# Training-Aware, Calibration-Free Model Reduction for Selective State Space Models

**Anonymous authors**
Paper under double-blind review

## Abstract

Modeling long sequences, such as text and audio, requires computationally efficient architectures; Mamba addresses this need using selective state-space models (SSMs). However, two structural dimensions intrinsic to these selective SSMs—the state order and the step-size scheduling dimension—directly affect per-step cost and parameter count. We introduce a training-aware framework that uses structural regularization during full-order training to enable joint reduction of both dimensions. A per-layer finite-horizon output-error bound motivates a scale-invariant, decay-aware state score and a factorized low-rank surrogate for the step-size map. After training, we perform one-shot reduction using state selection and truncated SVD to construct a model with the requested dimensions from the learned parameters alone, without calibration data or post-compression fine-tuning. On Selective Copying, SC09, and The Pile, the jointly reduced models outperform vanilla post-hoc reduction in reported point estimates at matched reduction levels. In isolated state reduction, our calibration-free method outperforms the adapted calibration-based baselines on Selective Copying and SC09 and remains competitive on The Pile.

## 1 Introduction

Deep state space models (deep SSMs), including S4 (Gu et al., 2022b), S4D (Gu et al., 2022a), S5 (Smith et al., 2023), and Mamba (Gu & Dao, 2024), process sequences with a computational cost linear in the sequence length, unlike standard self-attention (Vaswani et al., 2017). Among these, Mamba introduces a selective SSM whose coefficients depend on the input, enabling selective retention and forgetting while preserving efficient recurrent computation (Gu & Dao, 2024). Mamba-based models have been applied in control (Li et al., 2025), genomics (Schiff et al., 2024), and chemistry (Brazil et al., 2024). Subsequent variants, Mamba-2 (Dao & Gu, 2024) and Mamba-3 (Lahoti et al., 2026), further develop this line through structured SSM formulations and algorithms.

The selective SSM is a central sequence-processing component of Mamba. Despite Mamba's linear scaling with sequence length, the cost of this component still depends on its internal structural dimensions. In the Mamba-1 parameterization (Gu & Dao, 2024), two such dimensions are the state order $N$, which determines the number of recurrent states per channel, and the step-size scheduling dimension $R$, commonly called `dt_rank` in Mamba-1 implementations, which determines the factorization width used to generate the input-dependent step sizes. Large $N$ or $R$ therefore increases the structural parameter count and per-step computational cost, motivating their reduction.

Existing state-reduction methods for deep SSMs use criteria based on linear time-invariant (LTI) models or calibration statistics, and some do not explicitly model Mamba's input-dependent dynamics (Padhy et al., 2026; Chahine et al., 2026; Menezes & Kyrillidis, 2026). To our knowledge, these methods do not provide a parameter-only framework that directly accounts for these dynamics, jointly reduces both dimensions, is calibration-free at reduction time, and requires no post-reduction optimization.

Focusing on the Mamba-1 parameterization of the selective SSM (Gu & Dao, 2024), hereafter referred to simply as Mamba, we propose a training-aware framework that uses structural regularization during full-order training to facilitate subsequent reduction. After training, the framework constructs reduced models from learned parameters alone; the reduction is calibration-free at reduction time and



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.

requires no post-reduction fine-tuning. Our output-error analysis separates an input-dependent factor from an input-independent, model-dependent structural term, as illustrated in Figure 1, and uses the latter to derive parameter-only reduction criteria. State reduction retains the state coordinates with the largest parameter-only scores, which are scale-invariant and account for both input–output coupling and the decay rates of the diagonal state dynamics. Reduction of the step-size scheduling dimension uses truncated SVD of the learned linear step-size map.

Our contributions are summarized as follows:

- We formulate the joint reduction of the state order $N$ and step-size scheduling dimension $R$, and develop a training-aware framework that constructs reduced selective SSMs from learned parameters alone; the reduction is calibration-free at reduction time and requires no post-reduction fine-tuning.
- An output-error analysis that separates an input-dependent factor from model-dependent structural terms guides our design of two structure-inducing surrogates: a scale-invariant coupling-and-dynamics regularizer for state reduction and a factorized low-rank surrogate for the step-size map.
- We evaluate the proposed framework on Selective Copying (Gu & Dao, 2024), SC09 (Warden, 2018), and language modeling on The Pile (Gao et al., 2020), showing that the jointly reduced models outperform vanilla post-hoc reduction at matched reduction levels and that the proposed state reduction is competitive with our post-hoc adaptation of CompreSSM and our Mamba-1 adaptation of GHOST.

## 2 Related Work

### 2.1 General Compression Methods for Deep SSMs

Compression methods for deep models include pruning, quantization, and low-rank factorization (Cheng et al., 2018). These methods have also been adapted to deep SSMs. Q-S5 studies quantization for S5 (Abreu et al., 2024), MambaQuant develops post-training quantization for Mamba (Xu et al., 2025), and Mamba-Shedder removes redundant architectural components (Muñoz et al., 2025). However, these approaches do not directly reduce the state order or step-size scheduling dimension of Mamba selective SSMs, which are the structural dimensions considered in this work.

### 2.2 MOR for Deep SSMs

Model order reduction (MOR) is well established for LTI systems and has been extended to linear time-varying (LTV), linear parameter-varying (LPV), parametric, and nonlinear settings (Antoulas et al., 2001; Antoulas, 2005; Gugercin et al., 2008; Lang et al., 2016; Benner et al., 2015; 2019; Baur et al., 2014). Recent deep-SSM work includes post-training balanced or modal reduction, $\mathcal{H}^2$-based reduction, and training-time methods that promote reducible representations (Ezoe & Sato, 2024; Gwak et al., 2024; Bittner et al., 2025; Padhy et al., 2026; Sakamoto & Sato, 2025a;b; Forgione et al., 2024; Schwerdtner et al., 2025; Chahine et al., 2026). These formulations mainly target fixed LTI dynamics, as in S4/S5-type models (Gu et al., 2022b; Smith et al., 2023).

In a Mamba selective SSM, the step size and the input and output coefficients depend on the input (Gu & Dao, 2024). Thus, for a fixed input sequence, the model can be viewed as a finite-horizon LTV system. However, the overall input–output map remains nonlinear because the system coefficients themselves change with the input. Consequently, applying conventional LTV or LPV model reduction typically requires either representative input trajectories or a prescribed scheduling domain (Lang et al., 2016; Benner et al., 2015). Existing approaches for reducing selective SSMs differ in how they handle input dependence and whether they require calibration data. AIRE-Prune derives its pruning criterion from an impulse-response energy formulation for LTI systems and does not explicitly account for Mamba's input-dependent LTV/LPV dynamics (Padhy et al., 2026). CompreSSM uses input-averaged LTI surrogates constructed from calibration data to guide reduction (Chahine et al., 2026), while other approaches assess state importance using observed input trajectories or empirical statistics (Menezes & Kyrillidis, 2026; Bhattacharya, 2026). To the best of our knowledge, no prior method both explicitly accounts for this input-dependent structure and reduces Mamba's state order without requiring calibration data at reduction time or post-reduction optimization.

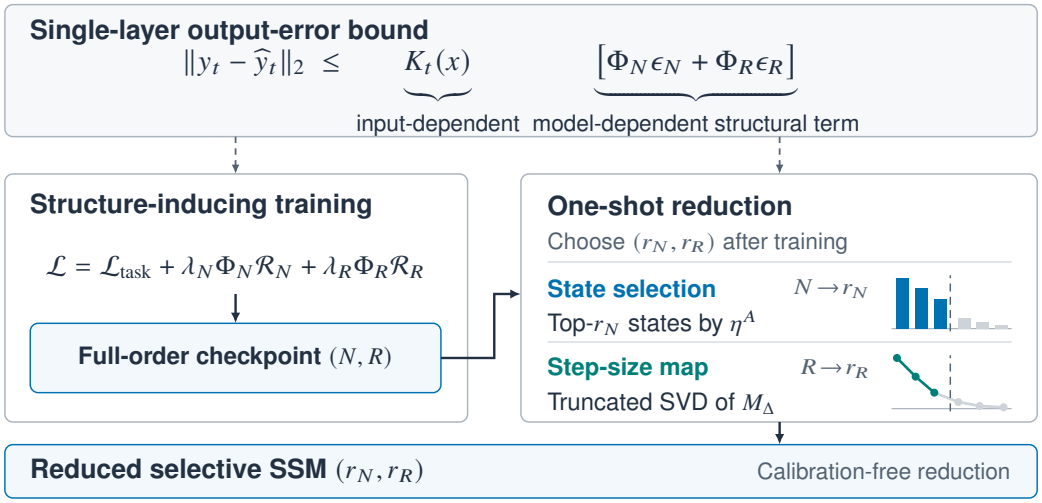



Figure 1: Overview of the proposed framework. The single-layer error bound guides structural regularization and one-shot reduction. After training, one checkpoint yields reduced models at chosen dimensions $(r_N, r_R)$; the reduction is calibration-free at reduction time and requires no post-reduction fine-tuning. Dashed arrows denote design guidance; solid arrows denote the workflow.

## 3 Preliminaries

We consider the selective SSM within a Mamba block (Gu & Dao, 2024). In particular, we focus on the selective SSM core, since the surrounding convolution, gating, and projection layers are unchanged by our reduction. Let $x_t \in \mathbb{R}^D$ denote its input at time step $t$, where $D$ is the SSM channel dimension. The step size is generated by $\rho_t = W_\rho x_t$ and $\Delta_t = \text{softplus}(W_\Delta \rho_t + b_\Delta)$, with softplus applied elementwise. The other input-dependent vectors are $B_t = W_B x_t$ and $C_t = W_C x_t$. Here $\rho_t \in \mathbb{R}^R$, $W_\rho \in \mathbb{R}^{R \times D}$, $W_\Delta \in \mathbb{R}^{D \times R}$, and $W_B, W_C \in \mathbb{R}^{N \times D}$. For each channel $d = 1, \ldots, D$, the selective SSM evolves as

$$\begin{cases} h_{t,d} = \bar{A}_{t,d} h_{t-1,d} + \bar{B}_{t,d} x_{t,d}, \\ y_{t,d} = C_t^\top h_{t,d} + D_d^{\text{skip}} x_{t,d}. \end{cases} \tag{1}$$

where $h_{t,d}, B_t, C_t \in \mathbb{R}^N$ and $h_{0,d} = 0$. The scalar $D_d^{\text{skip}} \in \mathbb{R}$ is a learned skip coefficient. The continuous-time state matrix is $A_d = \text{diag}(a_{d,1}, \ldots, a_{d,N}) \in \mathbb{R}^{N \times N}$ with $a_{d,n} < 0$. In the Mamba-1 implementation, $\bar{A}_{t,d} = \exp(\Delta_{t,d} A_d)$ and $\bar{B}_{t,d} = \Delta_{t,d} B_t$. The expression for $\bar{B}_{t,d}$ is the first-order approximation to the zero-order-hold formula stated in Gu & Dao (2024).

We target two internal dimensions for reduction: the state order $N$, which determines the number of dynamical states per channel, and the step-size scheduling dimension $R$, which is the dimension of $\rho_t$. With $M_\Delta := W_\Delta W_\rho \in \mathbb{R}^{D \times D}$, we have $\Delta_t = \text{softplus}(M_\Delta x_t + b_\Delta)$ and $\text{rank}(M_\Delta) \leq R$. Thus, $R$ specifies the factorization dimension used to generate $\Delta_t$ and upper bounds the rank of $M_\Delta$. Hereafter, "step-size map" denotes the linear map $x_t \mapsto M_\Delta x_t$, before bias addition and softplus.

For a sequence of length $L$, Mamba evaluates the input-dependent recurrence using selective scan, with a computational cost that is linear in $L$ (Gu & Dao, 2024). The state dynamics and the generation of $B_t$ and $C_t$ require $O(LDN)$ operations, while the factorized map $D \to R \to D$ that generates $\Delta_t$ requires $O(LDR)$ operations. Hence, the total computational cost of these structural components scales as $O(LD(N + R))$. The corresponding number of parameters in these structural components, excluding bias and skip terms, is $P_{\text{struct}}(N, R) = D(3N+2R)$. Reducing $N$ decreases the state-related cost, whereas reducing $R$ decreases the cost of generating $\Delta_t$. Other components of a Mamba block, such as the input and output projections and local convolution, are unaffected by these reductions.

We therefore aim to reduce the computational cost and the parameter count of a full-order Mamba model $\mathcal{M}_{N,R}(\theta)$, where $\theta$ denotes its trainable parameters, while preserving its selective SSM structure and minimizing degradation in task performance. In particular, for prescribed target dimensions $(r_N, r_R)$, we construct a reduced model $\widehat{\mathcal{M}}_{r_N, r_R}$ from the learned parameters alone; the reduction is calibration-free at reduction time and requires no post-reduction optimization, including fine-tuning.

# 4 Structure-Inducing Model Reduction for Mamba Selective SSMs

Building on the selective SSM formulation in Section 3, we first define reduced-order selective SSMs with target dimensions $(r_N, r_R)$. We then derive a joint finite-horizon output-error bound and use its structural terms to design training surrogates. After training, these surrogates enable one-shot reduction to the chosen target dimensions; the reduction is calibration-free at reduction time and requires no post-reduction fine-tuning.

## 4.1 Reduced-Order Selective SSMs

Let $1 \le r_N \le N$ and $1 \le r_R \le \min\{R, D\}$. Let $\mathcal{S} \subseteq \{1, \dots, N\}$ be a retained state set with $|\mathcal{S}| = r_N$. We denote its selection matrix by $P_{\mathcal{S}} \in \mathbb{R}^{r_N \times N}$; its rows are the standard basis vectors indexed by $\mathcal{S}$, so left multiplication by $P_{\mathcal{S}}$ extracts the retained coordinates. Thus, we reduce the state order via

$$\widehat{A}_d = P_{\mathcal{S}} A_d P_{\mathcal{S}}^{\top}, \qquad \widehat{W}_B = P_{\mathcal{S}} W_B, \qquad \widehat{W}_C = P_{\mathcal{S}} W_C. \tag{2}$$

Furthermore, for the linear step-size map, let $\widehat{M}_\Delta \in \mathbb{R}^{D \times D}$ satisfy

$$\operatorname{rank}(\widehat{M}_\Delta) \le r_R, \qquad \widehat{M}_\Delta = \widehat{W}_\Delta \widehat{W}_\rho, \tag{3}$$

where $\widehat{W}_\Delta \in \mathbb{R}^{D \times r_R}$ and $\widehat{W}_\rho \in \mathbb{R}^{r_R \times D}$. The reduced step size is then $\widehat{\Delta}_t = \operatorname{softplus}\left(\widehat{M}_\Delta x_t + b_\Delta\right)$.

Substituting $(\widehat{A}_d, \widehat{W}_B, \widehat{W}_C, \widehat{W}_\Delta, \widehat{W}_\rho)$ for $(A_d, W_B, W_C, W_\Delta, W_\rho)$ in equation 1 gives the reduced selective SSM. State truncation preserves the diagonal transition structure, while the low-rank factors preserve the input-dependent parameterization of the step-size map. Consequently, the reduced model retains the structure required by Mamba's selective scan and can be evaluated using the same algorithm as the full-order model.

For selective SSMs, the computational cost decreases from $O(LD(N + R))$ to $O(LD(r_N + r_R))$. The corresponding number of structural parameters decreases from $P_{\text{struct}}(N, R)$ to $P_{\text{struct}}(r_N, r_R) = D(3r_N + 2r_R)$, excluding order-independent biases and skip parameters.

Since state truncation preserves $a_{d,n} < 0$ and softplus ensures $\widehat{\Delta}_{t,d} > 0$, the reduced selective SSM remains pointwise Schur stable; Appendix C.1 provides the formal statement and proof.

## 4.2 Joint Reduction Error Analysis

To identify which states and rank components can be removed while controlling the output error, we compare the full-order selective SSM with the reduced model defined in Section 4.1. Specifically, we derive an upper bound on the difference between their outputs $y_t = (y_{t,d})_{d=1}^D$ and $\widehat{y}_t = (\widehat{y}_{t,d})_{d=1}^D$ when both models receive the same input sequence $x_1, \dots, x_L$.

For each state $n$, the error bound below motivates the state importance score

$$\eta_n^A(\tau) := (1 + \beta_n/\tau) q_n, \tag{4}$$

where $q_n := \|(W_B)_{n,:}\|_2 \|(W_C)_{n,:}\|_2$, $\beta_n := \left(\sum_{d=1}^D |a_{d,n}|^{-2}\right)^{1/2}$, and $\tau > 0$ is a fixed hyperparameter. Let

$$\epsilon_N(\mathcal{S}; \tau) := \sum_{n \notin \mathcal{S}} \eta_n^A(\tau), \qquad \epsilon_R(\widehat{M}_\Delta) := \|M_\Delta - \widehat{M}_\Delta\|_F \tag{5}$$

denote the state-reduction residual and the rank-reduction residual, respectively.

**Theorem 1** (Joint structural reduction error). *Let $x_1, \dots, x_L$ be any finite input sequence, let $\mathcal{S}$ be any retained state set, and let $\widehat{M}_\Delta$ satisfy equation 3 and $\|\widehat{M}_\Delta\|_F \le \|M_\Delta\|_F$. Assume zero initial states and $a_{d,n} < 0$ for all $d, n$. Let $y_t$ and $\widehat{y}_t$ denote the outputs of the full-order and reduced selective SSMs, respectively. Then, for every $t \le L$,*

$$\|y_t - \widehat{y}_t\|_2 \le K_t(x) \left[\Phi_N \epsilon_N(\mathcal{S}; \tau) + \Phi_R \epsilon_R(\widehat{M}_\Delta)\right], \tag{6}$$

*where $X_t := \max_{1 \le k \le t} \|x_k\|_2$ and $K_t(x) := X_t^3 (1 + X_t)^2$, and $\Phi_N, \Phi_R$ are the input-independent coefficients given in equation 7 below.*

The first term bounds the contribution from state truncation. The state-reduction residual $\epsilon_N(\mathcal{S};\tau)$ given in equation 5 sums the scores of discarded states. Each score combines the input–output coupling $q_n$ with the state-dynamics factor $\beta_n$. The quantity $q_n$ is unchanged by the state-coordinate rescaling $(W_B)_{n,:} \mapsto c_n(W_B)_{n,:}$ and $(W_C)_{n,:} \mapsto c_n^{-1}(W_C)_{n,:}$ for $c_n \neq 0$. The factor $\beta_n$ is larger when the diagonal modes decay slowly. Increasing $\tau$ emphasizes $q_n$, whereas decreasing it places more weight on slow modes through $\beta_n$.

The second term bounds the additional error from approximating $M_\Delta$. The coefficients $\Phi_N$ and $\Phi_R$ weight the state- and rank-reduction residuals, respectively, with $\Phi_R$ accounting for the effect of rank reduction over the horizon $L$. Writing $b_0 := \|\operatorname{softplus}(b_\Delta)\|_2$, $m_\Delta := \|M_\Delta\|_F$, $Q := \sum_{n=1}^{N} q_n$, and $A_Q := \max_{1\le d\le D} \sum_{n=1}^{N} |a_{d,n}| q_n$, the two coefficients are

$$\Phi_N := \tau + b_0 + m_\Delta, \qquad \Phi_R := LQ + \frac{L(L-1)}{2}(b_0 + m_\Delta)A_Q. \tag{7}$$

Using the full-order sums $Q$ and $A_Q$ makes $\Phi_R$ independent of the retained state set $\mathcal{S}$.

The bound separates the input-dependent factor $K_t(x)$ from the bracketed structural term. For fixed $L$, this structural term is independent of the input values and depends only on the learned parameters, the fixed hyperparameter $\tau$, and the reduction choices. Thus, the reduction choices can be determined from the learned parameters alone and are calibration-free at reduction time. Appendix C.2 provides the proof.

For a fixed full-order model and horizon, reducing either residual tightens the corresponding term of the bound. For fixed $r_N$, $\epsilon_N$ is minimized by retaining the $r_N$ states with the largest $\eta_n^A(\tau)$ defined in equation 4, while for fixed $r_R$, the truncated SVD of $M_\Delta$ minimizes $\epsilon_R(\widehat{M}_\Delta)$ among approximations of rank at most $r_R$. These parameter-based structural quantities can also be shaped during full-order training. Section 4.3 develops the resulting structure-inducing objective using the model-dependent weights $\Phi_N$ and $\Phi_R$.

The analysis applies to a single selective SSM layer and does not provide an end-to-end guarantee on the task loss. For multi-layer Mamba models, we apply the resulting training and reduction procedures layerwise and evaluate their end-to-end effect empirically.

### 4.3 Error-Bound-Aware Structure-Inducing Training

The error bound in Theorem 1 suggests reducing the state- and rank-reduction residuals $\epsilon_N(\mathcal{S};\tau)$ and $\epsilon_R(\widehat{M}_\Delta)$. We introduce parameter-based surrogates that encourage these residuals to remain small across different reduction levels without fixing the target dimensions $(r_N, r_R)$ during training.

For state reduction, we use the surrogate $\mathcal{R}_N := N^{-1}\sum_{n=1}^{N} \eta_n^A(\tau)$. Since $\eta_n^A(\tau) \ge 0$, $\epsilon_N(\mathcal{S};\tau) \le N\mathcal{R}_N$ for every $\mathcal{S}$. This $\ell_1$-type regularizer encourages small nonnegative state scores, with the factor $N$ absorbed into $\lambda_N$ below.

For a target rank $r_R$, the minimum achievable rank-reduction residual is

$$\epsilon_R^{\min}(r_R) := \min_{\operatorname{rank}(\widehat{M}_\Delta)\le r_R} \|M_\Delta - \widehat{M}_\Delta\|_F = \left(\sum_{i>r_R} \sigma_i(M_\Delta)^2\right)^{1/2}. \tag{8}$$

Here, $\sigma_i(M_\Delta)$ denotes the $i$th singular value of $M_\Delta$, and the singular values are ordered from largest to smallest. The minimum is attained by the truncated SVD.

To encourage rapid singular-value decay without computing an SVD at every optimization step, we use the factorized surrogate

$$\mathcal{R}_R(W_\Delta, W_\rho) = \sum_{j=1}^{R} \left(\|(W_\Delta)_{:,j}\|_2^2 + \|(W_\rho)_{j,:}\|_2^2 + \varepsilon\right)^p, \qquad 0 < p \le 1, \tag{9}$$

where $\varepsilon > 0$ is a small numerical constant. This group penalty is motivated by factorized Schatten-type regularization (Giampouras et al., 2020) and encourages the product $M_\Delta = W_\Delta W_\rho$ to have low effective rank. It is a surrogate motivated by the bound rather than the exact tail in equation 8. As shown in Appendix C.3, for every $r_R$, it bounds the minimum rank-reduction residual as follows:

$$\epsilon_R^{\min}(r_R) \le \frac{1}{2} r_R^{1/2-1/p} \mathcal{R}_R(W_\Delta, W_\rho)^{1/p}. \tag{10}$$

We use the following bound-motivated surrogate objective:

$$\boxed{\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_N \Phi_N \mathcal{R}_N + \lambda_R \Phi_R \mathcal{R}_R}, \tag{11}$$

where $\mathcal{L}_{\text{task}}$ is the task-specific prediction loss and $\lambda_N, \lambda_R \geq 0$ are regularization weights. For $p < 1$, the rank penalty uses $\Phi_R \mathcal{R}_R$ as a surrogate for the bound term involving $\Phi_R \mathcal{R}_R^{1/p}$, avoiding the vanishing gradient scaling induced by the outer power as $\mathcal{R}_R$ becomes small. The factors $\Phi_N$ and $\Phi_R$ in equation 7 retain the model-dependent weights from the bound.

For a multi-layer Mamba model, we average these regularization terms over selective SSM layers, using layer-specific quantities and a global task loss. All terms are computed from the full-order parameters, so training does not require a separate reduced model. The hyperparameter $\tau$ is fixed during training and reused after training. Section 4.4 then uses the learned state scores and step-size-map structure to construct the reduced model in one shot for a chosen $(r_N, r_R)$.

### 4.4 One-Shot Structural Reduction

After training, we reduce the state order using only the learned model parameters. Let $\text{TopIdx}_{r_N}$ return the indices of the $r_N$ largest entries for $\eta_n^A(\tau)$ defined in equation 4, with ties broken arbitrarily, and set

$$\mathcal{S}^\star := \text{TopIdx}_{r_N}\left(\{\eta_n^A(\tau)\}_{n=1}^N\right). \tag{12}$$

Because $\epsilon_N(\mathcal{S};\tau)$ sums the scores of discarded states and $\Phi_N$ is independent of $\mathcal{S}$, retaining the $r_N$ largest scores minimizes the state term $\Phi_N \epsilon_N(\mathcal{S};\tau)$ among all sets of size $r_N$. Note that this does not imply that the selection minimizes the actual output error or task loss.

Let $M_\Delta = U\Sigma V^\top$ be the SVD, where $\Sigma = \text{diag}(\sigma_1(M_\Delta), \ldots, \sigma_D(M_\Delta))$. We retain its leading $r_R$ singular components and choose the factors as

$$\widehat{M}_\Delta := U_{r_R} \Sigma_{r_R} V_{r_R}^\top, \qquad \widehat{W}_\Delta := U_{r_R} \Sigma_{r_R}, \qquad \widehat{W}_\rho := V_{r_R}^\top. \tag{13}$$

By the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), this gives $\epsilon_R(\widehat{M}_\Delta) = \epsilon_R^{\min}(r_R)$. The truncated approximation also satisfies $\|\widehat{M}_\Delta\|_F \leq \|M_\Delta\|_F$. Thus, this choice minimizes the rank-reduction residual in the bound.

Using $\mathcal{S}^\star$ and the factors in equation 13 to construct the reduced model from Section 4.1 allows a single full-order checkpoint with dimensions $(N, R)$ to be reduced to any chosen dimensions $(r_N, r_R)$. This reduction is calibration-free at reduction time and requires no post-reduction fine-tuning.

## 5 Experiments

We evaluate the proposed method using four research questions. **RQ1** asks whether the state order $N$ and step-size scheduling dimension $R$ can be reduced jointly while preserving task performance and improving execution efficiency. **RQ2** evaluates the proposed state regularization and selection pipeline across several target orders. **RQ3** evaluates truncated-SVD reduction of the step-size map across several target ranks. Finally, **RQ4** asks whether structure-inducing training produces the structures targeted by the error bound: concentrated state scores and low-rank step-size maps.

### 5.1 Experimental Setup

We evaluate the method on Selective Copying, audio modeling on SC09 (Warden, 2018), and language modeling on The Pile (Gao et al., 2020). We report test-set accuracy for Selective Copying, bits per byte (BPB) for SC09, and perplexity (PPL) for The Pile using custom contiguous splits of a preshuffled Pythia shard. Higher accuracy is better, whereas lower BPB and PPL are better.

We compare vanilla full models trained without the proposed structural regularizers, structurally regularized full models, Direct-small models trained from scratch at the target dimensions, post-hoc reduction controls, our post-hoc adaptation of CompreSSM (Chahine et al., 2026), and our Mamba-1 adaptation of GHOST (Menezes & Kyrillidis, 2026). In equation 11, state-only and rank-only regularization use $\lambda_R = 0$ and $\lambda_N = 0$, respectively. Our proposed reduction uses top-$\eta^A$ state selection

Table 1: Joint reduction of the state order and step-size scheduling dimension. M and A denote the moderate (75% retained) and aggressive (50% retained) settings. Random-control rows report mean ± sample standard deviation over four compression seeds applied to one frozen seed-1 checkpoint. Rows labeled Ours use a jointly regularized full model with both proposed deterministic reductions.

| Training source | State reduction | Rank reduction | Level | Sel. Copy Acc. ↑ | SC09 BPB ↓ | Pile PPL ↓ |
|---|---|---|---|---|---|---|
| Vanilla full model | – | – | Full | 1.0000 | 1.8523 | 17.26 |
| Jointly regularized full model | – | – | Full | 1.0000 | 1.8548 | 17.35 |
| Direct-small | – | – | M | 1.0000 | 1.8519 | 17.55 |
| Direct-small | – | – | A | 1.0000 | 1.8530 | 17.63 |
| Vanilla full model | top-$\eta^A$ | SVD | M | 0.8503 | 2.3357 | 20.90 |
| Vanilla full model | top-$\eta^A$ | SVD | A | 0.4987 | 3.4205 | 26.87 |
| Jointly regularized full model | Random | SVD | M | 0.2481±0.2133 | 2.1020±0.1085 | 44.68±47.11 |
| Jointly regularized full model | Random | SVD | A | 0.1048±0.0505 | 2.5097±0.1768 | 210.55±361.61 |
| Jointly regularized full model | top-$\eta^A$ | Random factor coordinates | M | 0.2981±0.1306 | 1.8685±0.0009 | 23.96±1.38 |
| Jointly regularized full model | top-$\eta^A$ | Random factor coordinates | A | 0.1047±0.0172 | 1.9125±0.0009 | 40.64±5.07 |
| **Ours** | **top-$\eta^A$** | **SVD** | **M** | **1.0000** | **1.8619** | **19.42** |
| **Ours** | **top-$\eta^A$** | **SVD** | **A** | **1.0000** | **1.8860** | **21.45** |

Table 2: Measured inference efficiency of the physically reduced joint models. M and A retain 75% and 50% of both $N$ and $R$. Speedup is measured against the full vanilla model at the same batch size $B^*$. Memory entries report reductions relative to the full model.

| Task / phase | $B^*$ | Speedup ↑ | | Peak alloc. reduction ↑ | Cache reduction ↑ |
|---|---|---|---|---|---|
| | | M | A | M / A | M / A |
| Selective Copying | 64 | 1.229× | 1.545× | 5.23 / 10.47% | – |
| SC09 | 16 | 1.083× | 1.222× | 0.27 / 0.55% | – |
| The Pile prefill | 16 | 1.092× | 1.255× | 0.18 / 0.36% | – |
| The Pile decode | 16 | 0.996× | 1.015× | 3.17 / 5.39% | 22.2 / 44.4% |

from equation 12 and step-size-map SVD from equation 13; it is calibration-free at reduction time and requires no post-reduction fine-tuning. The moderate and aggressive joint settings retain 75% and 50%, respectively, of both structural dimensions; Figure 2 additionally evaluates more aggressive isolated state- and rank-reduction settings. Appendix A specifies the data splits, architectures, target dimensions, training and coefficient-selection protocols, reduction controls, and seeds.

### 5.2 Joint Reduction Performance

Table 1 addresses RQ1 by reporting results for reduction that is calibration-free at reduction time on all three tasks. Performance retention under moderate and aggressive reduction varies across tasks. On all three tasks at both reduction levels, the proposed joint reduction yields lower reported task error than either Random control applied to the same jointly regularized checkpoint. Moreover, although the jointly regularized full-order models are slightly worse than their vanilla counterparts before reduction except for Selective Copying, their reduced models consistently retain performance better than vanilla post-hoc reduction at the same reduction levels. As discussed in RQ4 and Appendix B.1, this improved reducibility is associated with the reduction-friendly structures induced by regularization.

Note that direct-small models are competitive with or better than the reduced models. The benefit studied here is therefore the ability to reuse a trained full model while remaining calibration-free at reduction time, not superiority over training a model from scratch at the target size.

Table 2 reports measured inference efficiency of the reduced models. At the aggressive setting, batched throughput improves by up to 1.545×, while Pile decoding remains essentially throughput-neutral but reduces recurrent-plus-convolution cache memory by 44.4%. Detailed measurement protocols are provided in Appendix A.

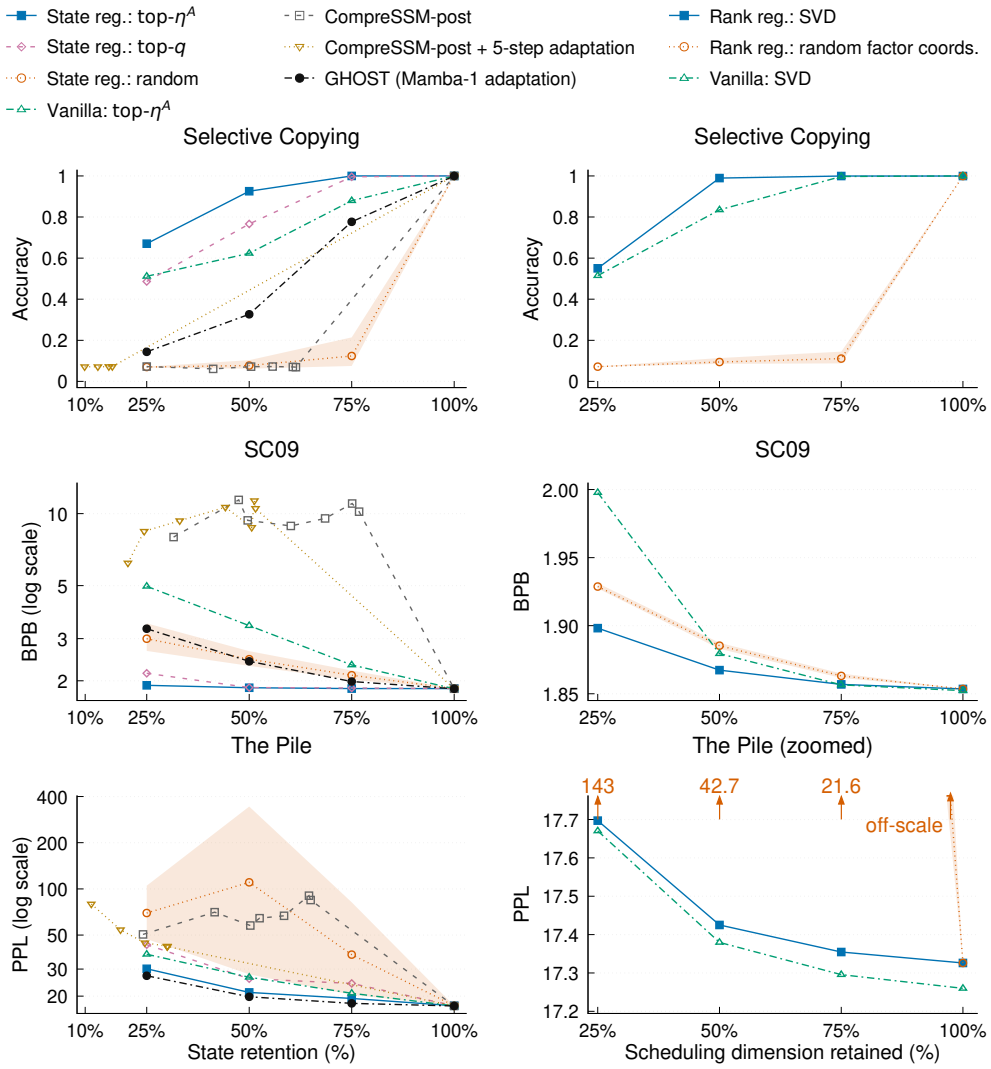



Figure 2: Isolated reduction on Selective Copying, SC09, and The Pile (rows): state order (left) and step-size-map rank (right). The proposed state and rank curves use state-only and rank-only regularized checkpoints, respectively. In each panel, the other structural dimension remains at full order. Random curves show means over four compression seeds; shaded bands span their observed ranges. Retention is reported as a percentage. The Pile rank panel uses a narrower vertical range; arrows mark off-scale Random means.

### 5.3 Isolated State and Step-Size-Map Rank Reduction

**State reduction (RQ2).** At every tested reduced order in Figure 2 (left), state-regularized top-$\eta^A$ reduction gives a better reported point estimate than all three baselines that are calibration-free at reduction time: vanilla post-hoc top-$\eta^A$ reduction, Random selection, and top-$q$. The latter two use the same state-only checkpoint as the proposed method. Note that the top-$q$ baseline retains the states with the largest $q_n$ in each layer, omitting the dynamics factor in $\eta_n^A$. These gains hold even though the regularization coefficient is selected without a compression target.

We further compare against our post-hoc adaptation of CompreSSM and our Mamba-1 adaptation of GHOST; detailed adaptation and evaluation protocols are provided in Appendix A. CompreSSM-post applies the CompreSSM reduction principle to a pretrained checkpoint, and a second variant addi-

tionally performs five update–reduction cycles. For our CompreSSM-post variants, the horizontal axis reports the sum of effective state counts across channels divided by the corresponding full-order total. Although the evaluated retention fractions do not always align exactly, the proposed method gives better reported point estimates than both variants at the nearest available retention levels on all three tasks. Note that the five-step variant on Selective Copying retains approximately 9.9–16.6% of states, outside the proposed method's 25–75% range, so these comparisons should not be interpreted as exactly matched-retention results.

We adapt GHOST's calibration-based state-importance score, originally developed for Mamba-2, to Mamba-1. At the shared 25%, 50%, and 75% state-retention levels, the proposed method gives better reported point estimates than our Mamba-1 adaptation of GHOST on Selective Copying and SC09, whereas our adaptation achieves lower PPL on The Pile. Unlike our CompreSSM-post variants and Mamba-1 adaptation of GHOST, our method is calibration-free at reduction time.

**Step-size-map rank reduction (RQ3).** The right column of Figure 2 separates the effect of SVD from that of rank-regularized training. SVD gives a better point estimate than Random factor-coordinate truncation at every tested reduced rank. Rank-regularized training provides additional gains on Selective Copying and at the more aggressive SC09 settings.

On The Pile, however, rank-regularized SVD gives slightly worse absolute PPL than vanilla SVD at all three tested reduced ranks. The selected regularization reduces the scale of the step-size map but induces little additional singular-value concentration (Appendix B.1). Stronger regularization produces clearer low-rank structure at the cost of worse full-order PPL (Appendix B.2). This suggests a task-dependent trade-off between low-rank structure and predictive performance. Nevertheless, the proposed joint reduction achieves lower Pile PPL than vanilla post-hoc joint reduction (Table 1).

**Learned reduction structure (RQ4).** The selected checkpoints also exhibit the structures targeted by the error bound: state scores become substantially more concentrated and the step-size-map SVD tails decrease on Selective Copying and SC09, whereas both effects are weaker on The Pile. Further analysis of these learned structures is provided in Appendix B.1.

## 6 Conclusion

We introduced a training-aware, structure-preserving reduction framework for Mamba selective SSMs that jointly reduces the state order $N$ and the step-size scheduling dimension $R$. A per-layer bound on finite-horizon output error motivates structure-inducing surrogates that encourage removable state coordinates and a low-rank step-size map during full-order training. After training, top-$\eta^A$ selection and truncated SVD construct reduced models at requested dimensions $(r_N, r_R)$ from the learned parameters alone; the reduction is calibration-free at reduction time and requires no post-reduction fine-tuning.

Empirically, the jointly reduced models achieve better reported point estimates than vanilla post-hoc reduction on all three tasks, while also providing measurable inference savings. In isolated state reduction, the proposed method achieves better reported point estimates than our post-hoc adaptation of CompreSSM (Chahine et al., 2026) at the nearest evaluated retention levels on all three tasks and than our Mamba-1 adaptation of GHOST (Menezes & Kyrillidis, 2026) at matched levels on Selective Copying and SC09; the latter achieves lower PPL on The Pile. Our method, unlike these adaptations, is calibration-free at reduction time. Direct-small models remain competitive; the main practical benefit is reusing a single trained full-order checkpoint across multiple target sizes.

The underlying reduction principle is not specific to Mamba-1 (Gu & Dao, 2024) and is potentially applicable to the structured selective SSMs used in Mamba-2 (Dao & Gu, 2024) and Mamba-3 (Lahoti et al., 2026). Future work includes validating such extensions empirically and further investigating task-dependent requirements on the effective scheduling dimension $R$, as suggested by the weaker rank reducibility observed on The Pile. We also plan to develop a layerwise error analysis that accounts for error propagation across depth. Such an analysis could reveal layer-dependent reduction sensitivity and enable nonuniform allocation of state and scheduling dimensions across layers, potentially improving accuracy under a fixed parameter budget.

## AI Use Statement

We used ChatGPT (OpenAI) to assist with theoretical and proof development, experimental design and implementation, literature retrieval and discovery, interpretation of results, and manuscript drafting and language editing. Mathematical content was independently checked by the authors through manual derivation, experimental suggestions and code were reviewed and tested, and AI-assisted writing and interpretations were checked against the underlying results and literature. The authors take responsibility for the final content.

## References

S. Abreu, J. E. Pedersen, K. M. Heckel, and A. Pierro. Q-S5: Towards quantized state space models. In *ICML 2024 Workshop on Next Generation of Sequence Modeling Architectures*, 2024.

A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*, volume 6 of *Adv. Des. Control*. SIAM Publications, Philadelphia, PA, 2005. ISBN 9780898715293. doi: 10.1137/1.9780898718713.

A. C. Antoulas, D. C. Sorensen, and S. Gugercin. A survey of model reduction methods for large-scale systems. *Contemp. Math.*, 280:193–219, 2001.

U. Baur, P. Benner, and L. Feng. Model order reduction for linear and nonlinear systems: A system-theoretic perspective. *Arch. Comput. Methods Eng.*, 21(4):331–358, 2014. doi: 10.1007/s11831-014-9111-2.

P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.*, 57(4):483–531, 2015. doi: 10.1137/130932715.

P. Benner, X. Cao, and W. Schilders. A bilinear $\mathcal{H}_2$ model order reduction approach to linear parameter-varying systems. *Adv. Comput. Math.*, 45:2241–2271, 2019. doi: 10.1007/s10444-019-09695-9.

R. Bhattacharya. An exact instrument for state usage in selective state-space models, and the input-driven migration it reveals. Preprint 2607.11796, arXiv, 2026. URL `https://arxiv.org/abs/2607.11796`.

M. Bittner, D. Schnöll, D. Dallinger, M. Wess, and A. Jantsch. Pruning state space models with model order reduction for efficient raw audio classification. In *33rd European Signal Processing Conference (EUSIPCO)*, pp. 271–275, 2025. doi: 10.23919/EUSIPCO63237.2025.11226287.

E. V. Brazil, E. Soares, V. Y. Shirasuna, R. Cerqueira, D. Zubarev, and K. Schmidt. A Mamba-based foundation model for chemistry. In *NeurIPS 2024 Workshop on AI for Accelerated Materials Design*, 2024. URL `https://openreview.net/forum?id=HTgCs0KSTl`.

M. Chahine, P. Nazari, D. Rus, and T. K. Rusch. The curious case of in-training compression of state space models. In *International Conference on Learning Representations*, 2026.

Y. Cheng, D. Wang, P. Zhou, and T. Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136, 2018. doi: 10.1109/MSP.2017.2765695.

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*, 2024. URL `https://arxiv.org/abs/2405.21060`.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. doi: 10.1007/BF02288367.

H. Ezoe and K. Sato. Model compression method for S4 with diagonal state space layers using balanced truncation. *IEEE Access*, 12:116415–116427, 2024. doi: 10.1109/ACCESS.2024.3446642.

M. Forgione, M. Mejari, and D. Piga. Model order reduction of deep structured state-space models: A system-theoretic approach. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pp. 8620–8625, 2024. doi: 10.1109/CDC56724.2024.10886865.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The Pile: An 800GB dataset of diverse text for language modeling. Preprint 2101.00027, arXiv, 2020. URL https://arxiv.org/abs/2101.00027.

P. Giampouras, R. Vidal, A. Rontogiannis, and B. D. Haeffele. A novel variational form of the Schatten-$p$ quasi-norm. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

A. Gu, K. Goel, A. Gupta, and C. Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983, 2022a. doi: 10.52202/068431-2607.

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

S. Gugercin, A. C. Antoulas, and C. Beattie. $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.*, 30(2):609–638, 2008. doi: 10.1137/060666123.

M. Gwak, S. Moon, J. Ko, and P. 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.

Aakash Lahoti, Kevin Y. 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://arxiv.org/abs/2603.15569.

N. Lang, J. Saak, and T. Stykel. Balanced truncation model reduction for linear time-varying systems. *Math. Comput. Model. Dyn. Syst.*, 22(4):267–281, 2016. doi: 10.1080/13873954.2016.1198386.

Z. Li, M. Han, and X. Yin. MamKO: Mamba-based Koopman operator for modeling and predictive control. In *International Conference on Learning Representations*, 2025. URL https://mlanthology.org/iclr/2025/li2025iclr-mamko/.

M. Menezes and A. Kyrillidis. GHOST: Unmasking phantom states in Mamba2 via grouped hidden-state output-aware selection & truncation. Preprint 2602.11408, arXiv, 2026. URL https://arxiv.org/abs/2602.11408.

J. P. Muñoz, J. Yuan, and N. 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, 2025. doi: 10.18653/v1/2025.naacl-long.195.

A. P. Padhy, F. F. Camacho, and S. Mukhopadhyay. AIRE-Prune: Asymptotic impulse-response energy for state pruning in state space models. In *International Conference on Learning Representations*, 2026.

H. Sakamoto and K. Sato. Compression method for deep diagonal state space model based on $\mathcal{H}_2$ optimal reduction. *IEEE Contr. Syst. Lett.*, 9:2043–2048, 2025a. doi: 10.1109/LCSYS.2025.3591023.

H. Sakamoto and K. Sato. A deep state-space model compression method using upper bound on output error. Preprint 2510.14542, arXiv, 2025b. URL https://arxiv.org/abs/2510.14542.

Y. Schiff, C. H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

P. Schwerdtner, J. Berman, and B. Peherstorfer. Hankel singular value regularization for highly compressible state space models. In *Advances in Neural Information Processing Systems*, 2025.

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

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. Preprint 1804.03209, arXiv, 2018. URL `https://arxiv.org/abs/1804.03209`.

Z. Xu, Y. Yue, X. Hu, Z. Yuan, Z. Jiang, Z. Chen, J. Yu, C. Xu, S. Zhou, and D. Yang. MambaQuant: Quantizing the Mamba family with variance aligned rotation methods. In *International Conference on Learning Representations*, 2025.

# A Experimental Setup

Experiments were conducted in PyTorch on NVIDIA A100 and H100 GPUs (one GPU per run), with all inference-efficiency measurements performed on an NVIDIA H100 NVL GPU.

## A.1 Common Experimental Protocols

**Model dimensions and target dimensions.** Table 3 gives the exact model sizes used in the main experiments. The base model width $d_{\text{model}}$ differs from the selective SSM channel dimension $D = e\, d_{\text{model}}$. Every block uses expansion $e = 2$ and convolution width 4. SC09 is a five-stage sequence U-Net with eight Mamba blocks per stage, pooling strides $(4, 4)$, and stage widths $d_{\text{model}} = [64, 128, 256, 128, 64]$. Its $D$ and $R$ entries follow the same down–center–up order. Parameter counts include the structural parameters, embeddings, output heads, and all order-independent components. Selective Copying and The Pile tie the input and output embeddings, whereas SC09 uses separate projections.

Table 3: Model dimensions and trainable parameter counts. Direct-small M and A models differ from the corresponding full-order models only in $(N, R)$.

| Task | Size | Blocks | $d_{\text{model}}$ | $e$ | $D$ | $N$ | $R$ | Parameters |
|---|---|---|---|---|---|---|---|---|
| Selective Copying | Full | 2 | 64 | 2 | 128 | 64 | 16 | 109,504 |
| | M | 2 | 64 | 2 | 128 | 48 | 12 | 95,168 |
| | A | 2 | 64 | 2 | 128 | 32 | 8 | 80,832 |
| SC09 | Full | 40 | [64, 128, 256, 128, 64] | 2 | [128, 256, 512, 256, 128] | 16 | [4, 8, 16, 8, 4] | 6,249,808 |
| | M | 40 | [64, 128, 256, 128, 64] | 2 | [128, 256, 512, 256, 128] | 12 | [3, 6, 12, 6, 3] | 6,073,680 |
| | A | 40 | [64, 128, 256, 128, 64] | 2 | [128, 256, 512, 256, 128] | 8 | [2, 4, 8, 4, 2] | 5,897,552 |
| The Pile | Full | 24 | 768 | 2 | 1536 | 32 | 48 | 130,904,832 |
| | M | 24 | 768 | 2 | 1536 | 24 | 36 | 129,135,360 |
| | A | 24 | 768 | 2 | 1536 | 16 | 24 | 127,365,888 |

**Data and optimization.** Selective Copying uses sequences of length $L = 4096$ with 16 selected symbols drawn from 14 data symbols. The full vocabulary contains these 14 data symbols plus a noise symbol and a query symbol, for 16 symbols in total. Training batches are generated online. We generate the validation and test sets independently using seeds 10000 and 20000, respectively. Each set contains 100 batches of 64 examples. We train for 400,000 steps with AdamW, a batch size of 64, a constant learning rate $10^{-4}$, zero weight decay, and gradient clipping at 1.0. Accuracy is token accuracy over the 16 queried output positions, not sequence-level exact match. The checkpoint maximizes validation accuracy, breaking ties in favor of lower loss.

SC09 uses the official Speech Commands split restricted to the ten digit classes. The resulting train, validation, and test splits contain 31,158, 3,643, and 4,107 waveforms, respectively. Mono 16-kHz audio is encoded using 8-bit $\mu$-law and padded to 16,000 samples; padded positions are excluded from the loss. The five stages operate at horizons $L = [16000, 4000, 1000, 4000, 16000]$, respectively, with each value shared by the stage's eight blocks. The model is trained for 200,000 steps with AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$), a batch size of 16, peak learning rate $2 \times 10^{-3}$, 10% linear warmup followed by cosine decay to zero, weight decay 0.1, gradient clipping at 0.1, and BF16 arithmetic. We select the checkpoint with the lowest validation negative log-likelihood (NLL); validation and test use the full splits. In the BPB formula, NLL is summed over the $T_{\text{unmasked}}$ unmasked byte positions: $\text{BPB} = \text{NLL}/(T_{\text{unmasked}} \log 2)$.

The Pile experiment uses a preshuffled Pythia shard tokenized with the GPT-NeoX tokenizer and pinned to revision `bac79b6820adb34e451f9a02cc1dc7cd920febf0`. Non-overlapping contiguous regions provide 1,228,800 training sequences, 25,000 validation sequences, and 25,000 test sequences, each of length $L = 2048$. This gives 2.516 billion training tokens. We train for 4,800 optimizer steps with AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$), a micro-batch size of 16, 16-way gradient accumulation (256 sequences per update), peak learning rate $10^{-3}$, 10% linear warmup followed by cosine decay to $10^{-5}$, weight decay 0.1, gradient clipping at 1.0, and BF16 arithmetic. We select the checkpoint with the lowest validation NLL; validation and test each contain 51.2M tokens. PPL is the exponential of the mean per-token negative log-likelihood.

**Evaluation protocol.** We report test-set accuracy for Selective Copying (higher is better), BPB for SC09, and PPL for The Pile (lower is better for both). M and A retain 75% and 50%, respectively, of both structural dimensions; the proposed isolated reductions additionally evaluate 25% retention. Results for our adapted baselines are plotted at the state-retention fractions produced by their respective reduction settings, detailed below.

State reduction follows equation 12 in every selective SSM layer, and rank reduction uses the truncated SVD in equation 13 for $M_\Delta = W_\Delta W_\rho$. We compare vanilla, state-only, rank-only, and jointly regularized full-order models; Direct-small models; deterministic reduction of vanilla checkpoints; Random controls; our post-hoc adaptation of CompreSSM; and our Mamba-1 adaptation of GHOST. The Random controls sample state or factor coordinates. In particular, random rank reduction samples factor coordinates rather than singular directions. For the proposed isolated reductions and their internal controls, we construct the reduced model at every target dimension from the same frozen source checkpoint. The state curves also include top-$q$, which retains the states with the largest $q_n = \|(W_B)_{n,:}\|_2 \|(W_C)_{n,:}\|_2$ in each layer, as a baseline that omits the dynamics factor in $\eta_n^A$.

All full models and Direct-small models use training seed 1. Random controls use four compression seeds on each frozen seed-1 source: $\{41001, 41002, 41003, 41004\}$ for Selective Copying and $\{1, 2, 3, 4\}$ for SC09 and The Pile. The Random controls in the Pile 1× ablation also use compression seeds $\{1, 2, 3, 4\}$. Reported standard deviations measure compression-subset variation; Direct-small and deterministic reductions have one reported value.

### A.2 Structural Regularization and Coefficient Selection

**Structural hyperparameters.** We set $\tau = 1$ and use $\varepsilon = 10^{-8}$ to smooth the rank penalty. When computing $\beta_n$ and $A_Q$ in $\Phi_R$, the implementation replaces $|a_{d,n}|$ by $\max(|a_{d,n}|, 10^{-12})$ to avoid numerical underflow; this does not constrain the learned dynamics. This safeguard was inactive for all reported checkpoints, so the implemented quantities match their theoretical definitions. The Schatten exponent is $p = 1/2$ for Selective Copying and SC09 and $p = 1/4$ for The Pile.

**Target-independent coefficient selection.** For each task and structural axis $a \in \{N, R\}$, the reference coefficient $\lambda_{a,0}$ (written $\lambda_0$ below) is obtained by rounding $0.01\mathcal{L}_{\text{ref}}/S_a(\theta_0)$, where $S_N = \text{mean}_\ell[\Phi_{N,\ell}\mathcal{R}_{N,\ell}]$ and $S_R = \text{mean}_\ell[\Phi_{R,\ell}\mathcal{R}_{R,\ell}]$ are the unweighted structural terms at initialization $\theta_0$. We use $\mathcal{L}_{\text{ref}} = \log K$, where $K$ is the number of output classes or vocabulary items. For a validation metric $m$, relative degradation is

$$d(m) = \begin{cases} (m_0 - m)/m_0, & \text{accuracy (higher is better),} \\ (m - m_0)/m_0, & \text{BPB or PPL (lower is better),} \end{cases}$$

where $m_0$ is the vanilla full-order result. For each regularization term, we train models with coefficients $\{1, 0.1, 0.01\}\lambda_0$ while setting the other structural coefficient to zero. We retain candidates with $d \le 0.01$ and select the largest feasible coefficient. If no candidate satisfies this condition, the procedure terminates without selecting a coefficient. Table 4 reports every candidate; each entry is "validation metric / degradation." Bold entries are selected. The vanilla baselines are 1.000000 accuracy, 1.930612 BPB, and 17.208456 PPL, respectively.

Table 4: Full-order, target-independent coefficient selection.

| Task | Axis | Reference $\lambda_0$ | 1×: metric / $d$ | 0.1×: metric / $d$ | 0.01×: metric / $d$ | Selected $\lambda$ |
|---|---|---|---|---|---|---|
| Selective Copying | State | $9 \times 10^{-3}$ | 0.073301 / 92.6699% | **0.999961 / 0.0039%** | 0.999990 / 0.0010% | $9 \times 10^{-4}$ |
| Selective Copying | Rank | $4 \times 10^{-14}$ | **0.998965 / 0.1035%** | 0.999990 / 0.0010% | 1.000000 / 0.0000% | $4 \times 10^{-14}$ |
| SC09 | State | $5 \times 10^{-3}$ | **1.933061 / 0.1269%** | 1.931181 / 0.0295% | 1.930950 / 0.0175% | $5 \times 10^{-3}$ |
| SC09 | Rank | $4 \times 10^{-13}$ | **1.931781 / 0.0605%** | 1.931085 / 0.0245% | 1.931103 / 0.0254% | $4 \times 10^{-13}$ |
| The Pile | State | $3 \times 10^{-3}$ | 17.813528 / 3.5161% | 17.406342 / 1.1499% | **17.241848 / 0.1940%** | $3 \times 10^{-5}$ |
| The Pile | Rank | $2 \times 10^{-13}$ | 17.842123 / 3.6823% | 17.437911 / 1.3334% | **17.273444 / 0.3777%** | $2 \times 10^{-15}$ |

In both Figures 2 and 3, the proposed method uses the selected state-only checkpoint for the state panels and the selected rank-only checkpoint for the rank panels. After the two independent selections, we train a new jointly regularized full-order model from scratch with both selected coefficients. Its validation metric and degradation are 0.999961/0.0039% for Selective Copying, 1.933019/0.1247%

for SC09, and 17.296900/0.5140% for The Pile; therefore, all jointly regularized checkpoints also satisfy the 1% condition. We use validation data only to select these full-order models. For our proposed method, reduction of each frozen checkpoint is then calibration-free at reduction time and uses no post-reduction validation, fine-tuning, or test-set feedback.

### A.3 Baselines and Reduction Controls

**Direct-small models and reduction controls.** Direct-small models use the same data, optimizer, number of steps, and checkpoint rule as the corresponding full-order model; only the dimensions in Table 3 change. Random state reduction samples, in each block, a uniform subset of existing state coordinates without replacement. Random factor-coordinate truncation samples $\mathcal{J} \subset \{1, \dots, R\}$ with $|\mathcal{J}| = r$ uniformly without replacement and uses $\widehat{M}_\Delta = (W_\Delta)_{:,\mathcal{J}}(W_\rho)_{\mathcal{J},:}$. In contrast, truncated SVD gives the best rank-$r$ approximation of $M_\Delta = W_\Delta W_\rho$ in the Frobenius norm. Our proposed reductions and the deterministic and Random controls are applied once without post-reduction fine-tuning.

**Our adaptations of CompreSSM and GHOST.** Both of our adapted baselines start from the same vanilla seed-1 full-order checkpoint used for the vanilla post-hoc controls.

CompreSSM (Chahine et al., 2026) performs balanced-truncation-based state reduction during training. Its Mamba variant constructs per-channel stationary LTI surrogates and applies balanced truncation to these surrogates. Because our setting concerns post-training reduction of an already trained full model, we adapt this procedure by applying the reduction once to the frozen pretrained checkpoint; we denote this baseline CompreSSM-post. For both CompreSSM-post variants, calibration forward passes use FP32 with the same fixed 128 calibration batches (calibration seed 50001), the stationary LTI surrogates are constructed in FP64, and balanced truncation is computed on the CPU. Our implementation includes numerical stabilization. We additionally evaluate a five-step variant of CompreSSM-post. Starting from the full checkpoint, we perform five update–calibration–reduction cycles. A fresh AdamW optimizer is initialized once without loading the original optimizer state, and its state is maintained across all five cycles. The learning rate is fixed to $10^{-5}$. Thus, four updates occur after the first reduction and no update occurs after the fifth (final) reduction. The five-step CompreSSM-post variant is the only reported baseline that performs optimization updates during the reduction procedure.

GHOST (Menezes & Kyrillidis, 2026) is designed for Mamba-2 and scores recurrent states using empirical input-to-state and state-to-output statistics collected from calibration inputs. We adapt its score to Mamba-1 by reconstructing the input-dependent Mamba-1 recurrence for every inner channel $d$ and state coordinate $n$, and scoring

$$S_n = \sum_{s,t,d} h_{s,t,d,n}^2 C_{s,t,n}^2,$$

where $h_{s,t,d,n}$ is the Mamba-1 recurrent state induced by $A = -\exp(A_{\log})$, input-dependent $\Delta$, and the $B$ projection. In contrast to Mamba-2's grouped state structure, we aggregate this score over all Mamba-1 inner channels. We use 128 fixed calibration sequences (seed 50001), processed one sequence at a time in FP32. For Selective Copying, these are length-4096 sequences generated from a separate seed under the same data-generating distribution. For SC09, they are 128 fixed 16000-sample clips selected from the training split. For The Pile, they are 128 length-2048 sequences sampled without replacement from the training range, with no overlap with validation or test token ranges. We then perform one sequential layerwise pruning pass, recalibrating after each layerwise pruning step, without any post-reduction fine-tuning.

### A.4 Inference-Efficiency Measurements

We measure inference efficiency for the physically truncated joint models corresponding to the Full, M, and A settings in Table 3. Measurements are performed on an H100 NVL using eager execution. Selective Copying uses FP32, whereas SC09 and The Pile use FP32 weights with BF16 autocast. The causal-convolution extension is unavailable in these measurements, so its native fallback is included.

For Selective Copying, SC09, and Pile prefill, the benchmark batch sizes $B^*$ are 64, 16, and 16, respectively. Pile decoding uses batch size 16 with a 512-token prompt followed by 512 recurrent decoding steps. Standalone prefill computes logits for all input tokens.

Batched speedup compares the full and reduced models at identical batch size, sequence length, and inference phase. Using paired measurements over ten outer rounds, we compute the per-round median ratios and report their geometric mean. Peak allocated memory is measured at $B^*$, and for decoding we additionally report the total recurrent-plus-convolution cache size.

Timings include the full model forward pass, including the output head, but exclude model loading, host–device transfer, tokenization, sampling, and the one-shot reduction procedure. They therefore measure model inference rather than complete application latency.

# B Additional Experimental Results

## B.1 Learned Structure for Model Reduction

We quantify the structures learned before reduction using two normalized concentration measures. The cumulative state-score mass is $C_k := \sum_{i=1}^{k} \eta^A_{(i)} / \sum_{i=1}^{N} \eta^A_{(i)}$, where scores are sorted in descending order; we set $C_k = 1$ when $\sum_n \eta^A_n = 0$. The relative SVD tail is $\rho_r := \sqrt{\sum_{j>r} \sigma_j(M_\Delta)^2} / \sqrt{\sum_j \sigma_j(M_\Delta)^2}$, with $\rho_r = 0$ when $\|M_\Delta\|_F = 0$. These measures determine the state-reduction residual $(1 - C_k) \sum_n \eta^A_n$ for top-$k$ state selection and the rank-reduction residual $\epsilon^{\min}_R(r) = \rho_r \|M_\Delta\|_F$ for rank-$r$ truncated SVD, respectively. They thus quantify the structural terms in equation 6, but do not directly guarantee task performance.

Figure 3 aggregates these measures across layers at common retained fractions. Each layer's state and rank curves, parameterized by $k/N_\ell$ and $r/R_\ell$, are linearly interpolated onto the grids $\{j/N_{\max}\}_{j=1}^{N_{\max}}$ and $\{j/R_{\max}\}_{j=1}^{R_{\max}}$, respectively, then averaged with equal layer weights; $N_{\max}$ and $R_{\max}$ are the largest layer dimensions within each task. Outside a layer's grid, its nearest endpoint value is held constant; no integer rounding is used. This interpolation matters for SC09, where $R_\ell \in \{4, 8, 16\}$, whereas its equal state orders yield coincident state grids.

State regularization increases score concentration substantially on Selective Copying and SC09, but only modestly on The Pile. At 50% state retention, the layer-mean $C_{N/2}$ increases from 0.760 to 0.928, from 0.780 to 0.988, and from 0.746 to 0.780, respectively, under the selected coefficients.

Rank regularization reduces the relative SVD tails on Selective Copying and SC09. At 50% rank retention, the layer-mean $\rho_r$ decreases from 0.219 to 0.020 on Selective Copying and from 0.301 to 0.186 on SC09. In the final SC09 rank-only checkpoint, 18 of 40 blocks also have $\sigma_{\max}(M_\Delta) = 0$ at stored numerical precision.

On The Pile, rank regularization reduces the overall magnitude of the step-size map but induces little additional low-rank concentration. At 50% rank retention, $\rho_r$ changes only from 0.343 to 0.341, while the median $\|M_\Delta\|_F$ decreases from 39.9 to 25.1. This limited concentration is consistent with the weaker additional benefit of rank-regularized training on The Pile compared with Selective Copying and SC09.

## B.2 The Pile Coefficient-Strength Ablation

The target-independent rule selects $\lambda_N = 3 \times 10^{-5}$ and $\lambda_R = 2 \times 10^{-15}$ for The Pile, 0.01 times their reference values. To test whether the weak separation in Figure 3 reflects these small coefficients, we repeat the isolated analysis using checkpoints trained with the reference coefficients in Figure 4. The state-only checkpoint uses $\lambda_N = 3 \times 10^{-3}$ and the rank-only checkpoint uses $\lambda_R = 2 \times 10^{-13}$, with the other coefficient set to zero. This isolated ablation does not use a jointly regularized 1× checkpoint.

At 50% state retention, $C_{N/2}$ is 0.746 for the vanilla model, 0.780 for the selected 0.01× state-only model, and 0.961 for the 1× state-only model. At 50% rank retention, $\rho_{R/2}$ is 0.343, 0.341, and 0.214 for the vanilla, selected 0.01× rank-only, and 1× rank-only models, respectively. Compared with the

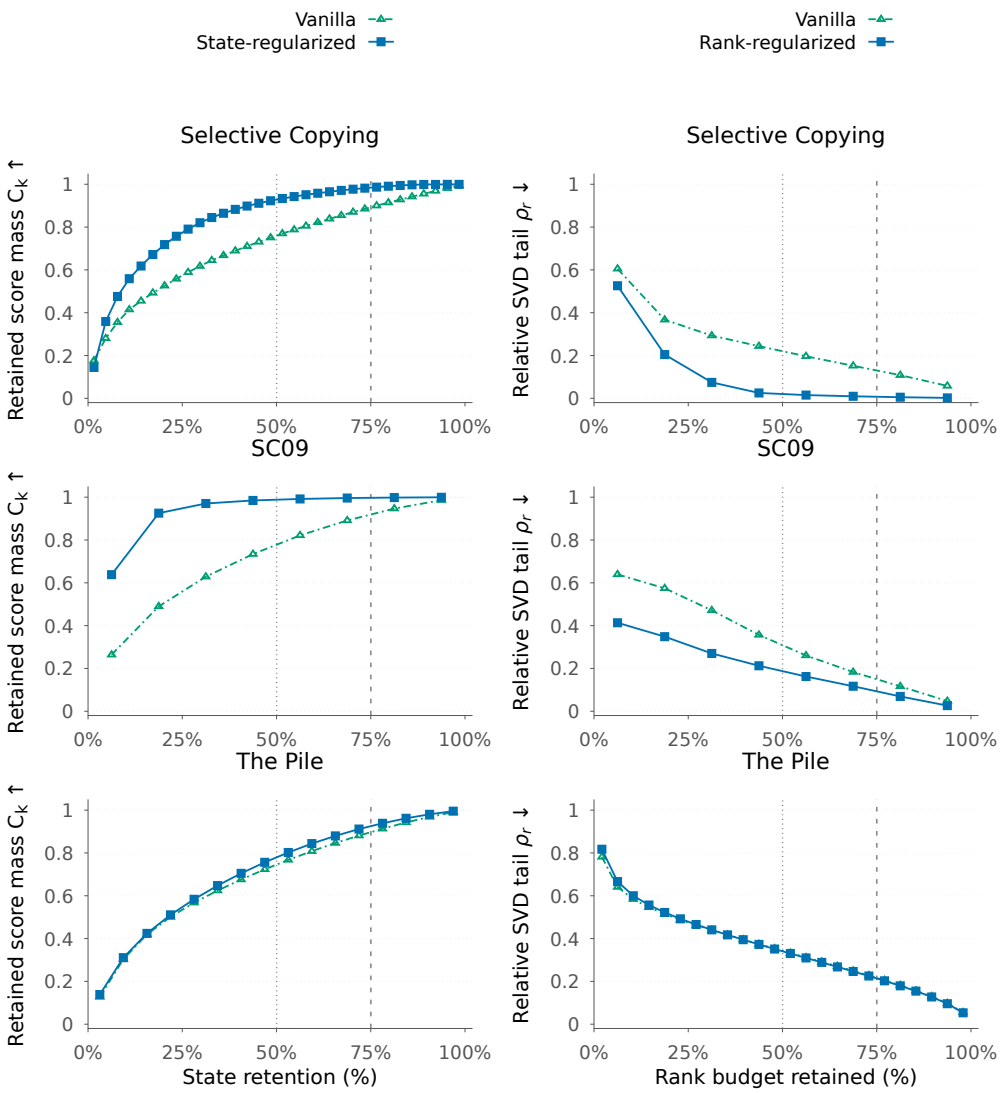



Figure 3: Structures learned before reduction. The left column shows cumulative state-score mass $C_k$ as a function of state retention (higher is better); the right shows the relative Frobenius norm of the discarded SVD tail, $\rho_r$, as a function of rank retention (lower is better). The two columns use state-only and rank-only checkpoints, respectively. Curves are averaged across layers within each checkpoint. Vertical lines mark 50% and 75% retention and are not post-reduction measurements.

selected 0.01× checkpoints in Figure 3, the 1× checkpoints in Figure 4 exhibit substantially stronger structural separation.

As shown in Figure 4, this stronger structure comes at a cost to full-order test performance: test PPL is 17.260 for the vanilla model, 17.868 for the 1× state-only model, and 17.898 for the 1× rank-only model. At 25% and 50% retention, the 1× state-only model outperforms both the vanilla model and the selected 0.01× state-only model, with PPL values of 25.36 and 20.90, respectively. At 75% retention, it still outperforms the vanilla model but not the selected 0.01× model. For rank reduction, SVD reduction of the 1× rank-only model yields PPL values of 18.29, 17.97, and 17.92 at 25%, 50%, and 75% retention, versus 17.67, 17.38, and 17.30 for vanilla post-hoc SVD and 17.70, 17.42, and 17.35 for the selected model.

The ablation therefore suggests a fit–compressibility trade-off: stronger coefficients produce more concentrated structure but do not consistently lower PPL.

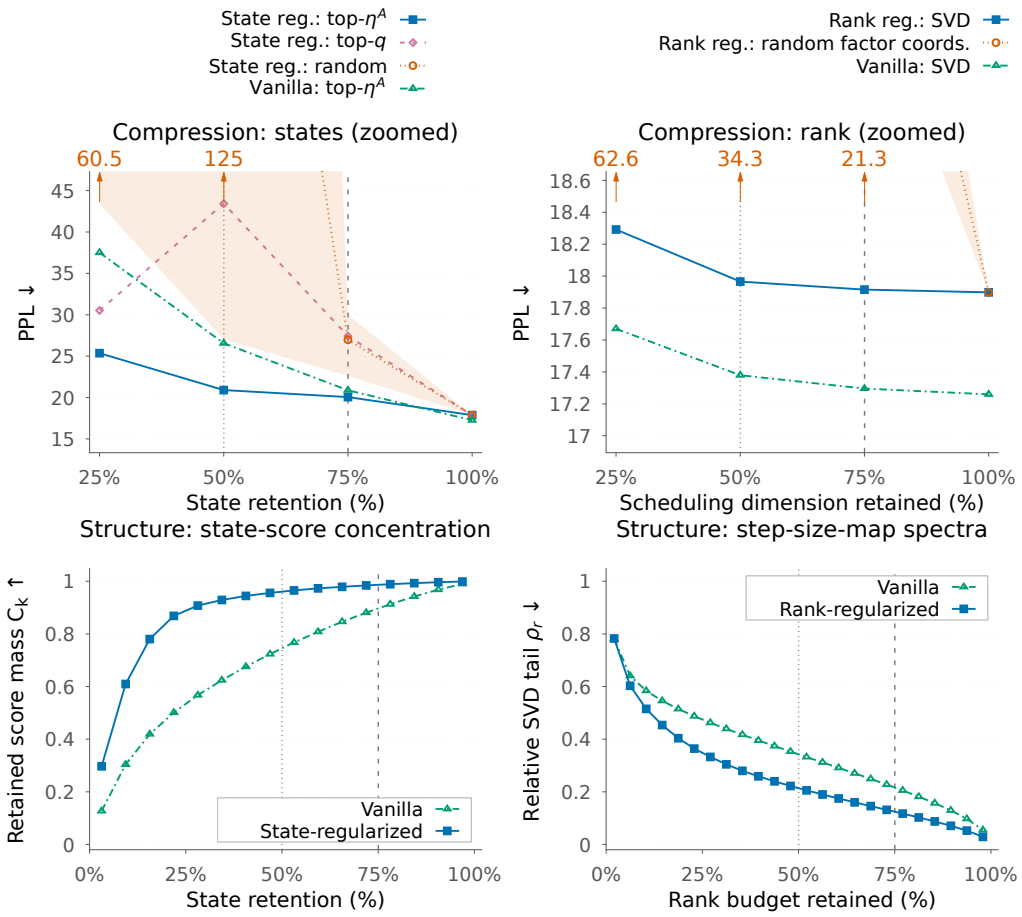



Figure 4: The Pile with reference (1×) regularization. The top row reports results for isolated state reduction and step-size-map SVD reduction; the bottom row reports the corresponding state-score mass and relative SVD-tail profiles. State and rank panels use state-only and rank-only checkpoints, respectively; the corresponding vanilla curves are the same as those in Figures 2 and 3. Random curves show means over four compression seeds; off-scale means are labeled numerically.

# C Additional Theoretical Details

## C.1 Pointwise Stability of the Reduced Model

**Proposition 1** (Pointwise stability preservation)**.** *Suppose that $a_{d,n} < 0$ for all $d$ and $n$. Then, for any retained state set $\mathcal{S}$ and any $\widehat{M}_\Delta$ satisfying equation 3,*

$$\rho\Big(\exp(\widehat{\Delta}_{t,d}\widehat{A}_d)\Big) < 1 \qquad \text{for all } t, d,$$

*where $\rho(\cdot)$ denotes the spectral radius.*

*Proof.* By equation 2, every diagonal entry $\widehat{a}_{d,n}$ of $\widehat{A}_d$ satisfies $\widehat{a}_{d,n} < 0$. Moreover, $\widehat{\Delta}_{t,d} = \mathrm{softplus}((\widehat{M}_\Delta x_t)_d + b_{\Delta,d}) > 0$ for every $t$ and $d$. Since $\widehat{A}_d$ is diagonal, each eigenvalue of $\exp(\widehat{\Delta}_{t,d}\widehat{A}_d)$ is $\exp(\widehat{\Delta}_{t,d}\widehat{a}_{d,n}) \in (0, 1)$; its spectral radius is therefore less than one, proving Proposition 1. □

### C.2 Proof of Theorem 1

*Proof.* Fix any $t \le L$. For each channel $d$, let

$$\delta_{t,d}^{\max} := \max_{1\le k\le t} \Delta_{k,d}, \qquad u_t := \left(\sum_{d=1}^{D} (\delta_{t,d}^{\max})^2\right)^{1/2}.$$

For convenience, write $E_\Delta := M_\Delta - \widehat{M}_\Delta$. We split the proof into the state-truncation error and the error introduced by step-size-map reduction.

Let $y_t^{\mathcal{S}}$ denote the output of the intermediate model that retains only the states in $\mathcal{S}$ but uses the original step-size map $M_\Delta$. By equation 2, its state parameters satisfy $A_d^{\mathcal{S}} = \widehat{A}_d$, $W_B^{\mathcal{S}} = \widehat{W}_B$, and $W_C^{\mathcal{S}} = \widehat{W}_C$, while it uses the original $\Delta_t$. By the triangle inequality,

$$\|y_t - \widehat{y}_t\|_2 \le \underbrace{\|y_t - y_t^{\mathcal{S}}\|_2}_{E_N(t)} + \underbrace{\|y_t^{\mathcal{S}} - \widehat{y}_t\|_2}_{E_R(t)}. \tag{14}$$

For a fixed channel $d$ and state $n$, define

$$\kappa_{t,d,n} := \sum_{k=1}^{t} \Delta_{k,d} \exp\left(a_{d,n} \sum_{j=k+1}^{t} \Delta_{j,d}\right).$$

Since $a_{d,n} < 0$, let $\lambda_{d,n} := |a_{d,n}| > 0$.

We first show that

$$\kappa_{t,d,n} \le \delta_{t,d}^{\max} + \frac{1}{|a_{d,n}|}. \tag{15}$$

Suppressing the indices $d, n$, the recursion for $\kappa_t := \kappa_{t,d,n}$ is

$$\kappa_t = \Delta_t + \mathrm{e}^{-\lambda\Delta_t}\kappa_{t-1}, \qquad \kappa_0 = 0.$$

Let $\delta_t^{\max} := \max_{1\le k\le t} \Delta_k$. The case $t = 1$ is immediate because $\kappa_1 = \Delta_1 = \delta_1^{\max} \le \delta_1^{\max} + \lambda^{-1}$. For $t \ge 2$, assume inductively that $\kappa_{t-1} \le \delta_{t-1}^{\max} + \lambda^{-1}$. If $\Delta_t \ge \delta_{t-1}^{\max}$, then $\delta_t^{\max} = \Delta_t$ and

$$\kappa_t \le \Delta_t + \mathrm{e}^{-\lambda\Delta_t}\left(\Delta_t + \frac{1}{\lambda}\right) \le \Delta_t + \frac{1}{\lambda} = \delta_t^{\max} + \frac{1}{\lambda},$$

where the second inequality follows from $(1+z)\mathrm{e}^{-z} \le 1$ for $z \ge 0$. If $\Delta_t < \delta_{t-1}^{\max}$, then $\delta_t^{\max} = \delta_{t-1}^{\max}$ and

$$\kappa_t \le \Delta_t + \mathrm{e}^{-\lambda\Delta_t}\left(\delta_t^{\max} + \frac{1}{\lambda}\right) \le \delta_t^{\max} + \frac{1}{\lambda},$$

because $\delta_t^{\max} \ge \Delta_t$ and

$$\Delta_t \le \left(1 - \mathrm{e}^{-\lambda\Delta_t}\right)\left(\Delta_t + \frac{1}{\lambda}\right) \le \left(1 - \mathrm{e}^{-\lambda\Delta_t}\right)\left(\delta_t^{\max} + \frac{1}{\lambda}\right),$$

where the first inequality is equivalent to $\mathrm{e}^z \ge 1 + z$ with $z = \lambda\Delta_t$. This proves equation 15 by induction.

With $\bar{A}_{t,d} = \exp(\Delta_{t,d}A_d)$, $\bar{B}_{t,d} = \Delta_{t,d}B_t$, and zero initial states, the contribution of state $n$ to channel $d$ is

$$y_{t,d,n} = C_{t,n} \sum_{k=1}^{t} \Delta_{k,d} \exp\left(a_{d,n} \sum_{j=k+1}^{t} \Delta_{j,d}\right) B_{k,n}x_{k,d}.$$

From $B_k = W_B x_k$ and $C_t = W_C x_t$, Cauchy–Schwarz gives

$$|B_{k,n}| \le X_t\|(W_B)_{n,:}\|_2, \qquad |C_{t,n}| \le X_t\|(W_C)_{n,:}\|_2, \qquad |x_{k,d}| \le X_t,$$

and hence

$$|y_{t,d,n}| \le X_t^3 q_n \kappa_{t,d,n}.$$

Therefore, by equation 15 and the triangle inequality,

$$\|y_{t,n}\|_2 \le X_t^3 q_n \left[\sum_{d=1}^{D} \left(\delta_{t,d}^{\max} + |a_{d,n}|^{-1}\right)^2\right]^{1/2} \le X_t^3 q_n \left(u_t + \beta_n\right).$$

For any fixed $\tau > 0$,

$$u_t + \beta_n \le (u_t + \tau)\left(1 + \frac{\beta_n}{\tau}\right),$$

hence

$$\|y_{t,n}\|_2 \le X_t^3 (u_t + \tau)\eta_n^A(\tau).$$

State truncation removes the contributions of states $n \notin \mathcal{S}$ and leaves the retained diagonal dynamics unchanged. Therefore,

$$E_N(t) \le X_t^3 (u_t + \tau)\epsilon_N(\mathcal{S}; \tau). \tag{16}$$

Because the softplus function is 1-Lipschitz, the following componentwise bound holds:

$$\Delta_{k,d} \le \mathrm{softplus}(b_{\Delta,d}) + |(M_\Delta x_k)_d|.$$

Taking the maximum over $k \le t$ and then the Euclidean norm over channels gives

$$\begin{aligned} u_t &\le \|\,\mathrm{softplus}(b_\Delta)\|_2 + X_t \|M_\Delta\|_F \\ &= b_0 + X_t m_\Delta. \end{aligned}$$

Consequently,

$$u_t + \tau \le (1 + X_t)\left(\tau + b_0 + m_\Delta\right) = (1 + X_t)\Phi_N.$$

Substituting this into equation 16 yields

$$E_N(t) \le X_t^3 (1 + X_t)\Phi_N \epsilon_N(\mathcal{S}; \tau). \tag{17}$$

We next bound the additional error caused by replacing $M_\Delta$ with $\widehat{M}_\Delta$ on the retained states. Since softplus is 1-Lipschitz, we have

$$|\Delta_{k,d} - \widehat{\Delta}_{k,d}| \le |(E_\Delta x_k)_d| \le X_t \|(E_\Delta)_{d,:}\|_2. \tag{18}$$

Moreover, the assumption $\|\widehat{M}_\Delta\|_F \le \|M_\Delta\|_F$ implies

$$\widehat{\Delta}_{k,d} \le b_0 + X_t m_\Delta =: \overline{\Delta}_t \qquad \text{for all } k \le t, d.$$

For $n \in \mathcal{S}$, let $h_{k,d,n}^{\mathcal{S}}$ and $\widehat{h}_{k,d,n}$ denote the corresponding states of the intermediate and final reduced models, and define

$$e_{k,d,n} := h_{k,d,n}^{\mathcal{S}} - \widehat{h}_{k,d,n}.$$

Since $0 < \mathrm{e}^{a_{d,n}\widehat{\Delta}_{j,d}} \le 1$ and the initial state is zero, the triangle inequality gives

$$|\widehat{h}_{k-1,d,n}| \le \sum_{\ell=1}^{k-1} \widehat{\Delta}_{\ell,d} |B_{\ell,n}||x_{\ell,d}| \le (k-1)\overline{\Delta}_t X_t^2 \|(W_B)_{n,:}\|_2.$$

Furthermore, the mean value theorem and $a_{d,n} < 0$ give

$$\left|\mathrm{e}^{a_{d,n}\Delta_{k,d}} - \mathrm{e}^{a_{d,n}\widehat{\Delta}_{k,d}}\right| \le |a_{d,n}||\Delta_{k,d} - \widehat{\Delta}_{k,d}|.$$

Subtracting the two state recursions therefore yields

$$\begin{aligned} |e_{k,d,n}| \le\ & |e_{k-1,d,n}| \\ & + |\Delta_{k,d} - \widehat{\Delta}_{k,d}| \|(W_B)_{n,:}\|_2 X_t^2 \left[1 + (k-1)\overline{\Delta}_t |a_{d,n}|\right]. \end{aligned}$$

Summing from $k = 1$ to $t$ and using equation 18 gives

$$|e_{t,d,n}| \le X_t^3 \|(E_\Delta)_{d,:}\|_2 \|(W_B)_{n,:}\|_2 \left[t + \frac{t(t-1)}{2}\overline{\Delta}_t |a_{d,n}|\right]. \tag{19}$$

Since the skip connections are identical, they cancel in the output difference. Using $|C_{t,n}| \le X_t \|(W_C)_{n,:}\|_2$ and equation 19,

$$|y^{\mathcal{S}}_{t,d} - \widehat{y}_{t,d}| \le X_t^4 \|(E_\Delta)_{d,:}\|_2 \left[ t \sum_{n\in\mathcal{S}} q_n + \frac{t(t-1)}{2} \overline{\Delta}_t \sum_{n\in\mathcal{S}} |a_{d,n}| q_n \right]$$
$$\le X_t^4 \|(E_\Delta)_{d,:}\|_2 \left[ LQ + \frac{L(L-1)}{2} \overline{\Delta}_t A_Q \right].$$

Taking the Euclidean norm over channels gives

$$E_R(t) \le X_t^4 \left[ LQ + \frac{L(L-1)}{2} \overline{\Delta}_t A_Q \right] \|E_\Delta\|_F.$$

Since $\overline{\Delta}_t = b_0 + X_t m_\Delta$,

$$LQ + \frac{L(L-1)}{2} \overline{\Delta}_t A_Q = LQ + \frac{L(L-1)}{2} (b_0 + X_t m_\Delta) A_Q$$
$$\le (1 + X_t) \left[ LQ + \frac{L(L-1)}{2} (b_0 + m_\Delta) A_Q \right]$$
$$= (1 + X_t) \Phi_R.$$

Hence,

$$E_R(t) \le X_t^4 (1 + X_t) \Phi_R \epsilon_R(\widehat{M}_\Delta). \tag{20}$$

Combining equation 17 and equation 20 with equation 14 yields

$$\|y_t - \widehat{y}_t\|_2 \le X_t^3 (1 + X_t) \Phi_N \epsilon_N(\mathcal{S}; \tau) + X_t^4 (1 + X_t) \Phi_R \epsilon_R(\widehat{M}_\Delta).$$

Finally,

$$X_t^3 (1 + X_t) \le X_t^3 (1 + X_t)^2$$

and

$$X_t^4 (1 + X_t) \le X_t^3 (1 + X_t)^2$$

for every $X_t \ge 0$. Therefore,

$$\|y_t - \widehat{y}_t\|_2 \le X_t^3 (1 + X_t)^2 \left[ \Phi_N \epsilon_N(\mathcal{S}; \tau) + \Phi_R \epsilon_R(\widehat{M}_\Delta) \right],$$

which proves equation 6. □

### C.3 Relation Between the Rank Tail and Factorized Regularizer

For $0 < p \le 1$, define $\|M\|_{\mathcal{S}_p} := \left( \sum_{i=1}^{D} \sigma_i(M)^p \right)^{1/p}$. Applying the standard tail bound to the singular values gives

$$\epsilon_R^{\min}(r_R) \le r_R^{1/2 - 1/p} \|M_\Delta\|_{\mathcal{S}_p}. \tag{21}$$

For any $M \in \mathbb{R}^{D\times D}$ with $\operatorname{rank}(M) \le R$, the variational characterization of (Giampouras et al., 2020) gives

$$\|M\|_{\mathcal{S}_p}^p = \min_{\substack{U\in\mathbb{R}^{D\times R}, V\in\mathbb{R}^{D\times R} \\ UV^\top = M}} \frac{1}{2^p} \sum_{j=1}^{R} \left( \|U_{:,j}\|_2^2 + \|V_{:,j}\|_2^2 \right)^p.$$

Setting $U = W_\Delta$ and $V = W_\rho^\top$ and using equation 9 with $\varepsilon > 0$, we obtain

$$\|M_\Delta\|_{\mathcal{S}_p}^p \le \frac{1}{2^p} \mathcal{R}_R(W_\Delta, W_\rho). \tag{22}$$

Combining equation 21 and equation 22 proves equation 10. Thus, $\mathcal{R}_R$ is a target-independent surrogate that controls the minimum rank-reduction residual for every $r_R$. After training, the truncated SVD in equation 13 achieves that rank-reduction residual.