SURCO: SURVIVOR COMPENSATION FOR RECURRENT STATE PRUNING OF STATE SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

State-space language models replace a growing KV cache with a fixed-size recurrent state, but this state can still become a substantial inference cost at scale. Recent state-pruning methods show that the recurrent state dimension can be reduced considerably while retaining useful model quality, making state pruning a direct lever for reducing inference cost. We study an orthogonal question: once a pruning mask and state budget are fixed, can the surviving states be re-optimized to improve accuracy? We introduce SURCO, a training-free framework that augments recurrent-state pruning with survivor compensation and layer-wise state allocation. Using the exact additive decomposition of Mamba-3 outputs, SURCO constructs an output-space Gram matrix from realized state responses, ranks pruning units by conditional necessity, and, after the complete pruning mask is fixed, solves a closed-form least-squares problem that reweights the surviving readouts to reconstruct the dense layer response. The resulting scalar gains are folded into the existing readout parameters and add no decode-time operation. At 50% state removal on Mamba-3 SISO/MIMO, SURCO outperforms by 2-6% accuracy (varying model sizes) improvement from prior pruning methods. We further evaluate the complete method across Mamba-3 MIMO/SISO and Mamba-2 on long context. When the reduced state is physically materialized, 50% state removal on Mamba-3 SISO-443M yields $1.72\times$ higher batch-256 throughput and 38% lower peak memory. Thus, SURCO improves the quality retained at a fixed recurrent-state budget without giving up the systems benefits of state pruning.

1 INTRODUCTION

State-space models (SSMs) offer an attractive alternative to attention for autoregressive inference: instead of retaining a KV cache that grows with sequence length, a recurrent SSM summarizes the past in a fixed-size state (Gu & Dao, 2024; Dao & Gu, 2024; Lahoti et al., 2026). Fixed-size, however, does not mean negligible. The commonly used state dimension has increased eightfold, from 16 in Mamba-1 to 128 in standard Mamba-2 configurations. With head dimension $P = 64$ and expansion factor two, the Mamba-2 recurrent state contains $256d$ elements per layer for model width d, equivalent to the KV cache of 128 tokens in a conventional multi-head Transformer of the same width (Chen et al., 2024). This cost is particularly visible in at higher batch size. After 8-bit weight quantization, Q-Mamba reports that recurrent states account for 79.6% of memory for Mamba2-2.7B at batch size 128 (Tianqi et al., 2025). Similarly, in Mamba-3 BF16 kernels, doubling the state dimension from 64 to 128 increases reported latency by $1.43\times$ for SISO and $1.36\times$ for MIMO (Lahoti et al., 2026). Therefore, a direct memory–latency tradeoff, not merely an architectural constant.

This has motivated a broader effort to compress SSM language models. Redundancy has been attacked at several orthogonal scales: individual weights can be sparsified, internal channels can be removed, and entire Mamba or feed-forward components can sometimes be dropped. Mamba-Shedder, for example, studies component removal at coarse granularity, while other work adapts conventional structured pruning criteria such as WANDA and Taylor-based sensitivity to SSM-based LLMs (Muñoz et al., 2025; Ghattas et al., 2025). These approaches reduce parameters and computation, but they do not all reduce the resource that is unique to recurrent inference. Removing weights can make a model smaller; *reducing the state dimension directly shrinks the working memory that must be retained for every active sequence and touched at every decode step*.

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.

This observation has led to an othogonal state-specific pruning methods. LAST (Gwak et al., 2024) uses a system-theoretic modal criterion to remove low-impact states , while AIRE-Prune (Padhy et al., 2026) scores states through long-horizon impulse-response energy; more recent work GHOST (Menezes & Kyrillidis, 2026) estimates state importance from forward-pass controllability and observability statistics .Together, these works establish that pretrained SSMs contain substantial removable recurrent capacity. Also, reducing the state dimension directly serves for high throughput and low peak memory(Menezes & Kyrillidis, 2026). This holds for Mamba-3 as well as we see later in section 4.3. We therefore take the systems motivation for state pruning as established and focus on a complementary question: *for a fixed recurrent-state budget, how much of the pruning-induced quality loss is actually recoverable by kept states?*

The dominant formulation in this literature is nevertheless a *selection* problem: assign each state an importance score, remove low-ranked states, and inherit the surviving states unchanged. Recent response-space analysis further shows that selective SSM outputs admit exact per-mode decompositions and that Gram structure can expose interactions between state contributions (Bhattacharya, 2026). These advances improve our ability to decide *what to remove*, but leave a complementary question largely open: *after a pruning mask is fixed, what should the surviving state realization become to improve model performance?*

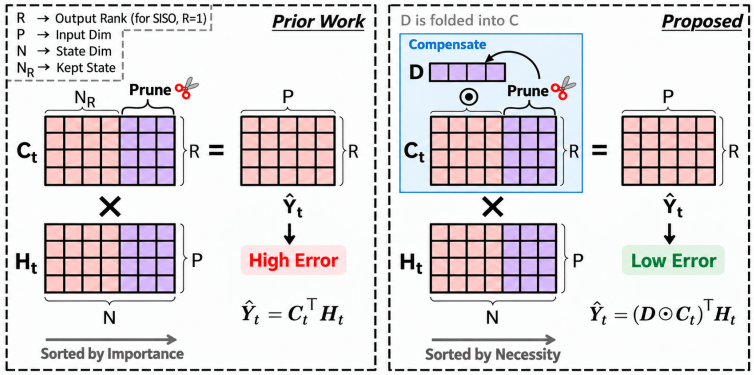


Figure 1: **State selection versus reduced-realization optimization.** Conventional state pruning removes low-ranked states and leaves the surviving states unchanged. SURCO additionally refits the surviving readouts so that they absorb response components that would otherwise be discarded.

We introduce **SURCO**, a training-free framework that explicitly optimizes this reduced state realization. For each Mamba-3 layer, SURCO collects realized state-to-output responses on a small calibration set and forms an output-space Gram matrix capturing both response energy and functional overlap. Conditional necessity provides a one-shot ranking of admissible pruning units, from which the complete removal set is chosen for a requested state budget. Only after this pruning mask is fixed do we perform *survivor compensation*: a single closed-form least-squares solve computes scalar readout gains that make the collective response of the survivors best approximate that of the dense layer. These gains are folded into the existing Mamba-3 readout parameters and introduce no additional decode-time operation. Within-layer replaceability does not, however, determine how much recurrent capacity each layer should retain. We therefore treat global state-budget allocation separately, using measured end-to-end distortion to distribute a fixed total state budget across layers.

On Mamba-3 SISO-1.5B at 50% state removal, we compared to LAST, AIRE-Prune, and GHOST, our method SURCO, outperforms average performance by 2-6% percent.

Our contributions are:

- **Reduced-realization state pruning.** We formulate post-training state compression as two distinct problems—which recurrent states survive and how the surviving realization should change—and introduce closed-form scalar survivor compensation for selective SSMs.
- **Response-aware local reduction with separate global allocation.** We use realized state-to-output geometry to measure conditional necessity within a layer, while end-to-end distortion determines how a fixed recurrent-state budget is distributed across layers.
- **Controlled and deployable validation.** We isolate compensation under fixed masks from multiple independent selectors, evaluate across Mamba-3 SISO/MIMO and Mamba-2, and

show that physical state compaction converts recurrent-state redundancy into lower serving memory and higher high-batch decode throughput.

The central implication is that state pruning should not end when a mask is chosen: *compression must optimize not only which state functions remain, but also the realization formed by those survivors.*

2 BACKGROUND

2.1 SELECTIVE SSMS AND MAMBA-3

A state-space model (SSM) represents a sequence through a recurrent state rather than an explicit token cache. In its discrete form,

$$\mathbf{h}_t = \overline{\mathbf{A}}_t \mathbf{h}_{t-1} + \overline{\mathbf{B}}_t \mathbf{x}_t, \qquad \mathbf{y}_t = \mathbf{C}_t^\top \mathbf{h}_t. \tag{1}$$

Selective SSMs make the recurrence input dependent—in Mamba, the step size and the input/output projections are functions of the current token—allowing the model to selectively write, retain, and read information (Gu & Dao, 2024). Mamba-2 further simplifies the state transition to a scalar-times-identity form, which enables the state-space duality formulation and efficient matrix-multiplication algorithms (Dao & Gu, 2024).

Mamba-3 retains this selective recurrent structure while introducing an exponential–trapezoidal discretization, complex/rotary state dynamics, and a multi-input multi-output (MIMO) parameterization (Lahoti et al., 2026). In the equivalent rotary frame, its SISO recurrence for one head can be written as

$$\mathbf{H}_t = \alpha_t \mathbf{H}_{t-1} + \beta_t \overline{\mathbf{B}}_{t-1} \mathbf{x}_{t-1}^\top + \gamma_t \overline{\mathbf{B}}_t \mathbf{x}_t^\top, \qquad \mathbf{y}_t = \mathbf{H}_t^\top \overline{\mathbf{C}}_t, \tag{2}$$

where $\mathbf{H}_t \in \mathbb{R}^{N \times P}$, $\overline{\mathbf{B}}_t, \overline{\mathbf{C}}_t \in \mathbb{R}^N$, and $\mathbf{x}_t, \mathbf{y}_t \in \mathbb{R}^P$. The trapezoidal coefficients are $\alpha_t = \exp(\Delta_t A)$, $\beta_t = (1 - \lambda_t)\Delta_t \alpha_t$, $\gamma_t = \lambda_t \Delta_t$, with data-dependent $\lambda_t \in [0, 1]$. For MIMO rank R, $\overline{\mathbf{B}}_t, \overline{\mathbf{C}}_t \in \mathbb{R}^{N \times R}$ and $\mathbf{X}_t \in \mathbb{R}^{P \times R}$, giving

$$\mathbf{H}_t = \alpha_t \mathbf{H}_{t-1} + \beta_t \overline{\mathbf{B}}_{t-1} \mathbf{X}_{t-1}^\top + \gamma_t \overline{\mathbf{B}}_t \mathbf{X}_t^\top, \qquad \mathbf{Y}_t = \mathbf{H}_t^\top \overline{\mathbf{C}}_t. \tag{3}$$

The persistent state remains $N \times P$; MIMO enriches how that state is written and read rather than enlarging it. Crucially for state reduction, the readout is additive along the state axis: if $\mathbf{h}_{t,i}^\top$ and $\mathbf{c}_{t,i}^\top$ denote row i of $\mathbf{H}_t$ and $\overline{\mathbf{C}}_t$, then

$$\mathbf{Y}_t = \sum_{i=1}^{N} \mathbf{h}_{t,i} \mathbf{c}_{t,i}^\top, \tag{4}$$

with SISO recovered at $R = 1$.

2.2 GRAM MATRICES

Given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^D$, collect them as $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$. Their (weighted) Gram matrix under a positive-semidefinite metric $\mathbf{Q} \succeq 0$ is

$$\mathbf{G} = \mathbf{V}^\top \mathbf{Q} \mathbf{V}, \qquad G_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle_{\mathbf{Q}} = \mathbf{v}_i^\top \mathbf{Q} \mathbf{v}_j. \tag{5}$$

For $\mathbf{Q} = \mathbf{I}$, this is the standard Gram matrix. Its diagonal contains squared norms, $G_{ii} = \|\mathbf{v}_i\|_{\mathbf{Q}}^2$, while off-diagonal entries measure pairwise alignment. Moreover,

$$\mathbf{a}^\top \mathbf{G} \mathbf{a} = \|\mathbf{V}\mathbf{a}\|_{\mathbf{Q}}^2 \geq 0, \tag{6}$$

so $\mathbf{G}$ is positive semidefinite and compactly represents the geometry of the vector set. In our method, the vectors will be realized state-to-output contributions, so the Gram matrix records both their individual energy and their functional overlap.

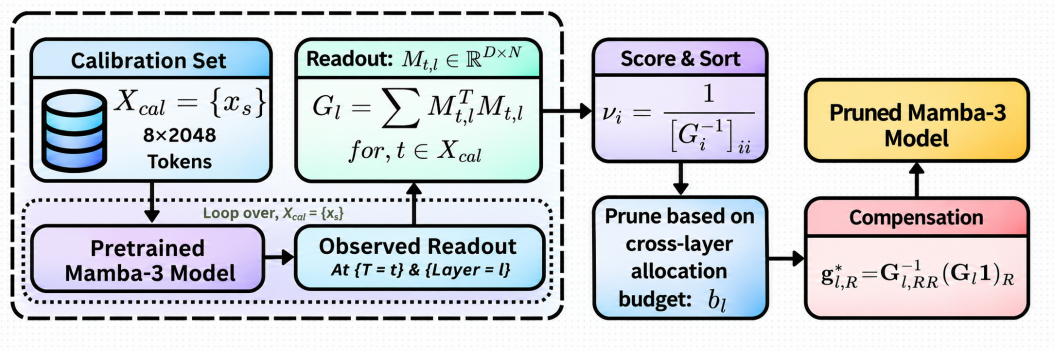


Figure 2: **SURCO framework overview.**

3 Method: Conditional State Reduction

We formulate post-training recurrent-state compression as a reduced-realization problem. A pruning mask determines *which* state functions remain; SURCO additionally optimizes *how* the surviving functions realize the original layer response. The method has three stages (Figure 2): (i) estimate a response geometry from a small calibration set and rank admissible pruning units within each layer, (ii) choose the removal set required by a given state budget and compensate the resulting survivor set once, and (iii) allocate a fixed global state-removal budget across layers using end-to-end predictive distortion. No gradients or parameter fine-tuning are used.

3.1 Calibration response geometry

Mamba-3 readout and pruning units. We use the Mamba-3 notation from Section 2: tensors are bold and scalars are unbolded. For one head at layer ℓ, the recurrent state is $\mathbf{H}_{\ell,t} \in \mathbb{R}^{N\times P}$ and, for MIMO rank R, the readout is $\mathbf{C}_{\ell,t} \in \mathbb{R}^{N\times R}$. The background writes the output in the Mamba-3 layout as $\mathbf{Y}_{\ell,t} = \mathbf{H}_{\ell,t}^\top \mathbf{C}_{\ell,t}$. Equivalently, writing the same readout in the conventional SSM order gives:

$$\mathbf{Y}_{\ell,t}^\top = \mathbf{C}_{\ell,t}^\top \mathbf{H}_{\ell,t} = \sum_{i=1}^{N} \mathbf{c}_{\ell,t,i}\mathbf{h}_{\ell,t,i}^\top, \tag{7}$$

where $\mathbf{c}_{\ell,t,i}^\top \in \mathbb{R}^{1\times R}$ and $\mathbf{h}_{\ell,t,i}^\top \in \mathbb{R}^{1\times P}$ are row i of $\mathbf{C}_{\ell,t}$ and $\mathbf{H}_{\ell,t}$, respectively. Thus each state coordinate contributes an $R \times P$ rank-one term, and SISO is the special case $R = 1$. All derivations below therefore apply to SISO and MIMO without a separate case analysis.

For a non-rotary state, one pruning unit is one state coordinate. For the rotary part of Mamba-3, the two real coordinates implementing a complex mode form one indivisible unit and are retained or removed together. Let $\mathcal{U}_\ell$ be the resulting set of m_ℓ pruning units. The readout contribution of a unit $u \in \mathcal{U}_\ell$ is the sum of the terms in equation 7 for $i \in u$.

How the calibration set is used. Let $\mathcal{X}_{\text{cal}} = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$ be a small set of calibration sequences; its exact size is specified in Section 4.2. We run the frozen dense model on these sequences under teacher forcing. For every calibration sequence s, token position t, and SSM layer ℓ, we record the contribution of each pruning unit at the monitoring point *after the layer output projection*. Conditioned on a fixed token, the Mamba-3 gate/rank-reduction coefficients and output projection are fixed, so this path is linear in the additive SSM readout. Consequently, the projected per-unit contributions still sum exactly to the dense state-dependent layer response.

Denote the observed post-projection contribution of unit u by $\mathbf{m}_{\ell,t,u}^{(s)} \in \mathbb{R}^D$ and collect the m_ℓ unit responses at that calibration position as

$$\mathbf{M}_{\ell,t}^{(s)} = \left[\mathbf{m}_{\ell,t,1}^{(s)} \ \cdots \ \mathbf{m}_{\ell,t,m_\ell}^{(s)}\right] \in \mathbb{R}^{D\times m_\ell}, \qquad \mathbf{y}_{\ell,t}^{(s)} = \mathbf{M}_{\ell,t}^{(s)}\mathbf{1}. \tag{8}$$

Here D is the monitored layer-output dimension. We then accumulate the per-layer Gram matrix

$$\mathbf{G}_\ell = \sum_{s=1}^{S}\sum_{t} \left(\mathbf{M}_{\ell,t}^{(s)}\right)^\top \mathbf{M}_{\ell,t}^{(s)}. \tag{9}$$

This is exactly the Gram matrix obtained by concatenating each unit's observed response over calibration sequences, tokens, heads, and output coordinates. Because the responses are recorded after the output projection, the geometry in equation 9 is already the reconstruction geometry used by SURCO; no additional metric $\mathbf{Q}$ is required. The diagonal $G_{\ell,uu}$ measures the realized output energy of unit u, while $G_{\ell,uv}$ measures functional overlap between units u and v.

3.2 Conditional necessity and within-layer selection

Energy alone measures how strongly a unit appears in the dense realization, but not whether its function is needed once the other unit responses are taken into account. For each admissible pruning unit $u \in \mathcal{U}_\ell$, consider representing its calibration response using all other units. The minimum residual energy is

$$\nu_{\ell,u} = G_{\ell,uu} - \mathbf{G}_{\ell,u,-u}\mathbf{G}^{-1}_{\ell,-u,-u}\mathbf{G}_{\ell,-u,u} = \frac{1}{(\mathbf{G}^{-1}_\ell)_{uu}}, \tag{10}$$

where the second equality holds when $\mathbf{G}_\ell$ is nonsingular. We call $\nu_{\ell,u}$ the *conditional necessity* of unit u: it is the squared norm of the component of its realized response outside the span of the other unit responses. Hence a unit can have large response energy but low necessity when its function is largely reproducible by the remaining units. A step-by-step Schur-complement derivation is given in Appendix A.1.

Normalization across pruning-unit sizes. A non-rotary unit contains one physical state coordinate, whereas a rotary unit contains the two real coordinates of one complex mode. Because the rotary pair is scored through its *joint* readout contribution and one pruning decision removes two physical coordinates, while a non-rotary decision removes one, we normalize the score by the physical cost of the unit. We therefore rank by conditional necessity per removed physical state coordinate,

$$s_{\ell,u} = \frac{\nu_{\ell,u}}{c_u}, \qquad c_u = |u| = \begin{cases} 1, & \text{non-rotary coordinate,} \\ 2, & \text{rotary pair.} \end{cases} \tag{11}$$

Thus a rotary unit is divided by two for ranking, not split into two independent candidates: both coordinates always receive the same unit score and are kept or removed together. This normalization makes the ordering comparable in terms of estimated distortion per removed physical state coordinate.

Budgeted set selection. For a requested layer budget b_ℓ measured in physical recurrent-state coordinates, we sort the units once in ascending order of $s_{\ell,u}$ and take the lowest-scoring units until their cumulative physical-coordinate cost reaches b_ℓ. Denoting this removal set by $\mathcal{P}_\ell(b_\ell) \subseteq \mathcal{U}_\ell$, we therefore require $\sum_{u\in\mathcal{P}_\ell(b_\ell)} c_u = b_\ell$ and consider only attainable budgets under the indivisible rotary pairs. The survivor set is then $\mathcal{R}_\ell(b_\ell) = \mathcal{U}_\ell \setminus \mathcal{P}_\ell(b_\ell)$. The ranking determines the complete pruning mask for the requested budget, and survivor compensation is solved once for that final mask. In implementation, the Gram solves used to compute conditional necessity are stabilized with a scale-normalized ridge term, replacing the relevant Gram block by $\mathbf{G} + \lambda\mathbf{I}$ where λ is very small just for numerical stability; equation 10 states the underlying unregularized quantity.

3.3 Survivor compensation

The budgeted selection step fixes the complete pruning mask and therefore the response subspace retained by the layer. We then compensate the survivors once for this final mask. For a fixed survivor set $\mathcal{R}$, define the calibration reconstruction objective

$$\mathcal{L}_\ell(\mathcal{R}, \mathbf{g}) = \frac{1}{2}\sum_{s=1}^{S}\sum_{t}\left\|\mathbf{M}^{(s)}_{\ell,t}\mathbf{1} - \mathbf{M}^{(s)}_{\ell,t,\mathcal{R}}\mathbf{g}\right\|_2^2. \tag{12}$$

Minimizing equation 12 gives

$$\boxed{\mathbf{g}^\star_{\ell,\mathcal{R}} = \mathbf{G}^{-1}_{\ell,\mathcal{R}\mathcal{R}}(\mathbf{G}_\ell\mathbf{1})_\mathcal{R}} \tag{13}$$

for the unregularized problem. This is the Euclidean projection of the dense post-projection calibration response onto the span of the surviving unit responses. Plain pruning is the feasible choice $\mathbf{g} = \mathbf{1}$; therefore the optimum in equation 13 cannot increase the calibration reconstruction error for a

fixed mask. The normal-equation proof and the corresponding fixed-mask dominance statement are provided in Appendix A.1. The implementation again adds only the small ridge term used to stabilize the solve.

Zero-overhead deployment. The compensation coefficient is one scalar g_u per pruning unit. Under the MIMO readout convention in equation 7, for every coordinate $i \in u$,

$$g_u \, \mathbf{c}_{\ell,t,i} \mathbf{h}_{\ell,t,i}^\top = (g_u \mathbf{c}_{\ell,t,i}) \mathbf{h}_{\ell,t,i}^\top. \tag{14}$$

Thus the same operation applies for every MIMO rank R, with SISO recovered at $R = 1$. In Mamba-3, the dynamic $\mathbf{C}_t$ readout is normalized at the original full state width before the surviving coordinates are gathered. We therefore apply g_u after this full-width normalization and fold it into the existing C readout scale and bias for every $i \in u$:

$$\boldsymbol{\gamma}'_i = g_u \boldsymbol{\gamma}_i, \qquad \mathbf{b}'_{C,i} = g_u \mathbf{b}_{C,i}. \tag{15}$$

The full-width normalization statistics are unchanged, and no separate gain kernel or decode-time elementwise operation is introduced. For a rotary pair, the same scalar multiplies both real coordinates, i.e., $g_u \mathbf{I}_2$, and therefore commutes with the corresponding 2×2 rotation.

3.4 Across-layer state-budget allocation

The within-layer criterion in Section 3.2 measures replaceability among responses inside one layer. Its numerical scale is not a reliable measure of how much an error at that layer changes the final model prediction. We therefore separate *local reduction*, determined by the response geometry above, from *global allocation*, determined by end-to-end predictive distortion.

Let $\mathcal{M}$ be the dense model and $\mathcal{M}_{\ell,b}$ the model formed by reducing only layer ℓ by b physical state coordinates using the normalized ranking in equation 11, selecting the corresponding removal set, and then compensating its final survivor set; every other layer is left dense. We reuse the same teacher-forced calibration sequences and compare the dense and reduced vocabulary distributions at identical token positions, $\mathbf{p}_t = \mathrm{softmax}(\mathbf{o}_t^{\mathcal{M}})$ and $\mathbf{q}_t^{(\ell,b)} = \mathrm{softmax}(\mathbf{o}_t^{\mathcal{M}_{\ell,b}})$. For the set $\mathcal{T}_{\mathrm{cal}}$ of evaluated calibration positions, we define

$$D_\ell(b) = \frac{1}{|\mathcal{T}_{\mathrm{cal}}|} \sum_{t \in \mathcal{T}_{\mathrm{cal}}} D_{\mathrm{KL}}\left(\mathbf{p}_t \,\|\, \mathbf{q}_t^{(\ell,b)}\right), D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) = \sum_v p_v \log \frac{p_v}{q_v}. \tag{16}$$

Thus $D_\ell(0) = 0$, and larger values indicate that reducing layer ℓ has a larger effect on the dense model's predictive distribution. We evaluate $D_\ell(b)$ only at budgets feasible under the pruning units above. If δ_ℓ is the next feasible removal increment, its measured marginal cost is $d_{\ell,b} = D_\ell(b + \delta_\ell) - D_\ell(b)$.

For a global removal budget B, the layerwise distortion curves define the allocation surrogate

$$\min_{\{b_\ell\}} \sum_{\ell=1}^{L} D_\ell(b_\ell) \qquad \text{s.t.} \qquad \sum_{\ell=1}^{L} b_\ell = B. \tag{17}$$

We approximately solve equation 17 by greedy water filling: starting from the dense model, assign the next feasible removal increment to the layer with the smallest measured $d_{\ell,b}$ and continue until budget B is exhausted. Layers with little effect on the predictive distribution therefore surrender more recurrent capacity, while KL-sensitive layers retain more. The additive objective in equation 17 is used only as an allocation surrogate—single-layer distortions need not add when many layers are reduced simultaneously—so the final jointly pruned model is always evaluated end-to-end.

Overall, SURCO uses the calibration set only to estimate response geometry, rank budgeted pruning units, solve one closed-form compensation problem for each final mask considered, and measure layerwise allocation costs. No downstream evaluation task is used for calibration, and no gradients or fine-tuning are required. The offline computation is dominated by the calibration rollout and small $m_\ell \times m_\ell$ Gram solves; after materialization, inference uses only the physically reduced recurrent state.

4 EXPERIMENTS

We evaluate SURCO along four axes: end-to-end quality across the Mamba-3 model family, retrieval under recurrent-state compression, cross-architecture transfer to Mamba-2, and decode efficiency after physical state compaction. We then isolate survivor compensation by applying it to fixed masks produced by several independent pruning methods.

4.1 EXPERIMENTAL SETUP

Models and baselines. We evaluate the released Mamba-3 SISO/MIMO checkpoints at 443M, and 1.5B parameters (Lahoti et al., 2026). Retrieval is evaluated at 1.5B, following the scale used for retrieval in Mamba-3. To test generality beyond Mamba-3, we additionally evaluate Mamba-2-780M (Dao & Gu, 2024). We compare against Random, LAST, AIRE-Prune, and GHOST (Gwak et al., 2024; Padhy et al., 2026; Menezes & Kyrillidis, 2026). Each baseline is reproduced with its own published score and allocation rule. All methods are one-shot and training-free. Full Mamba-3 models accuracy and PPL are shown for reference.

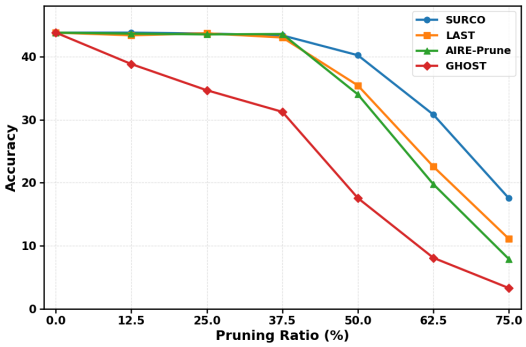


Figure 3: **Pruning Sweep:** Mamba-3 SISO 443M - LAMBADA Accuracy

Evaluation. We use the complete language-model suite reported for Mamba-3 (Lahoti et al., 2026): FineWeb-Edu perplexity, LAMBADA perplexity and accuracy, HellaSwag, PIQA, ARC-Easy, ARC-Challenge, WinoGrande, and OpenBookQA. For retrieval, we use the same benchmark families as Mamba-3: SWDE, SQuAD, FDA, TriviaQA, Natural Questions, and DROP, together with NIAH-Single-1/2/3. We report every task in the main tables rather than only an aggregate score. Our retrieval evaluation uses cloze-formatted real-world tasks and evaluates NIAH at 1K, 2K, and 4K contexts. No downstream task is used for calibration.

4.2 CALIBRATION

All methods use the same FineWeb-Edu calibration sample: eight sequences, or 16,384 tokens in total. Statistics for all methods are obtained from one shared recurrent rollout per layer. Held-out FineWeb-Edu evaluation is disjoint from calibration, and the same frozen pruned model is subsequently evaluated on every downstream and retrieval task. For Mamba-3, $d_{\text{state}} = 128$ contains 32 rotary pairs and 64 non-rotary coordinates; rotary pairs are pruned as indivisible two-coordinate units. Pruning ratios therefore always denote the fraction of *physical recurrent-state coordinates* removed.

4.3 MAIN RESULTS

Language modeling across scale, Mamba-3. Table 1 follows the language-model evaluation format of Mamba-3 (Lahoti et al., 2026) and reports the complete benchmark lm-suite at a single, challenging operating point: 50% recurrent-state removal. This keeps the comparison focused on the regime in which the pruning methods meaningfully separate. Across the nominal Mamba-3 SISO/MIMO scales (440M and 1.5B), SURCO gives the strongest pruned-model average. It is also consistently among the strongest methods in FineWeb-Edu and LAMBADA perplexity. The result is therefore stable across scale rather than being driven by one checkpoint. Also, SURCO outperforms prior work with a bigger margin on MIMO compared to SISO, making it effective method for Mamba-3. Results on Mamba-3 SISO/MIMO 187M and 888M is in Appendix A.7.

Retrieval at 1.5B, Mamba-3. Following the Mamba-3 retrieval evaluation, Table 2 reports both the six real-world retrieval datasets and the complete NIAH grid. We use 37.5% state removal, for which the full retrieval comparison is available for every baseline. The contrast with the language-model suite is pronounced. At the matched 37.5% budget, SURCO is best on all real-world retrieval datasets and on every nontrivial NIAH cell in the table. This shows SURCO is effective on long context unlike prior methods (LAST, AIRE-Prune and GHOST).

Table 1: **Downstream language modeling for Mamba-3 SISO/MIMO at 50% state removal.** Among pruned models, best results are **bolded** and second best are underlined.

Mamba-3		FW-Edu	LAMBADA		HellaS.	PIQA	ARC-E	ARC-C	WinoGr.	OBQA	Avg.
Size	Method	ppl↓	ppl↓	acc↑	acc_n↑	acc↑	acc↑	acc_n↑	acc↑	acc↑	acc↑
SISO - 440M	Full	13.57	9.56	43.3	49.9	70.3	68.9	35.8	55.7	26.2	50.02
	Random	87.17	259.62	6.2	37.5	61.6	49.5	26.8	51.5	17.6	35.90
	LAST	15.89	14.55	35.4	48.3	**69.8**	**68.1**	33.8	54.5	26.0	47.99
	AIRE-Prune	15.31	15.16	33.8	49.3	69.5	66.0	32.8	54.7	**27.2**	47.62
	GHOST	18.45	49.17	17.6	47.7	69.1	67.0	**34.8**	55.0	24.2	45.05
	SURCO	**14.32**	**11.15**	**40.3**	**49.6**	69.2	67.0	34.1	**56.2**	**27.2**	**49.09**
SISO - 1.5B	Full	11.09	6.96	48.0	55.7	73.9	76.5	42.8	61.0	29.0	55.28
	Random	19.17	36.06	18.5	48.9	70.8	69.3	32.1	52.5	22.0	44.71
	LAST	12.26	10.87	38.1	53.5	73.6	73.7	38.8	57.5	28.2	51.91
	AIRE-Prune	12.38	10.83	38.1	54.1	**74.1**	74.2	39.8	56.7	27.8	52.12
	GHOST	13.05	23.72	24.1	53.1	73.8	**76.1**	**42.1**	56.0	28.2	50.50
	SURCO	**11.73**	**9.37**	**42.5**	**55.0**	73.3	73.3	**42.1**	**59.4**	**29.8**	**53.65**
MIMO - 440M	Full	13.64	9.04	44.7	50.5	70.6	70.2	40.1	57.1	24.6	51.11
	Random	30.25	74.95	16.4	43.5	65.1	57.2	28.4	51.2	17.4	40.41
	LAST	32.80	51.86	21.0	44.3	65.8	58.6	29.4	51.4	21.4	41.70
	AIRE-Prune	34.70	64.69	18.3	44.7	66.2	58.1	30.4	52.6	21.4	41.68
	GHOST	25.34	40.90	22.6	45.7	67.4	61.4	28.8	53.4	21.0	42.90
	SURCO	**15.02**	**10.79**	**42.0**	**50.7**	**69.1**	**67.9**	**34.8**	**53.9**	**25.0**	**49.05**
MIMO - 1.5B	Full	10.83	6.79	50.4	57.1	74.0	75.8	40.5	60.9	30.6	55.61
	Random	18.68	41.32	20.4	48.7	69.7	66.7	31.4	53.6	22.6	45.41
	LAST	15.71	32.48	24.3	52.4	71.3	73.9	37.1	56.8	29.0	49.26
	AIRE-Prune	15.67	34.17	23.5	52.5	71.2	73.3	35.8	57.1	29.0	48.93
	GHOST	13.99	29.40	25.3	52.3	72.0	74.2	37.5	56.0	29.6	49.56
	SURCO	**11.47**	**8.26**	**45.8**	**56.2**	**73.7**	**76.0**	**42.5**	**59.9**	**31.6**	**55.08**

Table 2: **Retrieval at 37.5% state removal on Mamba-3 SISO-1.5B.** Dense Mamba-3 is shown for reference. Best pruned results are **bolded**, and second best are underlined.

Model (1.5B)	SWDE	SQuAD	FDA	TQA	NQ	DROP	NIAH-Single-1			NIAH-Single-2			NIAH-Single-3		
Context length							1024	2048	4096	1024	2048	4096	1024	2048	4096
Full	40.2	40.2	11.0	25.2	14.0	30.5	100.0	99.3	43.3	99.3	80.7	33.3	82.0	54.7	15.3
LAST	26.8	38.2	9.0	24.5	11.5	27.3	20.0	2.7	0.0	98.0	64.7	15.3	57.3	31.3	6.0
AIRE-Prune	24.5	37.0	9.2	24.0	11.8	27.5	3.3	0.0	0.0	97.3	60.7	14.0	57.3	30.0	6.0
GHOST	20.2	27.3	5.5	24.5	11.2	25.8	0.0	0.0	0.0	66.0	19.3	2.0	12.0	1.3	0.0
SURCO	**37.0**	**37.8**	**13.0**	**27.3**	**12.8**	**28.2**	**64.0**	**10.7**	0.0	**99.3**	**66.7**	**21.3**	**77.3**	**44.0**	**12.7**

Depth-wise pruning trajectory Interestingly, SURCO exhibits a structured depth-wise pruning trajectory as we sweep pruning ratio: it prunes later layers first(25%), then increasingly shifts pruning toward earlier layers (37.5%), while the middle layers are preserved until more pruning regime (50%) fig.4. (More details, Appendix A.9, Fig. 5).

Generalization to Mamba-2 Table 3 applies the same evaluation suite to Mamba-2-780M. SURCO remains effective despite the architectural change at 50%, SURCO retains 50.40, compared with 48.16 for GHOST and 47.4 for LAST/AIRE-Prune. Additional Results on Mamba-2 780M 25% pruning ratio is in Appendix A.7.

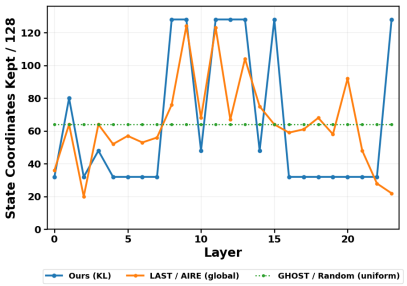


Figure 4: **Layer Allocation:** Mamba-3 SISO 1.5B at 50% state pruning budget

Table 3: **Generalization to Mamba-2-780M.** Among pruned models within each pruning budget, best results are **bolded** and second best are underlined.

Mamba-2-780M		FW-Edu	LAMBADA		HellaS.	PIQA	ARC-E	ARC-C	WinoGr.	OBQA	Avg.
Cut	Method	ppl↓	ppl↓	acc↑	acc_n↑	acc↑	acc↑	acc_n↑	acc↑	acc↑	acc↑
0%	Full	13.64	3.63	60.2	53.3	71.9	60.4	33.4	59.0	25.2	51.91
50%	Random	19.92	21.97	20.6	47.9	68.3	55.3	26.4	54.6	21.4	41.90
	LAST	16.55	10.05	36.0	**52.3**	71.2	59.5	31.1	**59.0**	22.8	47.41
	AIRE-Prune	16.55	10.02	35.9	**52.3**	71.2	59.5	31.1	**59.0**	22.8	47.40
	GHOST	14.35	6.91	41.9	51.3	**71.3**	59.5	**32.4**	57.3	23.4	48.16
	SURCO	**13.85**	**4.31**	**55.5**	52.1	71.2	**60.2**	31.4	58.6	**23.8**	**50.40**

Table 4: **Fixed-mask survivor compensation on Mamba-3 SISO-1.5B at 37.5% state removal.** For each method, the pruning mask and layer allocation are held fixed. The second row adds only SURCO's survivor compensation. Real-world retrieval uses 2048-token contexts; NIAH is evaluated at 1K, 2K, and 4K. Higher is better.

	Real-world retrieval (2K Context)							**NIAH-Single-1**				**NIAH-Single-2**				**NIAH-Single-3**			
Method	SWDE	SQuAD	FDA	TQA	NQ	DROP	Avg.	1K	2K	4K	Avg.	1K	2K	4K	Avg.	1K	2K	4K	Avg.
LAST	26.8	38.8	9.0	24.5	11.5	27.3	23.0	3.0	2.7	0.0	1.9	98.0	64.7	15.3	59.3	57.3	31.3	6.0	31.5
+ Comp.	32.2	40.0	10.5	27.3	11.8	32.2	**25.7**	4.0	3.1	0.0	2.4	98.7	72.7	25.3	65.6	68.0	39.3	9.3	38.9
AIRE-Prune	24.5	37.2	9.2	24.0	11.8	27.5	22.4	3.3	0.0	0.0	1.1	97.3	60.7	14.0	57.3	57.3	30.0	6.0	31.1
+ Comp.	32.8	39.2	11.2	25.2	12.5	31.5	**25.4**	1.3	1.3	0.0	0.9	99.3	68.0	18.0	61.8	69.3	38.7	8.7	38.9
GHOST	20.2	27.3	5.5	24.5	11.2	25.8	19.1	0.0	0.0	0.0	0.0	66.0	19.3	2.0	29.1	12.0	1.3	0.0	4.4
+ Comp.	26.8	35.0	6.0	25.8	12.8	28.0	**22.4**	0.0	0.0	0.0	0.0	86.7	46.7	6.0	46.5	31.3	10.0	1.3	14.2
SURCO	30.5	36.0	11.0	24.2	10.5	28.0	23.4	41.3	5.1	0.0	15.1	97.3	60.0	16.7	58.0	74.0	42.0	12.0	42.7
+ Comp.	37.0	37.0	13.0	27.3	12.8	28.2	**25.9**	64.0	10.7	0.0	24.9	99.3	66.7	21.3	62.4	77.3	44.0	12.7	44.7

Ablation: survivor compensation The central mechanism can be isolated without changing the pruning mask or layer budget. For each selector, we take the method exactly as produced by its own score and allocation rule and then add only our closed-form survivor compensation. Table 4 reports the fully populated matched-mask experiment at 37.5% state removal on Mamba-3 SISO-1.5B. Compensation improves both real-world recall and prose NIAH for every selector, showing that the effect is complementary to state selection rather than a consequence of choosing a different mask. This implies SURCO is not limited to its own method but our contribution also enhances the performance of other methods.

Hardware efficiency We finally materialize the pruned Mamba-3 SISO/MIMO models so that the recurrent state is physically smaller rather than masked. Measurements use one NVIDIA H100 80GB, bfloat16 weights, CUDA graphs, and greedy decoding. State pruning is primarily a serving-batch optimization: the benefit grows as per-sequence recurrent-state traffic becomes a larger fraction of memory traffic. At 50% state removal,at batch 256, for SISO throughput increases by $1.72\times$ on 443M and $1.86\times$ on 1.5B, while peak allocated memory falls by 38% and 36%, respectively. Similarily, for MIMO throughput increases by $1.67\times$ for 444M and $1.78\times$ for 1.5B.

5 Conclusion

We introduced SURCO, a training-free framework for recurrent-state pruning that separates *which states are retained* from *how the retained states are realized to improve performance at fixed state pruning ratio*. By combining response-space state selection, layer-wise budget allocation, and closed-form survivor compensation, SURCO improves quality at a fixed pruning budget without adding decode-time operations. Experiments across Mamba-3 SISO/MIMO and Mamba-2 show that SURCO consistently outperforms prior works at fixed pruning masks, while physically reducing the recurrent state preserves the expected memory and throughput benefits of state pruning.

REFERENCES

Raktim Bhattacharya. An exact instrument for state usage in selective state-space models, and the input-driven migration it reveals. *arXiv preprint arXiv:2607.11796*, 2026. doi: 10.48550/arXiv.2607.11796. URL https://arxiv.org/abs/2607.11796.

Yingfa Chen, Xinrong Zhang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Stuffed mamba: State collapse and state capacity of RNN-based long-context modeling. *arXiv preprint arXiv:2410.07145*, 2024. URL https://arxiv.org/abs/2410.07145.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Tamer Ghattas, Michael Hassid, and Roy Schwartz. On pruning state-space LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18800–18814. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.950. URL https://aclanthology.org/2025.emnlp-main.950/.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Minseon Gwak, Seongrok Moon, Joohwan Ko, and PooGyeon Park. Layer-adaptive state pruning for deep state space models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0339. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/14730e0dd6ac1c4a5765310909fd51b1-Abstract-Conference.html.

Aakash Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8abd2043b71a074278d5f687947bff9c-Abstract-Conference.html.

Michael Menezes and Anastasios Kyrillidis. GHOST: Unmasking phantom states in mamba2 via grouped hidden-state output-aware selection & truncation. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.11408.

Juan Pablo Muñoz, Jinjie Yuan, and Nilesh Jain. Mamba-shedder: Post-transformer compression for efficient selective structured state space models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3851–3863. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.195. URL https://aclanthology.org/2025.naacl-long.195/.

Apurba Prasad Padhy, Fernando Camacho, and Saibal Mukhopadhyay. AIRE-prune: Asymptotic impulse-response energy for state pruning in state space models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JLUxTl1Um6.

Chen Tianqi, Yuanteng Chen, Peisong Wang, Weixiang Xu, Zeyu Zhu, and Jian Cheng. Q-mamba: Towards more efficient mamba models via post-training quantization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10594–10610. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.551. URL https://aclanthology.org/2025.findings-acl.551/.

AI USE STATEMENT

In this work, we used generative AI tools for editing and polishing the text, as well as for assisting with software implementation. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for mathematical proofs, assist in writing proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative or thematic data analysis, or interpret experimental results.

The use of generative AI for generating synthetic datasets, translation, or dataset cleaning and reformatting was not applicable to this work. We additionally used generative AI tools to create or edit software code, summarize or analyze existing literature, and improve the readability and presentation of the manuscript.

All AI-assisted work was reviewed by the authors. LLM-generated code was independently verified and tested for correctness, and all AI-assisted manuscript refinements were checked for technical correctness and accuracy. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

A APPENDIX

A.1 DERIVATIONS FOR SURCO

This appendix expands the algebra used in Section 3. We state the derivations for one layer and suppress the layer index ℓ for readability. All results use the post-output-projection calibration responses from Section 3.1, so the reconstruction norm is Euclidean.

A.2 CALIBRATION LEAST SQUARES AND THE GRAM MATRIX

At calibration position (s,t) let $\mathbf{M}_t^{(s)} \in \mathbb{R}^{D\times m}$ contain the per-unit responses and let $\mathbf{y}_t^{(s)} = \mathbf{M}_t^{(s)}\mathbf{1}$. For survivor set $\mathcal{R}$, the reconstruction objective is

$$\mathcal{L}(\mathcal{R},\mathbf{g}) = \frac{1}{2}\sum_{s,t}\left\|\mathbf{M}_t^{(s)}\mathbf{1} - \mathbf{M}_{t,\mathcal{R}}^{(s)}\mathbf{g}\right\|_2^2. \tag{18}$$

Expanding the quadratic and defining $\mathbf{G} = \sum_{s,t}(\mathbf{M}_t^{(s)})^\top\mathbf{M}_t^{(s)}$ gives

$$\mathcal{L}(\mathcal{R},\mathbf{g}) = \text{const} - \mathbf{g}^\top(\mathbf{G}\mathbf{1})_\mathcal{R} + \frac{1}{2}\mathbf{g}^\top\mathbf{G}_{\mathcal{R}\mathcal{R}}\mathbf{g}. \tag{19}$$

Thus, for the response-reconstruction problem, the accumulated calibration responses enter the algebra only through the small Gram matrix.

A.3 CONDITIONAL NECESSITY

To measure the necessity of unit u, write the Gram matrix after permuting u to the first position as

$$\mathbf{G} = \begin{bmatrix} G_{uu} & \mathbf{G}_{u,-u} \\ \mathbf{G}_{-u,u} & \mathbf{G}_{-u,-u} \end{bmatrix}.$$

Approximating the response of unit u by a linear combination $\mathbf{a}$ of the remaining responses gives residual energy

$$\frac{1}{2}\sum_{s,t}\left\|\mathbf{m}_{t,u}^{(s)} - \mathbf{M}_{t,-u}^{(s)}\mathbf{a}\right\|_2^2 = \frac{1}{2}\left(G_{uu} - 2\mathbf{a}^\top\mathbf{G}_{-u,u} + \mathbf{a}^\top\mathbf{G}_{-u,-u}\mathbf{a}\right). \tag{20}$$

The minimizer is $\mathbf{a}^\star = \mathbf{G}_{-u,-u}^{-1}\mathbf{G}_{-u,u}$, yielding

$$\nu_u = G_{uu} - \mathbf{G}_{u,-u}\mathbf{G}_{-u,-u}^{-1}\mathbf{G}_{-u,u}. \tag{21}$$

By the block-inverse identity, the Schur complement satisfies $(\mathbf{G}^{-1})_{uu} = 1/\nu_u$, hence $\nu_u = 1/(\mathbf{G}^{-1})_{uu}$ when $\mathbf{G}$ is nonsingular.

A.4 NORMALIZATION OF MIXED PRUNING-UNIT SIZES

The conditional necessity ν_u is computed for an admissible pruning unit. For a non-rotary unit this is one real state coordinate, whereas for a rotary unit the response is the joint contribution of the two real coordinates of one complex mode. Let $c_u = |u| \in \{1, 2\}$ denote this physical-coordinate cost. SURCO therefore uses

$$s_u = \frac{\nu_u}{c_u} \tag{22}$$

for within-layer ranking. In particular, a rotary pair uses $s_u = \nu_u/2$. This does not treat its two coordinates as independent candidates: the pair is represented by one column in the unit-response geometry, receives one score, and is always retained or removed jointly. Dividing by c_u expresses the estimated irreducible response energy per physical state coordinate, making the ranking comparable between one-coordinate and two-coordinate units.

For a requested physical-coordinate budget b, units are ordered by increasing s_u and the lowest-scoring units are selected until the corresponding attainable cumulative cost equals b. This produces the complete removal set $\mathcal{P}(b)$ and survivor set $\mathcal{R}(b) = \mathcal{U} \setminus \mathcal{P}(b)$. No compensation solve is performed between individual selections.

A.5 CLOSED-FORM SURVIVOR COMPENSATION

Differentiating equation 19 with respect to $\mathbf{g}$ gives

$$\nabla_{\mathbf{g}} \mathcal{L} = \mathbf{G}_{\mathcal{R}\mathcal{R}}\mathbf{g} - (\mathbf{G}\mathbf{1})_{\mathcal{R}}.$$

Setting the gradient to zero yields

$$\mathbf{g}^{\star}_{\mathcal{R}} = \mathbf{G}^{-1}_{\mathcal{R}\mathcal{R}}(\mathbf{G}\mathbf{1})_{\mathcal{R}}. \tag{23}$$

Since plain pruning corresponds to the feasible point $\mathbf{g} = \mathbf{1}$, optimal compensation satisfies

$$\mathcal{L}(\mathcal{R}, \mathbf{g}^{\star}_{\mathcal{R}}) \leq \mathcal{L}(\mathcal{R}, \mathbf{1}).$$

This guarantee concerns the unregularized calibration reconstruction objective. The implementation adds a scale-normalized ridge term only to stabilize the linear solve.

A.6 WHY NO ADDITIONAL OUTPUT METRIC IS NEEDED

Suppose a pre-projection unit response is $\mathbf{r}_u$ and the monitored output is obtained through a linear map $\mathbf{P}$, so that $\mathbf{m}_u = \mathbf{P}\mathbf{r}_u$. Computing the Gram matrix at the monitored point gives

$$\mathbf{m}_u^{\top}\mathbf{m}_v = \mathbf{r}_u^{\top}\mathbf{P}^{\top}\mathbf{P}\mathbf{r}_v.$$

Thus a weighted pre-projection geometry with $\mathbf{Q} = \mathbf{P}^{\top}\mathbf{P}$ is exactly equivalent to ordinary Euclidean geometry after projection. SURCO records the responses after the output projection directly, so the main text uses $\mathbf{G} = \sum \mathbf{M}^{\top}\mathbf{M}$ and does not introduce an additional $\mathbf{Q}$.

A.7 ADDITIONAL MAMBA-3 RESULTS

Table 5 & 6 have the complete Mamba-3 result across 4 model sizes for the FineWeb-Edu and lm-harness benchmarks

A.8 ADDITIONAL MAMBA-2 RESULTS

A.9 LAYER-WISE PRUNING ALLOCATION

Figure 5 shows how the retained state budget is distributed across layers of SISO-1.5B as pruning increases. A notable pattern emerges for SURCO: pruning is initially concentrated in the later layers, then progressively extends toward the earlier layers, while the middle layers remain comparatively protected until the most aggressive pruning regime. This structured progression differs from the approximately uniform allocation of random pruning and indicates a clear depth-dependent variation in state redundancy.

Table 5: **Downstream language modeling at 50% state removal.** Full Mamba-3 models are shown for reference. Random denotes uniform random state removal (median over three seeds). Baselines are LAST (Gwak et al., 2024), AIRE-Prune (Padhy et al., 2026), and GHOST (Menezes & Kyrillidis, 2026). Among pruned models, best results are **bolded** and second best are underlined.

Mamba-3 SISO		FW-Edu	LAMBADA		HellaS.	PIQA	ARC-E	ARC-C	WinoGr.	OBQA	Avg.
Size	Method	ppl↓	ppl↓	acc↑	acc_n↑	acc↑	acc↑	acc_n↑	acc↑	acc↑	acc↑
	Full	17.18	17.05	33.7	44.7	67.8	62.5	31.8	52.8	21.2	44.93
180M	Random	1,065.84	26,685	0.0	27.7	53.8	27.2	26.4	50.7	16.4	28.95
	LAST	22.20	25.65	29.0	**44.0**	66.5	60.4	26.8	50.9	**21.2**	42.67
	AIRE-Prune	22.32	34.00	26.1	43.1	66.0	60.7	29.8	50.1	20.6	42.34
	GHOST	23.40	27.89	29.6	43.2	66.6	59.6	**30.1**	51.6	**21.2**	43.13
	SURCO	**18.96**	**19.62**	**31.4**	43.2	**66.9**	**61.8**	**30.1**	**52.5**	20.8	**43.80**
	Full	13.57	9.56	43.3	49.9	70.3	68.9	35.8	55.7	26.2	50.02
440M	Random	87.17	259.62	6.2	37.5	61.6	49.5	26.8	51.5	17.6	35.90
	LAST	15.89	14.55	35.4	48.3	**69.8**	**68.1**	33.8	54.5	26.0	47.99
	AIRE-Prune	15.31	15.16	33.8	49.3	69.5	66.0	32.8	54.7	**27.2**	47.62
	GHOST	18.45	49.17	17.6	47.7	69.1	67.0	**34.8**	55.0	24.2	45.05
	SURCO	**14.32**	**11.15**	**40.3**	**49.6**	69.2	67.0	34.1	**56.2**	**27.2**	**49.09**
	Full	11.89	7.57	46.5	54.8	73.9	74.7	39.5	57.1	29.6	53.73
880M	Random	25.64	53.07	15.8	44.0	67.2	61.4	29.4	52.3	19.0	41.50
	LAST	13.23	**9.40**	**43.2**	52.3	72.6	72.3	37.1	**56.3**	29.0	51.81
	AIRE-Prune	13.18	11.77	37.7	52.1	72.3	**72.8**	37.1	55.8	29.2	51.01
	GHOST	14.16	23.48	24.9	51.8	**72.7**	72.6	**38.1**	54.9	28.4	49.06
	SURCO	**12.65**	9.59	41.2	**53.6**	72.4	72.6	36.8	55.3	**30.6**	**51.97**
	Full	11.09	6.96	48.0	55.7	73.9	76.5	42.8	61.0	29.0	55.28
1.5B	Random	19.17	36.06	18.5	48.9	70.8	69.3	32.1	52.5	22.0	44.71
	LAST	12.26	10.87	38.1	53.5	73.6	73.7	38.8	57.5	28.2	51.91
	AIRE-Prune	12.38	10.83	38.1	54.1	**74.1**	74.2	39.8	56.7	27.8	52.12
	GHOST	13.05	23.72	24.1	53.1	73.8	**76.1**	**42.1**	56.0	28.2	50.50
	SURCO	**11.73**	**9.37**	**42.5**	**55.0**	73.3	73.3	**42.1**	**59.4**	**29.8**	**53.65**

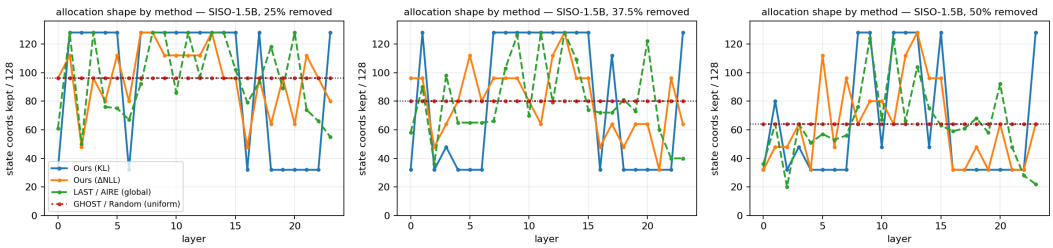


Figure 5: **Layer-wise state allocation for SISO-1.5B.** Number of retained state coordinates per layer at 25%, 37.5%, and 50% pruning. The dotted line denotes uniform allocation under the same global budget. SURCO exhibits a structured depth-wise progression: pruning is concentrated first in later layers, then extends toward earlier layers, while the middle layers remain comparatively protected as the pruning budget increases.

Table 6: **Downstream language modeling for Mamba-3 MIMO at 50% state removal.** Full models are shown for reference. Random denotes uniform random state removal (median over three seeds). Baselines are LAST (Gwak et al., 2024), AIRE-Prune (Padhy et al., 2026), and GHOST (Menezes & Kyrillidis, 2026). Among pruned models, best results are **bolded** and second best are underlined.

Mamba-3 MIMO		FW-Edu	LAMBADA		HellaS.	PIQA	ARC-E	ARC-C	WinoGr.	OBQA	Avg.
Size	Method	ppl↓	ppl↓	acc↑	acc_n↑	acc↑	acc↑	acc_n↑	acc↑	acc↑	acc↑
180M	Full	16.97	16.17	34.1	45.7	68.2	62.3	29.4	49.4	22.2	44.47
	Random	55.55	454.19	5.1	35.3	60.2	47.4	23.4	49.9	14.8	33.72
	LAST	35.81	145.99	10.3	41.7	63.9	55.8	27.4	48.3	17.6	37.85
	AIRE-Prune	33.01	148.19	10.1	42.6	64.7	56.8	26.1	50.7	17.6	38.37
	GHOST	30.67	48.67	20.3	43.3	64.9	57.4	**28.8**	50.7	17.2	40.36
	SURCO	**20.32**	**21.79**	**30.8**	**44.3**	**67.0**	**60.2**	27.1	**50.9**	**19.8**	**42.87**
440M	Full	13.64	9.04	44.7	50.5	70.6	70.2	40.1	57.1	24.6	51.11
	Random	30.25	74.95	16.4	43.5	65.1	57.2	28.4	51.2	17.4	40.41
	LAST	32.80	51.86	21.0	44.3	65.8	58.6	29.4	51.4	21.4	41.70
	AIRE-Prune	34.70	64.69	18.3	44.7	66.2	58.1	30.4	52.6	21.4	41.68
	GHOST	25.34	40.90	22.6	45.7	67.4	61.4	28.8	53.4	21.0	42.90
	SURCO	**15.02**	**10.79**	**42.0**	**50.7**	**69.1**	**67.9**	**34.8**	**53.9**	**25.0**	**49.05**
880M	Full	11.70	7.46	47.8	55.9	73.8	74.4	43.5	59.2	29.0	54.81
	Random	24.96	41.35	20.2	46.9	68.4	60.4	28.8	53.0	19.2	42.97
	LAST	22.22	55.80	18.0	50.1	70.5	64.0	30.1	54.8	23.6	44.43
	AIRE-Prune	23.84	68.61	15.8	48.7	70.3	62.8	31.4	55.1	23.4	43.94
	GHOST	18.71	41.15	20.2	50.9	70.7	66.0	32.1	53.1	24.0	45.28
	SURCO	**12.81**	**9.63**	**42.6**	**54.3**	**73.4**	**73.0**	**40.8**	**57.5**	**27.6**	**52.75**
1.5B	Full	10.83	6.79	50.4	57.1	74.0	75.8	40.5	60.9	30.6	55.61
	Random	18.68	41.32	20.4	48.7	69.7	66.7	31.4	53.6	22.6	45.41
	LAST	15.71	32.48	24.3	52.4	71.3	73.9	37.1	56.8	29.0	49.26
	AIRE-Prune	15.67	34.17	23.5	52.5	71.2	73.3	35.8	57.1	29.0	48.93
	GHOST	13.99	29.40	25.3	52.3	72.0	74.2	37.5	56.0	29.6	49.56
	SURCO	**11.47**	**8.26**	**45.8**	**56.2**	**73.7**	**76.0**	**42.5**	**59.9**	**31.6**	**55.08**

Table 7: **Generalization to Mamba-2-780M.** Among pruned models within each pruning budget, best results are **bolded** and second best are underlined.

Mamba-2-780M		FW-Edu	LAMBADA		HellaS.	PIQA	ARC-E	ARC-C	WinoGr.	OBQA	Avg.
Cut	Method	ppl↓	ppl↓	acc↑	acc_n↑	acc↑	acc↑	acc_n↑	acc↑	acc↑	acc↑
0%	Full	13.64	3.63	60.2	53.3	71.9	60.4	33.4	59.0	25.2	51.91
25%	Random	20.88	6.37	44.9	50.7	71.8	57.0	31.8	57.0	22.8	48.21
	LAST	208.99	15.03	43.9	51.9	70.5	**60.9**	31.4	58.5	24.2	48.76
	AIRE-Prune	208.99	15.03	43.9	51.9	70.5	60.7	31.4	58.5	24.2	48.74
	GHOST	13.78	4.61	52.6	52.6	71.2	60.2	**34.4**	**59.4**	23.6	50.57
	SURCO	**13.66**	**3.67**	**59.6**	**53.1**	**71.9**	60.2	32.8	59.2	**25.0**	**51.68**
50%	Random	19.92	21.97	20.6	47.9	68.3	55.3	26.4	54.6	21.4	41.90
	LAST	16.55	10.05	36.0	**52.3**	71.2	59.5	31.1	**59.0**	22.8	47.41
	AIRE-Prune	16.55	10.02	35.9	**52.3**	71.2	59.5	31.1	**59.0**	22.8	47.40
	GHOST	14.35	6.91	41.9	51.3	**71.3**	59.5	**32.4**	57.3	23.4	48.16
	SURCO	**13.85**	**4.31**	**55.5**	52.1	71.2	**60.2**	31.4	58.6	**23.8**	**50.40**