MAMBA-B3: BLOCK-STRUCTURED, BOUNDED, AND BIJECTIVE FREQUENCY SSM FOR VISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision state-space models (SSMs) typically use real-valued diagonal transitions, which efficiently propagate information but cannot represent rotational dynamics that naturally arise in periodic textures, oriented structures, and motion. We introduce Mamba-B3, a vision SSM that combines three complementary modifications to the state transition dynamics. B1 replaces independent real-valued channels with block-structured decay–rotation dynamics and introduces low-rank cross-frequency coupling within each head, allowing frequency components to interact during state evolution. B2 controls the resulting non-normal dynamics through a spectral budget and a block-wise reachability regularizer, providing a sufficient condition for stable state propagation while improving input conditioning. B3 uses bilinear discretization, which, within the one-parameter family of methods, is uniquely able to preserve imaginary-axis modes on the unit circle while maintaining a one-to-one mapping between continuous rotation rates and discrete angles. We evaluate Mamba-B3 on ImageNet-1K, ADE20K, COCO, Something-Something V2, and Breakfast. Across image classification, dense prediction, and video recognition, Mamba-B3 consistently improves over the corresponding vision-Mamba baselines, while systematic component and interaction ablations support the roles of coupling, stability conditioning, and discretization. These results suggest that explicitly modeling and controlling frequency dynamics is a promising direction for more expressive visual SSMs.

1 INTRODUCTION

State space models (SSMs) have become a practical alternative to transformers for visual perception, offering linear-time sequence modeling with inductive biases suited to spatial and temporal structure (Gu et al., 2022b; Gu & Dao, 2024). Vision-adapted instantiations (Zhu et al., 2024; Liu et al., 2024; Li et al., 2024) are competitive with attention-based architectures in classification, dense prediction, and video understanding. These commonly used formulations inherit one limitation from language modeling: the state transition matrix is diagonal and real-valued, with continuous eigenvalues on the negative real axis and discrete ones in the unit interval, so it cannot rotate the hidden state. Textures and oriented features repeat, and video adds frequency-dependent motion; encoding a repeating pattern means tracking visual features at their signal frequency, which a matrix that scales each channel independently cannot do at any depth or width. Recent theoretical work exposes complementary state tracking limitations of conventional SSMs: Grazzi et al. (2025) show limitations associated with non-negative spectra and recover state tracking by admitting negative eigenvalues, while Merrill et al. (2024) establish broader expressivity limits for SSM state tracking. Ran-Milo et al. (2024) prove that a real diagonal SSM matching a complex one typically requires exponentially large, unlearnable parameter values.

Mamba-3 (Lahoti et al., 2026) resolves the deficit for language by introducing complex-valued transitions equivalent to a block-diagonal matrix of 2×2 rotation sub-blocks. It does not study the vision-specific implications of complex transitions; our contributions address the following obstacles:

(i) Cross-frequency coupling. A complex block-diagonal $\mathbf{A}$ makes the per-head state a bank of independent *frequency channels*, one 2×2 block per rotation rate ω_h. Mamba-3's MIMO variant raises the rank of the input and output projections, but $\mathbf{A}$ itself stays block-diagonal, so no channel

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.

reads from another while the state moves across the image. Vision needs this: a texture is a fast repeating pattern whose strength varies slowly across a region, and reading it means combining a fast channel with a slow one. We keep the block-structured dynamics and add low-rank coupling

$$\mathbf{A} = \mathrm{blockdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_H) + \mathbf{U}\mathbf{V}^\top, \qquad \mathbf{U}, \mathbf{V} \in \mathbb{R}^{N \times R},\ R \ll N, \tag{1}$$

at $\mathcal{O}(NR)$ cost, with N the per-head state dimension and $H = N/b$ blocks. It couples the frequency channels inside one head, inside $\mathbf{A}$ rather than at the projections, which is the primary difference from Mamba-3 (Lahoti et al., 2026).

S4 (Gu et al., 2022b) used a normal-plus-low-rank $\mathbf{A}$ for a different reason, diagonalizing HiPPO-LegS, and diagonal only variants matched it (Gu et al., 2022a; Gupta et al., 2022), so selective SSMs (Gu & Dao, 2024; Dao & Gu, 2024) dropped the term. Our $\mathbf{U}\mathbf{V}^\top$ shares that algebra and none of that motivation: under a real diagonal $\mathbf{A}$ with time-invariant Δ there are no frequency channels to couple and the operator cannot leave the unit interval, but once the diagonal part is a bank of rotations and both ω_h and Δ are token-dependent, the coupling is the only path between channels and the only route by which the operator can grow, which is why it now needs a budget the LTI setting never did.

(ii) Stability and conditioning. Eigenvalues near or above one lead to poorly conditioned gradients over long scans, and over long sequences, the input map can separately lose its grip on the state, which negatively affects tasks where the deciding evidence arrives late. The coupling is the only route by which the operator can grow, so a spectral budget on it alone keeps the continuous matrix dissipative and bounds the discrete operator at every step size; and the full-state objective is ill-conditioned at realistic state sizes, so we condition each rotation block separately. With the task loss these give $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{stab}} + \mathcal{L}_{\mathrm{cond}}$.

(iii) Bijective frequency mapping. An undamped oscillation sits on the imaginary axis of the s-plane and must land on the unit circle after discretization. Within the one-parameter family $\mathcal{F}_\gamma$ of one-step schemes, only the bilinear transform does this for every rate and step size. Zero-order hold (ZOH), the default in prior vision SSMs, lies outside the family and holds the magnitude, but two rates that differ by $2\pi/\Delta$ give the same discrete block under ZOH, so different continuous rotation rates can induce the same discrete operator; under bilinear every rate gives a distinct block.

2 Related Work

Vision SSMs. Structured SSMs were reestablished by Gu et al. (2022b); Gu & Dao (2024) added selective gating and Dao & Gu (2024) introduced structured state space duality, both with diagonal transitions. Vision Mamba (Zhu et al., 2024) and VMamba (Liu et al., 2024) adapted these to images, VideoMamba (Li et al., 2024) to video, and MambaVision (Hatamizadeh & Kautz, 2025) to hybrids. Scan-order work (Huang et al., 2024; Yang et al., 2024), non-causal duality (Shi et al., 2025), native 2D recurrences (Baty et al., 2024) and adaptive scans (Li et al., 2025) raise hierarchical accuracy at this scale to 83.7 to 84.0, but all modify the scan or its causality, not the eigenvalue structure of the transition, and are orthogonal to ours.

Register tokens (Darcet et al., 2024; Wang et al., 2025) address high-norm artifacts, a symptom, rather than the eigenvalue magnitudes we constrain, and Lu et al. (2025) identify historical decay and element contradiction in Mamba's token-processing dynamics for video understanding. We compare against isotropic and hierarchical SSM backbones at matched scale.

Diagonal plus low-rank transitions. S4 (Gu et al., 2022b) writes $\mathbf{A}$ as normal-plus-low-rank, uses the bilinear transform, and inverts by Woodbury; the correction exists to make HiPPO-LegS diagonalisable, not to couple anything, and Gu et al. (2022a) and Gupta et al. (2022) showed that a plain diagonal $\mathbf{A}$ matches it, after which selective SSMs (Gu & Dao, 2024; Dao & Gu, 2024) kept the diagonal and dropped the correction. Our $\mathbf{U}\mathbf{V}^\top$ shares that algebra and none of that motivation, and Section 1 states the conditions under which the term stops being redundant.

Complex dynamics, stability, and conditioning. Grazzi et al. (2025), Ran-Milo et al. (2024) and Orvieto et al. (2024) prove the expressivity limit of non-negative real spectra; Lahoti et al. (2026) addresses it for languages with block-diagonal rotations, and their MIMO variant raises the rank of $\mathbf{B}$

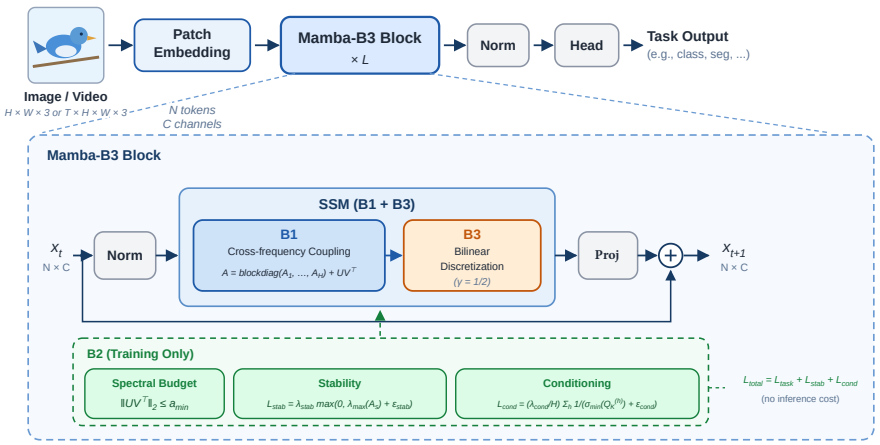


Figure 1: Mamba-B3 inside a ViM/VideoMamba block. B1 adds the low-rank coupling $\mathbf{UV}^\top$ to a block-diagonal rotation transition, and B3 discretizes the coupled transition with the bilinear transform. B2 acts only during training, through the spectral projection of $\mathbf{UV}^\top$ and the losses $\mathcal{L}_{\text{stab}}$ and $\mathcal{L}_{\text{cond}}$.

and $\mathbf{C}$ while $\mathbf{A}$ stays block-diagonal, whereas $\mathbf{UV}^\top$ couples channels inside $\mathbf{A}$. Dao & Gu (2024) and Lahoti et al. (2026) rely on QK-normalization as an implicit stabilizer; LRU (Orvieto et al., 2023) and StableSSM (Wang & Li, 2024) constrain the eigenvalues of a normal operator, which does not bound a non-normal one. We use spectral normalization (Miyato et al., 2018) of the coupling, a symmetric-part regularizer grounded in Willems (1972), and a per-block conditioning penalty built from finite-horizon Gramian theory (Chen, 1999; Kailath, 1980). Control-theoretic analyses of deep SSMs exist (Alonso et al., 2025), as diagnostics rather than training-time constraints; Gramian-based conditioning regularization has not appeared in a vision SSM backbone.

Discretization. Most vision SSMs adopt ZOH. Ibrahim et al. (2026) compares schemes on a real-diagonal backbone and reports bilinear outperforming ZOH, with higher-order holds stronger still; that study asks which scheme performs best, not which a given eigenvalue parameterization admits, which is the question this parameterization raises. Lahoti et al. (2026) combines an exponential state transition with a generalized trapezoidal approximation of the state-input integral; our analysis instead concerns the continuous-to-discrete eigenvalue mapping itself, exact across the full imaginary axis within the family $\mathcal{F}_\gamma$ of Eq. (7). Rusch & Rus (2025) preserves oscillations with an implicit-explicit integrator for a second-order system, outside the family Eq. (7) defines.

3 Methodology

3.1 Preliminaries and Block-Structured Dynamics

Figure 1 shows where each component enters the block. We model visual sequences with $\dot{\mathbf{h}}(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t)$, $\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t)$, $\mathbf{h}(t) \in \mathbb{R}^N$; discretization with step Δ gives $\mathbf{h}_{t+1} = \overline{\mathbf{A}}\mathbf{h}_t + \overline{\mathbf{B}}\mathbf{x}_t$. An overline marks a discretized quantity throughout. We use Eq. (1) with blocks $\mathbf{A}_h \in \mathbb{R}^{b\times b}$, $N = Hb$, each a continuous-time generator in decay-rotation form,

$$\mathbf{A}_h = a_h\mathbf{I} + \omega_h\mathbf{J}, \qquad \mathbf{J} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad a_h < 0, \tag{2}$$

with eigenvalues $a_h \pm i\omega_h$, where the decay rate a_h and the rotation rate ω_h are produced by data-dependent projections and $a_h < 0$ is enforced by construction. Discretization turns each block into an operator $\overline{\mathbf{A}}_h$ of magnitude r_h and angle $\theta_h = \arg\overline{\mathbf{A}}_h$, equal to $2\arctan(\omega_h\Delta/2)$ only when $a_h = 0$; Table 5 reports the measured value at convergence, and $\overline{\mathbf{A}}_h$ always denotes the discretization of the isolated block. The bilinear transform sends the open left half-plane into the open unit disk, so $a_h < 0$ gives $r_h < 1$: an uncoupled block is a contraction by construction, and the coupling is the only route by which the operator can grow.

3.2 Stability and Conditioning Constraints

Spectral normalization. Bounding $\overline{\mathbf{A}}$ directly is awkward, because it is input-dependent and rebuilt at every token. The decay rates carry a floor, $a_h \le -a_{\min} < 0$, so only the coupling can make $\|\overline{\mathbf{A}}\|_2$ exceed one. Write $\mathbf{A}_s = (\mathbf{A} + \mathbf{A}^\top)/2$, which governs energy growth through $\frac{d}{dt}\|\mathbf{h}\|_2^2 = 2\mathbf{h}^\top \mathbf{A}_s \mathbf{h}$. Since $\lambda_{\max}(\mathbf{A}_s) \le \max_h a_h + \|\mathbf{U}\mathbf{V}^\top\|_2$, a spectral budget on the coupling alone keeps the whole matrix dissipative, which is sufficient for non-expansion at every step size.

$$\|\mathbf{U}\mathbf{V}^\top\|_2 \;\le\; a_{\min} \quad\Longrightarrow\quad \mathbf{A}_s \;\preceq\; \mathbf{0}, \tag{3}$$

Expanding both sides of Eq. (6) gives $\|(\mathbf{I}+\frac{\Delta}{2}\mathbf{A})\mathbf{x}\|_2^2 - \|(\mathbf{I}-\frac{\Delta}{2}\mathbf{A})\mathbf{x}\|_2^2 = 2\Delta\,\mathbf{x}^\top\mathbf{A}_s\mathbf{x} \le 0$, so Eq. (3) gives $\|\overline{\mathbf{A}}\|_2 \le 1$ at *every* $\Delta > 0$, with nothing checked per token. The budget is enforced as spectral normalization of $\mathbf{U}\mathbf{V}^\top$ (Miyato et al., 2018) after each optimizer step. Its single hyperparameter is the floor $a_{\min}$, which sets the longest memory an uncoupled block can hold. The operator norm is the right quantity because $\overline{\mathbf{A}}$ is non-normal once the coupling is active, and eigenvalue magnitude alone does not bound it (Yu et al., 2024).

Symmetric-part regularization. The projection keeps $\lambda_{\max}(\mathbf{A}_s)$ at or below zero, but at zero the state neither grows nor decays. A soft penalty on the same scalar adds a margin,

$$\mathcal{L}_{\text{stab}} = \lambda_{\text{stab}}\,\max\big(0,\;\lambda_{\max}(\mathbf{A}_s) + \varepsilon_{\text{stab}}\big), \qquad 0 < \varepsilon_{\text{stab}} < a_{\min}. \tag{4}$$

It is zero when $\lambda_{\max}(\mathbf{A}_s) \le -\varepsilon_{\text{stab}}$ and active in the band between $-\varepsilon_{\text{stab}}$ and zero that the projection allows. Where it is zero, $\|\overline{\mathbf{A}}\|_2$ is strictly below one at every step size, so the state and the gradients through it decay along the scan rather than merely staying bounded. Its gradient reaches the coupling and the decay rates but not the rotation rates: the symmetric part of a block is $a_h\mathbf{I}$, which carries no ω_h, and with $\mathbf{U} = \mathbf{V} = \mathbf{0}$ we have $\lambda_{\max}(\mathbf{A}_s) = \max_h a_h \le -a_{\min} < -\varepsilon_{\text{stab}}$ and $\mathcal{L}_{\text{stab}}$ is identically zero. Rotation is energy-neutral, and the penalty does no work unless the coupling is active.

Per-block conditioning penalty. The natural object to measure is the reachability matrix $\mathbf{Q}_K = [\overline{\mathbf{B}}, \overline{\mathbf{A}\mathbf{B}}, \ldots, \overline{\mathbf{A}}^{K-1}\overline{\mathbf{B}}] \in \mathbb{R}^{N\times K}$, with $\overline{\mathbf{B}} \in \mathbb{R}^{N\times 1}$ per channel as in Mamba, but on the full state it does not work: for $K < N$ its K singular values describe only the subspace reached so far, and the smallest of them collapses. At $N = 128$ we measure $\sigma_{\min}$ falling from $1.45{\times}10^{-2}$ at $K{=}2$ to $8.3{\times}10^{-18}$ at $K{=}32$ with no recovery, so $\lambda/(\sigma_{\min} + \varepsilon)$ saturates and the gradient dies. It also conflicts with the contraction of Eqs. (3)-(4): for $K < N$, $\sigma_{\min}(\mathbf{Q}_K) \le \|\mathbf{Q}_K\mathbf{e}_K\|_2 = \|\overline{\mathbf{A}}^{K-1}\overline{\mathbf{B}}\|_2$ by definition, so the objective is capped by a quantity the contraction shrinks. We therefore evaluate reachability per rotation block and average,

$$\mathbf{Q}_K^{(h)} = \big[\overline{\mathbf{B}}_h,\, \overline{\mathbf{A}}_h\overline{\mathbf{B}}_h,\, \ldots,\, \overline{\mathbf{A}}_h^{K-1}\overline{\mathbf{B}}_h\big], \qquad \mathcal{L}_{\text{cond}} = \frac{\lambda_{\text{cond}}}{H}\sum_{h=1}^{H}\frac{1}{\sigma_{\min}(\mathbf{Q}_K^{(h)}) + \varepsilon_{\text{cond}}}. \tag{5}$$

With $b = 2$ and $\overline{\mathbf{B}}_h \in \mathbb{R}^{2\times 1}$, $\mathbf{Q}_K^{(h)} \in \mathbb{R}^{2\times K}$ is wide at every horizon we use; $\mathbf{Q}_K^{(h)}(\mathbf{Q}_K^{(h)})^\top$ is a sum of positive semi-definite terms, so $\sigma_{\min}$ is non-decreasing in K, and at $N = 128$ the median rises from $4.7{\times}10^{-3}$ to $1.0{\times}10^{-1}$ between $K{=}2$ and $K{=}32$. We average over blocks rather than take the worst, which is nearly always a block with angle near zero whose columns are rescalings of the first, hence rank one; $\sigma_{\min}$ and $|\theta_h|$ have rank correlation 0.83. Blocking removes the vanishing gradient but not the tension with the contraction, since the block Gramian grows with r_h; λ_{cond} sets the trade-off, and $\mathcal{L}_{\text{cond}}$ maintains conditioning rather than restoring it. $\mathbf{Q}_K^{(h)}$ is the frozen linearization of the isolated block at one cached token with the coupling removed, so $\mathcal{L}_{\text{cond}}$ conditions that block and not the coupled scan; $\rho(\overline{\mathbf{A}})$ of the coupled operator is reported in Table 5 as a diagnostic.

3.3 Bilinear Discretization

Under ZOH the discrete block is $\exp(\mathbf{A}_h\Delta)$, so a rotation rate ω becomes the angle $\omega\Delta$; since angles differing by 2π coincide, the rates ω and $\omega + 2\pi/\Delta$ produce the same block, the pole-domain form of aliasing (Åström & Wittenmark, 2013; Oppenheim & Schafer, 2010). With ω_h and Δ both

data-dependent, the map from rate to operator is non-injective: the same ω_h can be a fast rotation at one token and a slow or reversed one at the next. Bilinear maps the imaginary axis onto the unit circle exactly once, with angle strictly increasing in $\omega\Delta$, so the rate is identifiable and never folds. The blocks at risk are those near angle π, the fine textures that repeat every few tokens, and because Δ varies per token no statistic of the angle distribution can certify that a rate is never folded; bilinear removes the possibility, so B3 is an exclusion result and Section 5.6 tests only its observable consequence. Substituting $s = \frac{2}{\Delta}\frac{z-1}{z+1}$ gives

$$\overline{\mathbf{A}} = \big(\mathbf{I} - \tfrac{\Delta}{2}\mathbf{A}\big)^{-1}\big(\mathbf{I} + \tfrac{\Delta}{2}\mathbf{A}\big), \qquad \overline{\mathbf{B}} = \big(\mathbf{I} - \tfrac{\Delta}{2}\mathbf{A}\big)^{-1}\Delta\mathbf{B}, \tag{6}$$

with $|z| = 1$ exactly on the imaginary axis and $\theta = 2\arctan(\omega\Delta/2)$. As in S4 (Gu et al., 2022b), the Woodbury identity replaces the $N \times N$ inverse with the 2×2 blocks and one $R \times R$ solve, rebuilt per token because Δ, a_h and ω_h are: applying $\mathbf{U}\mathbf{V}^\top$ costs $\mathcal{O}(NR)$, obtaining $\overline{\mathbf{A}}$ costs $\mathcal{O}(NR^2 + R^3)$.

Bilinear was a numerical convenience in S4 (Gu et al., 2022b) and DSS (Gupta et al., 2022); once the transition is complex-valued the choice is no longer free. For the one-parameter family $\mathcal{F}_\gamma$ of one-step schemes (the θ-method, written with γ because θ is the rotation angle), which evaluates $\mathbf{A}$ at a γ-weighted mix of the two interval endpoints, requiring a purely imaginary eigenvalue to land on the unit circle for every ω and Δ gives

$$|\overline{A}(\gamma)|^2 = \frac{1 + (1-\gamma)^2\omega^2\Delta^2}{1 + \gamma^2\omega^2\Delta^2} = 1 \iff (1-\gamma)^2 = \gamma^2 \iff \gamma = \tfrac{1}{2}, \tag{7}$$

Any other γ moves imaginary-axis eigenvalues off the circle at every step size: for $\gamma < \frac{1}{2}$ the mode expands, for $\gamma > \frac{1}{2}$ it decays. Bilinear is therefore the unique member of $\mathcal{F}_\gamma$ mapping the imaginary axis *bijectively* onto the unit circle minus $z = -1$; its angle warps the rate axis but is invertible, $\omega = (2/\Delta)\tan(\theta/2)$, so the warp is absorbed into the learned ω_h. ZOH lies outside the family, since $\exp(\mathbf{A}\Delta)$ is not rational, and is excluded on identifiability rather than admissibility. Eq. (7) is the classical unit-modulus property of the θ-method (Wanner & Hairer, 1996) and the bilinear map is Tustin's (Tustin, 1947); what changes here is that, with ω_h and Δ token-dependent, the property decides the discretization rather than leaving it a hyperparameter. The result does not cover higher-order schemes, which Ibrahim et al. (2026) show can be more accurate at greater cost.

4 Experiments

Mamba-B3 replaces the selective SSM layer in vision-adapted Mamba backbones, ViM-S (Zhu et al., 2024) for image tasks and VideoMamba-S (Li et al., 2024) for video, with everything else unchanged, so ablation variants differ only in the internal dynamics. Each dataset stresses a property the theory predicts. SSv2 (Goyal et al., 2017) carries the full 2^3 ablation: order-sensitive temporal reasoning probes the rotational expressivity B3 preserves, and late-arriving evidence probes conditioning. ImageNet-1K (Deng et al., 2009) at two resolutions tests stability under longer scans. ADE20K (Zhou et al., 2019) and COCO (Lin et al., 2014) target the coupling, since textured segmentation and multi-scale localization both require frequency channels to combine. Breakfast (Kuehne et al., 2014) tests whether late-arriving evidence still reaches the state. The coupling breaks the elementwise associative scan and the coupled path runs as a sequential recurrence, so its cost is not captured by FLOPs and we report measured images/second at both resolutions and peak memory at 224.

Table 1 reports that cost on a single H100. It is concentrated almost entirely in B1, because the coupling is the only component that changes the scan; B2 acts through the loss and a projection and has no inference cost, and B3 adds one small solve. Rank R is therefore the efficiency knob: R=1 recovers half of the throughput lost at R=2, at 0.2M fewer parameters.

For classification, a larger evaluation resolution makes the spatial sequence longer. Because the 384 model is fine-tuned from the 224 checkpoint with the operator otherwise unchanged, the change in accuracy measures how the trained operator adapts to a longer scan. Resolution also changes image content, so the column is a length test only up to that confound. The prediction is that variants with B2 active tolerate the longer scan at least as well as those without it.

Textured segmentation must hold the phase of a repeating pattern across a long scan and combine a fast channel with a slow one to track how the pattern's strength varies; a boundary is a burst aligned in

Table 1: Throughput and peak memory on a single H100, batch 128, fp16. Memory is at 224.

Variant	Params (M)	GFLOPs	img/s @224	img/s @384	Peak mem. (GB)
Baseline (real diag., ZOH)	26.0	5.1	1320	473	8.5
B3 only	26.0	5.2	1294	461	8.6
B2 only	26.0	5.1	1320	473	8.5
B1 only	26.4	5.7	952	348	9.8
Mamba-B3 (Full)	26.4	5.8	940	334	10.3
Mamba-B3 (Full), R=1	26.2	5.4	1130	402	9.2

Table 2: ImageNet-1K. Top-1@384 is fine-tuned from the 224 checkpoint; the final column is the change from the longer scan.

Model	Params (M)	GFLOPs	Top-1@224	Top-1@384	$384-224$
Isotropic SSM backbones					
ViM-S	26.0	5.1	80.3	–	–
LocalVim-S	28.0	4.8	81.2	–	–
Mamba-R-S	28.0	9.9	81.1	–	–
MambaScope-S	28.0	4.3	81.2	–	–
PlainMamba-L2	25.7	8.1	81.6	–	–
Hierarchical SSM backbones					
VMamba-T	30.0	4.9	82.6	–	–
VSSD-T	24.0	4.5	83.7	–	–
M2D-T	26.5	4.7	84.0	–	–
Baseline (real diag., ZOH)	26.0	5.1	80.3	80.8	+0.5
Baseline, parameter-matched	26.4	5.5	80.6	81.2	+0.6
B1 only	26.4	5.7	81.1	81.4	+0.3
B1 + B2	26.4	5.7	82.0	83.2	+1.2
B1 + B3	26.4	5.8	82.4	83.1	+0.7
Mamba-B3 (Full)	26.4	5.8	83.0	**84.4**	**+1.4**

phase across many rates. Object size is spatial frequency seen another way, so multi-scale localization calls on the same mechanism. Our detection rows use Mask R-CNN with the $3\times$ multi-scale schedule of Liu et al. (2021), as do the hierarchical comparators; Table 3 gives the results for dense prediction.

Breakfast (Kuehne et al., 2014) contains 1,712 videos across 10 cooking activities, averaging 2.3 minutes with roughly six sub-actions per video. Many classes share identical opening sub-actions, with the discriminating sub-actions appearing only in the final segments, so a model that loses input responsiveness partway through never registers the distinguishing evidence. We fine-tune from Kinetics-400 (Kay et al., 2017) following Li et al. (2024) with 32 frames at 224×224, re-evaluate on the first 50% and 75% of sampled frames, and report all three accuracies against the end-to-end VideoMamba models of Li et al. (2024) and the feature-based ViS4mer (Islam & Bertasius, 2022) and S5+LSMCL (Wang et al., 2023) in Table 4 to compare and contrast our model on long horizon video recognition.

Configuration. All variants use coupling rank R=2 unless noted, per-head state N=128 with b=2 (H=64 rotation blocks when B1 is active; none when it is not), decay floor $a_{\min}$=1.0, margin $\varepsilon_{\text{stab}}$=0.05, λ_{stab}=1.0, λ_{cond}=0.01, horizon K=8 and $\varepsilon_{\text{cond}}=10^{-3}$. $\mathcal{L}_{\text{cond}}$ is averaged over eight cached tokens per layer and each block term is clamped at $1/\varepsilon_{\text{cond}}$. $\mathbf{U}$ is initialized to zero and $\mathbf{V}$ at $\mathcal{N}(0, 0.02^2)$, so training starts from the block-diagonal transition; ω_h is initialized log-uniform on $[0.1/\bar{\Delta},\ 3/\bar{\Delta}]$ with $\bar{\Delta}$ the per-layer mean of the Mamba Δ initialization, and Δ follows the soft-plus projection of Gu & Dao (2024); $\|\mathbf{U}\mathbf{V}^\top\|_2$ is computed exactly from the $R\times R$ Gram matrices. Optimizer, schedule, augmentation and epoch budget follow Zhu et al. (2024) for classification and Li et al. (2024) for video, unchanged across variants; Tables 1 to 4 report a single seed, Table 5 report five per row, and Table 6 report twenty per cell with seeds, data order, and augmentation streams shared across cells, which is what makes the permutation test paired.

Table 3: ADE20K semantic segmentation with a UperNet head (Xiao et al., 2018) and COCO object detection (Lin et al., 2014).

	ADE20K			COCO			
Model	Params (M)	mIoU (SS)	mIoU (MS)	Params (M)	AP	AP_{50}	AP_{75}
Isotropic SSM							
ViM-S	46	44.9	–	–	–	–	–
MambaScope-S	48	45.5	–	–	–	–	–
PlainMamba-L2	55	46.8	–	53	46.0	66.9	50.1
Hierarchical SSM							
VSSD-T	53	47.9	48.7	44	48.8	70.4	53.4
VMamba-T	62	47.9	48.8	50	48.8	70.4	53.5
Spatial-Mamba-T	57	48.6	49.4	46	49.3	70.7	54.3
DAMamba-T	55	50.3	51.2	45	50.4	71.4	55.5
Baseline	46	44.8	45.7	46	46.4	67.6	50.5
Mamba-B3 (Full)	48	49.0	50.0	47	49.4	71.0	54.0

Table 4: Breakfast long-horizon activity recognition. Top-1 when the model sees the first 50%, 75% and 100% of sampled frames. The discriminating sub-actions fall in the final quartile.

Model	Backbone	Params (M)	50%	75%	100%	final-quartile gain
ViS4mer	Swin-B + SSM	–	–	–	88.2	–
S5+LSMCL	Swin-B + SSM	–	–	–	90.7	–
VideoMamba-Ti$_{f32}$	VideoMamba-Ti	7	–	–	94.3	–
VideoMamba-S$_{f32}$	VideoMamba-S	26	–	–	95.3	–
Mamba-B3, B2 inactive	VideoMamba-S	26	77.4	87.2	95.6	$+8.4$
Mamba-B3 (Full)	VideoMamba-S	26	77.0	87.5	**97.1**	$+\mathbf{9.6}$

5 RESULTS

5.1 COMPUTATIONAL COST

Table 1 shows B2 contributes zero parameters and zero FLOPs, as expected for a projection applied after the optimizer step. B3 contributes 0.1 GFLOPs and no parameters, the cost of the per-token solution of Eq. (6), with B1 off this is one scalar division per channel, and with B1 active it is the 2×2 block inverses and the single $R\times R$ solve. B1 contributes 0.4M parameters and 0.6 GFLOPs, which is 86% of the full model's 0.7 GFLOP increment over the baseline. Reducing the coupling rank to $R{=}1$ lowers the full model from 5.8 to 5.4 GFLOPs. The throughput columns separate the components to assess each impact on efficiency of the Mamba-B3 model. B2 only matches the baseline at 1320 img/s, the projection runs after the optimizer step and the loss terms are not evaluated at inference. B3 costs 2% (1294), B1 costs 28% (952). That is the price of the coupled recurrence; the associative scan requires the composition of two transitions to stay in the same structured family, and the product of two diagonal-plus-rank-R matrices is generically not diagonal-plus-rank-R, so the coupled path is evaluated token by token. The complete model runs at 940 img/s, 29% below the baseline and 1% below B1 only, so B3 does not add anything measurable once B1 is active.

5.2 IMAGE CLASSIFICATION

Mamba-B3 is isotropic; constant token resolution and channel width at every layer, with no convolutional stem and no downsampling stages. Three results follow from Table 2. (i) The full model attains 83.0 at 224 and 84.4 at 384, exceeding the baseline by 2.7 and 3.6 respectively, and leading every isotropic backbone at this scale: $+1.4$ over PlainMamba-L2 (Yang et al., 2024) at 28% fewer FLOPs, $+1.8$ over MambaScope-S (Liu et al., 2025), $+1.9$ over Mamba-R-S (Wang et al., 2025), $+1.8$ over LocalVim-S (Huang et al., 2024) and $+2.7$ over ViM-S (Zhu et al., 2024). Against hierarchical backbones it clears VMamba-T (Liu et al., 2024) by 0.4 at 3.6M fewer parameters but trails VSSD-T (Shi et al., 2025) by 0.7 and M2D-T (Baty et al., 2024) by 1.0; both add a convolutional stem and four

Table 5: Exclusive 2^3 ablation on SSv2. $\rho(\overline{\mathbf{A}})$ is the spectral radius at convergence, $\sigma_{\min}$ and $|\theta|$ are means over blocks from Eq. (5) and Eq. (2); $|\theta|$ is reported as a location statistic for the rate distribution. Rank 1 indicates $\mathcal{L}_{\text{cond}}$ is degenerate. All variants replace the selective SSM layer of VideoMamba-S (26M parameters) and are evaluated at 16×224^2.

Variant	B1	B2	B3	Top-1	$\rho(\overline{\mathbf{A}})$	$\sigma_{\min}$	rank	$\lvert\theta\rvert$
Published, IN-1K pretrain, 16×224^2								
VideoMamba-Ti (7M)	–	–	–	66.0	–	–	–	–
VideoMamba-S (26M)	–	–	–	67.6	–	–	–	–
Baseline	×	×	×	67.6	0.93	$\equiv 0$	1	0.00
B1 only	✓	×	×	68.2	1.03	0.021	2	0.19
B2 only	×	✓	×	67.6	0.93	$\equiv 0$	1	0.00
B3 only	×	×	✓	67.9	0.92	$\equiv 0$	1	0.00
B1 + B2	✓	✓	×	68.9	0.96	0.074	2	0.26
B1 + B3	✓	×	✓	69.4	0.99	0.029	2	0.21
B2 + B3	×	✓	✓	67.9	0.92	$\equiv 0$	1	0.00
Stability sub-ablation (within B2)								
B1+B3, spectral norm only	✓	○	✓	69.6	0.96	0.031	2	0.22
B1+B3, conditioning only	✓	○	✓	69.9	0.98	0.069	2	0.27
Mamba-B3 (Full)	✓	✓	✓	**70.4**	0.95	**0.081**	2	**0.29**

Table 6: Discretization interaction on SSv2. Twenty seeds per cell; p from a one-sided paired permutation test.

State structure	Discretization	Top-1	Effect of bilinear
Diagonal real	ZOH	67.6 ± 0.21	$+0.3$
Diagonal real	Bilinear	67.9 ± 0.19	
Complex block	ZOH	68.2 ± 0.34	$+\mathbf{1.2}$
Complex block	Bilinear	69.4 ± 0.28	
Interaction $\delta = 1.2 - 0.3 = +\mathbf{0.9}$		(ratio $4.0\times$, $p < 10^{-6}$)	

resolution stages, and we make no claim against that class on this task. (ii) The parameter-matched baseline isolates capacity from architecture. At the same 26.4M parameters it attains 80.6, 2.4 below the full model, so the gain is not attributable to parameter count. (iii) The $384-224$ column tests the stability prediction of Section 4. The two rows with B2 active gain 1.2 and 1.4 from the longer sequence; the four rows without B2 gain between 0.3 and 0.7, and B1 only gains least at 0.3. The effect is concentrated at the longer scan: B1+B3 exceeds B1+B2 by 0.4 at 224 but not at 384, a swing of 0.5 that B3 alone does not supply.

5.3 Dense Prediction

Table 3 shows the full model attains 49.0 mIoU (SS) and 50.0 (MS), 4.2 and 4.3 above the baseline on ADE20K, 2.2 above the strongest isotropic SSM, PlainMamba-L2 (Yang et al., 2024), and 1.1 (SS) above VMamba-T (Liu et al., 2024) and VSSD-T (Shi et al., 2025) at 5 to 14M fewer parameters. On COCO under Mask R-CNN with the $3\times$ multi-scale schedule it attains 49.4 AP, 3.0 above the baseline, with AP_{50} and AP_{75} rising by 3.4 and 3.5, and 0.6 above both VMamba-T and VSSD-T, at 3M fewer parameters than the former and 3M more than the latter. The isotropic comparison on COCO is incomplete rather than favorable: no isotropic SSM reports this schedule, and PlainMamba-L2 is quoted at $1\times$ with a ViTAdapter neck, which is not comparable. Table 3 reports the full model against the baseline only; the component ablation is carried by SSv2 in Table 5. The magnitude of the gains is consistent with the mechanism of Section 4.

5.4 Long-Horizon Video Recognition

The primary result of Table 4 is full-video Top-1: the full model attains 97.1, 1.5 above the B2-inactive variant and 1.8 above the unmodified VideoMamba-S under the same protocol (Li et al., 2024). The two variants are within 0.4 of each other through 75% of frames (77.0 against 77.4, then 87.5 against 87.2) and separate in the final quartile, where the B2-inactive variant gains 8.4 and the full model 9.6; that 1.2 accounts for most of the 1.5 margin. This is the pattern the per-block conditioning argument predicts, since the discriminating sub-actions fall in the last quartile. Both variants clear the unmodified backbone, but by 1.8 against 0.3: the coupling alone recovers little of what the conditioning term delivers, consistent with the $\rho = 0.99$ of the B1+B3 row in Table 5.

5.5 Exclusive Ablations

Table 5 activates each subset of components with the rest at baseline (real diagonal $\mathbf{A}$, ZOH). With B1 off, B2 has nothing to act on (Section 3.2), so *B2 only* and *B2+B3* are structural null controls and the design reduces to $\{\text{B1, B3}\}$ with B2 crossed on B1; $\circ$ marks a partially active B2.

The table satisfies four predictions. (i) B1 alone is unstable: $\rho(\overline{\mathbf{A}}) = 1.03$ is the only value above one, and every row combining B1 with B2 returns below it (0.96 for B1+B2, 0.95 for Full). (ii) B2 is inert without B1: the null controls reproduce Baseline and B3 only exactly (67.6, 67.9), with $\sigma_{\min} \equiv 0$. (iii) The full model exceeds every pair, 70.4 against 69.4 for B1+B3. (iv) The sub-components of B2 separate on the diagnostics: spectral normalization alone leaves $\sigma_{\min}$ at 0.031, level with B1+B3 (0.029), while the conditioning term raises it to 0.069 alone and 0.081 with both, so the gain in $\sigma_{\min}$ is attributable to $\mathcal{L}_{\text{cond}}$.

The single-component gains of B1 (0.6) and B3 (0.3) sum to 0.9, but B1+B3 gains 1.8; the 0.9 excess is the interaction of Section 5.6. $\sigma_{\min}$ is comparable only among rows with $\rho < 1$: B1 only reaches 0.021 at $\rho = 1.03$, outside the stable region, the trade-off of Section 3.2 rather than a failure of the penalty. The baseline is our own training run of the unmodified layer and reproduces the published VideoMamba-S figure at 16 frames (67.6); the full model exceeds it by 2.8 at the same 26M (Li et al., 2024).

5.6 Discretization Interaction

The uniqueness result predicts that bilinear should beat ZOH by more when $\mathbf{A}$ has a complex block structure than when it is real-diagonal, where there is no rotational geometry to preserve:

$$\big[\text{Acc}(\text{B1+B3}) - \text{Acc}(\text{B1 only})\big] > \big[\text{Acc}(\text{B3 only}) - \text{Acc}(\text{Baseline})\big], \tag{8}$$

tested with a one-sided paired permutation test over twenty seeds per cell, whose attainable floor is 2^{-20}; a null or reversed interaction would not overturn the uniqueness result, which is an exclusion rather than a performance claim. Table 6 confirms Eq. (8). Bilinear adds 0.3 over ZOH on the real-diagonal state and 1.2 on the complex-block state, an interaction of $\delta = +0.9$ (ratio $4.0\times$). The seed standard deviations are informative in their own right: the two complex-block cells show 0.34 and 0.28 against 0.21 and 0.19 for the diagonal cells. Both complex-block cells run B1 without B2, at $\rho = 1.03$ under ZOH and $\rho = 0.99$ under bilinear, so both sit at or just inside the boundary the spectral budget defines; the elevated variance is consistent with the instability exposed by the B1-only condition and motivates the role of B2.

6 Conclusion

Mamba-B3 revives the diagonal-plus-low-rank transition of S4 (Gu et al., 2022a;b) in a setting where the low-rank term does work that the diagonal cannot: coupling frequency channels inside a complex block-structured state under token-dependent Δ. The three contributions of this work are not independent additions. The dependencies are supported at three levels: B2 is structurally degenerate without B1; the spectral budget provides a sufficient stability condition for the coupled dynamics; and, within $\mathcal{F}_\gamma$, bilinear discretization is uniquely compatible with imaginary-axis preservation. The ablations of Section 5.5 support each dependency in the direction claimed. Among isotropic vision SSMs at 25 to 28M parameters, Mamba-B3 leads every backbone on each task where a comparable isotropic result exists.

AI USE STATEMENT

Generative AI tools were used for LaTeX formatting, grammar and clarity revision of author-written text, and for checking internal consistency of numbers between tables and prose. The architecture, derivations, code, experiments and all reported numbers were produced by the authors; every AI-suggested edit was verified against the source before inclusion. No AI tool was used to generate results, figures of results, or citations.

REPRODUCIBILITY STATEMENT

The block-structured coupling, the stability and conditioning constraints, and the bilinear discretization are specified in Section 3, with every architectural and loss equation stated there (Eqs. 1–7). The complete hyperparameter set, coupling rank R, per-head state size N, block size b, decay floor $a_{\min}$, stability margin $\varepsilon_{\text{stab}}$, λ_{stab}, λ_{cond}, conditioning horizon K, and $\varepsilon_{\text{cond}}$, is given in the Configuration paragraph of Section 4, and every variant modifies only the selective SSM layer of the stated ViM-S and VideoMamba-S backbones with all other training settings left at their published values. All theoretical claims are proved in full in Appendix A: the bilinear map and the $\mathcal{F}_\gamma$ uniqueness result (Lemma 1, Proposition 1), the spectral budget and the contraction-at-every-step-size guarantee (Lemma 2, Proposition 2), and per-block reachability (Proposition 3), each with its assumptions ($a_h < 0$, block size $b = 2$, and the frozen-operator linearization) stated explicitly. The exclusive 2^3 ablation protocol, including the two degenerate cells that serve as null controls, is described in Section 5.5, and the one-sided paired permutation test with its seed count and attainable p-value floor is derived in the Seed Counts subsection of Appendix A.6. All five datasets (ImageNet-1K, ADE20K, COCO, Something-Something V2, and Breakfast) are public, and the evaluation protocols, the UperNet segmentation head, the Mask R-CNN multi-scale detection schedule, and Kinetics-400 pretraining with fine-tuning, are specified with citations in Section 4.

REFERENCES

Carmen Amo Alonso, Jerome Sieber, and Melanie N. Zeilinger. State space models as foundation models: A control theoretic overview. In *2025 American Control Conference (ACC)*, 2025. doi: 10.23919/ACC63710.2025.11107969.

Karl J. Åström and Björn Wittenmark. *Computer-Controlled Systems: Theory and Design*. Dover Publications, 2013.

Enis Baty, Alejandro Hernández Díaz, Chris Bridges, Rebecca Davidson, Steve Eckersley, and Simon Hadfield. Mamba2d: A natively multi-dimensional state-space model for vision tasks. *arXiv preprint arXiv:2412.16146*, 2024.

Chi-Tsong Chen. *Linear System Theory and Design*. Oxford University Press, 3rd edition, 1999.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ztn8FCR1td.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Riccardo Grazzi, Julien Siems, Jörg K.H. Franke, Arber Zela, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *The Thirteenth*

International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=UvTo3tVBk2.

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.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in neural information processing systems*, 35:35971–35983, 2022a.

Albert Gu, Karan Goel, and Christopher Re. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022b. URL https://mlanthology.org/iclr/2022/gu2022iclr-efficiently/.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in neural information processing systems*, 35:22982–22994, 2022.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25261–25270, 2025. doi: 10.1109/CVPR52734.2025.02352.

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. LocalMamba: Visual State Space Model with Windowed Selective Scan. In *European Conference on Computer Vision Workshops*, pp. 12–22, 2024. doi: 10.1007/978-3-031-91979-4_2. URL https://mlanthology.org/eccvw/2024/huang2024eccvw-localmamba/.

Fady Ibrahim, Guangjun Liu, and Guanghui Wang. Beyond zoh: Advanced discretization strategies for vision mamba. In *Proceedings of the 2026 International Conference on Pattern Recognition (ICPR)*, 2026.

Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision (ECCV)*, 2022.

Thomas Kailath. *Linear Systems*. Prentice Hall, 1980.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Hildegard Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, 2014.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HwCvaJOiCj.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pp. 237–255. Springer, 2024.

Tanzhe Li, Caoshuo Li, Jiayi Lyu, Hongjuan Pei, Baochang Zhang, Taisong Jin, and Rongrong Ji. Damamba: Vision state space model with dynamic adaptive scan. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, 2025. doi: 10.52202/085713-1326. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/38b23e2328096520e9c889ae03e372c9-Paper-Conference.pdf.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Shanhui Liu, Rui Xu, and Yunke Wang. Mambascope: Coarse-to-fine scoping for efficient vision mamba. *arXiv preprint arXiv:2512.00647*, 2025.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ZgtLQQR1K7.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Hui Lu, Albert A. Salah, and Ronald Poppe. Snakes and ladders: Two steps up for videomamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24234–24244, October 2025.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=QZgo9JZpLq.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

Béla Sz Nagy, Ciprian Foias, Hari Bercovici, and László Kérchy. *Harmonic analysis of operators on Hilbert space*. Springer Science & Business Media, 2010.

Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Pearson, 3rd edition, 2010.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://openreview.net/forum?id=M3Yd3QyRG4.

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L Smith. Universality of linear recurrences followed by non-linear projections: Finite-width guarantees and benefits of complex eigenvalues. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=47ahBl70xb.

Yuval Ran-Milo, Eden Lumbroso, Edo Cohen-Karlik, Raja Giryes, Amir Globerson, and Nadav Cohen. Provable benefits of complex parameterizations for structured state space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=h15RyEj151.

T. Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GRMfXcAAFh.

Yuheng Shi, Mingjia Li, Minjing Dong, and Chang Xu. Vssd: Vision mamba with non-causal state space duality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10819–10829, October 2025.

Arnold Tustin. A method of analysing the behaviour of linear systems in terms of time series. *Journal of the Institution of Electrical Engineers-Part IIA: Automatic Regulators and Servo Mechanisms*, 94(1):130–142, 1947.

Feng Wang, Jiahao Wang, Sucheng Ren, Guoyizhe Wei, Jieru Mei, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. Mamba-reg: Vision mamba also needs registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14944–14953, June 2025.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6387–6397. IEEE, 2023.

Shida Wang and Qianxiao Li. StableSSM: Alleviating the curse of memory in state-space models through stable reparameterization. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=nMN5hNZMQK.

Gerhard Wanner and Ernst Hairer. *Solving ordinary differential equations II*, volume 375. Springer Berlin Heidelberg New York, 1996.

Jan C. Willems. Dissipative dynamical systems Part I: General theory. *Archive for Rational Mechanics and Analysis*, 45:321–351, 1972.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024.

Annan Yu, Arnur Nigmetov, Dmitriy Morozov, Michael W. Mahoney, and N. Benjamin Erichson. Robustifying state-space models for long sequences via approximate diagonalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DjeQ39QoLQ.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=YbHCqn4qF4.

A Appendix

A.1 Proofs

No claim in the main text depends on this appendix; it supplies the work behind Sections 3.1 to 3.3. Each derivation is written from standard facts about complex numbers, quadratic forms, and matrix powers. Each result below restates or specializes a classical fact. Proposition 1 is the stability-function characterization of the θ-method, in which $\gamma = \frac{1}{2}$ is the unique member with unit modulus on the imaginary axis and the unique second-order member (Wanner & Hairer, 1996); Lemma 1 is the Tustin map (Tustin, 1947; Oppenheim & Schafer, 2010); Proposition 2 is the correspondence between dissipative generators and contractive Cayley transforms (Nagy et al., 2010; Willems, 1972); and Proposition 3 follows from the Kalman rank condition and the Cayley–Hamilton theorem (Kailath, 1980; Chen, 1999). The contribution lies in their joint application to a selective SSM in which a_h, ω_h and Δ vary per token: under this parameterization, the discretization determines whether a learned rotation rate is identifiable, the block-rotation structure confines all growth of the operator to $\mathbf{U}\mathbf{V}^\top$ so that the budget acts on the coupling alone, and reachability becomes a trainable per-block objective.

A.2 Bilinear Discretization

A continuous oscillation at rate ω is the eigenvalue $i\omega$ on the imaginary axis. A discretization maps it to a point z in the complex plane, and two things must hold: $|z| = 1$, since $|z| < 1$ damps the mode and $|z| > 1$ inflates it, and distinct rates must map to distinct points, since rates sharing an image cannot be distinguished afterwards. The bilinear substitution $s = \frac{2}{\Delta}\frac{z-1}{z+1}$, solved for z, is the scalar map

$$z = f(s) = \frac{1 + cs}{1 - cs}. \tag{9}$$

Lemma 1 (Bilinear map)**.** *For $\Delta > 0$: (i) f maps the imaginary axis one-to-one onto the unit circle minus $z = -1$, with angle $\theta = 2\arctan(\omega\Delta/2) \in (-\pi, \pi)$; (ii) f maps every eigenvalue with negative real part strictly inside the unit circle; (iii) f is one-to-one on the left half-plane, so the map from (a_h, ω_h) to $\overline{\mathbf{A}}_h$ is one-to-one for every a_h; (iv) the discretized block $\overline{\mathbf{A}}_h$ is a rotation by θ_h scaled by $r_h < 1$ when $a_h < 0$, and $\theta_h = 2\arctan(\omega_h\Delta/2)$ exactly when $a_h = 0$.*

Proof. (i) Substituting $s = i\omega$ and writing $u = c\omega$,

$$f(i\omega) = \frac{1 + iu}{1 - iu}, \qquad |f(i\omega)| = \frac{|1 + iu|}{|1 - iu|} = \frac{\sqrt{1 + u^2}}{\sqrt{1 + u^2}} = 1, \tag{10}$$

because the numerator and denominator are complex conjugates and conjugates have equal magnitude. The angle of a quotient is the difference of the angles, and $1 \pm iu$ have angles $\pm\arctan u$, so

$$\theta = \arctan(u) - (-\arctan(u)) = 2\arctan(\omega\Delta/2), \tag{11}$$

which is the angle stated in Section 3.3 (Oppenheim & Schafer, 2010). Since $\arctan$ is strictly increasing with range $(-\pi/2, \pi/2)$, the map $\omega \mapsto \theta$ is strictly increasing with range $(-\pi, \pi)$ and inverse $\omega = \frac{2}{\Delta}\tan(\theta/2)$: every angle comes from exactly one rate, and $\theta = \pm\pi$ (the point $z = -1$) is approached only as $\omega \to \pm\infty$.

(ii) For a general eigenvalue $s = \sigma + i\omega$,

$$|f(s)|^2 = \frac{|1 + cs|^2}{|1 - cs|^2} = \frac{(1 + c\sigma)^2 + c^2\omega^2}{(1 - c\sigma)^2 + c^2\omega^2}. \tag{12}$$

The two sides differ only in the sign in front of $c\sigma$. When $\sigma < 0$, the numerator's first term is the smaller one, so $|f(s)| < 1$: a decaying continuous mode becomes a contracting discrete one, at every step size.

(iii) f is a ratio of two linear functions of s, and setting $f(s_1) = f(s_2)$ and cross-multiplying gives $s_1 = s_2$ directly, so f never sends two eigenvalues to the same point. Since $\overline{\mathbf{A}}_h$ is determined by $f(a_h + i\omega_h)$, two parameter pairs producing the same discrete block must coincide, so damping does not reintroduce the ambiguity that Section 3.3 removes.

(iv) The block $\mathbf{A}_h = a_h\mathbf{I} + \omega_h\mathbf{J}$ with $\mathbf{J}^2 = -\mathbf{I}$ behaves exactly as the complex number $a_h + i\omega_h$: sums, products and inverses of such blocks match sums, products and inverses of the corresponding numbers. Applying Eq. (6) to the block therefore returns the block corresponding to $f(a_h + i\omega_h)$, whose polar form $r_h e^{i\theta_h}$ is a rotation by θ_h scaled by r_h. Then $r_h < 1$ for $a_h < 0$ by (ii), which is the statement of Section 3.1 that an uncoupled block is a contraction by construction, and setting $\sigma = a_h = 0$ in Eq. (12) and $u = c\omega_h$ in Eq. (11) recovers $\theta_h = 2\arctan(\omega_h\Delta/2)$, the boundary case. □

Proposition 1 (Uniqueness within $\mathcal{F}_\gamma$)**.** *The one-step scheme with weight $\gamma \in [0, 1]$ on the right endpoint of the interval is*

$$\overline{\mathbf{A}}(\gamma) = (\mathbf{I} - \gamma\Delta\mathbf{A})^{-1}\big(\mathbf{I} + (1 - \gamma)\Delta\mathbf{A}\big) : \tag{13}$$

$\gamma = 0$ is forward Euler, $\gamma = 1$ is backward Euler, $\gamma = \frac{1}{2}$ is Eq. (6). The scheme keeps every purely imaginary eigenvalue on the unit circle, for all ω and Δ, if and only if $\gamma = \frac{1}{2}$. It is then the unique member mapping the imaginary axis bijectively onto the circle minus $z = -1$, and the unique second-order member.

Proof. On the eigenvalue $i\omega$, Eq. (13) is the scalar $\big(1 + i(1 - \gamma)\omega\Delta\big)/\big(1 - i\gamma\omega\Delta\big)$, and taking squared magnitudes of numerator and denominator gives

$$|\overline{A}(\gamma)|^2 = \frac{1 + (1 - \gamma)^2\omega^2\Delta^2}{1 + \gamma^2\omega^2\Delta^2}, \tag{14}$$

which is Eq. (7). The numerator carries the weight left on the interval, the denominator the weight moved to the right endpoint, and the magnitude is one exactly when the two balance: setting Eq. (14) to one and clearing denominators gives $(1 - \gamma)^2\omega^2\Delta^2 = \gamma^2\omega^2\Delta^2$, and for any $\omega\Delta \neq 0$ this forces $(1 - \gamma)^2 = \gamma^2$, whose only solution in $[0, 1]$ is $\gamma = \frac{1}{2}$. Conversely at $\gamma = \frac{1}{2}$ the ratio is one for

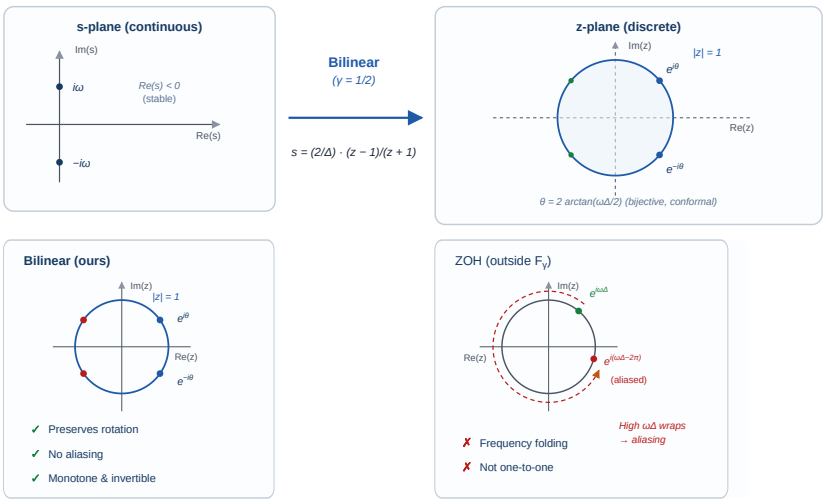


Figure 2: Discretization of an imaginary-axis eigenvalue $\pm i\omega$. The bilinear transform ($\gamma = \frac{1}{2}$) maps the imaginary axis one-to-one onto the unit circle minus $z = -1$, with $\theta = 2\arctan(\omega\Delta/2) \in (-\pi, \pi)$. ZOH also lands on the unit circle, but with $\theta = \omega\Delta$, so the rates ω and $\omega + 2\pi/\Delta$ reach the same point and cannot be distinguished.

every ω and Δ at once. For $\gamma < \frac{1}{2}$ the numerator dominates and the oscillation grows; for $\gamma > \frac{1}{2}$ the denominator dominates and it decays; the ratio differs from one at every nonzero $\omega\Delta$, so neither error is confined to large steps. Bijectivity at $\gamma = \frac{1}{2}$ is Lemma 1(i), and no other member sends the axis into the circle at all.

For the order claim, expand both maps for a scalar eigenvalue λ in powers of Δ:

$$\overline{A}(\gamma) = 1 + \lambda\Delta + \gamma\lambda^2\Delta^2 + \mathcal{O}(\Delta^3), \qquad e^{\lambda\Delta} = 1 + \lambda\Delta + \tfrac{1}{2}\lambda^2\Delta^2 + \mathcal{O}(\Delta^3), \tag{15}$$

using the geometric series $(1 - \gamma\lambda\Delta)^{-1} = 1 + \gamma\lambda\Delta + \gamma^2\lambda^2\Delta^2 + \cdots$. The Δ^2 coefficients agree only at $\gamma = \frac{1}{2}$. □

Remark 1 (Zero-order hold)**.** *ZOH sets* $\overline{\mathbf{A}} = \exp(\mathbf{A}\Delta)$*, which is not a ratio of polynomials in* $\Delta\mathbf{A}$*, so ZOH is not in the family and Proposition 1 does not exclude it. Nor does it fail the magnitude condition:* $|e^{i\omega\Delta}| = 1$ *exactly, so ZOH never damps an oscillation and never inflates one. Its failure is the second condition. Its angle is* $\theta_{\mathrm{ZOH}} = \omega\Delta$*, which grows without bound, so the rates* ω *and* $\omega + 2\pi/\Delta$ *arrive at the same point of the circle and are indistinguishable afterwards (Åström & Wittenmark, 2013). The bilinear angle Eq. (11) instead approaches* π *without reaching it and is invertible, so the rate axis is stretched but never folded, and because* ω_h *is learned rather than externally specified, the model absorbs the stretch into the parameter. The same absorption does not rescue ZOH: a periodic map folded into a learned parameter leaves the parameter unidentifiable. ZOH loses rate identity, not amplitude, which is the exclusion of Section 3.3 (Figure 2)*

A.3 Stability

Growth can enter at two points: the discrete operator $\overline{\mathbf{A}}$ can amplify the state over the scan, or the continuous matrix $\mathbf{A}$ can be growing on its own, in which case no step size repairs it. The useful direction is that controlling the continuous matrix once bounds the discrete operator at every step size. The controlling object is the symmetric part $\mathbf{A}_s = (\mathbf{A} + \mathbf{A}^\top)/2$, because for any matrix the quadratic form sees only the symmetric part: $\mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top\mathbf{A}_s\mathbf{x}$, since the skew part cancels against itself.

Lemma 2 (Spectral budget, Eq. (3))**.** *With blocks as in Eq. (2) and* $a_h \le -a_{\min} < 0$*,*

$$\lambda_{\max}(\mathbf{A}_s) \;\le\; \max_h a_h + \|\mathbf{U}\mathbf{V}^\top\|_2 \;\le\; -a_{\min} + \|\mathbf{U}\mathbf{V}^\top\|_2, \tag{16}$$

so $\|\mathbf{U}\mathbf{V}^\top\|_2 \le a_{\min}$ *gives* $\mathbf{A}_s \preceq \mathbf{0}$*, and* $\|\mathbf{U}\mathbf{V}^\top\|_2 \le a_{\min} - \varepsilon_{\mathrm{stab}}$ *gives* $\mathcal{L}_{\mathrm{stab}} = 0$*.*

Proof. $\mathbf{J}$ is skew-symmetric, so the symmetric part of each block $a_h\mathbf{I} + \omega_h\mathbf{J}$ is just $a_h\mathbf{I}$, and

$$\mathbf{A}_s = \mathrm{blockdiag}(a_1\mathbf{I}, \ldots, a_H\mathbf{I}) + \tfrac{1}{2}\big(\mathbf{U}\mathbf{V}^\top + \mathbf{V}\mathbf{U}^\top\big). \tag{17}$$

For any unit vector $\mathbf{x}$, the first term contributes at most $\max_h a_h$ to $\mathbf{x}^\top\mathbf{A}_s\mathbf{x}$ and the second contributes at most its spectral norm, which is at most $\|\mathbf{U}\mathbf{V}^\top\|_2$; maximizing over $\mathbf{x}$ gives Eq. (16). The two implications follow from $a_h \le -a_{\min}$ and the hinge in Eq. (4). □

Two facts stated in Section 3.2 are immediate from Eq. (17). The rotation rates enter $\mathbf{A}$ only through $\mathbf{J}$, which drops out of the symmetric part, so neither the budget nor $\mathcal{L}_{\mathrm{stab}}$ touches any ω_h: rotation is energy-neutral. And with $\mathbf{U} = \mathbf{V} = \mathbf{0}$ the bound is $\max_h a_h \le -a_{\min} < -\varepsilon_{\mathrm{stab}}$, so $\mathcal{L}_{\mathrm{stab}}$ is identically zero without the coupling.

Proposition 2 (Contraction at every step size)**.** *If $\mathbf{A}_s \preceq \mathbf{0}$ then $\|\overline{\mathbf{A}}\|_2 \le 1$ for every $\Delta > 0$. If $\mathbf{A}_s \preceq -\varepsilon_{\mathrm{stab}}\mathbf{I}$, which is where $\mathcal{L}_{\mathrm{stab}} = 0$, the inequality is strict, so for the operator frozen at one token both $\|\mathbf{h}_t\|_2$ and the gradient norm $\|\partial\mathbf{h}_t/\partial\mathbf{h}_0\|_2$ decay geometrically, and along a scan with token-dependent Δ_t the product of the per-token operators has norm below one.*

Proof. Expand the two squared norms that define the bilinear map; the $\|\mathbf{x}\|_2^2$ and $c^2\|\mathbf{A}\mathbf{x}\|_2^2$ terms are common to both and cancel, leaving only the cross terms, which see the symmetric part:

$$\|(\mathbf{I} + c\mathbf{A})\mathbf{x}\|_2^2 - \|(\mathbf{I} - c\mathbf{A})\mathbf{x}\|_2^2 = 4c\,\mathbf{x}^\top\mathbf{A}_s\mathbf{x} = 2\Delta\,\mathbf{x}^\top\mathbf{A}_s\mathbf{x} \;\le\; 0, \tag{18}$$

the identity quoted in Section 3.2. The same expansion shows $\mathbf{x}^\top(\mathbf{I} - c\mathbf{A})\mathbf{x} \ge \|\mathbf{x}\|_2^2$, so $\mathbf{I} - c\mathbf{A}$ is invertible and $\overline{\mathbf{A}}$ is well defined. Writing $\mathbf{y} = (\mathbf{I} - c\mathbf{A})\mathbf{x}$, Eq. (18) says $\|\overline{\mathbf{A}}\mathbf{y}\|_2 = \|(\mathbf{I} + c\mathbf{A})\mathbf{x}\|_2 \le \|\mathbf{y}\|_2$ for every $\mathbf{y}$, so $\|\overline{\mathbf{A}}\|_2 \le 1$ at any $\Delta > 0$. Under the margin the right side of Eq. (18) is at most $-2\Delta\,\varepsilon_{\mathrm{stab}}\|\mathbf{x}\|_2^2 < 0$, so the inequality is strict and $\|\overline{\mathbf{A}}\|_2 < 1$. Applying the operator t times multiplies the bound t times, which gives the geometric decay of both the state and the gradient $\partial\mathbf{h}_t/\partial\mathbf{h}_0 = \overline{\mathbf{A}}^t$; and because the hypothesis on $\mathbf{A}_s$ does not involve Δ, each per-token operator along the scan satisfies its own bound, so the product does too. The spectral norm is the right quantity here rather than the eigenvalues: the coupling makes $\overline{\mathbf{A}}$ non-normal, and a non-normal operator can have all eigenvalues inside the unit circle yet still amplify vectors for many steps (Yu et al., 2024). □

Remark 2 (What $\mathcal{L}_{\mathrm{stab}}$ bounds)**.** *Along a continuous trajectory, $\frac{d}{dt}\|\mathbf{h}\|_2^2 = 2\mathbf{h}^\top\mathbf{A}_s\mathbf{h} \le 2\lambda_{\max}(\mathbf{A}_s)\|\mathbf{h}\|_2^2$, so $\lambda_{\max}(\mathbf{A}_s)$ is an upper bound on the growth rate and $\mathbf{A}_s \prec \mathbf{0}$ is dissipativity in the sense of Willems (1972), with the state energy $\|\mathbf{h}\|_2^2$ itself as the storage function. It is only an upper bound: a non-normal $\mathbf{A}$ can be stable while $\lambda_{\max}(\mathbf{A}_s) > 0$, so the constraint is sufficient, not necessary, and the main text claims no more. The budget sits on the continuous matrix because a bound imposed on $\overline{\mathbf{A}}$ holds only at the step size it was checked, while the hypothesis of Proposition 2 holds at all of them; the same budget also bounds the ZOH operator, since $\|e^{\mathbf{A}\Delta}\|_2 \le e^{\lambda_{\max}(\mathbf{A}_s)\Delta}$, so this is a property of the architecture, not an argument for B3, which rests on Proposition 1 alone.*

A.4 Per-Block Conditioning

A stable model can still stop listening: if the directions the input can push the state into collapse to a line, late evidence has nowhere to go. Reachability measures this. Throughout, $\overline{\mathbf{A}}_h$ is the isolated block of Lemma 1(iv) frozen at one token, a rotation by θ_h scaled by r_h, with $\overline{\mathbf{B}}_h \in \mathbb{R}^{2\times 1}$ fixed, and $\mathbf{W}_K^{(h)} = \mathbf{Q}_K^{(h)}(\mathbf{Q}_K^{(h)})^\top$ with $\mathbf{Q}_K^{(h)}$ from Eq. (5).

Proposition 3 (Block reachability)**.** *(i) The block is K-step controllable exactly when $\sigma_{\min}(\mathbf{Q}_K^{(h)}) > 0$, and the least input energy reaching a target $\mathbf{h}^*$ is at most $\|\mathbf{h}^*\|_2^2/\sigma_{\min}(\mathbf{Q}_K^{(h)})^2$. (ii) For $K \ge b$ the reachable set stops growing, and $\sigma_{\min}(\mathbf{Q}_K^{(h)})$ never decreases as K grows. (iii) At fixed θ_h and $\overline{\mathbf{B}}_h$, $\sigma_{\min}(\mathbf{Q}_K^{(h)})$ never decreases as r_h grows. (iv) If $\omega_h = 0$ the block matrix has rank one for every K, $\sigma_{\min} = 0$, and the block's term in $\mathcal{L}_{\mathrm{cond}}$ carries no gradient to $\overline{\mathbf{B}}_h$ or a_h.*

Proof. (i) Unrolling the recurrence from $\mathbf{h}_0^{(h)} = \mathbf{0}$ collects the inputs into one product,

$$\mathbf{h}_K^{(h)} = \sum_{k=0}^{K-1} \overline{\mathbf{A}}_h^{\,k}\,\overline{\mathbf{B}}_h\,u_{K-1-k} = \mathbf{Q}_K^{(h)}\mathbf{u}, \tag{19}$$

with $\mathbf{u}$ the stacked scalar inputs. Every target is reachable exactly when this linear map is onto, that is when $\mathbf{Q}_K^{(h)}$ has full row rank, equivalently $\sigma_{\min} > 0$ (Chen, 1999). Among all inputs reaching a given target, the least-norm one has energy $\mathbf{h}^{*\top}(\mathbf{W}_K^{(h)})^{-1}\mathbf{h}^*$ (Kailath, 1980), and bounding the inverse by its largest eigenvalue, with $\lambda_{\min}(\mathbf{W}_K^{(h)}) = \sigma_{\min}(\mathbf{Q}_K^{(h)})^2$ for a wide matrix, gives the stated bound: small $\sigma_{\min}$ means some direction costs enormous input energy, and $\sigma_{\min} = 0$ means it cannot be reached at all.

(ii) A $b \times b$ matrix satisfies its own characteristic polynomial (Cayley–Hamilton), so $\overline{\mathbf{A}}_h^b$ is a combination of lower powers and every column of $\mathbf{Q}_K^{(h)}$ beyond the b-th already lies in the span of the first b: the reachable set is the same for all $K \geq b$, which is why a short check speaks to long sequences. For the singular value, adding one column adds one outer product,

$$\mathbf{W}_{K+1}^{(h)} = \mathbf{W}_K^{(h)} + \mathbf{q}\mathbf{q}^\top, \qquad \mathbf{q} = \overline{\mathbf{A}}_h^K \overline{\mathbf{B}}_h, \tag{20}$$

and adding a positive semidefinite term can only raise $\mathbf{x}^\top \mathbf{W}\mathbf{x}$ for every $\mathbf{x}$, hence can only raise $\lambda_{\min}$. This is the statement in Section 3.2 that the block Gramian is a sum of positive semidefinite terms and $\sigma_{\min}$ is non-decreasing in K.

(iii) Because the block is a scaled rotation, $\overline{\mathbf{A}}_h^k = r_h^k \mathbf{R}(k\theta_h)$ with $\mathbf{R}$ a rotation, and rotations preserve norms, so

$$\mathbf{W}_K^{(h)} = \sum_{k=0}^{K-1} r_h^{2k}\, \mathbf{R}(k\theta_h)\, \overline{\mathbf{B}}_h \overline{\mathbf{B}}_h^\top\, \mathbf{R}(k\theta_h)^\top. \tag{21}$$

Each summand is a fixed positive semidefinite matrix scaled by r_h^{2k}, which grows with r_h, so the whole sum, and with it $\lambda_{\min}$, never decreases as r_h grows. This is the monotonicity in r_h that Section 3.2 identifies as the residual tension with the contraction: the Gramian improves as $r_h \to 1$, the direction the stability constraints move away from, and λ_{cond} sets the trade-off.

(iv) With $\omega_h = 0$ the rotation disappears, $\overline{\mathbf{A}}_h = r_h \mathbf{I}$, so every column of $\mathbf{Q}_K^{(h)}$ is $r_h^k \overline{\mathbf{B}}_h$, a rescaling of the first. The matrix has rank one for every K, its smallest singular value is exactly zero, and no extra horizon repairs it: the input can only ever push the state along a single fixed direction. Changing $\overline{\mathbf{B}}_h$ or a_h preserves this collinearity, so $\sigma_{\min}$ stays at zero and the block's term contributes no gradient to those parameters. This is the rank-one degeneracy reported in Table 5 and the reason $\mathcal{L}_{\text{cond}}$ averages over blocks rather than taking the worst block, which is nearly always one with θ_h near zero. □

Part (ii) holds for the operator frozen at one token with the coupling removed, as Section 3.2 specifies, not for the time-varying coupled scan; this is the usual linearization for gated recurrences, its tightness is not measured, and the long-horizon argument of Section 4 rests on the certified property, collapse of the reachable subspace, which does not depend on sequence length.

Remark 3 (Why not the full state)**.** *On the undecomposed state, $\mathbf{Q}_K \in \mathbb{R}^{N \times K}$ is tall for $K < N$, and the smallest singular value of a tall matrix is at most the norm of any one column:*

$$\sigma_{\min}(\mathbf{Q}_K) \;\leq\; \|\overline{\mathbf{A}}^{K-1}\overline{\mathbf{B}}\|_2 \;\leq\; \|\overline{\mathbf{A}}\|_2^{K-1}\, \|\overline{\mathbf{B}}\|_2, \tag{22}$$

the cap quoted in Section 3.2: the objective is bounded by a quantity that Proposition 2 shrinks geometrically, so the conditioning objective and the stability constraints pull against each other by construction. Successive columns also align with the dominant direction of $\overline{\mathbf{A}}$, which is why the measured $\sigma_{\min}$ at $N = 128$ falls from 1.45×10^{-2} at $K = 2$ to 8.3×10^{-18} at $K = 32$ and the gradient of $\lambda/(\sigma_{\min} + \varepsilon)$ vanishes. Normalizing the columns removes the shrinking but not the alignment; a log-determinant objective rewards a volume that Eq. (3) deliberately shrinks. Splitting into $b = 2$ blocks avoids all of this: each block matrix is wide from $K = 2$, so Eq. (22) does not apply, and by Eq. (20) further columns can only help. $\mathcal{L}_{\text{cond}}$ maintains conditioning; it does not restore it once lost.

A.5 Cost of the Coupled Operator

Write $\mathbf{M} = \mathbf{I} - c\,\text{blockdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_H)$, which is block-diagonal with 2×2 blocks and is inverted blockwise in $\mathcal{O}(Nb)$. The Woodbury identity turns the full inverse into the blockwise one plus a

rank-R correction,

$$(\mathbf{I} - c\mathbf{A})^{-1} = \mathbf{M}^{-1} + c\,\mathbf{M}^{-1}\mathbf{U}\big(\mathbf{I}_R - c\,\mathbf{V}^\top\mathbf{M}^{-1}\mathbf{U}\big)^{-1}\mathbf{V}^\top\mathbf{M}^{-1}, \tag{23}$$

where the $R \times R$ inverse exists because the full one does (Proposition 2). Forming $\mathbf{M}^{-1}\mathbf{U}$ costs $\mathcal{O}(NR)$, the $R\times R$ product $\mathbf{V}^\top\mathbf{M}^{-1}\mathbf{U}$ costs $\mathcal{O}(NR^2)$, and its inverse $\mathcal{O}(R^3)$, which is the per-token $\mathcal{O}(NR^2 + R^3)$ of Section 3.3. Applying $\overline{\mathbf{A}}$ to a vector then costs $\mathcal{O}(Nb)$ for the blocks plus $\mathcal{O}(NR)$ for the correction, $\mathcal{O}(Nb + NR)$ in total; no dense $N \times N$ operator is ever formed, and everything is rebuilt per token because Δ, a_h and ω_h are.

A.6 Seed Counts

A one-sided paired permutation test over n seeds admits 2^n sign assignments, so the smallest attainable p-value is 2^{-n}: 0.031 at $n = 5$ and $2^{-20} \approx 10^{-6}$ at $n = 20$. Eq. (8) is a difference of two paired differences; with equal per-cell variance σ^2, its variance is $4\sigma^2$ against $2\sigma^2$ for a single paired difference, so matching the standard error takes twice the seeds. The four interaction cells of Table 6 carry twenty each, so the reported $p = 2^{-20}$ is the floor the test can return.