Kacper Piskorski, and Slim Frikha. Falcon-H1: A family of hybrid-head language models redefining efficiency and performance. *arXiv preprint arXiv:2507.22448*, 2025.

A Leakage-free ablation controls

Section 4 measures SSM-only recall under a leakage-free replacement of attention. Figure 5 gives the full control on Granite and Nemotron-H. Replacing each attention output with an input-independent constant, the mean attention output over random-token calibration sequences (*random-token mean*), leaves gap-graded SSM recall above chance, distinctly on Granite and weakly on Nemotron-H. A counterfactual patch from a different recall sequence (*cf-patch*) cannot encode the current answer but does carry generic recall-sequence structure. It gives a higher Granite peak (0.42 versus 0.25 for the random-token mean) and so sits between the conservative random-token mean and the leaky per-sequence broadcast (0.61). We report the random-token mean throughout as the conservative measure. Wilson 95% intervals on the leakage-free (random-token mean) peak are Granite-4.0-h-tiny 0.25 $[0.16, 0.37]$, Nemotron-H-8B 0.08 $[0.03, 0.17]$, and Nemotron-H-4B 0.06 $[0.02, 0.15]$. The parallel-in-block and shared-attention models (Falcon-H1-1.5B, Zamba2-1.2B) sit at 0.00. The broadcast inflates the peak by $2.4\times$ (Granite) to $4.0\times$ (Nemotron-H-8B) relative to the random-token mean, while the range is essentially unchanged (half-peak crossing without chance correction: Granite $7.9 \to 8.0$, Nemotron-H-8B $5.3 \to 6.0$): the leakage inflates magnitude, not range.

B Selectivity versus mean timescale

Section 5 isolates input-dependence from the mean timescale with a within-model control on the two-key task. Starting from a trained selective SSM, at evaluation only we replace every head's

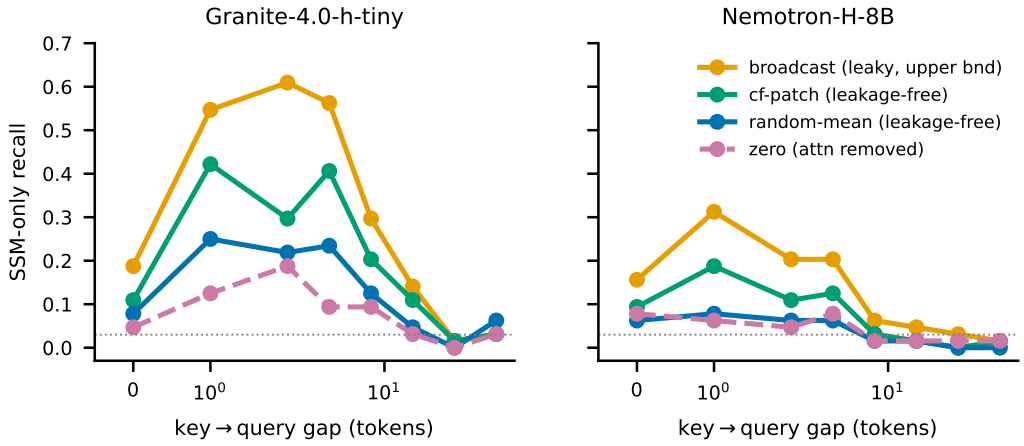

Figure 5: SSM-only recall persists under leakage-free replacement of attention on Granite and Nemotron-H. The leakage-free random-token mean (*random-mean* in the legend) leaves gap-graded SSM recall above chance (dotted line), distinctly on Granite and weakly on Nemotron-H. A *cf-patch* from a different recall sequence cannot encode the current answer but carries generic recall-sequence structure, so it yields higher recall than the random-token mean. *Zero* replaces every attention output with zero. The per-sequence *broadcast* (orange) is a leaky upper bound: it overstates peak recall 2.4–4.0× but tracks the same range.

timestep with that head's mean Δ, measured once on separate calibration batches. This removes input-dependence while holding the model's average timestep, and because the constant is a fixed calibration mean, the control uses no statistics of the evaluated sequence. In both configurations peak recall falls below the query-blind baseline of 0.5 (d_{model}=160, depth 5: 0.28, per seed 0.21, 0.29, 0.34; d_{model}=128, depth 4: 0.35, per seed 0.34, 0.31, 0.42), so no range can be defined against the query-blind baseline (n=6 models; Figure 6). Because the average timestep is held fixed, the recall needs the token alignment of Δ, not the mean decay.

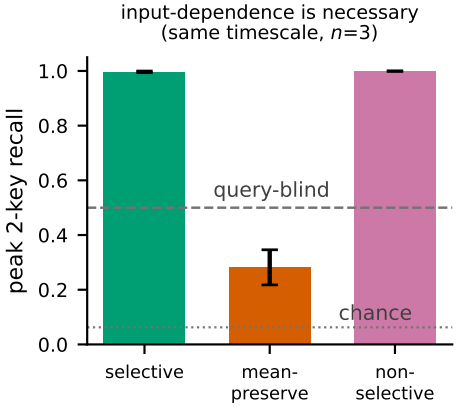

Figure 6: Holding the mean timestep fixed does not preserve the SSM's recall. Replacing a trained selective two-key SSM's Δ_t at evaluation with each head's fixed calibration mean (*mean-preserve*, which holds the average timestep) drops peak recall below the query-blind baseline of 0.5 (dashed), while the selective and non-selective models both recall near-perfectly at short gaps. Mean ± sample standard deviation over three seeds (d_{model}=160, depth 5).

C ANSWER-INJECTION DEPTH CONTROL

Section 6 tests whether the distance result explains the reported depth preferences. We freeze two-key SSM backbones, selective and non-selective (three seeds each), and add the correct value's output-head weight vector (its row of the readout matrix), scaled by 12, to the residual stream at the query position after the first d SSM layers. Depth $d{=}0$ is before all five layers, and $d{=}5$ is immediately before the readout. Injecting at any depth from 1 to 5 gives nearly the same recall (largest spread 0.03), so beyond the first layer recall is nearly independent of the depth at which the correct answer enters. Recall of the injected answer falls with the gap at every depth, including injection immediately before the readout, because it competes with the backbone's own long-gap prediction. It falls faster for the short-range backbone (gap-128 recall 0.13 non-selective versus 0.58 selective, averaged over depths 1–5). Only depth 0 differs. An answer injected before all five layers is recovered 0.09–0.19 less often than at depths 1–5 in the selective backbone; in the non-selective one it is recovered up to 0.08 less often at short gaps but up to 0.07 more often at the longest gaps. This first-layer deficit points toward disfavored initial placement, but it is small and specific to injection before the first layer, and it does not account for the reported depth preferences.

D SPECTRAL REACH VERSUS RECALL RANGE

A candidate predictor of g_{50} is a linear-memory summary of the SSM: the decay rate of every SSM head gives a horizon, on the intuition that a slower-decaying state retrieves further. On the two checkpoints for which we extract every layer's decay rates (Granite-4.0-h-tiny, Nemotron-H-8B), we find no reliable relationship, for two reasons. First, the two g_{50} estimates are noisy and overlapping: Granite's is 7.6 $[1.8, 11.8]$, and Nemotron-H-8B's SSM-only peak is low (0.08), so its g_{50} (5.7 $[0.7, 7.5]$) is imprecise. With $n{=}2$, such estimates cannot be ordered with confidence. Second, the decay horizons order the two models opposite to their range estimates. Each head h has decay rate $\rho_h = -A_h\,\mathbb{E}[\Delta_h]$ and horizon $1/\rho_h$ tokens, where $A_h < 0$ is its state parameter and $\mathbb{E}[\Delta_h]$ its mean timestep over the tokens of 16 recall prompts. Over the heads of all SSM layers, Nemotron-H-8B has the longer reach whether it is read as the reciprocal of the minimum, 10th-percentile, or median decay rate ($1/\mathrm{median}_h\,\rho_h$ is 11.7 versus 5.0 tokens), while its recall-range estimate is the shorter one. A slower-decaying state does not imply a longer recall range here. With $n{=}2$, this comparison does not establish a general relationship between spectral reach and recall range. Section 5 tests selectivity through controlled interventions. A second frozen-model predictor, aimed at individual heads, also fails. We replace one attention head's output at a time with its mean over positions and measure the drop in the answer token's log-probability at the query (32 multi-pair recall prompts; 48 heads on Granite, 128 on Nemotron-H). The proposed score multiplies the positive part of a head's retrieval attention (its attention mass at the query on the correct value, minus the mean on distractor values and the mass on the first token) by its layer's deficit (one minus the logit-lens probability of the answer in the layer's input). Its Spearman correlation with the ablation damage (average ranks for ties) is 0.31 on Granite and -0.14 on Nemotron-H. Direct logit attribution of each head to the answer gives 0.53 and 0.23, and attention-sink mass gives -0.22 and 0.34.

E EXPERIMENTAL DETAILS

Controlled testbeds. Each pure-SSM model is a stack of Mamba-2 layers (per-head scalar A) with an embedding, per-layer root-mean-square normalization (RMSNorm), a final norm, and a linear output head. There is no attention. The selectivity toggle zeros only the Δ_t rows of the input projection, W_Δ (leaving B_t, C_t input-dependent). The Δ-init scale multiplies the initialized W_Δ; it has no effect on non-selective models, whose W_Δ is zero. All controlled experiments use the two-key associative probe over a 16-token vocabulary: the two keys and two values are distinct, and fillers are drawn from the remaining 12 tokens. We train with AdamW (learning rate $3{\times}10^{-3}$, weight decay 0.01, gradient clip 1.0) at batch size 64 for 2500 steps on gaps $\{0, \ldots, 4\}$. Recall is the two-key exact-match rate over 8 fresh batches per gap at gaps $\{0, 1, 2, 4, 8, 16, 24, 32, 48, 64, 96, 128\}$, and g_{50} is chance-corrected against the query-blind baseline ($c{=}1/2$) unless stated otherwise. Because the decay is a per-head scalar shared across state dimensions, we evaluate the scan in its exact closed (Mamba-2 duality) form rather than as a sequential loop; against the loop, logits agree to within $6{\times}10^{-5}$, predictions are identical, and attention-layer gradients agree to $2{\times}10^{-6}$ relative.

Throughout, $\pm$ denotes the sample standard deviation over seeds. Every evaluation-time intervention uses per-layer, per-head constants measured once on 24 separate calibration batches of the unmodified model (the mean Δ over all positions for mean-preserve; the mean over content positions for the filler interventions). The mechanism, answer-injection, and mean-preserve experiments use d_{model}=160, depth 5, d_{state}=64, head dimension 32, Δ-init scale 2, three seeds; the mean-preserve and content-zero controls add d_{model}=128, depth 4, Δ-init 1.

Staged placement. The native-backbone experiment (Section 6) uses five architectures,

$$(d_{\text{model}}, d_{\text{state}}, \text{depth}, \Delta\text{-init}) \in \{(160, 32, 5, 2),\ (160, 64, 6, 1),\ (160, 48, 4, 2),\ (160, 64, 5, 2),\ (128, 64, 4, 1)\},$$

each trained selective and non-selective with three seeds (30 backbones). One two-head causal attention layer is inserted after the third SSM layer of the frozen backbone and trained for 1400 steps on gaps $\{0, \ldots, 128\}$. Attention *onset* is the smallest gap at which the added layer raises recall over the frozen SSM by at least 0.2. A backbone enters the analysis when its SSM peak exceeds 0.5 and both g_{50} and onset are below 128 tokens (28 of 30: one selective backbone has onset 128 at g_{50}=103, and in one non-selective backbone the attention layer never reaches the 0.2 benefit). Each predictor is a least-squares linear rule fit on four architectures and evaluated on the fifth; the paired bootstrap resamples backbones. The area summary integrates $\max(r(g) - \frac{1}{2}, 0)$ over the evaluated gaps by the trapezoidal rule, and g_{80}, unlike g_{50}, is not chance-corrected. An unregularized linear baseline in width, state size, depth, and Δ-init gives 20.4 (19.4 with the selectivity flag), but it fits five coefficients (six with the flag) from four training architectures and is therefore underidentified. The ridge (penalty chosen by inner leave-one-architecture-out) and single-feature baselines of Section 6 are identified, except the width-only baseline in the fold that holds out the only d_{model}=128 architecture. The evaluation-time intervention design varies the range within three of these architectures (three seeds each) by interpolating every layer's timestep toward its fixed calibration mean, $\Delta'_t = \bar{\Delta} + \alpha(\Delta_t - \bar{\Delta})$ with $\alpha \in \{0.6, 0.72, 0.8, 0.86, 0.9, 0.94, 0.97\}$, or by raising filler Δ toward the content-level constant with weight $\lambda \in \{0.15, \ldots, 1\}$. The test fits an onset rule on the α family (62 conditions from nine selective backbones) and predicts the λ family (51 conditions) and the nine native non-selective backbones: mean absolute error 16.9 versus 11.8 for a constant on the λ family, and 10.1 versus 10.9 on the native backbones. Within the α family onset still rises with g_{50} (r=0.67). On intervened backbones the added attention delivers less at long gaps (recall 0.54 at gap 48 in the filler-targeted family, against 0.70 for the unmodified non-selective backbones in the same experiment), so onset tracks the attention's shortfall. Repeating the intervention design with the attention layer after the last SSM layer, so that no intervened layer follows it, removes the downstream confound, but the added layer then reaches the benefit threshold in only 4 of 126 conditions, too few to evaluate the test.

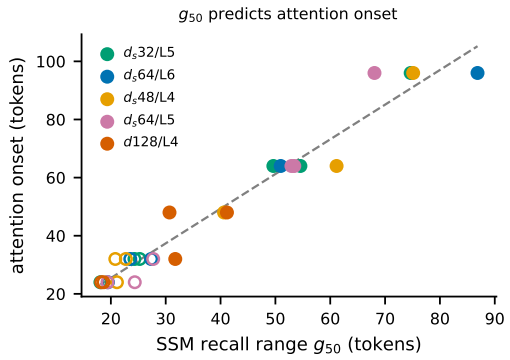

Figure 7: Onset of an attention layer trained on each frozen backbone (smallest gap with benefit $\geq$0.2) versus its range g_{50}, for the 28 natively trained backbones of Section 6. Filled: selective; open: non-selective; dashed: least-squares fit. Legend: d_s=d_{state}, L= SSM layers, d=d_{model} (160 if omitted).

Validity checks for the onset result. These analyses use the SSM-only and hybrid recall curves of the 30 backbones. Because hybrid recall cannot exceed 1, a benefit of at least 0.2 requires SSM-

only recall of at most 0.8, so onset cannot precede the first evaluated gap at which SSM recall falls to 0.8. This bound holds in all 28 backbones; onset equals it in 54% of them and lies at most one evaluated gap beyond it in all. Used directly, the bound predicts onset with a mean absolute error of 7.1 tokens (9.1 when fit on held-out architectures like g_{50}), against 3.9 for g_{50}, which correlates with the bound at $r{=}0.96$. The comparison differs in both interpolation and fitting, so the lower prediction error alone does not establish a mechanism beyond the recall ceiling. At the onset gap the added attention brings recall to 0.91 on average (minimum 0.72). Onset is therefore where the SSM's recall first leaves room for a 0.2 benefit, or the next evaluated gap, and g_{50} summarizes where that happens. Table 2 gives held-out errors per architecture and at other thresholds. Permuting g_{50} among backbones with the same selectivity setting and repeating the held-out fit gives a mean absolute error of 14.6 (95% interval $[10.2, 18.7]$ over 2000 permutations, none at or below 3.9), so g_{50} carries information beyond the selectivity flag.

Table 2: Held-out mean absolute error of onset (tokens). Top: per held-out architecture at threshold 0.2. Bottom: onset thresholds 0.1, 0.2, 0.3, on all backbones included at that threshold and on the 23 included at all three.

Held out (d_{model}, d_{state}, depth, Δ-init)	n	g_{50}	recall area	width + flag
$(160, 32, 5, 2)$	6	2.5	2.6	7.9
$(160, 48, 4, 2)$	6	4.9	5.6	10.9
$(160, 64, 6, 1)$	5	3.5	2.9	8.2
$(128, 64, 4, 1)$	5	4.2	3.4	15.9
$(160, 64, 5, 2)$	6	4.6	5.3	9.2
Thresholds 0.1/0.2/0.3 (n=30/28/23)		5.9/3.9/10.0	6.2/4.0/9.2	9.5/10.3/11.5
Same 23 backbones		4.8/3.4/10.0	4.8/3.1/9.2	5.7/5.5/11.5

Pretrained models. Single-key prompts are templated sentences, `The` k `is` v`.`, then g fillers, then `The` k `is`, with keys, values, and fillers drawn from a list of single-token words (65 for Granite); fillers may repeat the key or value (Section 3). Recall is the exact-match rate of the greedy next token. The leakage-free ablation of Section 4 uses 64 prompts per gap; the Δ_t interventions and controls of Section 7 use 256 prompts per gap with paired bootstrap intervals over prompts, and the query swap and per-sequence mean are evaluated on the identical prompts of the calibrated interventions. The per-sequence mean records each layer's native per-sequence, per-head mean timestep in an unmodified pass and replays it in a second pass on the same batches. Computing each layer's mean from intervened activations instead, a different control, gives a Granite range of 11.3 tokens natively and 3.0 under the control. The two-key selection test uses 256 contexts per gap, `The` k_1 `is` v_1`.` `The` k_2 `is` v_2`.` followed by fillers that exclude all four, with both keys queried on the same context. The SSM-only condition replaces every attention-layer output with the leakage-free input-independent constant (Appendix A). For each layer, the constant is the mean output over two 16-token calibration sequences whose token ids are drawn uniformly from 100 (inclusive) to $\min(8000, \text{vocabulary size})$ (exclusive), one fixed draw per model. We use the kernel-free PyTorch implementation of the Mamba layers throughout. Chance is estimated from the long-gap floor (about 0.02 for Granite and at most 0.01 for Nemotron-H); for the ranges of Section 7 it is the query-swapped SSM-only recall at the two longest gaps (0.018 for Granite). These 256-prompt ranges differ from the 64-prompt estimate of Table 1 (7.6 $[1.8, 11.8]$ tokens), whose interval overlaps them.

Calibration draws. All reported intervals resample evaluation prompts with the replacement constant held fixed, so they are conditional on the calibration draw. We evaluate the Section 7 native and per-sequence-mean conditions under five fresh draws per model, on the same prompts and with each draw's own query-swap chance floor. The redrawn constants lie at a mean relative distance of 0.36–0.38 from the constant used in the main results on Granite and 0.60–0.86 on Nemotron-H. The findings persist across all five draws. On Granite the native SSM-only peak is 0.23–0.29. The per-sequence mean shrinks the range from 12.0–13.3 to 2.9–3.4 tokens, with non-overlapping intervals in every draw, and lowers recall at gaps 4 and 8 by 0.10–0.12 and 0.13–0.16 (every interval above zero). It leaves the peak without a significant decrease (paired differences -0.055 to 0.016). On Nemotron-H the native peak is 0.06–0.13, and the per-sequence mean reduces it by 0.035–0.094, every interval above zero. The other SSM-only results use the original draw only. Granite's calibration forward pass is not bitwise deterministic (constants from the same seed differ by up to 2.6%),

and reruns on identical prompts differ in recall by up to 0.012; on Nemotron-H they differ by at most one prompt in 256.

Analysis specification. The following analyses had success criteria fixed before their runs. The mechanism comparisons of Section 5, the native-backbone onset test against the constant and linear-architecture baselines, and the query-swap control met them, the last on both models. The evaluation-time intervention test did not, and its variant with attention after the last SSM layer had too few conditions to evaluate. The per-sequence-mean experiment, with peak recall as its endpoint, met its criterion on Nemotron-H but not on Granite. The calibration-draw criteria, including Granite's range contraction, were met in all five fresh draws, which use the same recall prompts. Three sets of analyses were chosen after inspecting results: the Granite range and gap-4 and gap-8 comparisons under the per-sequence mean, the matched-pair comparison of Section 6, and the ridge and single-feature architecture baselines, which complement the underidentified linear baseline.

Table 3: Frozen-weight Δ_t interventions on the templated single-key probe (256 prompts per gap). SSM-only peak with Wilson 95% interval; effect on the SSM-only peak (native peak minus intervention peak, from exact counts) with paired bootstrap 95% interval; clean (attention-on) recall at gap 64. †Per-sequence mean from a separate run that records native timesteps and replays them on the same prompts; its row reports recall and effect at the native peak gap.

Model	Intervention	SSM-only peak	Effect on peak	Clean, gap 64
Granite-4.0-h-tiny	native	0.25 $[0.20, 0.31]$	–	1.00
	zero-proj	0.00 $[0.00, 0.01]$	0.25 $[0.22, 0.31]$	0.96
	mean-matched	0.13 $[0.09, 0.18]$	0.12 $[0.07, 0.18]$	0.97
	permute	0.12 $[0.09, 0.17]$	0.13 $[0.08, 0.19]$	0.72
	per-seq. mean†	0.22 at gap 1	0.04 $[-0.02, 0.09]$	1.00
Nemotron-H-8B	native	0.11 $[0.07, 0.15]$	–	1.00
	zero-proj	0.00 $[0.00, 0.01]$	0.11 $[0.08, 0.14]$	0.97
	mean-matched	0.04 $[0.02, 0.07]$	0.07 $[0.04, 0.10]$	0.71
	permute	0.01 $[0.00, 0.03]$	0.09 $[0.07, 0.13]$	0.96
	per-seq. mean†	0.02 at gap 1	0.08 $[0.05, 0.12]$	1.00

F Robustness of the range effect

The selectivity-set range of Section 5 holds in a second configuration (d_{model}=128, depth 4, d_{state}=64, Δ-init 1, three seeds): selective g_{50} is 34.5 ± 5.8 against non-selective 18.8 ± 0.7, and the mean-preserve control again drops peak recall below the query-blind baseline (Appendix B). The separation does not depend on the chance correction. Against the looser vocabulary baseline (c=$1/16$) the main configuration gives selective g_{50} 108.1 (per seed 128.0, 111.6, 84.6; one seed reaches the window) and non-selective 32.2, a wider ratio than under the primary query-blind baseline. A protocol variant in which fillers could repeat keys and values detected no graded dose-response between measured selectivity and g_{50} within the selective setting, so we report the effect as a contrast between input-dependent gating and none rather than a graded law.

G Filler suppression sets the range

Section 5 tests the filler-suppression account by intervening on the timestep Δ_t by token role in trained selective two-key SSMs (three seeds; content positions are the two keys, two values, and query; filler positions are the intervening tokens). The raised value is each head's content-position mean Δ, measured once on separate calibration batches, so no intervention uses statistics of the evaluated sequence. We quantify filler effects using g_{50} and long-gap recall. Peak recall occurs at short gaps and remains at 0.99–1.00 under every filler intervention. Raising filler Δ to the content level collapses g_{50} from 58.1 ± 8.6 to 4.6 ± 2.2 (per seed 2.8, 4.0, 7.1) and gap-64 recall from 0.71 to 0.11. Setting filler Δ near zero (10^{-3}) extends the range to the full 128-token window in every seed (gap-128 recall 1.00, against 0.46 natively). Admitted filler tokens therefore erode the binding, and suppressing their timestep carries the range. Because Δ drives both the decay $\bar{A}$= $\exp(\Delta A)$

and the write $\bar{B}{=}\Delta B$, we reimplement the diagonal scan with separate timesteps for the two paths and raise filler Δ in one at a time. The decay path carries the collapse (g_{50} 3.7 ± 1.7, gap-64 recall 0.05), while the write path leaves most of the range (g_{50} 46.8 ± 12.6; per seed 35.9, 43.8, 60.6; gap-64 recall 0.63). The range effect is thus filler-induced *forgetting* far more than filler overwriting. Content gating is separately necessary: zeroing the content-position Δ removes the state write, and peak recall falls to 0.18 (per seed 0.19, 0.16, 0.17), leaving no chance-corrected range to define.