SUPPORT GEOMETRY BOUNDS DELAYED COPYING IN COMPACT SEQUENCE MIXERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A sequence mixer can have a receptive field longer than a copying task's required delay and still have no access to the token it must recover. We identify this failure through the reachable lag set R, obtained by Minkowski addition of the architectural taps in a residual stack. For finite-bandwidth linear time-invariant mixers, the network's Jacobian support is contained in R. On delayed copying with independent payload tokens, a required lag outside R therefore limits expected accuracy to the optimal payload-marginal baseline, regardless of the learned weights. We test this exclusion while holding support diameter fixed at $Dw = 64$ under a shared, parameter-matched backbone. Schedules that miss the required lag 54 remain at chance; schedules that cover it exceed 99.9% copying accuracy. Replacing one sparse layer with dense taps restores copying without increasing diameter, whereas coprime dilation rates help only when their finite tap sets cover the required lag. These results separate architectural access from soft influence energy and establish a training-free failure screen for compact mixers on tasks with explicit lag requirements.

1 INTRODUCTION

A long receptive field is a natural starting point when choosing a sequence model for long-range recall. Stack bandwidth-w layers to depth D until Dw reaches the relevant delay, and the target distance may appear to be available. This reasoning underlies the appeal of dilated temporal convolutions: dilation expands the receptive field without a proportional increase in kernel size (Yu & Koltun, 2016). For a task that requires a particular past token, however, the farthest reachable position is only part of the specification. A model can reach beyond that token while systematically skipping its position.

Delayed copying turns this distinction into a direct test. An independent payload is followed by a blank interval, a marker, and a scored recall span. Every output in that span must reproduce one specified payload token at a known lookback ℓ_{req}. Seeing nearby tokens cannot reveal an independent target. A sparse dilation schedule can therefore pass $Dw \geq \ell_{\mathrm{req}}$ and still be unable to solve the task, regardless of training. The missing information is the geometry inside the receptive field: which lags can occur along paths through the stack?

The same diameter can hide opposite outcomes. Our central comparison holds the support diameter at $Dw = 64$ and the required lag at 54. Two layers with dilation-8 taps reach only multiples of eight, so they miss the target lag even though the diameter extends past it. Under the same residual backbone and parameter budget, a dense window reaches 54 and copies at more than 99.9% accuracy, while the sparse schedule remains at chance. Replacing just one sparse layer with dense taps also restores copying at the same diameter. The contrast persists when the dense layer is placed first or second. Together, these fixed-diameter controls isolate access to the required lag from the size of the enclosing receptive field.

A support set, rather than a single radius. Let T_ℓ contain the temporal lags admitted by mixer layer ℓ, including lag zero for the residual. For fixed linear time-invariant (LTI) taps, the relevant set is $R = T_1 + \cdots + T_D$, where addition is the Minkowski sum. Pointwise normalization and feed-forward layers introduce no new temporal connections. Consequently, the Jacobian of the full residual stack has support only on R. Combining this structural statement with payload independence

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.

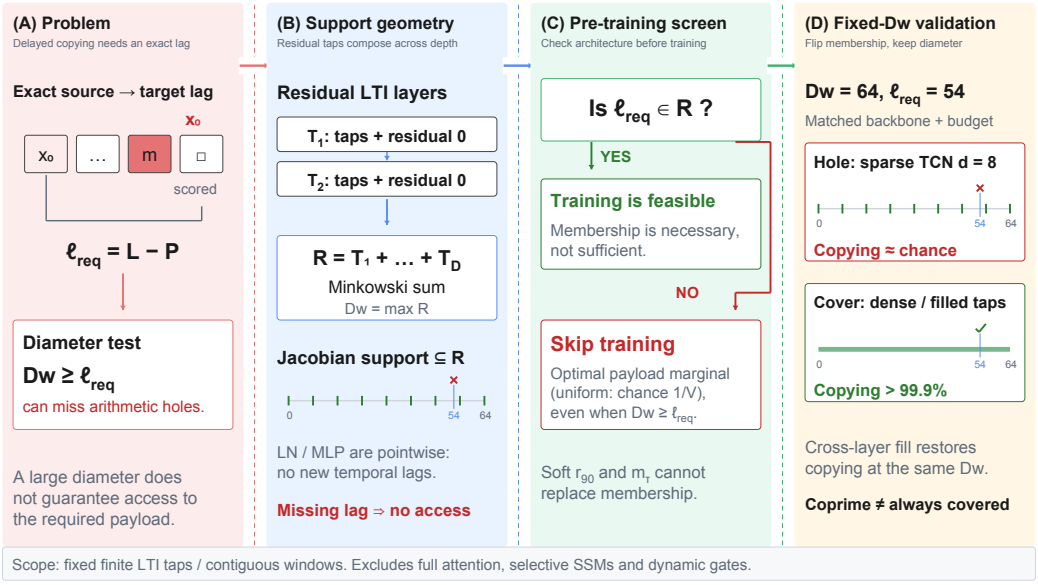


Figure 1: Support-geometry mechanism and pre-training screen. **(A)** Delayed copying requires the exact lag $\ell_{\text{req}} = L - P$; diameter alone can hide arithmetic holes. **(B)** Architectural taps compose as the Minkowski sum $R = T_1 + \cdots + T_D$, which bounds Jacobian support. **(C)** Missing the required lag bounds expected copying accuracy by the optimal payload-marginal baseline; membership is necessary, not sufficient for training success. **(D)** At fixed $Dw = 64$ and $\ell_{\text{req}} = 54$, hole schedules remain at chance while covering and filled schedules succeed in the reported experiments. The hard screen is restricted to fixed finite LTI taps and contiguous windows; soft diagnostics do not replace membership.

yields a task-level exclusion: if $\ell_{\text{req}} \notin R$, expected copying accuracy cannot exceed the optimal payload-marginal baseline. For a uniform payload vocabulary of size V, that baseline is $1/V$. The exclusion holds for every setting of the learned weights, including when $\max R \geq \ell_{\text{req}}$.

Membership is a necessary condition, not a sufficient condition for learning. Its value is that failure of the test is decisive and can be checked from the architecture before training. In contrast, a large Jacobian-energy radius describes where influence accumulates, without certifying access to the exact source position. Our controlled experiments expose this distinction: hole and fill schedules can have the same reported soft reach but opposite copying outcomes. The mechanism and its operational use are summarized in Figure 1; Table 1 separates the information provided by diameter and membership.

Table 1: What the two architectural checks resolve. Passing either check is a necessary condition for copying, not a guarantee of successful training.

Controlled change	Diameter check	Reachable-set check
Sparse versus dense taps at fixed Dw	Same outer bound	Can separate missing and covered target lags
Large soft reach on a sparse grid	Says nothing about holes	Excludes copying if $\ell_{\text{req}} \notin R$
Cross-layer fill at fixed Dw	Outer bound unchanged	Detects restored access to the target lag

Why this comparison matters. Dilation gridding and support composition are established ideas in convolutional modeling (Yu & Koltun, 2016; Wang et al., 2018; Gao et al., 2023). We turn that structure into a falsifiable pre-training statement about a sequence task, and test the statement while holding the usual diameter criterion fixed. Merely increasing a kernel would change both support and diameter. Here the hole-versus-cover, cross-layer fill, and coprime controls distinguish the specific

membership mechanism. A coprime rate pair can still miss 54 because its tap sets are finite; it is the bounded sum, not coprimeness alone, that predicts the observed separation.

Synthetic recall benchmarks such as Zoology and MAD establish useful comparisons among trained mixers (Arora et al., 2024b; Poli et al., 2024). This supplies an architectural check before such comparisons. We use delayed copying because the required lag and the target's independence from other inputs are both explicit. This permits an accuracy bound instead of a correlation between a diagnostic and a benchmark score. We evaluate the same support reasoning on finite LTI mixers and fixed contiguous attention windows; broader mixer and task comparisons remain useful for identifying the boundary between access, optimization, and learned recall.

Contributions. Our contributions are three connected results:

1. A reachable-lag support result for residual LTI stacks and a positionwise copying exclusion under independent payloads, valid even when the required lag is smaller than the support diameter (Section 3.2).
2. Controlled tests at fixed $Dw = 64$ that separate holes from coverage and restore copying through cross-layer fill. Covering and non-covering coprime schedules test the exact membership prediction (Sections 5.1–5.2).
3. A direct pre-training screening procedure that computes support from tap metadata and distinguishes hard exclusions from soft energy diagnostics (Section 3.4).

2 Related work

Receptive fields and dilation geometry. Receptive-field composition is a standard tool for sizing convolutional sequence models: depth, kernel width, and dilation determine the largest accessible lag (Yu & Koltun, 2016). Theoretical support and effective receptive fields answer different questions; the latter measure where influence is concentrated within the architectural support (Luo et al., 2016). Hybrid dilated convolutions address gridding by interleaving sampling patterns (Wang et al., 2018), and RF-Next searches receptive-field configurations (Gao et al., 2023). We use the same path-support composition. Our task-level consequence is: with independent payloads and a known delay, a missing path gives an accuracy bound before training. The fixed-diameter experiments isolate this consequence from receptive-field growth, while the coprime controls distinguish finite-set coverage from a rate-selection heuristic.

Recall in efficient sequence models. Zoology and MAD use synthetic tasks to measure recall and compare mixing mechanisms (Arora et al., 2024b; Poli et al., 2024). BASED studies the recall–throughput tradeoff (Arora et al., 2024a); Taipan combines state-space processing with selective attention for long-context modeling (Nguyen et al., 2024). Convolutional and all-MLP alternatives also make information retention a central design question (Poli et al., 2023; Cui et al., 2024; Badger & Roland, 2026). These studies primarily assess what trained models recover. We make an earlier question explicit: whether the computation graph admits the source-to-output dependency required by a particular copying task. It complements trained comparisons by distinguishing an absent dependency from one that training has failed to exploit.

Heterogeneous mixing and structured operators. Recent heterogeneous mixers motivate the same design question from an adaptive setting. Decision MetaMamba studies heterogeneous selective SSMs for offline reinforcement learning and motivates denser mixing when sparse updates omit important earlier states (Kim et al., 2024). The Latin Square construction studies how placement and composition balance a heterogeneous mixer stack (Kim et al., 2026). These works are complementary: for fixed causal tap sets, the support sum used here is commutative, while the learned nonlinear functions and optimization can still depend on layer placement. Kalman Linear Attention treats sequence processing as parallel Bayesian filtering and state tracking (Shaj et al., 2026); its adaptive state is outside our fixed finite-LTI exclusion. Hydra likewise uses generalized matrix mixers in a bidirectional state-space model (Hwang et al., 2024); bidirectionality changes the causal dependency geometry and is not covered by the delayed-copying result.

Operator structure and learned memory. Structured matrix and dynamical-systems views relate attention, recurrences, and state-space models (Sieber et al., 2024; Dao & Gu, 2024). Effective State-Size measures memory utilization through operator rank (Parnichkun et al., 2025), while mimetic initialization studies how initialization affects recall learning in state-space models (Trockman et al., 2024). Support geometry precedes these questions of capacity and optimization: it identifies which temporal dependencies are structurally absent. A reachable lag need not carry substantial influence at initialization, and a large influence radius need not contain the required lag. We use this distinction to interpret the energy diagnostics and retain the broader family comparisons in Appendix A.

3 Support geometry and the copying screen

3.1 Residual stacks and architectural taps

Let $F^{(D)}$ denote a causal stack of D pre-norm residual blocks. With $Z^{(0)} = X$, block ℓ has the form

$$U^{(\ell)} = Z^{(\ell-1)} + M_\ell(\mathrm{LN}(Z^{(\ell-1)})), \qquad Z^{(\ell)} = U^{(\ell)} + \mathrm{MLP}_\ell(\mathrm{LN}(U^{(\ell)})). \tag{1}$$

LayerNorm and the MLP act independently at each position. Token embeddings, fixed positional encodings, and the pointwise output head also introduce no temporal connections. The mixer M_ℓ determines which input times can affect a given output time.

Tap sets and bandwidth. We use architectural support: lags permitted by the computation, whether or not the corresponding learned weights or realized derivatives are nonzero. Include 0 in every tap set T_ℓ to account for residuals, and write $w_\ell = \max T_\ell$. A causal convolution with kernel size k and dilation d has $T_\ell = \{0, d, \dots, (k-1)d\}$ and $w_\ell = (k-1)d$. A fixed contiguous attention window with maximum lookback w has $T_\ell = \{0, \dots, w\}$. Left padding and causal masking can remove connections at a sequence boundary but cannot create additional lags. Full-context attention has no compact finite-w exclusion in the comparisons considered here.

3.2 Diameter and reachable support

Lemma 1 (Support diameter)**.** *Suppose each mixer is causal and its output at time t depends only on inputs at times s satisfying $0 \le t - s \le w$. If LayerNorm and MLP operations are pointwise in time and residuals add at the same time index, then, for every parameter setting,*

$$\frac{\partial F^{(D)}(X)_t}{\partial X_s} = 0 \quad \textit{whenever} \quad t - s > Dw. \tag{2}$$

With layer-dependent bandwidths, the bound is $\sum_{\ell=1}^{D} w_\ell$.

This outer bound follows by adding the largest lag contributed at each layer. It does not determine which lags within the bound have a path. For that question, the full tap sets must be composed.

Lemma 2 (Reachable lag set)**.** *Under the residual and pointwise assumptions of Lemma 1, suppose mixer ℓ is LTI with fixed architectural tap set T_ℓ. Then the Jacobian support of the stack is contained in*

$$R = T_1 + \cdots + T_D := \{a_1 + \cdots + a_D : a_\ell \in T_\ell\}, \qquad \frac{\partial F^{(D)}(X)_t}{\partial X_s} = 0 \textit{ if } t - s \notin R. \tag{3}$$

Why the sum is the right object. Each temporal path through a mixer contributes one lag from that layer's set. Composition adds these lags; pointwise operations preserve them, and residual paths contribute zero. Induction over depth therefore gives R. An absent path means that changing only X_s cannot change the output at t, for any inputs or weights. This is stronger than a zero derivative at one probe: cancellation or saturation may hide an existing path. Proofs and a claim–evidence map appear in Appendix C.

For D identical dilated convolutions, $R = \{0, d, \dots, D(k-1)d\}$. The diameter is $D(k-1)d$, yet most lags inside it may be missing. Contiguous taps instead give the full integer interval $R = \{0, \dots, \sum_\ell w_\ell\}$. For finite nonempty sets, $\max R = \sum_\ell \max T_\ell$; we use Dw for this support diameter in the matched experiments, including heterogeneous schedules.

Fixed windows and scope. Windowed softmax attention is not an LTI operator. Nevertheless, its fixed mask bounds every temporal dependency by the same contiguous tap set, so the path-support proof applies with $T_\ell = \{0, \dots, w_\ell\}$. The hard screen in this paper is restricted to finite LTI taps and such fixed contiguous windows. We do not apply the LTI statement to selective SSMs, content-dependent gates, or dynamic masks by substituting a tap set measured on one input. Their empirical diagnostic values are reported separately.

3.3 An accuracy bound for delayed copying

For a length-L sequence, set $P = \max(2, \lfloor L/6 \rfloor)$ and $m = L - P - 1$. Independent payload tokens $x_0, \dots, x_{P-1}$ occupy the prefix; a marker occupies position m and all remaining inputs are blank. At each scored position $t = m + 1 + j$, the target is x_j. Thus

$$\ell_{\mathrm{req}} = t - j = L - P. \tag{4}$$

The recall inputs remain blank: earlier targets are not fed back to the model. Let p_{pay} be the payload distribution and V its support size. This V counts payload symbols, excluding reserved special tokens (Appendix D.1).

Corollary 1 (Copying exclusion)**.** *Under the assumptions of Lemma 2 and the independent payload generator above, if $\ell_{\mathrm{req}} \notin R$, then, for any fixed learned parameters and each scored position $t = m + 1 + j$,*

$$\Pr\{\widehat{x}_t = x_j\} \le \max_c p_{\mathrm{pay}}(c). \tag{5}$$

The bound is $1/V$ for a uniform payload. The Bayes-optimal predictive distribution given the model's accessible inputs is p_{pay}, even when $\max R \ge \ell_{\mathrm{req}}$.

The output at t cannot depend on x_j when its lag is absent. All remaining inputs are independent of that target, so conditioning on anything the model can compute leaves its distribution unchanged. The bound concerns expected accuracy on fresh sequences, conditional on learned parameters; it neither requires the model to output the optimal marginal nor prevents finite-sample fluctuations around chance. For multiple required lags, apply the bound at each position whose lag is missing. Coverage of all required lags is necessary for perfect recall.

3.4 Computing and using the screen

Start with $R_0 = \{0\}$ and compute $R_\ell = \{r + a : r \in R_{\ell-1},\ a \in T_\ell\}$ through the stack. This finite-set recurrence needs only the tap metadata. Parallel branches contribute the union of their supports before composition; residuals contribute zero. The resulting test is $\ell_{\mathrm{req}} \in R_D$. A failed test rules out above-baseline copying; a passed test leaves learning performance unresolved. The supplement provides a support-set export utility for heterogeneous convolutional stacks.

Diagnostic lag and soft energy. The earlier screening protocols use $\tau(L) = L - P - 1$, one step smaller than the actual copying lag. We retain τ when describing those protocols; $Dw < \tau$ is sufficient for failure, but $Dw \ge \tau$ is not the exact membership test. For a probed operator g, influence energy is $E_{ts}(X) = \|\partial g(X)_t / \partial X_s\|_F^2$. The normalized reach r_{90} summarizes the nearest-first lag enclosing 90% of causal energy; m_τ summarizes energy in a narrow window around diagnostic lag τ. These quantities describe realized influence and depend on probe length, inputs, and weights. Their formal definitions and inactive-row handling are in Appendix B. Neither quantity establishes the structural membership used in Corollary 1.

4 Experimental design

We organize the experiments around three questions: whether insufficient diameter excludes copying, whether reachable-set membership distinguishes models at the same diameter, and whether changing taps across layers repairs a missing lag. The fixed-diameter suites provide the central comparison. Supplementary analyses report broader task and diagnostic comparisons under their own protocols.

Shared backbone and training. Mixers occupy the same two-block pre-norm residual model, with token embeddings, sinusoidal positions, a pointwise SwiGLU feed-forward block, and a tied output projection. Hidden width is selected for a target of one million parameters within 5%; consequently, width may differ between mixers. Both geometry suites use 2500 training steps, batch size 32, learning rate 3×10^{-4}, weight decay 0.01, and gradient clipping at 1.0. Evaluation averages 16 batches of 64 sequences. AdamW, the data generator, training length, and evaluation rule are shared. Parameter matching controls overall size while leaving FLOPs and optimization dynamics unconstrained.

Fixed-diameter geometry. At $L = 64$, the payload has length $P = 10$, so every scored prediction needs $\ell_{\mathrm{req}} = 54$. We hold $Dw = 64$ and compare five tap patterns: two sparse schedules missing 54, an even-strided schedule covering 54, a dense convolution, and contiguous windowed attention. Every model passes the diameter requirement, leaving membership in R as the proposed separator. Runs use seeds $\{11, 29\}$.

Cross-layer repair and coprime controls. A second suite retains $Dw = 64$ and the same task while varying the taps of the two layers separately. It compares the uniform dilation-8 hole with two orders of sparse-plus-dense fill, a covering coprime pair, and a non-covering coprime pair. Both layer orders matter as an experimental control: Minkowski addition predicts the same support in either order, although the resulting nonlinear networks need not have the same learned function. Seeds, budget tolerance, and training settings match the first suite.

Matched-length screening. The length-128 study trains and evaluates 29 mixer configurations on copying, using seeds $\{11, 29, 47\}$ and the same 2500-step budget. Eight configurations belong to the diagnostic calibration suite; the remaining 21 form the confirmation split. Here $\ell_{\mathrm{req}} = 107$ and the archived diagnostic lag is $\tau = 106$. We retain the original rules $Dw \geq \tau$, $r_{90} \geq 0.5$, and $m_\tau \geq \theta$ when reporting that study. Only finite-bandwidth configurations have a hard diameter exclusion; full-context and recurrent rows provide empirical context. The distinction between τ and ℓ_{req} is maintained throughout.

Soft diagnostics and reporting. The mass threshold $\theta = 5.423 \times 10^{-2}$ was frozen as the minimum initial mass among calibration mixers that later copied, using length-32 probes and length-64 copying labels. Calibration therefore has no false negatives by construction. The confirmation values, probe lengths, and estimator settings are reported in Appendices E.1–E.3. In the geometry tables, accuracy and r_{90} are means over the two reported seeds; the latter uses length-32 probes and is not a direct measurement of lag 54. The near-chance and near-perfect outcomes agree across the reported seeds. Saved configurations and result files retain failed runs as well as successful ones, and implementation checks cover shapes, finite outputs, and causality.

5 Results

5.1 Membership separates outcomes at fixed diameter

At $Dw = 64$ and $\ell_{\mathrm{req}} = 54$, every architecture in Table 2 passes the diameter test. The reachable-set test then separates their copying outcomes. Dilation-8 and dilation-4 schedules miss 54 and achieve 0.7% and 0.8% mean accuracy, respectively. The even-strided dilation-2 schedule, dense convolution, and windowed attention all contain 54 and exceed 99.9% before rounding. Membership separates the outcomes across all five configurations and both reported seeds. Figure 2 shows the corresponding support patterns.

The even-strided control is informative: its support is not dense, yet it solves the task because the required lag is even. The observed split therefore follows membership of the target lag, not support density. Conversely, $r_{90} = 0.73$ on both failing sparse schedules is already large on the normalized scale, so a broad soft receptive field does not establish access to the source token.

5.2 Cross-layer fill restores access without enlarging diameter

Table 3 changes the layerwise taps while retaining $Dw = 64$. A uniform pair of dilation-8 layers has $R = \{0, 8, \ldots, 64\}$ and remains at chance. Replacing either layer with a dense kernel-33 convolution

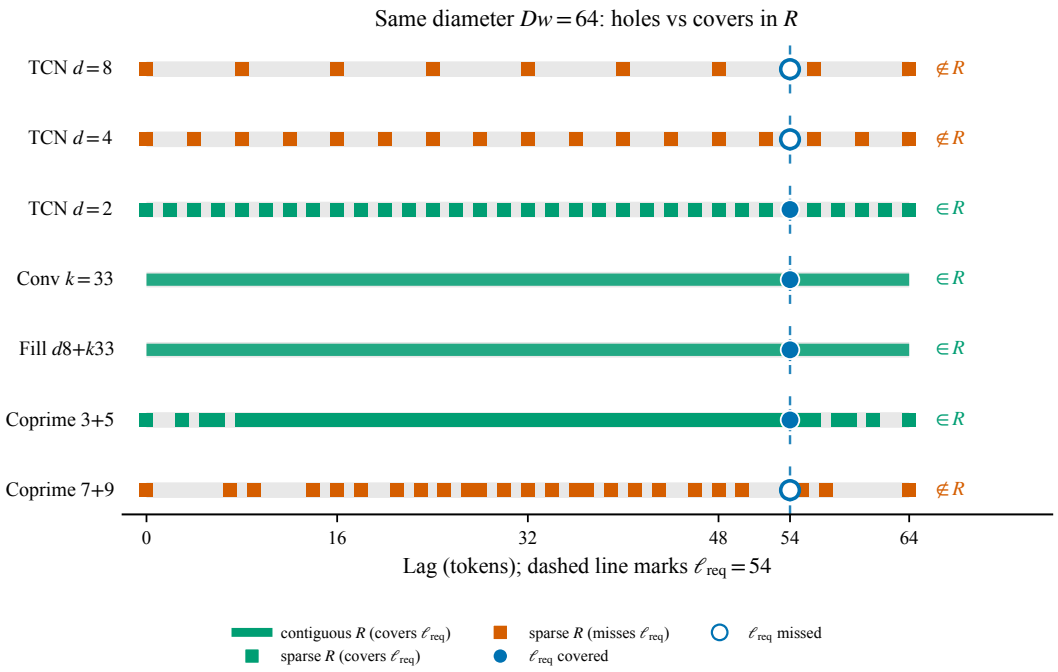


Figure 2: Reachable lag sets at fixed Dw=64. Dense green bars mark covering R; orange markers mark sparse grids with holes. Blue markers show whether $\ell_{\text{req}} = 54$ is covered. Heterogeneous fill and covering schedules restore membership without enlarging diameter.

yields coverage of 54 and near-perfect copying. Because each layer still has maximum lag 32, this repair changes the interior support without changing the diameter. Commutativity of the Minkowski sum gives both orders the same support; both also succeed in the reported runs.

This comparison also isolates support from measured influence. The uniform hole and sparse-first fill both have reported $r_{90} = 0.73$, but their copying accuracies are 0.7% and 100% after rounding. Reversing the fill order changes reach to 0.91 while retaining the same support and successful copying. Reach therefore distinguishes aspects of the realized operator that are not decisive for this exclusion test.

Coprimeness is not the criterion. For the coprime pair $(k, d) = (9, 3), (9, 5)$, the decomposition $54 = 3 \cdot 8 + 5 \cdot 6$ respects both tap ranges. For $(k, d) = (5, 7), (5, 9)$, no $a, b \in \{0, \ldots, 4\}$ solve $7a + 9b = 54$. Both pairs have diameter 64, yet only the first covers the target and copies. The finite coefficient bounds explain why coprimeness alone is insufficient.

Table 2: Fixed support diameter Dw=64. Reachable set R is the Minkowski sum of per-layer taps (Lemma 2). Copying reports rounded means over seeds $\{11, 29\}$.

Mixer	R (sketch)	$54 \in R$?	r_{90}	Copying
TCN k=5, d=8	$8\mathbb{Z} \cap [0, 64]$	**no**	0.73	0.7%
TCN k=9, d=4	$4\mathbb{Z} \cap [0, 64]$	**no**	0.73	0.8%
TCN k=17, d=2	$2\mathbb{Z} \cap [0, 64]$	**yes**	0.86	100%
Conv. k=33, d=1	$\{0, \ldots, 64\}$	**yes**	0.91	100%
Attn. window 32	$\{0, \ldots, 64\}$	**yes**	0.66	99.9%

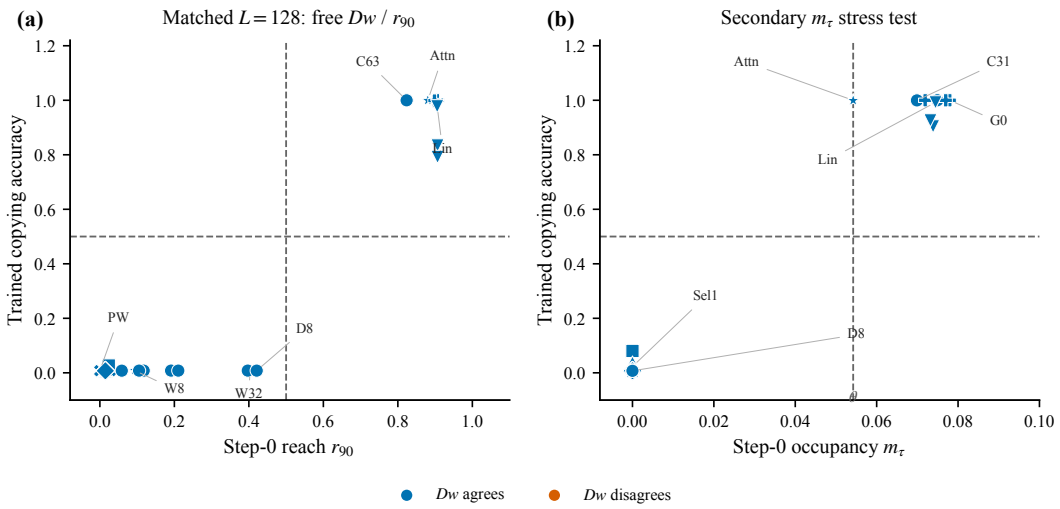


Figure 3: Initialization diagnostics and trained copying accuracy. (a) Matched length 128. (b) Diagnostic mass comparison. The 21 confirmation configurations comprise 7 finite-bandwidth, 5 full-context, and 9 undefined-bandwidth cases. Colors code the last two groups as positive and negative, respectively; only the finite-bandwidth subset is evaluated within the hard-support scope.

5.3 Diameter screening and soft diagnostics

At matched length 128, window-32 attention and kernel-31 convolution have $Dw = 64$ and 60, below the required lag 107, and remain at chance. Kernel-63 convolution has diameter 124 and copies successfully. The stored confirmation results agree with the diameter rule on all seven finite-bandwidth convolution/window configurations. The broader archived 21-configuration classifier also records complete agreement, but includes out-of-scope rows coded by convention; Figure 3 retains that original visualization. The fixed-diameter tests above provide the direct evidence for the finer reachable-set mechanism.

The soft diagnostics expose two distinct failure modes. On the length-32 probe / length-64 copying transfer grid, $r_{90} \geq 0.5$ agrees with 90.5% of confirmation labels, including false positives on intermediate or dilated kernels. On matched length 128, $m_\tau \geq \theta$ agrees with 95.2% but rejects kernel-63 convolution, which trains successfully. Its diagnostic mass is zero at every probed seed even though the architectural diameter exceeds the required lag. The recorded profiles locate influence at shorter lags (Appendix E.3). A soft energy measurement can therefore miss a usable path just as a broad energy radius can obscure a structural hole.

5.4 Broader recall and task comparisons

Table 4 gives the existing length-64 readout across mixer families. Long convolution and linear attention copy successfully; local controls remain near chance. Under the repaired MQAR protocol, softmax attention reaches 99.8%, whereas linear attention and long convolution remain near 27%.

Table 3: Same Dw=64, heterogeneous layer taps (seeds $\{11, 29\}$). Rounded seed means. Cross-layer fill and a covering coprime schedule restore copying; a non-covering coprime schedule does not.

Mixer taps	$54 \in R$?	r_{90}	Copying
$(k{=}5, d{=}8) + (k{=}5, d{=}8)$	**no**	0.73	0.7%
$(k{=}5, d{=}8) + (k{=}33, d{=}1)$	**yes**	0.73	100%
$(k{=}33, d{=}1) + (k{=}5, d{=}8)$	**yes**	0.91	100%
$(k{=}9, d{=}3) + (k{=}9, d{=}5)$	**yes**	0.77	100%
$(k{=}5, d{=}7) + (k{=}5, d{=}9)$	**no**	0.64	0.7%

Table 4: Length-64 readout after the adaptivity repair (Jacobian length 32). Ungated long convolution is the LTI control. MQAR is protocol-sensitive.

Mixer	Reach r_{90}	Adaptivity	Copying	MQAR
Softmax attention	0.76	0.41	**100.0%**	**99.8%**
Linear attention	0.91	0.05	**99.9%**	27.5%
Ungated long conv.	0.89	0.00	**100.0%**	26.9%
Gated long conv.	0.89	0.07	**100.0%**	26.8%
Selective SSM	0.07	0.17	2.2%	25.2%
Short conv. (k=5)	0.22	0.00	0.7%	14.7%
Position-wise	0.00	0.00	0.7%	1.7%

The same attention configuration reaches about 28% under the unrepaired protocol. These comparisons distinguish fixed-delay copying from associative retrieval and make the protocol part of the interpretation (Appendix F.3).

The appendix reports training dynamics, intervention tests, family-held-out prediction, selective-SSM tuning, and matched-budget enwik8; reach does not reliably rank enwik8 loss (Appendix H). Geometry is therefore used to certify absent task dependencies, not to predict all learned performance.

6 Discussion

Architectural access precedes learned use. The fixed-diameter comparisons distinguish missing access from learned performance when a path exists. Cross-layer fill restores access without enlarging diameter. Matched parameters do not equalize FLOPs or optimization, which affect learned use but cannot create an excluded path.

Designing for the required lags. Compute R before training, including residual lag zero: perfect recall requires every scored lag to be covered. Dense windows ensure coverage, but even-strided taps show that full density is unnecessary. Coprime rates still require a bounded-sum check; shifted or boundary-sensitive masks also require valid source indices (Appendix C.4).

Connecting support to adaptive memory. Infini-attention's compressive memory (Munkhdalai et al., 2024) and Taipan's selective attention (Nguyen et al., 2024) add pathways beyond local windows. Memory reads, writes, and routing separate path existence from retention and learned use; our screen addresses only the first under fixed support.

When visible context predicts the target. With correlated payloads, visible context may predict the target despite a missing source path. The ceiling becomes conditional Bayes accuracy rather than the marginal; independent copying makes them coincide (Appendix C.4).

The geometry suites use depth two, length 64, and two seeds; the broader screen reaches length 128. They do not establish success across depths, delays, or budgets. The proof holds at arbitrary depth, but success under coverage remains empirical. Selective SSMs, dynamic retrieval, and enwik8 require separate interpretations.

7 Conclusion

For delayed copying, the decisive structural question is whether the required lag is reachable. Diameter gives an outer bound; the Minkowski sum of layerwise tap sets identifies holes inside it. If the required lag is absent, expected accuracy on independent payloads cannot exceed the payload-marginal baseline, regardless of learned weights. Our fixed-diameter comparisons show the corresponding split: missing-lag schedules remain at chance, while covering schedules copy successfully. Cross-layer fill restores access without increasing the maximum lookback. The resulting support test can exclude configurations missing the required path before training, but coverage alone cannot guarantee learning; realized influence, retention, and optimization remain distinct questions.

AI USE STATEMENT

In this work, generative AI tools were utilized for grammatical polishing, sentence structure optimization, and academic expression refinement of the manuscript text, and the experimental code of this study was generated with the assistance of Cursor AI. No generative AI tools were employed for the design of research schemes, conduction of core experiments, analysis of original experimental data, and formulation of research conclusions. All figures and illustrations in this manuscript were manually created by the authors using PowerPoint without any generative AI involvement, which effectively avoids AI hallucination problems in visual content production. All other undisclosed AI-assisted tasks are not applicable to this study. Furthermore, large language models (LLMs) were adopted to assist in sorting out literature frameworks and optimizing the logical fluency of manuscript paragraphs. All AI-generated and AI-assisted contents in this work, including the Cursor AI-generated experimental code and polished manuscript text, have been fully reviewed, verified, and revised by multiple authors. Specifically, the authors manually checked all AI-polished text contents to eliminate logical errors, semantic deviations and potential plagiarism risks; the AI-generated experimental code has been jointly examined, tested and confirmed to be completely correct by multiple authors to ensure the accuracy and reproducibility of experimental results; all AI-optimized manuscript frameworks and narrative structures were repeatedly verified and adjusted according to the actual research content and experimental results of this study. The authors take full and ultimate responsibility for all contents of this manuscript, including all texts, research claims, experimental descriptions, code implementations, academic conclusions and visual figures produced or optimized with the assistance of generative AI tools or manually created by the authors.

ETHICS STATEMENT

This study concerns controlled sequence-model experiments using synthetic recall tasks and an existing text benchmark. It does not involve human participants or deploy a system for consequential decisions. The copying bound is task-specific; extrapolating it to practical reliability or language-model safety would require evidence beyond these experiments.

REPRODUCIBILITY STATEMENT

The main paper specifies the model family, task protocol, matched-budget comparison, and evaluation measures used in the study, while the appendix provides the assumptions, complete support arguments, data-generation description, notation, and diagnostic definitions needed to interpret the results. The supplementary materials contain the anonymous, self-contained experiment code archive, including the implementation, configuration files, data-generation and support-export utilities, recorded seeded results, and verification tests. Together, these materials provide the source code, formal details, and execution artifacts needed to reproduce the reported analyses without relying on undisclosed components.

REFERENCES

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*, 2024a.

Simran Arora et al. Zoology: Measuring and improving recall in efficient language models. *arXiv preprint arXiv:2312.04927*, 2024b.

Benjamin L. Badger and Ethan Roland. Toeplitz MLP mixers are low complexity, information-rich sequence models. *arXiv preprint arXiv:2605.06683*, 2026.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Chenwei Cui, Zehao Yan, Gedeon Muhawenayo, and Hannah Kerner. An all-MLP sequence modeling architecture that excels at copying. *arXiv preprint arXiv:2406.16168*, 2024.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

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

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, 2022.

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.

Daniel Y. Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W. Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. Monarch mixer: A simple sub-quadratic GEMM-based architecture. In *Advances in Neural Information Processing Systems*, 2023a.

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*, 2023b.

Shanghua Gao, Zhong-Yu Li, Qi Han, Ming-Ming Cheng, and Liang Wang. RF-next: Efficient receptive field search for convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2984–3002, 2023.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, 2017.

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

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

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

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

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, 2020.

Sukjun Hwang, Aakash Lahoti, Ratish Puduppully, Tri Dao, and Albert Gu. Hydra: Bidirectional state space models through generalized matrix mixers. In *Advances in Neural Information Processing Systems*, 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*, 2020.

T. Kim, Y. Hong, M. Kim, S. Choi, J. Jang, et al. Placement is free, composition is not: The latin square as a provably-balanced construction for heterogeneous sequence-mixer stacks. *arXiv preprint*, 2026.

W. Kim, C. Song, and H. Kim. Decision metamamba: Enhancing selective ssm in offline rl with heterogeneous sequence mixing. *arXiv preprint arXiv:2408.10517*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Kevin Y. Li, Asher Trockman, Ananda Theertha Suresh, and Ziteng Sun. Multi-mixer models: Flexible sequence modeling with shared representations. *arXiv preprint arXiv:2605.28769*, 2026.

Opher Lieber et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Toni J. B. Liu, Baran Zadeoğlu, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. Jacobian scopes: Token-level causal attributions in LLMs. *arXiv preprint arXiv:2601.16407*, 2026.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2016.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Jun He, Liang Gui, and Graham Neubig. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022.

Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZN4mXk8D2E.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Chien Van Nguyen, Huy Huu Nguyen, Thang Pham, Ruiyi Zhang, Hanieh Deilamsalehy, Puneet Mathur, Ryan A. Rossi, Trung Bui, Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. Taipan: Efficient and expressive state space language models with selective attention. *arXiv preprint arXiv:2410.18572*, 2024.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gülçehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, volume 202, pp. 26670–26698. PMLR, 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

TaekHyun Park, Yongjae Lee, Daesan Park, Dohee Kim, and Hyerim Bae. JustDense: Just using dense instead of sequence mixer for time series analysis. *arXiv preprint arXiv:2508.09153*, 2025.

Rom Parnichkun, Neehal Tumma, Armin W. Thomas, Alessandro Moro, Qi An, Taiji Suzuki, Atsushi Yamashita, Michael Poli, and Stefano Massaroli. Quantifying memory utilization with effective state-size. In *International Conference on Machine Learning*, 2025.

Bo Peng, Eric Alcaide, Quentin Anthony, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

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*, volume 202, pp. 28043–28078. PMLR, 2023. URL https://proceedings.mlr.press/v202/poli23a.html.

Michael Poli et al. Mechanistic design and scaling of hybrid architectures. In *International Conference on Machine Learning*, 2024.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. doi: 10.1162/tacl_a_00353. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00353/96478.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, 2021.

V. Shaj, C. Barker, A. Scannell, A. Szecsenyi, et al. Kalman linear attention: Parallel bayesian filtering for efficient language modelling and state tracking. *arXiv preprint*, 2026.

Jerome Sieber, Carmen Amo Alonso, Alexandre Didier, Melanie N. Zeilinger, and Antonio Orvieto. Understanding the differences in foundation models: Attention, state space models, and recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4275. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f271a36160097fbdb06a9adeb1605343-Paper-Conference.pdf.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Shuxin Xia, Jilong Xue, et al. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Asher Trockman, Hrayr Harutyunyan, J. Zico Kolter, Sanjiv Kumar, and Srinadh Bhojanapalli. Mimetic initialization helps state space models learn to recall. *arXiv preprint arXiv:2410.11135*, 2024.

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

Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, 2018.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañon, Philip Pham, Anirudh Ravula, Wenhui Wang, Li Yang, and Sebastian Riedel. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

CONTENTS OF THE APPENDIX

The appendix separates background, formal support arguments, diagnostic protocols, additional results, and reproducibility materials. Page numbers refer to the first page of each section.

No.	Section	What it contains	Page
A	Extended related work	Long-context architectures, recurrent and state-space models, retrieval, heterogeneous mixers, path-support, and dilation context.	15
B	Notation and diagnostics	Symbols, support diameter, reachable sets, normalized reach, and lag-mass definitions.	17
C	Support proofs	Proofs of the diameter lemma, reachable-set lemma, and copying exclusion.	17
D	Task protocols and reporting	Copying generator, preregistration, result ledger, and matched-length details.	20
E	Screening analyses	Threshold calibration, confirmation split, sensitivity checks, and intervention records.	21
F	Recall and model comparisons	MQAR, selective-SSM tuning, family-held-out checks, and related task diagnostics.	23
G	Training dynamics and robustness	Copying curves, probe sensitivity, and robustness summaries.	25
H	Matched-budget enwik8	Language-model comparison and the limits of reach as a loss predictor.	25
I	Jacobian probe cost	Exact versus randomized Jacobian estimation time and memory.	26
J	Recovered symmetry algebra	Toeplitz mass, closure residual, and recovered algebra diagnostics.	26
K	Reproducibility artifacts	Anonymous code archive, stored evidence, verification commands, and scope.	27

A Extended Related Work

This appendix broadens the positioning around support geometry without changing the theorem's scope. The relevant distinction is between a model's architectural ancestors, the information carried along those ancestors, and the optimization needed to use them.

Attention and efficient long-context architectures. The Transformer introduced content-dependent all-to-all attention (Vaswani et al., 2017); its full-context dependency makes a finite missing-lag certificate vacuous unless a restricted mask is imposed. Efficient attention variants replace or approximate this pattern in several ways. Linear attention expresses attention as a recurrent feature state (Katharopoulos et al., 2020); Performers approximate softmax attention with random features (Choromanski et al., 2020); and Longformer and BigBird combine local and selected long-range patterns (Beltagy et al., 2020; Zaheer et al., 2020). These mechanisms motivate careful accounting of the actual control and read paths: a nominal window is not the complete support when a global summary or selected edge is present. Our fixed-window analysis applies only when the mask and its taps are fixed and known.

State-space and recurrent memory. HiPPO constructs recurrent memories through polynomial projections (Gu et al., 2020); S4 makes structured state-space computation practical for long sequences (Gu et al., 2021); and simplified S5 layers reduce the parameterization of state-space models (Smith et al., 2022). Resurrecting recurrent networks revisits stable recurrent dynamics for long sequences (Orvieto et al., 2023). These models can transport information through a state rather than expose a direct finite tap. The present result does not treat state capacity or stability as a support certificate: a state update may have a path to the output while still losing information, and a reachable direct lag may still be difficult to learn.

Retention, gating, and moving averages. Mega augments gated attention with moving-average memory (Ma et al., 2022), while RetNet presents retention mechanisms with parallel and recurrent computation modes (Sun et al., 2023). Such designs help separate the questions of access, retention, and utilization that motivate our Discussion. They are not counterexamples to the fixed-support theorem; they lie outside it whenever their gates or states depend on content or carry information through an unmodeled recurrent pathway.

Heterogeneous and adaptive sequence mixing. Decision MetaMamba explicitly studies heterogeneous selective SSM mixing in offline RL (Kim et al., 2024). Its motivation for dense mixing after important steps are omitted is closely related to our practical concern, but its learned selective setting is outside the fixed-support exclusion. The Latin Square construction addresses balanced composition and placement in heterogeneous sequence-mixer stacks (Kim et al., 2026). Our Minkowski sum is commutative for fixed causal tap sets, so layer order does not change structural membership; trained nonlinear functions and optimization can nevertheless change with order, as in Table 3.

Kalman Linear Attention frames efficient language modeling and state tracking as parallel Bayesian filtering (Shaj et al., 2026). Hydra uses bidirectional state-space models through generalized matrix mixers (Hwang et al., 2024); both illustrate why adaptive or non-causal operators require a different dependency analysis.

Sequence convolutions and sparse access. Convolutional sequence-to-sequence models established learned temporal kernels as an alternative to recurrent encoders (Gehring et al., 2017), while quasi-recurrent convolutional encoders combine local mixing with recurrent pooling. Sparse Transformers and Routing Transformers instead select a subset of attention edges, through fixed patterns or content-based routing (Child et al., 2019; Roy et al., 2021). These designs reinforce the need to distinguish an enclosing context size from the actual paths available to a target; the present theorem covers only fixed causal supports.

Long-range benchmarks and external memory. Long Range Arena evaluates efficient sequence models across tasks with different dependency patterns (Tay et al., 2021). Transformer-XL extends context through a recurrent segment-level memory (Dai et al., 2019), while Compressive Transformers retain compressed older states (Rae et al., 2020). Memorizing Transformers, REALM, and retrieval-

augmented generation add external retrieval or nonlocal memory (Wu et al., 2022; Guu et al., 2020; Lewis et al., 2020). These systems make long-range information available through memory or retrieval operations, so their dependency analysis must include writes, reads, and retrieval controls; a local tap-set sum alone is insufficient.

Additional state-space and recurrent views. Diagonal state-space constructions show that structured long-range dynamics can be represented with simpler parameterizations (Gupta et al., 2022; Gu et al., 2022), while gated state-space models explore adaptive or gated state updates (Mehta et al., 2023). H3 and RWKV provide further state-space or recurrent alternatives for language modeling (Fu et al., 2023b; Peng et al., 2023); Griffin combines gated linear recurrences with local attention (De et al., 2024). Linear Transformers are also related to fast-weight updates rather than fixed finite taps (Schlag et al., 2021). These methods are useful comparators for the boundary of the current result: a state or gate can preserve information across time without exposing a direct fixed-lag path, and can also introduce input-dependent dependencies that require a different support envelope.

Memory-efficient implementation. FlashAttention reduces the memory traffic of exact attention without changing the mathematical attention dependencies (Dao et al., 2022). This distinction is useful here: implementation efficiency and support geometry are separate axes. A kernel optimization can change wall-clock cost while leaving the reachable set unchanged.

Positioning. Across these families, a large context window, a recurrent state, a global summary, or a structured operator can provide useful long-range behavior without sharing the fixed finite tap assumptions. The contribution of this paper is narrower and testable: under fixed causal support and independent payloads, membership of the required lag in the reachable set is a training-free necessary condition. The broader literature supplies the architectural cases where one should instead enumerate state, control, and memory-read dependencies or use empirical diagnostics.

Architecture comparisons and hybrids. Zoology and MAD compare mixers after training on synthetic memory tasks (Arora et al., 2024b; Poli et al., 2024). BASED studies tradeoffs between recall and throughput (Arora et al., 2024a); Taipan couples selective state spaces with sparse attention for long-context efficiency (Nguyen et al., 2024). Operator taxonomies and hybrids continue to proliferate (Sieber et al., 2024; Dao & Gu, 2024; Poli et al., 2023; Gu & Dao, 2023; Lieber et al., 2024; Fu et al., 2023a; Cui et al., 2024). Toeplitz MLP mixers emphasize information retention and copying (Badger & Roland, 2026). Multi-mixer / sequence-axis hybrids switch mixing mechanisms within a sequence (Li et al., 2026), so layer-wise bandwidths should not be collapsed into one global w without care. JustDense asks when explicit sequence mixers are even needed versus dense substitutes in time series (Park et al., 2025).

Receptive fields, path support, and dilation design. Classical CNN/TCN work emphasizes theoretical receptive-field growth with depth, kernel size, and dilation (Yu & Koltun, 2016), and distinguishes theoretical from effective receptive fields (Luo et al., 2016). Composing causal LTI layers propagates support along paths whose lag sets Minkowski-add—the same algebra behind R. Hybrid dilated convolution designs choose rates so sparse sampling grids interleave and avoid gridding holes (Wang et al., 2018; Yu & Koltun, 2016); RF-Next searches dilation/pooling schedules to reduce blind spots (Gao et al., 2023). Relative to that literature, we do not search architectures: we give a free membership screen for delayed copying under residual LTI stacks, and show with same-Dw flips that diameter coverage is insufficient once holes appear. An RF-search schedule can use membership to rule out unsupported configurations before training. Coverage permits copying but does not guarantee it, so search and screening address different stages of design. Coprimeness helps only when the required lag lands in R. Effective State-Size links operator rank to recurrent state (Parnichkun et al., 2025); mimetic initialization can decide whether SSMs learn recall (Trockman et al., 2024)—orthogonal to the LTI screen. Jacobian Scopes attribute LLM predictions to tokens (Liu et al., 2026). We use operator-level Jacobians only as soft diagnostics alongside the exact geometry tests.

Positioning. The central comparison holds diameter fixed while changing whether a required lag is reachable. This connects classical path-support algebra to an expected-accuracy exclusion and tests that connection with controlled copying experiments. Broader operator coordinates describe

additional properties, but their predictive checks in this study do not establish a general architecture-ranking rule.

B NOTATION AND DIAGNOSTIC DEFINITIONS

Table 5: Notation used in the support argument and diagnostics.

Symbol	Meaning
$F^{(D)}$	Depth-D pre-norm residual stack
w, w_ℓ	Per-layer mixer bandwidth (max lag)
Dw / $\sum_\ell w_\ell$	Support diameter
T_ℓ	LTI tap set of layer ℓ
$R = T_1 + \cdots + T_D$	Reachable lag set (Minkowski sum)
ℓ_{req}	Required lookback on every scored token
$\tau(L)$	Diagnostic lag $L - P - 1$ for screens / m_τ
E_{ts}	Soft influence energy $\|\partial g_t / \partial X_s\|_F^2$
r_{90}	Median nearest-first 90% energy lag (normalized)
m_τ	Soft lag-mass heuristic at diagnostic τ

Reach r_{90}. For each active row t, let $e_{t,k} = E_{t,t-k}$. Define $C_t(q) = (\sum_{k=0}^{q} e_{t,k})/(\sum_{k=0}^{t} e_{t,k})$ and $q_t^\star = \min\{q : C_t(q) \geq 0.9\}$. Then $r_{90}(t) = q_t^\star / \max(1, t)$, and the reported r_{90} is the median over active rows and probes. Under Lemma 1, $q_t^\star \leq Dw$.

Diagnostic mass m_τ. For each active row with $t \geq \tau$, sum the causal energy at lags satisfying $|k - \tau| \leq \delta$, divide by the total causal energy in that row, and take the median across eligible rows. Probe-level summaries are averaged across samples. The default half-width is $\delta = \max(1, \lfloor L_{\mathrm{probe}}/32 \rfloor)$. Rows with total causal energy at most 10^{-12} are skipped; a probe with no eligible active rows returns zero. The normalized reach is also computed per probe and then averaged across probe samples.

C PROOFS OF SUPPORT LEMMAS

C.1 PROOF OF LEMMA 1 (SUPPORT DIAMETER)

Proof. Proceed by induction on depth D.

Base case $D = 1$. Write one residual block (up to the concrete pre-norm wiring in our implementation) as

$$U = X + \mathrm{mixer}(\mathrm{LN}(X)), \qquad F^{(1)}(X) = U + \mathrm{MLP}(\mathrm{LN}(U)).$$

LayerNorm and the MLP act diagonally in time, so they do not enlarge temporal support. By assumption (i), the mixer contributes only lags in $\{0, \ldots, w\}$. By assumption (iii), residuals add at equal times and therefore contribute lag 0. Hence $\partial F^{(1)}(X)_t / \partial X_s = 0$ whenever $t - s > w$.

Inductive step. Assume the claim for $F^{(D-1)}$, i.e. support diameter at most $(D-1)w$. Write $F^{(D)} = f^{(D)} \circ F^{(D-1)}$ where $f^{(D)}$ is one residual block of bandwidth w, and set $Z = F^{(D-1)}(X)$. The chain rule gives

$$\frac{\partial F_t^{(D)}}{\partial X_s} = \sum_u \frac{\partial f_t^{(D)}}{\partial Z_u} \frac{\partial F_u^{(D-1)}}{\partial X_s}.$$

The first factor vanishes unless $t - u \in \{0, \ldots, w\}$; the second vanishes unless $u - s \leq (D-1)w$. Any nonzero term therefore requires $t - s \leq Dw$.

Heterogeneous bandwidths. If layer ℓ has bandwidth w_ℓ, the same induction yields diameter $\sum_{\ell=1}^{D} w_\ell$ in place of Dw. □

Remark 1 (What the diameter does not claim)**.** *Lemma 1 is an exclusion outside* $[0, Dw]$*. It does not assert that every lag inside* $[0, Dw]$ *is attainable. That gap is exactly what Lemma 2 closes under LTI taps.*

C.2 Proof of Lemma 2 (reachable lag set)

Proof. Again induct on D.

Base case $D = 1$**.** By definition of the LTI tap set T_1, the mixer Jacobian can be nonzero only for lags in T_1. Pointwise LayerNorm/MLP layers are diagonal in time, and residuals contribute lag 0. Under the standing residual wiring we take $0 \in T_1$ (equivalently, absorb the residual into the tap set). Thus nonzero entries of $\partial F^{(1)}/\partial X$ occur only on $R = T_1$.

Inductive step. Assume nonzero entries of $\partial F^{(D-1)}/\partial X$ occur only for lags in $R_{D-1} := T_1 + \cdots + T_{D-1}$. For $F^{(D)} = f^{(D)} \circ F^{(D-1)}$ with $Z = F^{(D-1)}(X)$, the chain rule only pairs factors with $t - u \in T_D$ and $u - s \in R_{D-1}$. Therefore $t - s \in T_D + R_{D-1} = T_1 + \cdots + T_D$.

Identical dilations. If every layer has taps $T = \{0, d, \ldots, (k-1)d\}$, then by induction $R = \{0, d, \ldots, D(k-1)d\}$. Fixed-window attention has the same architectural support composition, although its mixer is not LTI. Its contiguous tap set is $T = \{0, \ldots, w\}$, hence $R = \{0, \ldots, Dw\}$ for uniform stacks. □

Remark 2 (Minkowski sum versus diameter)**.** *Always* $R \subseteq \{0, \ldots, \max R\}$ *and* $\max R = \sum_\ell w_\ell$ *when* $w_\ell = \max T_\ell$*. Equality* $R = \{0, \ldots, \max R\}$ *holds for contiguous taps, but fails for sparse dilations. Coprime dilations enlarge* R *inside a fixed diameter only insofar as the resulting Minkowski sum covers the required lag; coprimeness alone is not a substitute for the membership test* $\ell_{\mathrm{req}} \in R$*.*

C.3 Proof of Corollary 1

Proof. Under Section 3.3, every scored time $t = m + 1 + j$ must recover payload token x_j from input index j at lag $\ell_{\mathrm{req}} = L - P$. Write $X_{\setminus j}$ for all input tokens other than x_j. By construction of the generator, x_j is independent of $X_{\setminus j}$ and of the blank recall inputs. If $\ell_{\mathrm{req}} \notin R$, Lemma 2 excludes every computational path from input j to output t. Thus, for fixed learned parameters, the network output at t is a function of $X_{\setminus j}$ alone. This is a structural dependence statement, stronger than observing a zero derivative at one probe input. The Bayes-optimal predictor of x_j given any such function is therefore the payload marginal p_{pay}, with accuracy at most $\max_c p_{\mathrm{pay}}(c)$. When p_{pay} is uniform on a vocabulary of size V, this is chance accuracy $1/V$.

If the recall span induces several required lags $\{\ell_j\}$, apply the same argument at each scored time: vanishing coupling at any $\ell_j \notin R$ bounds expected accuracy at that position by the optimal marginal baseline. It does not assert that an arbitrary trained model outputs the Bayes-optimal distribution, nor that finite-sample accuracy equals its expectation.

If only the diameter bound is used, $Dw < \tau(L) = L - P - 1$ forces $Dw < \ell_{\mathrm{req}}$, hence the same vanishing Jacobian entry by Lemma 1. □

C.4 Boundary-aware support and task dependence

Structural dependence and boundaries. The support proof describes the computation graph, not a linear approximation at one input. For fixed parameters, changing an input outside the graph ancestors of output t leaves that output unchanged. Constant left padding adds no input-dependent ancestors. For a finite sequence, let $P_\ell(t)$ be the valid input positions admitted by layer ℓ at position t, including its residual. Starting from $A_0(t) = \{t\}$, propagate

$$A_\ell(t) = \bigcup_{u \in P_\ell(t)} A_{\ell-1}(u).$$

For fixed causal taps, $P_\ell(t) = \{t - a : a \in T_\ell,\ 0 \le t - a < L\}$. Thus boundaries can only remove paths from the unrestricted Minkowski support. For position-dependent shifted windows, the same index-level recurrence respects the actual masks; pooling all shifts into one lag set can overestimate access. No shifted-window experiment is reported.

Boundary check for the reported copying positions. For the causal, position-independent taps used here, the finite-sequence ancestor set is exactly $A_D(t) = \{t - r : r \in R,\ r \le t\}$. Indeed, for a decomposition $r = a_1 + \cdots + a_D$, every partial sum lies between 0 and r because all taps are nonnegative. Backtracking from t therefore stays between $t - r$ and t; constant left padding cannot remove a path whose source is a valid input. At length 64, scored times are $54 + j$ for $j = 0, \ldots, 9$; at length 128, they are $107 + j$ for $j = 0, \ldots, 20$. Thus boundary truncation changes none of the required-lag membership conclusions in these setups. This follows from the masks and indices, without an additional training run.

Several targets and correlated context. Apply the exclusion separately to each scored source–output pair. A missing lag bounds that position's expected accuracy, rather than forcing all positions to chance. More generally, if target Y_t is correlated with accessible inputs $X_{A_D(t)}$, the information available to the model gives the Bayes ceiling $\mathbb{E}[\max_c \Pr(Y_t = c \mid X_{A_D(t)})]$. For the independent copying generator and a missing source, this reduces to $\max_c p_{\text{pay}}(c)$. This conditional Bayes ceiling is at least the marginal optimum, since a predictor can always ignore the context; it is not a sub-chance bound. We do not estimate a conditional ceiling for the broader tasks; support alone does not quantify their correlations.

Content-dependent gates. A gate whose inputs all lie within the same fixed support cannot create a dependency outside that support. A general routing module, however, may itself read distant tokens. A valid support envelope must include those control dependencies, as well as the routed values and memory updates, over all admissible inputs. A mask observed on one input is therefore insufficient for a global exclusion. Such an envelope may be full-context and provide no useful missing-lag certificate; no probabilistic extension or dynamic-model guarantee is claimed here.

C.5 Claims, evidence, and interpretation

Table 6: Assumptions, conclusions, and the evidence reported in this paper.

Assumptions / object	Supported conclusion	Evidence / interpretation
Fixed taps or fixed windows; pointwise layers; residuals	Missing graph ancestors are excluded at every parameter setting.	Lemmas 1–2; any depth.
Independent payload; required lag outside R	Expected accuracy is at most $\max_c p_{\text{pay}}(c)$.	Corollary 1; a population bound, not exact finite-sample accuracy.
$D = 2$, $L = 64$, $Dw = 64$; two seeds	Reported holes remain near chance; covers exceed 99.9%.	Tables 2–3; success of covers is empirical.
Length-128 confirmation split	Diameter rule agrees on all 7 finite-bandwidth configurations.	The other 14 rows use classifier conventions, not this certificate.
Probed energy r_{90}, m_τ	Summaries of realized influence; neither establishes membership.	Geometry probes use length 32, while copying requires lag 54 at length 64.

Graph ancestry and attribution. In the unrolled computation graph, $A_D(t)$ is the set of input ancestors of output t. No directed path implies invariance to changing that input while holding the others fixed. This is a property of the model's graph, not a claim about causal relations in the data-generating process. Attribution methods such as Jacobian Scopes (Liu et al., 2026) study realized influence. A zero attribution can reflect the input or weights; absence from the architectural ancestor set holds for both. The support test complements attribution by checking whether a dependency is possible before assessing its learned strength.

What the matched controls identify. Cross-layer fill changes connectivity and can also change width, FLOPs, and optimization. The contrast establishes that diameter alone cannot explain the observed separation and that the missing-path schedules are excluded by the theorem. It does not identify support membership as the only determinant of accuracy among covering models. The recorded Jacobian timings in Appendix I measure diagnostic cost, not compute-normalized training

of the geometry suites. No corresponding training-time comparison or broader seed/depth sweep is reported.

D TASK PROTOCOLS AND REPORTING

D.1 COPYING GENERATOR

See Section 3.3 for the full index/mask definition. Briefly: payload length $P = \max(2, \lfloor L/6 \rfloor)$; marker at $m = L - P - 1$; recall span $\{m+1, \ldots, L-1\}$ carries targets equal to the payload; those positions are blank (zeros) in the input, so evaluation does not feed previous payload targets as inputs. Accuracy is mean exact-match on the masked recall span. The diagnostic $\tau(L) = m$ is used by m_τ and by the free screen $Dw \geq \tau$; the geometric lookback for every scored token is $L - P = \tau(L) + 1$. Train and eval lengths are equal unless a transfer protocol is explicitly labeled. The lag used by m_τ is always $\tau(L_{\text{probe}})$, independent of L_{train} when those differ, so the secondary L=32/L=64 result is a scale-transfer heuristic, not task-lag alignment.

The implementation uses 128 embedding symbols, reserving indices 0, 1, and 2. Payloads are sampled uniformly from indices 3 through 127. Accordingly, the payload vocabulary in Corollary 1 has $V = 125$ and its chance baseline is $1/125 = 0.8\%$. Reporting V as the payload support size keeps the theoretical expression separate from the embedding vocabulary. Fresh seeded batches generate the scored targets; the model never receives recall targets as subsequent inputs.

D.2 PRE-REGISTRATION AND REPORTING CHECKLIST

The machine-readable configurations included with this submission define the mixer grid, tasks, lengths, seeds, optimizer, matching tolerance, coordinate probe scale, candidate-search space, and selection score. Round 2 of task repair replaced the content-gated long convolution with an ungated long convolution as the LTI control after round 1 failed the adaptivity gate.

D.3 RESULT CHECKLIST

The ledger retains the outcomes of the registered diagnostic and architecture-prediction checks, including negative results. It separates those exploratory objectives from the fixed-diameter copying comparisons in the main text. Status labels refer to the recorded gates, not to formal guarantees for every mixer family.

Table 7: Recorded outcomes of the diagnostic and architecture-prediction checks.

Claim	Status
Coordinate stability (discrete atlas)	pass
Kernel/window knobs jointly move reach and copying	pass
SSM selection jointly moves adaptivity and MQAR	pass
Init τ-mass screen (matched L=128; secondary)	95.2% FN kernel-63 conv.
Init τ-mass transfer (L=32/L=64; dilation patch)	pass (100%)
Archived diameter classifier, matched L=128	pass (100%; n=21)
Init r_{90} on L=32/L=64 (historical; 90.5%)	pass
Matched length-128 r_{90} screen	pass (100%; n=21)
Matched length-128 archived diameter classifier	pass (100%; n=21)
Matched length-128 τ-mass screen	secondary (95.2%; FN kernel-63 conv.)
Adaptivity aligned with MQAR under Hyena gate	fail
Bottleneck / equivariance within-family laws	fail
Unnormalized ESS distinguishes SSM width; copying does not	pass
Recovered algebra Toeplitz mass: LTI vs attention	pass
Four-axis architecture-law conjunction	fail
Neighborhood interpolation of accuracy	fail
Neighborhood interpolation of loss	pass
Exclusion: no reach, no copying	pass
Missing-family / new architecture admission	fail
Length extrapolation of copying	fail
Early reach filters final copying (87.5% agreement)	pass
enwik8 Spearman(reach, −val. loss) at 1M / 10M / 100M	fail
Headline laws at $n = 6$ seeds	signs unchanged
Support-matched Dw=64 attn vs TCN (L64)	pass
Fixed-Dw reachable-set geometry suite	pass (5/5 predicted)
Heterogeneous same-Dw fill / coprime suite	pass (5/5 predicted)

The archived diameter classifier codes missing finite-bandwidth values as negative and full-context values as positive. Its $21/21$ agreement includes nine undefined-bandwidth rows, five full-context rows, and seven finite-bandwidth convolution/window rows. Only the last group has the finite-diameter interpretation, with agreement on all seven. Undefined bandwidth is an applicability flag, not a certificate of insufficient support. The stored outputs are unchanged by this distinction.

D.4 Matched-length copying occupancy

Copying at length 128 uses the same 29 occupancy mixers with seeds $\{11, 29, 47\}$, 2500 steps, and a 1M-parameter budget. The required lag is $\ell_{\mathrm{req}} = 107$ and the diagnostic lag is $\tau = 106$. Compact two-layer window-32 attention and kernel-31 convolution have $Dw \in \{64, 60\}$ and remain at chance, while kernel 63, gated long convolution, linear attention, and full attention copy. On the 21 confirmation mixers, frozen $r_{90} \geq 0.5$ and the archived diameter classifier agree at 100%, and frozen $m_\tau \geq \theta$ agrees at 95.2% with one false negative on kernel-63 convolution. The archived classifier uses the applicability conventions documented in the result ledger. The hard architectural test remains membership in the task-specific reachable set.

E Additional screening analyses

E.1 Raw m_τ values and rule comparison

The frozen threshold is $\theta = 5.423 \times 10^{-2}$. It was selected from length-32 initialization probes and length-64 copying labels. Calibration positives are full attention, kernel-31 convolution, and ungated long convolution; the stored threshold is their minimum mass minus a 10^{-6} numerical tolerance.

Matched L=128 confirmation (n=21). Both the archived diameter classifier (with the conventions above) and $r_{90} \geq 0.5$ achieve 100% agreement on confirmation labels. For $m_\tau \geq \theta$: TP= 5, TN= 15, FP= 0, FN= 1 (kernel-63 convolution); agreement is 95.2%. On the cleanest protocol m_τ is therefore slightly *stricter* than r_{90} / Dw, but not more accurate. Init probes for length 128 use seeds $\{11, 29, 47\}$; the kernel-63 cell has $m_\tau = 0$ at every seed while $r_{90} \approx 0.82$. Copying labels at length 128 were averaged over the same three seeds, and binary success/failure agrees on every seed.

Secondary L=32 probe / L=64 copying confirmation. Here $\tau = 26$ for the length-32 probe and $\delta = 1$, so rows $t \in \{26, \ldots, 31\}$ contribute. Agreement under m_τ is 100% (no FP/FN); under $r_{90} \geq 0.5$ it is 90.5%, with false positives on kernel-15 convolution and dilation-4 TCN. The probe uses lag 26 while copying trains at required lag 54 (diagnostic lag 53); we interpret this only as a dilation / scale-transfer heuristic. The comparison measures diagnostic transfer across lengths rather than energy at the target lag itself.

E.2 Estimator details and δ sensitivity

Any output index with $\sum_{s \leq t} E_{ts} \leq 10^{-12}$ is treated as inactive and skipped (not zero-filled) before the median. Exact Jacobians are used for the principal screen tables; production large-L scans use the randomized estimator after agreement checks (Table 11). A local exact-Jacobian sweep over $\delta \in \{1, 2, 4\}$, plus median versus mean aggregation, on a five-mixer slice spanning short and long convolution, dilated TCN, windowed attention, and full attention preserves the same pass/fail pattern relative to θ.

E.3 Autopsy of the kernel-63 m_τ miss

Exact-Jacobian profiles at initialization explain why matched m_τ can be zero while $Dw = 124 > \tau = 106$ and training still succeeds. On a length-128 mixer-only probe, every active row $t \geq \tau$ has zero fractional energy in the τ-window, while the median argmax lag is ≈ 49 (near mid-kernel), far from 106. Short-lag mass ($|\text{lag}| \leq 8$) is nontrivial (≈ 0.14). The same pattern appears for one- and two-block residual stacks at L=64 across seeds $\{11, 29, 47\}$. Thus the failure mode is initialization energy concentrated on shorter lags under Gaussian probes, a soft proxy miss rather than a violation of Lemma 1.

E.4 End-to-end screen cost

Prospective skip on matched length-128 copying trains only m_τ-positives, taking 0.083 GPU·h versus an estimated 0.34 GPU·h for the full grid when probes are cached. The practical recommendation differs: apply free $Dw \geq \tau$ first (zero probe cost). Only candidates that pass support but look like dilation-style r_{90} false positives warrant a Jacobian estimate. A new diagnostic probe incurs additional computation; a tap-set test does not require that probe. Table 11 reports the measured cost of exact and randomized estimators separately. Ratios to forward passes depend on family and implementation and should not be read as end-to-end training savings. The cached-probe timing experiment is an empirical accounting exercise, not a cost guarantee for a new model.

E.5 Support-matched Dw controls and geometry suite

Table 8 records continuum cells at matched (or near-matched) Dw=64; the geometry experiments in the main text explain the split between windowed attention and dilation-8 TCN. Table 2 is the controlled geometry suite: all five mixers use Dw=64, and whether $\ell_{\text{req}} = 54$ belongs to R depends on the dilation/kernel choice. At Dw=16 (window-8 attention, kernel-9 convolution, and dilation-2 TCN) every continuum cell fails both length protocols, as required when $Dw < \tau$.

Table 3 is the heterogeneous fill suite: Dw=64 and $\ell_{\text{req}} = 54$ are the same, but per-layer taps differ. Predicted membership matches copying on all five cells (uniform hole and non-covering coprime at chance; both fill orders and covering coprime at $> 99.9\%$).

Registered interventions. The recorded intervention signs (Table 9) persist across six seeds: kernel/window jointly move reach and copying; dilation can move reach without copying; selection

Table 8: Continuum cells at Dw=64 (and near match). Unrounded window-32 attention m_τ is $5.4232 \times 10^{-2} \geq \theta$.

Protocol	Mixer	Dw	m_τ	r_{90}	Copying
L=64	Attn. window 32	64	0.054232	0.90	99.9%
L=64	TCN dil. 8	64	0.000	0.73	0.7%
L=64	Conv. k=31	60	0.070	0.92	100%
L=128	Attn. window 32	64	0.000	0.40	0.9%
L=128	TCN dil. 8	64	0.000	0.42	0.8%

Table 9: Pre-registered intervention laws. Gate = frozen sign test. The table records all gate outcomes, including the linear-feature miss.

Law	Δ coord.	Δ acc.	Gate	Reading
Conv. kernel $3 \to 31$	$+0.82$	$+0.99$	pass	joint reach/copy
Attn. window $8 \to$ full	$+0.44$	$+0.99$	pass	joint reach/copy
TCN dilation $1 \to 8$	$+0.51$	$+0.00$	pass	reach only
SSM selection off$\to$on	$+0.20$	$+0.23$	pass	joint adapt/MQAR
Hyena gate off$\to$on	$+0.10$	$+0.00$	pass	gate $\neq$ MQAR
Gate on$\to$off (copying)	-0.07	$+0.00$	pass	ungated copy
SSM state $16 \to 128$	$+0.00$	$+0.00$	pass	bottleneck inert
Linear feature $8 \to$ full	$+0.00$	-0.25	fail	miss
Time coupling off$\to$on	$+0.00$	$+0.00$	pass	no change

and gating provide mixed evidence for a relationship between adaptivity and MQAR. The miss on linear features and the inert bottleneck/equivariance rows are reported for completeness.

E.6 Alternative r_{90} screen protocol

The diagnostic study uses a frozen mass threshold. Applying the additional rule $r_{90} \geq 0.5$ on the same dynamics protocol still yields 90.5% confirmation agreement at the length-64 labels (precision 0.78, recall 1.0) and 100% matched length-128 confirmation. Dilated short kernels and intermediate kernels remain the false-positive mode under r_{90}; m_τ is designed to suppress that mode.

E.7 Prospective skip GPU-hours

For matched length-128 copying (seed 11), the frozen τ-screen ($m_\tau \geq \theta$) is applied before training. Of 29 occupancy mixers, 22 are predicted negative (76%) and skipped; the seven predicted positives (full attention, three gated long convolutions, and three linear-attention widths) take 297 s (0.083 GPU·h). Extrapolating that positive-only wall-clock to the full grid gives 0.34 GPU·h, so the screen saves roughly 0.26 GPU·h ($\approx$ 76% of training time) under this protocol. The held length-128 labels show one false negative (kernel-63 convolution); all other true positives are covered. These figures exclude the cost of producing new probes. The full-grid time is extrapolated from the positive-only run rather than measured as a paired end-to-end baseline.

F Additional recall and model comparisons

F.1 Leave-one-family-out negative control

Family-held-out ridge regression provides a test of cross-family prediction. Accuracy $\Delta R^2 = 0.049$ with bootstrap interval $[0.022, 0.075]$ passes the sign test, but both coordinate and capacity R^2 are near zero (-0.003 vs -0.052). Loss $\Delta R^2 = -0.181$ fails, because coordinates are worse than capacity. Family-held-out regression therefore cannot replace the mixer-held-out occupancy screen in the main text.

Table 10: MQAR accuracy at length 64. Occupancy = default atlas protocol; repaired = adaptivity-repair protocol (seed 11).

Mixer	Occupancy MQAR	Repaired MQAR
Softmax attention	27.9%	99.8%
Linear attention	27.9%	27.5%
Ungated long conv.	27.9%	26.9%
Gated long conv.	28.2%	26.8%
Selective SSM	24.8%	25.2%
Short conv. (local)	13.3%	14.7%
Position-wise	1.7%	1.7%

F.2 Discrete-family in-distribution results

Averaged over five seeds and three training lengths, in-distribution copying accuracies are attention 99.97%, gated long convolution 100%, linear attention 99.2%, GRU 5.4%, selective SSM 2.4%, and position-wise / short convolution / TCN / diagonal SSM at chance ($\approx 0.8\%$). MQAR remains compressed on this un-repaired protocol (attention 16.9%; gated long convolution 11.6%), so Table 4 uses the repaired MQAR setting.

F.3 MQAR protocol contrast (unrepaired vs repaired)

Table 10 compares the repaired readout (Table 4) with same-budget occupancy MQAR under the default atlas protocol (length 64, no adaptivity repair). Attention is the only mixer that improves from weak occupancy MQAR ($\approx 28\%$) to near-perfect accuracy under repair (99.8%); LTI long-range and linear-attention cells remain in the mid-20% range under both protocols. The repair therefore reveals an attention-versus-LTI dissociation and shows that MQAR outcomes depend strongly on the task protocol.

Occupancy short-conv cell is kernel 3; repaired short-conv cell is kernel 5 from the calibration config. Their kernel sizes differ, so the row is a comparison of local controls rather than a matched-kernel intervention.

F.4 Held-out induction

Nine families, two seeds, 4000 steps, length 64. Mean induction accuracy is 4.8% for attention, 3.5% for linear attention, and at chance ($\approx 0.8\%$) for all other families. Spearman correlations with reach and adaptivity are weak (0.48 and 0.29). The task does not yet provide a usable second adaptivity readout at this compute, analogous to unrepaired MQAR on the v2 atlas. At this training budget the low accuracies do not support a useful ranking of retrieval capability.

F.5 Selective SSM tuning grid

A pre-registered micro-grid sweeps selective SSM state dimension $\{64, 128\}$, selection strength $\{0.5, 1.0\}$, and learning rate $\{3 \times 10^{-4}, 10^{-3}\}$ on copying and MQAR at length 64 (seed 11). The best copying cell uses state 128 and selection strength 1 at $\mathrm{lr} = 10^{-3}$ (peak accuracy 9.1%; grid-mean 6.4% for that mixer), versus chance-level copying for selection 0.5. The best MQAR on the grid is 27%. Relative to the paper default selective SSM (about 2% copying in Table 4), this grid improves copying but remains far below the successful covering configurations. Its narrow setting does not characterize selective SSMs under other initialization or optimization choices.

F.6 Bottleneck autopsy

We compare diagonal SSMs with state sizes 16 and 128 against full attention and a k=31 convolution as high-copying controls, using 500 copying steps. Unnormalized exact ESS at initialization is 15 for SSM-16, 75 for SSM-128, 325 for attention, and 1554 for conv-31. Randomized rank-32 ESS values are 15, 28, 28, and 32. The normalized bottleneck for the two SSMs is 0.0044 versus 0.0082.

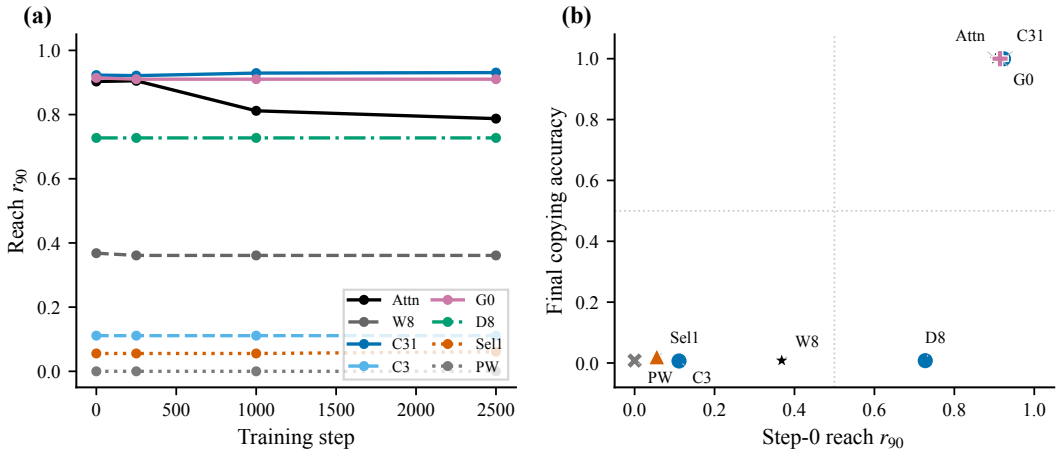


Figure 4: Copying dynamics. (a) Reach versus training step at diagnostic length 32 with Gaussian probes. (b) Step-0 reach versus final copying: long kernels, ungated long convolution, and full attention start already above the copying threshold; window-8 attention, short kernels, dilated convolution, selective SSM, and position-wise do not.

Exact and randomized estimates agree for the narrow SSM and diverge once ESS exceeds the sketch rank. Trained values are essentially the same as initialization for both SSMs, and copying stays at chance. Occupancy randomized ESS across state sizes 16, 32, 64, and 128 is 14.7, 23.8, 22.0, and 14.9, which is not monotone. The production sketch cannot replace exact ESS.

G Training dynamics and probe robustness

G.1 Training dynamics and probe robustness

We trained eight mixers with two seeds on the copying task for 2500 steps. We saved snapshots at steps 0, 250, 1000, and 2500, evaluated extra diagnostic lengths 64 and 128, and fitted both Gaussian and task-embedding probes at initialization and at the final checkpoint (Figure 4).

At step 0, the Gaussian and task-embedding probes agreed strongly on r_{90} (Spearman 0.93, mean absolute $\Delta \approx 0.086$ on eight mixers). Using $r_{90} \geq 0.5$ yielded the same 87.5% agreement with final copying labels under either probe kind, and no mixer showed disagreement between probe kinds on the cutoff. We therefore use Gaussian probes as the default screen, though we do not claim probe-kind invariance of m_τ (task m_τ was not logged in that dynamics suite).

G.2 Reach quantile and inactive rows

For the eight dynamics mixers at initialization, r_{85}, r_{90}, and r_{95} all give 87.5% screen agreement with final copying. Coding inactive Jacobian rows as skipped or as zero does not change the summary. The production definition skips rows with causal energy $\sum_{s \leq t} E_{ts} \leq 10^{-12}$. The 90% energy threshold is not a sharp cutoff on this grid.

H Matched-budget enwik8

Protocol details for Figure 5. Seven mixers, two seeds, 8000 steps, context 512. At ≈1M parameters, validation loss (mean over seeds): selective SSM 1.425, GRU 1.433, short conv 1.467, gated long conv 1.502, diagonal SSM 1.508, softmax attention 1.560, linear attention 1.923. The Spearman correlation between mixer-mean reach and negative validation loss was -0.46 ($p = 0.29$); adaptivity was 0.04 ($p = 0.94$). At ≈10M parameters, every mixer met the ±5% budget, with widths from 400 to 504. Validation loss: softmax attention 1.257, selective SSM 1.320, GRU 1.328, gated long conv 1.336, short conv 1.353, diagonal SSM 1.407, linear attention 1.444. The Spearman correlation for

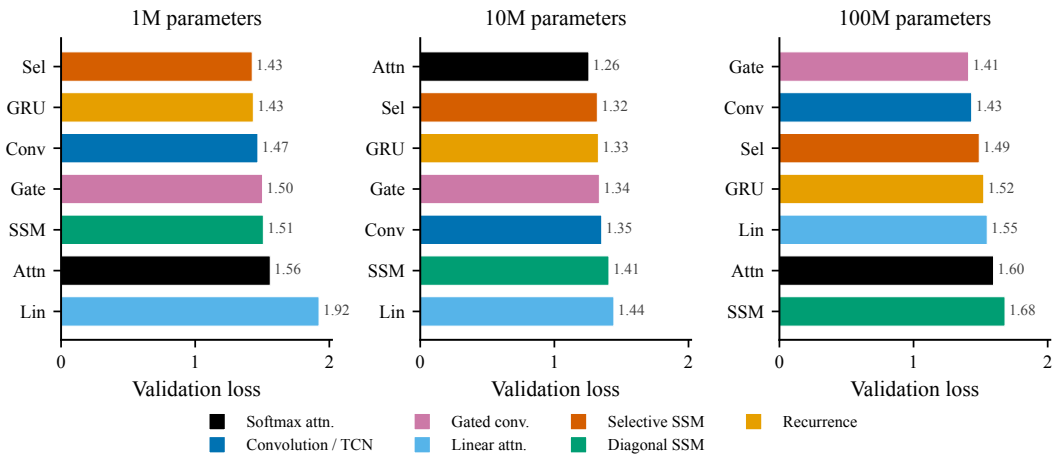


Figure 5: Matched-budget enwik8: reach does not rank validation loss.

Table 11: Jacobian cost at L=64, d=64 (one sample). The randomized path matches the production coordinate estimator.

Family	Fwd.	Rand.	Exact	Ratio	Peak MB
Position-wise	0.08	48.5	60.4	618	272
Convolution	0.16	142	223	900	344
Dilated conv.	0.13	139	306	1098	368
GRU	0.33	1498	7455	4565	303
Diagonal SSM	0.53	314	1141	593	463
Selective SSM	0.54	481	1614	893	719
Gated long conv.	0.26	207	429	802	527
Linear attention	0.59	333	1075	563	5393
Softmax attention	0.33	226	344	680	977

reach was -0.32 ($p = 0.48$); adaptivity was 0.67 ($p = 0.10$). At ≈100M parameters, every mixer met the $\pm 5\%$ budget, with widths from 1296 to 1648. Linear attention used 4-process DDP that splits the global batch of 16. Validation loss: gated long conv 1.410, short conv 1.433, selective SSM 1.490, GRU 1.523, linear attention 1.549, softmax attention 1.597, diagonal SSM 1.682. The Spearman correlation for reach was $+0.29$ ($p = 0.53$); adaptivity was 0.02 ($p = 0.97$). The association is not statistically resolved at any reported scale. The ranking of attention changes between the 10M and 100M fixed-step experiments; these runs do not isolate the cause of that change. Their consistent implication for this paper is that soft reach does not supply a reliable language-model ranking.

I Jacobian probe cost

Table 11 reports wall-clock time and peak memory for one exact block Jacobian versus the production randomized VJP/JVP estimator (rank 32, 8 probes, 8 output positions) on nine family representatives at length 64 and width 64, relative to a single forward pass. Exact Jacobians use much more memory on attention-like families; the randomized estimator stays close to forward-pass memory and is used for the broader coordinate scans; the principal initialization screen tables use exact Jacobians.

J Recovered symmetry algebra

Protocol: diagnostic length 32, four Gaussian probes, Hutchinson influence operator E, commutant form on $\mathfrak{gl}(L)$, $n_{\max} = 8$. Reported invariants of $\hat{c}_{ij}^{k}$: Toeplitz mass of E, nonabelianness, closure residual. Family means of Toeplitz mass: convolution 1.00, SSM 0.998, gated convolution 0.995, GRU 0.996, attention 0.47, linear attention 0.30. Dilation $1 \to 8$ keeps Toeplitz mass 1.00 while

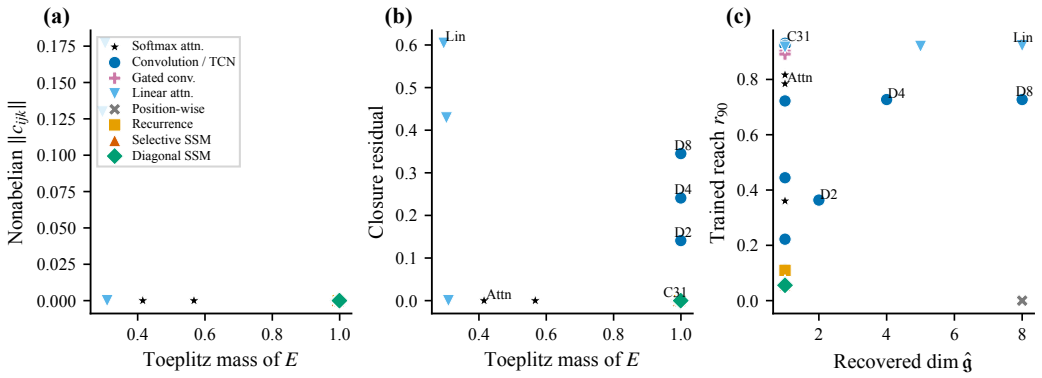


Figure 6: Invariants of the recovered time-mixing structure constants on the occupancy grid. Left: Toeplitz mass versus nonabelianness. Middle: dilation remains translation-equivariant while the closure residual grows. Right: recovered algebra dimension versus trained reach.

recovered $\dim \hat{\mathfrak{g}}$ goes $1 \to 8$ and closure residual $0 \to 0.35$. The plot does not classify all groups; discrete translations are the only G with an explicit group algebra at the Toeplitz mass values shown.

K Reproducibility Artifacts

The supplement `submission_source.zip` contains a self-contained `code/` directory. Table 12 maps paths relative to this directory; stored results can be inspected without training.

Table 12: Locations of the implementation and stored evidence in the anonymous code archive.

Component	Location
Residual model	`src/atlas/model.py`
Mixers and task generators	`src/atlas/mixers/core.py`; `src/atlas/tasks/synthetic.py`
Support geometry runs	`results/support_geometry_Dw64.csv`
Cross-layer fill runs	`results/support_geometry_hetero_Dw64.csv`
Geometry configurations	`configs/support_geometry_Dw64.yaml`; `configs/support_geometry_hetero_Dw64.yaml`
Saved screen comparison	`results/score_boost_analysis/init_screen_mtau_L128.csv`
Support-set export	`scripts/export_support_geometry.py`
Coordinate definitions	`src/atlas/metrics/coordinates.py`
Checks and analysis	`tests/`; `results/score_boost_analysis/`
Package verification	`scripts/verify_archive.py`; `MANIFEST.sha256.json`

Reproduction. Use Python 3.11 or later. From `code/`, install with `pip install -e ".[dev,paper]"`, verify the manifest with `python scripts/verify_archive.py`, and run `python -m pytest -q`. To rerun either central suite, use `atlas run --config` followed by its configuration path in Table 12. These configurations require CUDA; use a separate working copy to preserve the stored results.

Package scope. `README.md` documents the commands and evidence; `ENVIRONMENT.txt` records the code-check environment, not the original training environment. The archive includes corpus preparation and checksums, but no full corpus or trained checkpoints. The same archive also supplies the manuscript source and all original figures unchanged. No new experiment runs were added for this revision.