Causal Volterra Dynamics of Mamba

Ming-Ching Chang MCHANG2@ALBANY.EDU
Department of Computer Science, University at Albany, Albany, NY, USA

Ting Yu Tsai TTSAI2@ALBANY.EDU
Department of Computer Science, University at Albany, Albany, NY, USA

Davis Wertheimer DAVIS.WERTHEIMER@IBM.COM
IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

Felix X.-F. Ye† XYE2@ALBANY.EDU
Department of Mathematics & Statistics, University at Albany, Albany, NY, USA

Abstract

Transformers admit continuum descriptions based on mean-field interactions, but selective state space models fall into a different class. We show that the many-token limit of an input-conditioned, single-head SISO Mamba-3 block is a *causal Volterra equation* on the sphere with an explicit exponential memory kernel. The key reason is that Mamba's causal mask is chain-dependent: the influence of one token on another is transmitted through the full sequence of intermediate gates rather than a single pairwise weight. Under a constant-horizon scaling, we prove convergence to this limit, validate it numerically, and show that the same framework covers the SISO Mamba-3 rotation, with Mamba-2 as a special case. We also characterize the resulting memory-horizon interpolation and relate the kernel scale to pretrained Mamba-2 heads.

1. Introduction

Rigollet and collaborators [8, 11, 21] model transformer layers as interacting particle systems on the unit sphere, establishing that self-attention drives tokens toward clustering. For causal attention with $V = I$, Karagodin et al. [15] prove all tokens collapse to $x_1(0)$; metastability, quantitative mean-field rates, normalization effects, long-context scaling, noisy stationary states, circle synchronization, measure-to-measure views, homogenized multi-head limits, and optimization-inspired variants further enrich the picture [2, 4, 5, 9, 10, 16, 17, 20, 24]. A natural question is whether this particle-system framework extends to selective state space models [7, 12, 18]. We answer it for a single SISO Mamba-3 block [18], omitting the output gating and MIMO structure. Our Mamba-2 special case uses the same simplifications: it keeps the selective scan but omits the short 1D convolution used in the standard Mamba-2 block [7]; the formal reduction is given in Section 2 and Appendix A.

We study the dense-token limit $T \to \infty$ on a fixed continuous-depth horizon. In this limit, the transformer mean-field framework does not extend directly to selective SSMs. Transformer dense-token limits, including causal transformers, are governed by row-normalized attention kernels: a causal mask restricts attention to past tokens, but each row still defines a probability law over its admissible source tokens. Thus causal transformers lead to normalized non-anticipative attention operators, whereas selective SSMs have no row normalization and preserve sequence order through

† indicates corresponding author.

the scan. This difference changes the limiting mathematical object. After unrolling the selective scan, the influence of token j on token i is transported through all intermediate gates between j and i. Hence the continuum limit is not a softmax-normalized average over past tokens, as in causal attention, but an ordered trajectory field with an unnormalized causal memory operator. The resulting dense-token limit is a *causal Volterra integro-differential equation* over the SSM clock.

This token-mesh limit should also be distinguished from the continuous-time SSM that Mamba discretizes. The latter supplies the local scan generator before the sequence is refined; our limit instead sends the token mesh size to zero and identifies the causal memory operator induced by the unrolled scan. It is also distinct from the complementary long-depth limit at fixed T, studied by Vo et al. [22] for scalar tokens and by Nguyen et al. [19] for d-dimensional tokens on the sphere. The two limits need not commute. As a finite-time diagnostic, Proposition 3 shows that for any fixed first token independent of the weights, the Mamba vector field is generically nonzero at that token; we do not use this observation to make a long-time clustering or stationarity claim.

The contributions are: (i) we identify the dense-token limit of the SISO recurrence as a causal Volterra equation with an explicit memory kernel and $O(1/T)$ convergence rate (Theorem 1); (ii) we verify that the same Lipschitz-kernel theorem covers the SISO Mamba-3 rotation, with Mamba-2 as the $\theta \equiv 0$ reduction; and (iii) we characterize how the decay rate a interpolates between decoupled pointwise dynamics ($a \to \infty$, speed $O(1/a)$) and full-history causal transport ($a \to 0$), and estimate this memory scale across all 576 heads of pretrained Mamba-2-130M. The constant-horizon scaling is the non-degenerate dense-token regime in normalized coordinates, not a claim that fixed pretrained parameters extrapolate to arbitrary lengths. Section 4 summarizes the numerical checks behind the main claims.

2. Selective SSMs as Particle Systems

We formulate the continuous-depth dynamics of a single-head selective SSM block as a projected particle system, using the SISO Mamba-3 recurrence [18] as the general framework. Let $\bar{u}_i$ denote the input tokens at the start of the block. In the fixed-clock formulation analyzed here, the step sizes and two-tap gates are computed from $\bar{u}_i$ and then held fixed over continuous depth; the token states $u_i(t)$ evolve through the projected residual flow. The content projections B_j, C_i and the value map are still evaluated on the evolving states; only the mesh variables Δ_i, λ_i are held fixed. With RMSNorm weight $\gamma = \mathbf{1}$, normalization reduces to projection onto $\mathcal{S}^{d-1}_{\mathrm{rms}}$ (see Appendix A), so we model token representations $u_1, \dots, u_T \in \mathcal{S}^{d-1}_{\mathrm{rms}} := \{u \in \mathbb{R}^d : \|u\|^2 = d\}$. The block is parameterized by input and output projections $S_x \in \mathbb{R}^{P\times d}$, $S_o \in \mathbb{R}^{d\times P}$, gating projections $S_B, S_C \in \mathbb{R}^{N\times d}$, and the SiLU activation ($\mathrm{SiLU}(z) := z\,\sigma(z)$, where σ denotes the sigmoid), applied coordinatewise; we write $C_i := \mathrm{SiLU}(S_C u_i)$, $B_j := \mathrm{SiLU}(S_B u_j)$, and $P_u := I - uu^\top/d$ for the RMSNorm-consistent tangent-space projector on $\|u\|^2 = d$.

The underlying SSM evolves an N-dimensional state bank for each value channel; equivalently, with $x \in \mathbb{R}^P$ and $B \in \mathbb{R}^N$, write $h \in \mathbb{R}^{N\times P}$ and $\dot{h} = Ah + Bx^\top$. Its state matrix has complex diagonal entries $A + i\theta[m]$ for $m = 1, \dots, N/2$, with shared scalar decay $A < 0$ (we write $a := -A > 0$) and learnable frequencies $\theta[m] \in \mathbb{R}$. Discretization splits these into two ingredients (see Appendix A). The first is a block-diagonal rotation that accumulates data-dependent phases along the chain from token j to token i: $\mathcal{R}_{j\to i} = \mathrm{Block}\Big(R\Big(\sum_{k=j+1}^{i} \Delta_k\,\theta[m]\Big)\Big)_{m=1}^{N/2}$, where $R(\vartheta)$ is the rotation matrix by ϑ. The second is the *chain-dependent causal mask*, which combines the

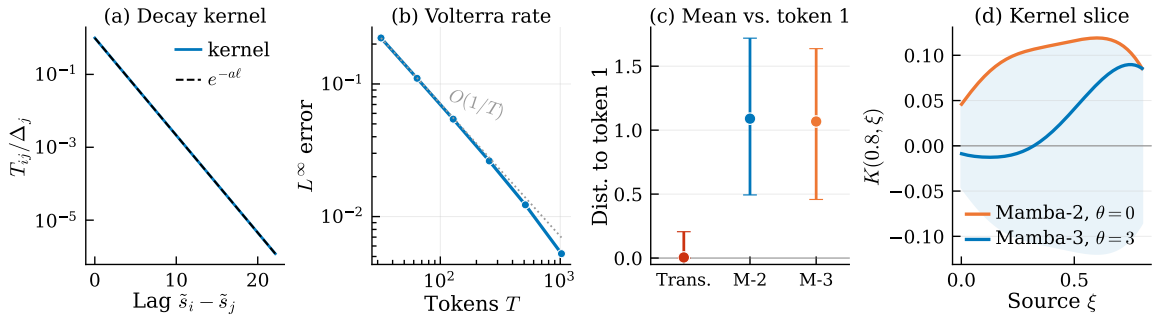

Figure 1: Empirical summary for the dense-token Volterra picture. (a) Pretrained Mamba-2-130M layer 0 head 7 exhibits an exponential chain-memory slice. (b) Synthetic fixed-clock particles ($d = 2$) converge to the Volterra equation at the predicted $O(1/T)$ rate. (c) Synthetic mean-to-token-1 distance (final normalized centroid to initial token 1): medians (IQRs) are 0.005, 1.090, and 1.068 for transformer, Mamba-2, and Mamba-3; only the transformer collapses to token 1. (d) Synthetic kernel slice at $x = 0.8$: the SISO Mamba-3 interaction ($\theta = 3$) changes sign, unlike the Mamba-2 reduction ($\theta = 0$).

exponential decay with the endpoint weights from the exponential-trapezoidal discretization [18]:

$$T_{ij} = \lambda_i \, \Delta_i \, \mathbf{1}_{\{j=i\}} \; + \; e^{-a(\tilde{s}_i - \tilde{s}_j)} \left[\lambda_j \, \Delta_j + (1 - \lambda_{j+1}) \, \Delta_{j+1}\right] \mathbf{1}_{\{j<i\}}. \tag{1}$$

where $\tilde{s}_i := \sum_{k=1}^{i} \Delta_k$, $\Delta_k := \mathrm{softplus}(S_\Delta \bar{u}_k + b) > 0$, and $\lambda_j := \sigma(S_\lambda \bar{u}_j) \in (0, 1)$ is an input-dependent gate. The diagonal entry is a pure self-interaction; the off-diagonal entries carry the chain-dependent decay $e^{-a(\tilde{s}_i - \tilde{s}_j)}$ with a bracket of endpoint weights that merge in the dense-token limit (see (5)). After normalization, the diagonal self term misses one endpoint cell; this $O(\delta_{\max})$ error is absorbed in the quadrature bound.

With these ingredients, the projected flow on $\mathcal{S}^{d-1}_{\mathrm{rms}}$ reads

$$\frac{du_i}{dt} = P_{u_i} \sum_{j=1}^{i} \langle C_i, \, \mathcal{R}_{j \to i} \, B_j \rangle \, T_{ij} \, S_o \, \mathrm{SiLU}(S_x u_j). \tag{2}$$

Setting $\lambda_j \equiv 1$ and $\theta \equiv 0$ recovers the Mamba-2 recurrence [7] (without its short 1D convolution on B, C): $\mathcal{R}_{j \to i} = I$ and the mask reduces to $T_{ij} = e^{-a(\tilde{s}_i - \tilde{s}_j)} \Delta_j \mathbf{1}_{\{j \le i\}}$, the 1-semiseparable structure of Dao and Gu [7]. Compared with the causal transformer ODE $\dot{x}_k = P_{x_k} Z_k^{-1} \sum_{j \le k} e^{\beta \langle x_k, x_j \rangle} V x_j$, the Mamba flow replaces a normalized pairwise attention weight by the content kernel $\langle C_i, \, \mathcal{R}_{j \to i} B_j \rangle$, the linear value $V x_j$ by the nonlinear value $S_o \, \mathrm{SiLU}(S_x u_j)$, and the fixed mask $\mathbf{1}_{\{j \le k\}}$ by the chain-dependent mask (1). These three changes are what turn the dense-token limit from a McKean–Vlasov/lifted attention limit into a Volterra limit. The following figure previews the mechanism and empirical checks used below. A side consequence of the nonlinear value map is that, for any fixed first token, token 1 is generically non-stationary at finite depth in Mamba, unlike the exact first-token stationarity of causal attention with $V = I$ [15].

3. Main Result: A Causal Volterra Limit

We derive the continuum limit of the fixed-clock particle system (2) as $T \to \infty$. Recall the cumulative positions $\tilde{s}_i = \sum_{k=1}^{i} \Delta_k$ from (1); let $S_T := \tilde{s}_T$ and define normalized coordinates

$s_i := \tilde{s}_i/S_T \in (0,1]$ with quadrature weights $\delta_i := \Delta_i/S_T$. The coordinate s_i is an *internal clock* of the SSM recurrence: each token k advances the clock by a data-dependent step δ_k, so tokens with larger gates are spaced farther apart on the internal timeline, while those with smaller gates cluster together. To interpret the discrete recurrence as a quadrature rule on a fixed interval, we normalize positions to $(0,1]$ via $s_i := \tilde{s}_i/S_T$ with weights $\delta_j := \Delta_j/S_T$ and rescale time by S_T. The decay kernel becomes $e^{-a_{\text{eff}}(s_i-s_j)}\,\delta_j$, where $a_{\text{eff}} := a\,S_T$ combines the learned decay with the discretization scale; the analogous effective frequencies $\theta_{\text{eff}}[m] := \theta[m]\,S_T$ govern Mamba-3's rotations. As $T \to \infty$ the mesh refines ($\delta_{\max} \to 0$) and the Riemann sum converges to a Volterra integral at rate $O(1/T)$. We write a and $\theta[m]$ for the effective quantities hereafter. For the convergence theorem we abstract the content kernel $\langle C_i,\, \mathcal{R}_{j\to i}B_j\rangle$ as a general kernel $\kappa_\theta\colon [0,1]^2 \times \mathcal{S}^{d-1}_{\text{rms}} \times \mathcal{S}^{d-1}_{\text{rms}} \to \mathbb{R}$, assumed bounded and jointly Lipschitz in the sense of (5). Empirically, across all 576 heads of pretrained Mamba-2-130M the median memory span is ≈ 7 tokens; at operating length $T = 128$ this is about 5% of the sequence. This finite-horizon behavior motivates the constant-horizon scaling as the dense-token analogue at operating lengths; with fixed learned rates, the normalized memory fraction shrinks as T grows and the limit becomes increasingly local. See Figure 1(a) and Appendix F.

Particle system. Under the constant-horizon scaling and fixed mesh, the particle system becomes

$$\partial_t u_i^{(T)} = P_{u_i^{(T)}}\, S_o \sum_{j\le i} \kappa_\theta(s_i, s_j;\, u_i^{(T)}, u_j^{(T)})\, e^{-a(s_i-s_j)}\,\delta_j\ \mathrm{SiLU}(S_x u_j^{(T)}), \tag{3}$$

where $\kappa_\theta : [0,1]^2 \times \mathcal{S}^{d-1}_{\text{rms}} \times \mathcal{S}^{d-1}_{\text{rms}} \to \mathbb{R}$ is a content kernel satisfying (5).

Theorem 1 (Causal Volterra limit for a fixed internal clock) *Under the constant-horizon scaling, suppose the normalized positions and weights are generated by the input-conditioned gates and held fixed over depth. With κ_θ bounded and jointly Lipschitz (see (5)), suppose $\delta^{(T)}_{\max} \to 0$ and $u^{(T)}(\cdot,0) \to u_0 \in \mathrm{Lip}([0,1];\mathcal{S}^{d-1}_{\text{rms}})$ uniformly. The SISO Mamba-3 kernel in (6) satisfies this condition with $L^{\text{pos}}_\kappa = \|\theta\|_\infty M_C M'_B$. Here $u^{(T)}(x,t) := u_i^{(T)}(t)$ denotes the piecewise-constant interpolant on cells $(s_{i-1}, s_i]$. Then this interpolant converges uniformly on $[0,1]\times[0,t_f]$ for any finite horizon $t_f > 0$ to the solution u of the causal Volterra integro-differential equation*

$$\boxed{\partial_t u(x,t) = P_{u(x,t)} \int_0^x e^{-a(x-\xi)}\,\kappa_\theta\big(x,\xi;\, u(x,t), u(\xi,t)\big)\, S_o\ \mathrm{SiLU}\big(S_x u(\xi,t)\big)\, d\xi, \quad x \in [0,1].} \tag{4}$$

A Grönwall estimate gives $\sup_{0\le t\le t_f}\|u^{(T)} - u\|_\infty \le e^{Lt_f}(\|u_0^{(T)} - u_0\|_\infty + t_f\,\omega(\delta^{(T)}_{\max}))$, where L and the quadrature modulus $\omega(\delta_{\max}) = O(\delta_{\max})$ depend only on the uniform bounds and Lipschitz constants in (5) and the displayed projection/value maps. In particular, for quasi-uniform meshes ($\delta_{\max} = O(1/T)$) with $\|u_0^{(T)} - u_0\|_\infty = O(\delta_{\max})$, the convergence rate is $O(\delta_{\max}) = O(1/T)$.

The proof (Appendix B) introduces piecewise-constant interpolants, bounds the quadrature error via Lipschitz continuity, and closes with a Grönwall argument.

Kernel regularity. The kernel κ_θ is assumed bounded ($|\kappa_\theta| \le M_\kappa$) and jointly Lipschitz:

$$|\kappa_\theta(x,\xi;u,v) - \kappa_\theta(x',\xi';u',v')| \le L^{\text{pos}}_\kappa\big(|x-x'| + |\xi-\xi'|\big) + L^{\text{state}}_\kappa\big(\|u-u'\| + \|v-v'\|\big). \tag{5}$$

Under the constant-horizon scaling, the Mamba-3 kernel is

$$\kappa_\theta(x,\xi;u,v) = \langle \mathrm{SiLU}(S_C u),\, \mathcal{R}_\theta(x-\xi)\ \mathrm{SiLU}(S_B v)\rangle, \qquad \mathcal{R}_\theta(\cdot) := \mathrm{Block}\big(R(\theta[m]\,\cdot)\big)_{m=1}^{N/2}. \tag{6}$$

Orthogonality of $\mathcal{R}_\theta$ gives boundedness; 1-Lipschitz continuity of $R(\cdot)$ gives $L_\kappa^{\mathrm{pos}} = \|\theta\|_\infty M_C M_B'$ with $M_C := \sup_u \|\,\mathrm{SiLU}(S_C u)\|$, $M_B' := \sup_u \|\,\mathrm{SiLU}(S_B u)\|$ (for Mamba-2, $\theta \equiv 0$ and $L_\kappa^{\mathrm{pos}} = 0$; see Appendix A). The Mamba-3 exponential-trapezoidal discretization has the same $O(\delta_{\max})$ limit when the input-conditioned gates are smooth and quasi-uniform, with $\lambda_{s+1} - \lambda_s = O(\delta_{\max})$ and $\delta_{s+1} - \delta_s = O(\delta_{\max}^2)$; the diagonal endpoint is one $O(\delta_{\max})$ cell. The MIMO extension requires a matrix-valued kernel. Equation (4) is structurally distinct from the McKean–Vlasov limits of transformer theory [4, 11, 17, 21]. The integral runs from 0 to x, not over all positions: the field at x depends on the causal history $\xi < x$ but not on $\xi > x$. The exponential kernel $e^{-a(x-\xi)}$ introduces a memory horizon $1/a$ with no analogue in symmetric pairwise attention interactions.

Remark (memory-horizon scaling). Let $u^{(a)}$ denote the solution of (4) with decay rate a, and define $\mathcal{G}(x,z) := P_z[\kappa_\theta(x,x;z,z)\, S_o\ \mathrm{SiLU}(S_x z)]$. As $a \to \infty$, $\|\partial_t u^{(a)}\|_\infty \le C/a$ and the unscaled dynamics freeze; on slow time at, the kernel concentrates at $\xi = x$ and gives the pointwise ODE $\partial_\tau v = \mathcal{G}(x,v)$ away from the boundary. As $a \to 0$, $e^{-a(x-\xi)} \to 1$ uniformly and the limit becomes full-history causal transport. Thus a interpolates between nearly local dynamics and full-history transport, unlike attention temperature β, which changes coupling strength rather than memory range. Appendix D gives the proof, the boundary-layer qualification for the slow-time local limit, and the effective-lag formula; Appendix Figure 4(a) confirms the $O(1/a)$ scaling numerically.

4. Experiments and Discussion

Figure 1 summarizes four checks: pretrained Mamba-2-130M layer 0 head 7 shows the triangular chain-memory kernel; the synthetic fixed-clock particle system (3) matches a Volterra reference at $M = 4096$, with L^∞ error falling from 0.222 at $T = 32$ to 0.005 at $T = 1024$; the mean-to-token-1 distance is essentially zero for the transformer but near one for Mamba-2/3; and the kernel slice contrasts the non-oscillatory Mamba-2 reduction with the sign-changing SISO Mamba-3 interaction induced by $A + i\theta$. Appendix E gives supplementary convergence details, dynamic-gate sanity checks, dynamical diagnostics, and empirical memory-horizon statistics.

Our analysis treats a fixed-clock SISO block, where the scan parameters are computed from the input sequence and held fixed along the continuous-depth flow; it does not yet cover fully dynamic output-gated or MIMO Mamba-3. Key open problems are moving-coordinate limits for multi-layer gates and long-time behavior of (4), where clustering or metastability may connect to depth-asymptotic token-flow results [19, 22].

The novelty is to identify the *Volterra limit* and *chain-dependent kernel* as dense-token signatures of selective SSMs. Related work clarifies nearby structures: Ali et al. [1] and Zimerman et al. [23] interpret Mamba and gated-linear RNNs as implicit causal attention; SSD [7] gives an algebraic semiseparable duality, with Hu et al. [14] isolating the structured state-space duality viewpoint; and Mamba-3 [18] adds complex-valued states and MIMO structure. Cirone et al. [6] develop rough-path foundations, Halloran et al. [13] prove Lyapunov stability, and Castin et al. [3] derive a Vlasov PDE for deep transformers with masked attention, but not the unnormalized scan kernel or chain-dependent SSM mask. Our contribution adds the *dense-token continuum limit*. By formulating the convergence theorem in terms of a general Lipschitz kernel (condition (5)), the result

covers the SISO Mamba-3 recurrence (complex-valued, with a rotation-modulated kernel) and its Mamba-2 reduction ($\theta \equiv 0$) without any change to the proof; the rank-R MIMO extension requires a matrix-valued kernel. The Mamba-3 kernel (6) introduces a qualitatively new feature: the effective interaction $e^{-a(x-\xi)}\langle C, \mathcal{R}_\theta(x-\xi)B\rangle$ is a damped oscillation, parametrized by the complex diagonal $A + i\theta$ of the underlying SSM. This frequency extension of the memory horizon has no transformer analogue and creates scale-selective memory across heads. More broadly, the Rigollet et al. program for transformer dynamics and mean-field attention [2, 4, 5, 8–11, 15–17, 20, 21, 24] can be extended beyond transformers, but the extension changes the mathematics. Classifying the continuum limits of different architectures, including gated linear attention and hybrid models, is a natural next step.

Acknowledgments

FY is grateful for partial support from seed funding by the Center for Emerging Artificial Intelligence Systems at the University at Albany.

References

[1] Ameen Ali Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of mamba models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1516–1534, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.76. URL https://aclanthology.org/2025.acl-long.76/.

[2] Krishnakumar Balasubramanian, Sayan Banerjee, and Philippe Rigollet. On the structure of stationary solutions to McKean–Vlasov equations with applications to noisy transformers. *arXiv preprint arXiv:2510.20094*, 2025.

[3] Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

[4] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025.

[5] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.05554.

[6] Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

[8] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[9] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

[10] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.

[11] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62:427–479, 2025.

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

[13] John T. Halloran, Manbir Gulati, and Paul F. Roysdon. Mamba state-space models are Lyapunov-stable learners. *Transactions on Machine Learning Research (TMLR)*, 2025.

[14] Jerry Yao-Chieh Hu, Xiwen Zhang, Ali ElSheikh, Weimin Wu, and Han Liu. On structured state-space duality. *arXiv preprint arXiv:2510.04944*, 2025. doi: 10.48550/arXiv.2510.04944. URL https://arxiv.org/abs/2510.04944.

[15] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[16] Nikita Karagodin, Shu Ge, Yury Polyanskiy, and Philippe Rigollet. Normalization in attention dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.22026.

[17] Hugo Koubbi, Borjan Geshkovski, and Philippe Rigollet. Homogenized transformers. *arXiv preprint arXiv:2604.01978*, 2026.

[18] 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 (ICLR)*, 2026.

[19] Trinh Tien Nguyen, Minh-Khoi Nguyen-Nhat, Duy-Tung Pham, Hoang-Son Do, Tan Minh Nguyen, and Thieu Vo. Token dynamics on spheres in mamba models, 2025. Submitted to ICLR 2026; OpenReview: https://openreview.net/forum?id=466gVY2sBQ.

[20] Yury Polyanskiy, Philippe Rigollet, and Andrew Yao. Synchronization of mean-field models on the circle. *arXiv preprint arXiv:2507.22857*, 2025.

[21] Philippe Rigollet. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2026. To appear in Proceedings of the International Congress of Mathematicians (ICM 2026).

[22] Thieu Vo, Duy-Tung Pham, Xin T. Tong, and Tan Minh Nguyen. Demystifying the token dynamics of deep selective state space models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=qtTIP5Gjc5. Spotlight.

[23] Itamar Zimerman, Ameen Ali Ali, and Lior Wolf. Explaining modern gated-linear RNNs via a unified implicit attention formulation. In *International Conference on Learning Representations (ICLR)*, 2025.

[24] Aleksandr Zimin, Yury Polyanskiy, and Philippe Rigollet. YuriiFormer: A suite of Nesterov-accelerated transformers. *arXiv preprint arXiv:2601.23236*, 2026.

Appendix organization. Appendix A derives the projected particle ODE from the SISO Mamba-3 recurrence and records the Mamba-2 reduction. Appendix B proves the Volterra convergence theorem. Appendices C–D prove the auxiliary token-1 and memory-horizon statements. The remaining appendices collect expanded/sanity-check figures and empirical horizon statistics.

Appendix A. Derivation of the Projected ODE from the SSM Recurrence

We derive the projected particle system (2) from the discrete Mamba-3 recurrence; the Mamba-2 case follows by setting $\lambda \equiv 1$, $\theta \equiv 0$.

Step 1: Discrete SSM recurrence. A single-head selective SSM block evolves a hidden state $h_t \in \mathbb{R}^{N\times P}$ via a scalar decay and data-dependent projections $B_t = \mathrm{SiLU}(S_B u_t) \in \mathbb{R}^N$, $C_t = \mathrm{SiLU}(S_C u_t) \in \mathbb{R}^N$, and value $x_t = \mathrm{SiLU}(S_x u_t) \in \mathbb{R}^P$. (In Mamba-2, the projections B_t, C_t are preceded by a short 1D convolution which we omit; in Mamba-3, this convolution is absent [18].) By Proposition 3.2.1 of Lahoti et al. [18], the complex SSM $\dot{h} = (A+i\theta)h + Bx^\top$ with exponential-trapezoidal discretization is equivalent to the real recurrence

$$h_t = \alpha_t\, R_t\, h_{t-1} + \beta_t\, (R_t B_{t-1})\, x_{t-1}^\top + \gamma_t\, B_t\, x_t^\top, \qquad y_t = C_t^\top h_t, \tag{7}$$

where $\alpha_t = e^{-a\Delta_t}$ is the scalar decay, $R_t = \mathrm{Block}(R(\Delta_t\theta[m]))_{m=1}^{N/2}$ is the block-diagonal rotation, $\gamma_t = \lambda_t\Delta_t$ and $\beta_t = (1-\lambda_t)\Delta_t\alpha_t$ are the two-tap weights, and $\lambda_t \in [0,1]$ is a data-dependent gate. Setting $\lambda_t \equiv 1$, $R_t = I$ recovers the Mamba-2 recurrence $h_t = e^{-a\Delta_t}h_{t-1} + \Delta_t B_t x_t^\top$.

Step 2: Unrolling to the causal mask. Unrolling from $h_0 = 0$ and defining cumulative positions $\tilde{s}_t := \sum_{k=1}^t \Delta_k$: the decay and rotation both telescope. Within each 2×2 block m, the rotations commute: $R(\Delta_{s+1}\theta[m])\cdots R(\Delta_t\theta[m]) = R(\sum_{k=s+1}^t \Delta_k\theta[m])$. This block-commutativity requires scalar A_t (shared across state dimensions within each block). The relative rotation from s to t is $\mathcal{R}_{s\to t} = \mathrm{Block}\big(R(\sum_{k=s+1}^t \Delta_k\theta[m])\big)_{m=1}^{N/2}$, and the output becomes

$$y_t = \sum_{s=1}^{t} \langle C_t,\, \mathcal{R}_{s\to t}\, B_s\rangle \underbrace{\Big[\lambda_s\Delta_s\, e^{-a(\tilde{s}_t-\tilde{s}_s)} + \mathbf{1}_{s<t}\,(1-\lambda_{s+1})\Delta_{s+1}\, e^{-a(\tilde{s}_t-\tilde{s}_s)}\Big]}_{=:\,T_{ts}\ (\text{cf. }(1))}\, x_s. \tag{8}$$

Setting $\lambda_s \equiv 1$ eliminates the second tap and recovers the Mamba-2 mask $T_{ts}^{\mathrm{M2}} = e^{-a(\tilde{s}_t-\tilde{s}_s)}\Delta_s$, the 1-semiseparable structure of Dao and Gu [7].

Step 3: Residual update and RMSNorm projection. This step is identical for both architectures. Each layer applies a residual update followed by RMSNorm:

$$u_i^{(\ell+1)} = \mathrm{RMSNorm}\big(u_i^{(\ell)} + \epsilon\, S_o\, y_i^{(\ell)}\big), \qquad \mathrm{RMSNorm}(z) = z\cdot\sqrt{d}\,/\,\|z\|,$$

where ϵ is the residual scaling and $S_o \in \mathbb{R}^{d\times P}$ is the output projection. Taylor-expanding for small ϵ:

$$u_i^{(\ell+1)} = u_i + \epsilon\, P_{u_i}(S_o y_i) + O(\epsilon^2), \qquad P_u := I - uu^\top/d.$$

Step 4: Continuous-depth ODE. Taking $\epsilon \to 0$ gives the projected ODE on $\mathcal{S}^{d-1}_{\mathrm{rms}}$:

$$\frac{du_i}{dt} = P_{u_i}\Big(S_o \sum_{j=1}^{i} \kappa_\theta(\tilde{s}_i, \tilde{s}_j;\, u_i, u_j)\, T_{ij}\ \mathrm{SiLU}(S_x u_j)\Big),$$

which is (2) with mask T_{ij} (1) and content kernel $\kappa_\theta(\tilde{s}_i, \tilde{s}_j; u_i, u_j) = \langle \mathrm{SiLU}(S_C u_i),\ \mathcal{R}_{j\to i}\ \mathrm{SiLU}(S_B u_j)\rangle$. For Mamba-2 ($\lambda \equiv 1$, $\theta \equiv 0$): $T_{ij} = e^{-a(\tilde{s}_i - \tilde{s}_j)}\Delta_j \mathbf{1}_{\{j\le i\}}$ and $\kappa_\theta = \langle \mathrm{SiLU}(S_C u_i),\ \mathrm{SiLU}(S_B u_j)\rangle$. The projection P_{u_i} ensures $\|u_i(t)\|^2 = d$ for all time.

Step 5: Continuum limit. Passing to normalized coordinates $s_i = \tilde{s}_i/S_T$, $\delta_i = \Delta_i/S_T$, under the constant-horizon scaling:

- The decay becomes $e^{-a(s_i - s_j)}$ and the quadrature weight becomes δ_j.
- The accumulated rotation angle becomes $\theta[m](s_i - s_j)$ under the raw frequency scaling $\theta_T[m] = \theta[m]/S_T$, where the displayed θ denotes the fixed effective frequency. Thus $\mathcal{R}_{j\to i} \to \mathcal{R}_\theta(x-\xi) := \mathrm{Block}(R(\theta[m](x-\xi)))_{m=1}^{N/2}$. For Mamba-2 ($\theta \equiv 0$), this is the identity.
- The normalized two-tap weights from (1) merge under the same Lipschitz/quasi-uniform mesh condition used for the quadrature estimate:

  $$\lambda_s \delta_s + (1-\lambda_{s+1})\delta_{s+1} = \delta_s + O(\delta^2_{\max}).$$

  Here $\lambda_{s+1} - \lambda_s = O(\delta_{\max})$ and $\delta_{s+1} - \delta_s = O(\delta^2_{\max})$ follow from the Lipschitz input field and smooth gates. Indeed, on a quasi-uniform token grid a smooth gate $\Delta_i = \mathrm{softplus}(S_\Delta \bar{u}_i + b)$ varies by $O(1/T)$ from cell to cell, and the normalization $S_T = \Theta(T)$ gives $\delta_{s+1} - \delta_s = (\Delta_{s+1} - \Delta_s)/S_T = O(1/T^2)$. Summing over $O(\delta^{-1}_{\max})$ cells gives an $O(\delta_{\max})$ contribution to the quadrature error. For the diagonal source cell $s = t$, the recurrence supplies only $\lambda_t \delta_t$ rather than the merged weight δ_t. The missing endpoint contribution $(1-\lambda_t)\delta_t$ is $O(\delta_{\max})$ for that single cell, so it is absorbed into the same global $O(\delta_{\max})$ quadrature modulus. For Mamba-2 ($\lambda \equiv 1$), the mask is already single-tap.

The particle system takes the form (3), and Theorem 1 gives convergence to (4) at rate $O(\delta_{\max})$. Thus the theorem applies directly to the fixed-mesh particle system; applying it to the unrolled SISO recurrence uses the additional smooth quasi-uniform gate condition and diagonal endpoint accounting above.

Verification of kernel regularity (condition (5)). Boundedness follows from $\|\mathcal{R}_\theta\|_{\mathrm{op}} = 1$ (orthogonality); the position-Lipschitz constant is $L^{\mathrm{pos}}_\kappa = \|\theta\|_\infty M_C M'_B$ (with M_C, M'_B as in (6)) since $\|R(\phi) - R(\psi)\|_{\mathrm{op}} \le |\phi - \psi|$; and $L^{\mathrm{state}}_\kappa$ is inherited from $\mathrm{SiLU}(S_C\cdot)$ and $\mathrm{SiLU}(S_B\cdot)$. For Mamba-2 ($\theta \equiv 0$): $L^{\mathrm{pos}}_\kappa = 0$.

Summary. The derivation chain is:

$$\text{SSM recurrence} \xrightarrow{\text{unroll}} \text{chain-dependent mask } T_{ij}\ (1) \xrightarrow{\text{residual + RMSNorm}} \text{projected ODE on } \mathcal{S}^{d-1}_{\mathrm{rms}} \xrightarrow{T\to\infty} \text{Volterra equation (4).}$$

Remark: oscillation-horizon scaling. The scaling $\theta_T = \theta/S_T$ is the rotational analogue of $a_T = a/S_T$: both ensure that accumulated per-step effects remain $O(1)$ in normalized coordinates. If instead $\theta[m] = O(1)$ (fixed frequencies), the accumulated phase $\theta[m]S_T(x-\xi) \to \infty$ and the nonzero-frequency components average out by the Riemann–Lebesgue lemma, leaving only any zero-frequency/DC components rather than the full Mamba-2 kernel.

Appendix B. Proof of Theorem 1

We establish convergence of the T-particle system (3) to the Volterra equation (4) under the constant-horizon scaling.

Step 1: Mesh-point comparison. Given T particles at normalized positions $0 < s_1 < \cdots < s_T \le 1$ with widths $\delta_i := s_i - s_{i-1}$ ($s_0 := 0$), we compare *at mesh points* $x = s_i$ to avoid cell-interior artifacts. The particle equation (3) evaluated at $x = s_i$ reads

$$\partial_t u_i^{(T)} = P_{u_i^{(T)}}\, S_o \underbrace{\sum_{j=1}^{i} \kappa_\theta(s_i, s_j;\, u_i^{(T)}, u_j^{(T)})\, e^{-a(s_i - s_j)}\, \delta_j\, \mathrm{SiLU}(S_x u_j^{(T)})}_{=:\,\mathcal{I}_T^i(t)}. \tag{9}$$

The Volterra equation (4) at $x = s_i$ reads $\partial_t u(s_i,t) = P_{u(s_i,t)}\, S_o\, \mathcal{I}[u](s_i,t)$ with

$$\mathcal{I}[u](s_i,t) := \int_0^{s_i} e^{-a(s_i-\xi)}\, \kappa_\theta\big(s_i, \xi;\, u(s_i,t), u(\xi,t)\big)\, \mathrm{SiLU}\big(S_x u(\xi,t)\big)\, d\xi.$$

The key observation is that $\mathcal{I}_T^i$ is a *right-endpoint Riemann sum* for $\mathcal{I}[u](s_i,\cdot)$, with quadrature points at $s_1, \ldots, s_i$ (the right endpoints of the partition intervals $(s_{j-1}, s_j]$) and weights $\delta_1, \ldots, \delta_i$.

Step 2: Regularity bounds. Since all tokens lie on the compact set $\mathcal{S}_{\mathrm{rms}}^{d-1}$:

- SiLU is Lipschitz with constant L_{SiLU}, and $\|\mathrm{SiLU}(S_x u)\| \le M_{\mathrm{SiLU}}$ for all $u \in \mathcal{S}_{\mathrm{rms}}^{d-1}$.
- By condition (5), $|\kappa_\theta| \le M_\kappa$ and κ_θ is jointly Lipschitz with constants L_κ^{pos} (in positions) and $L_\kappa^{\mathrm{state}}$ (in states).
- The projector satisfies $\|P_u v - P_{u'} v\| \le C_P \|u - u'\|\, \|v\|$.
- $e^{-a(x-\xi)} \le 1$ for $\xi \le x$, and $\xi \mapsto e^{-a(x-\xi)}$ is Lipschitz with constant a.

One admissible stability constant for the mesh-point comparison is

$$L := \|S_o\|(2L_\kappa^{\mathrm{state}} M_{\mathrm{SiLU}} + M_\kappa L_{\mathrm{SiLU}} \|S_x\|) + C_P \|S_o\| M_\kappa M_{\mathrm{SiLU}}. \tag{10}$$

Position-Lipschitz terms enter the quadrature error and the spatial Lipschitz constant D in Lemma 2.

Step 3: Quadrature error at mesh points. For a *Lipschitz* field $v \in \mathrm{Lip}([0,1];\mathcal{S}^{d-1}_{\mathrm{rms}})$ with Lipschitz constant Λ_v, the integrand $f_v(s_i,\xi) := e^{-a(s_i-\xi)}\kappa_\theta(s_i,\xi;\, v(s_i),v(\xi))\,\mathrm{SiLU}(S_x v(\xi))$ is Lipschitz in ξ with constant L_f depending on $a, L^{\mathrm{pos}}_\kappa, L^{\mathrm{state}}_\kappa, L_{\mathrm{SiLU}}, M_\kappa, M_{\mathrm{SiLU}}, \|S_x\|, \Lambda_v$. The right-endpoint Riemann sum error at $x = s_i$ satisfies

$$\left\|\mathcal{I}^i_T[v] - \mathcal{I}[v](s_i)\right\| \le L_f\,\delta_{\max}, \tag{11}$$

by the standard bound for Lipschitz integrands on a partition of mesh width $\delta_{\max}$.

Regularity requirement: the bound (11) requires v to be spatially Lipschitz. We verify this for the Volterra solution:

Lemma 2 (Lipschitz propagation via Banach-space route) *Let $M_B := \|S_o\| M_{\mathrm{SiLU}}$, $c_a := (1-e^{-a})/a$, $D := 2M_\kappa M_B + c_a L^{\mathrm{pos}}_\kappa M_B$, and $B_a := c_a(L^{\mathrm{state}}_\kappa M_B + C_P M_\kappa M_B)$. If $u_0 \in \mathrm{Lip}([0,1];\mathcal{S}^{d-1}_{\mathrm{rms}})$ with constant Λ_0, then the Volterra equation (4) admits a unique global solution $u \in C^1([0,\infty);C([0,1];\mathbb{R}^d)) \cap C([0,\infty);C([0,1];\mathcal{S}^{d-1}_{\mathrm{rms}}))$, and*

$$\mathrm{Lip}_x\, u(\cdot,t) \le e^{B_a t}\Lambda_0 + \frac{D}{B_a}(e^{B_a t} - 1).$$

(For Mamba-2, $L^{\mathrm{pos}}_\kappa = 0$ and $D = 2M_\kappa M_B$.)

Proof Define the vector field $F : C([0,1];\mathcal{S}^{d-1}_{\mathrm{rms}}) \to C([0,1];\mathbb{R}^d)$ by $(F[v])(x) := P_{v(x)}\,G[v](x)$, where $G[v](x) := \int_0^x e^{-a(x-\xi)}H(x,\xi;\, v(x),v(\xi))\,d\xi$ and $H(x,\xi;z,w) := \kappa_\theta(x,\xi;z,w)\,S_o\,\mathrm{SiLU}(S_x w)$. By condition (5), $\|H\|_\infty \le M_\kappa M_B$, so $\|G[v]\|_\infty \le c_a M_\kappa M_B$. For the state-Lipschitz bound:

$$\|G[v] - G[w]\|_\infty \le c_a(2L^{\mathrm{state}}_\kappa M_B + M_\kappa L_B)\|v - w\|_\infty,$$

where $L_B := \|S_o\| L_{\mathrm{SiLU}} \|S_x\|$. (The position-Lipschitz constant L^{pos}_κ does not enter here because v and w are evaluated at the same positions.) Adding the projector difference yields $\|F[v] - F[w]\|_\infty \le L_F\|v-w\|_\infty$ with $L_F = c_a(2L^{\mathrm{state}}_\kappa M_B + M_\kappa L_B + C_P M_\kappa M_B)$. Here L_F is the global Lipschitz constant used for the Picard–Lindelöf argument, whereas B_a in the lemma is the smaller constant governing spatial-modulus propagation. Since F is globally Lipschitz on the Banach space $C([0,1];\mathbb{R}^d)$, Picard–Lindelöf gives a unique global C^1 solution; the sphere constraint is preserved because P_u projects tangentially.

For the spatial modulus, fix $0 \le x < y \le 1$ with $h := y - x$ and write

$$\begin{aligned} G[v](y) - G[v](x) = &\underbrace{\int_0^x (e^{-a(y-\xi)} - e^{-a(x-\xi)})H(y,\xi;\, v(y),v(\xi))\,d\xi}_{\text{decay shift}} \\ &+ \underbrace{\int_0^x e^{-a(x-\xi)}\big(H(y,\xi;\, v(y),v(\xi)) - H(x,\xi;\, v(x),v(\xi))\big)\,d\xi}_{\text{kernel+field shift}} \\ &+ \underbrace{\int_x^y e^{-a(y-\xi)}H(y,\xi;\, v(y),v(\xi))\,d\xi}_{\text{new interval}}. \end{aligned}$$

The decay-shift and new-interval terms are together bounded by $2M_\kappa M_B\, h$. For the middle term, the H-difference is bounded by $(L_\kappa^{\mathrm{pos}} + L_\kappa^{\mathrm{state}}\Lambda_v)\, h \cdot M_B$ using both the position-Lipschitz constant (from $|y - x| = h$) and the state-Lipschitz constant (from $\|v(y) - v(x)\| \le \Lambda_v h$); integrated against $e^{-a(x-\xi)}$ this gives $c_a(L_\kappa^{\mathrm{pos}} + L_\kappa^{\mathrm{state}}\Lambda_v)\, M_B\, h$. More precisely, the field-shift part is $c_a L_\kappa^{\mathrm{state}} M_B\, \omega_v(h)$ while the position-shift part is $c_a L_\kappa^{\mathrm{pos}} M_B\, h$. Adding the projector difference: $\|F[v](y) - F[v](x)\| \le D\, h + B_a\, \omega_v(h)$. Grönwall on $\omega_{u(\cdot,t)}(h)$ gives the stated bound, with $D\, h$ as the inhomogeneous forcing. ∎

Let Λ_u denote the spatial Lipschitz bound from Lemma 2 at time t_f:

$$\Lambda_u := e^{B_a t_f}\Lambda_0 + \frac{D}{B_a}(e^{B_a t_f} - 1).$$

Define the raw quadrature error $\omega(\delta_{\max}) := L_f\, \delta_{\max}$, where L_f depends on Λ_u.

Step 4: Grönwall estimate at mesh points. Define the mesh-point error $e(t) := \max_{i=1,\dots,T} \|u_i^{(T)}(t) - u(s_i,t)\|$. Let $\Pi_T u$ be the piecewise-constant projection of the Volterra solution onto the mesh: $(\Pi_T u)(x) := u(s_i)$ for $x \in (s_{i-1}, s_i]$. Subtracting (9) from the Volterra equation at $x = s_i$ and adding/subtracting $\mathcal{I}_T^i[\Pi_T u]$ gives

$$\begin{aligned}
\partial_t(u_i^{(T)} - u(s_i)) &= P_{u_i^{(T)}} S_o\, \mathcal{I}_T^i[u^{(T)}] - P_{u(s_i)}\, S_o\, \mathcal{I}[u](s_i) \\
&= \underbrace{P_{u_i^{(T)}} S_o\big(\mathcal{I}_T^i[u^{(T)}] - \mathcal{I}_T^i[\Pi_T u]\big)}_{\text{(I) stability}} + \underbrace{P_{u_i^{(T)}} S_o\big(\mathcal{I}_T^i[\Pi_T u] - \mathcal{I}[u](s_i)\big)}_{\text{(II) quadrature/projection}} \\
&\quad + \underbrace{(P_{u_i^{(T)}} - P_{u(s_i)}) S_o\, \mathcal{I}[u](s_i)}_{\text{(III) projector}}.
\end{aligned} \tag{12}$$

Here $u^{(T)}(x,t) := u_i^{(T)}(t)$ for $x \in (s_{i-1}, s_i]$. Bounding each:

- (I): By Lipschitz dependence of $\mathcal{I}_T^i$ on both state arguments at fixed mesh positions, $\|\mathcal{I}_T^i[u^{(T)}] - \mathcal{I}_T^i[\Pi_T u]\| \le L' e(t)$ where $L' = 2L_\kappa^{\mathrm{state}} M_{\mathrm{SiLU}} + M_\kappa L_{\mathrm{SiLU}} \|S_x\|$.
- (II): Since $\mathcal{I}_T^i[\Pi_T u]$ is exactly the right-endpoint Riemann sum for the Lipschitz field $u(\cdot,t)$ in $\mathcal{I}[u](s_i)$, whose Lipschitz constant is bounded by Λ_u, applying (11) with $v = u(\cdot,t)$ gives $\|\mathcal{I}_T^i[\Pi_T u] - \mathcal{I}[u](s_i)\| \le L_f \delta_{\max}$. Thus this contribution is $\le \omega(\delta_{\max})$ with $\omega(\delta_{\max}) = O(\delta_{\max})$.
- (III): $\le C_P \|S_o\| M_\kappa M_{\mathrm{SiLU}}\, e(t)$.

Combining, with ω denoting the quadrature modulus, $e'(t) \le L\, e(t) + \omega(\delta_{\max})$, By Grönwall's inequality,

$$e(t) \le e^{Lt}\, e(0) + \frac{\omega(\delta_{\max})}{L}(e^{Lt} - 1) \le e^{Lt_f}\big(e(0) + t_f\, \omega(\delta_{\max})\big) \tag{13}$$

for all $t \in [0, t_f]$.

Step 5: Conclusion. Since $u^{(T)}(\cdot,0) \to u_0$ uniformly, $e(0) \to 0$. Since $\delta_{\max}^{(T)} \to 0$, $\omega(\delta_{\max}^{(T)}) \to 0$. Hence $e(t) \to 0$ uniformly on $[0, t_f]$.

For the rate: if $\delta_{\max} = O(1/T)$ (e.g. uniform partition) and u_0 is Lipschitz so that $e(0) = O(\delta_{\max})$, then $e(t) = O(\delta_{\max}) = O(1/T)$.

The piecewise-constant extension to all $x \in [0,1]$ then gives $\|u^{(T)}(x,t) - u(x,t)\| \le e(t) + \Lambda_u \delta_{\max} = O(\delta_{\max})$ uniformly. □

Appendix C. Token-1 Non-Stationarity

Proposition 3 (Token-1 non-stationarity) *Assume $d \ge 2$. In the causal transformer with $V = I$, token 1 is exactly stationary: $\dot{x}_1 = P_{x_1} x_1 = 0$. In the Mamba-2/3 system (2), draw the entries of (S_o, S_x, S_B, S_C) from any absolutely continuous distribution. Then for any fixed $u_1 \in \mathcal{S}_{\mathrm{rms}}^{d-1}$ independent of the weights, and for any fixed input-gate parameters (S_λ, S_Δ, b) whose first-token factors satisfy $\lambda_1, \Delta_1 > 0$, $\dot{u}_1 \neq 0$ almost surely.*

Proof Setting $i = 1$ in (2): $\dot{u}_1 = \alpha_1 P_{u_1}(S_o \,\mathrm{SiLU}(S_x u_1))$ with $\alpha_1 := \lambda_1 \Delta_1 \langle \mathrm{SiLU}(S_C u_1), \mathrm{SiLU}(S_B u_1) \rangle$. Stationarity $\dot{u}_1 = 0$ requires either $\alpha_1 = 0$ or $S_o \,\mathrm{SiLU}(S_x u_1) \in \mathrm{span}(u_1)$. We show each event has probability zero when the parameter entries are drawn from an absolutely continuous distribution.

Branch 1 ($\alpha_1 = 0$). The input-fixed factors $\lambda_1 = \sigma(S_\lambda \bar{u}_1)$ and $\Delta_1 = \mathrm{softplus}(S_\Delta \bar{u}_1 + b)$ are strictly positive. Thus we need $\langle \mathrm{SiLU}(S_C u_1), \mathrm{SiLU}(S_B u_1) \rangle = 0$. Write $S_B = (s_1^B, \ldots, s_N^B)^\top$ where each $s_k^B \in \mathbb{R}^d$ is a row, and define $\varphi_k := \mathrm{SiLU}(s_k^B \cdot u_1)$. Fix S_C and all rows of S_B except the first. The inner product $\langle \mathrm{SiLU}(S_C u_1), \mathrm{SiLU}(S_B u_1) \rangle = \sum_{k=1}^N c_k \varphi_k$ is affine in $\varphi_1 = \mathrm{SiLU}(s_1^B \cdot u_1)$. On the full-measure set where $c_1 := \mathrm{SiLU}((S_C)_1 \cdot u_1) \neq 0$ (since $\{s : s \cdot u_1 = 0\}$ is a hyperplane), the sum vanishes for at most one value of φ_1, call it φ_1^*. The level set $\{s_1^B : \mathrm{SiLU}(s_1^B \cdot u_1) = \varphi_1^*\}$ is contained in at most finitely many hyperplanes in $\mathbb{R}^d$ (since SiLU is C^1 with $\mathrm{SiLU}(z) = 0 \Leftrightarrow z = 0$ and $\mathrm{SiLU}'(z) \neq 0$ for all but finitely many z), hence has Lebesgue measure zero. By Fubini, $\varphi^{-1}(0)$ has measure zero in (S_B, S_C).

Branch 2 ($\alpha_1 \neq 0$, $S_o \,\mathrm{SiLU}(S_x u_1) \in \mathrm{span}(u_1)$). Since $\mathrm{SiLU}(z) = 0 \Leftrightarrow z = 0$, the event $z := \mathrm{SiLU}(S_x u_1) = 0$ requires every row of S_x to map u_1 to 0, a hyperplane condition on each row, hence probability zero under any absolutely continuous distribution on S_x. When $z \neq 0$, the linear map $S_o \mapsto S_o z$ is surjective ($\mathbb{R}^{d \times P} \to \mathbb{R}^d$), so $\{S_o : S_o z \in \mathrm{span}(u_1)\}$ is an affine subspace of codimension $d-1 \ge 1$, which has measure zero.

For the fixed token u_1, taking the union of these probability-zero events gives $\dot{u}_1 \neq 0$ almost surely. ■

Appendix D. Memory-Horizon Scaling

Proof [Proof of the memory-horizon remark] For bounded f, $\int_0^x e^{-a(x-\xi)} f(\xi)\, d\xi \le \|f\|_\infty / a$, giving $\|\partial_t u^{(a)}\| \le C/a$ and hence $\|u^{(a)}(\cdot,t) - u_0\|_\infty \le Ct/a$. For $a \to 0$, $|e^{-a(x-\xi)} - 1| \le a$ on $0 \le \xi \le x \le 1$, so Grönwall gives convergence to the equation with the exponential removed. For the slow-time local limit, write

$$\partial_\tau v^{(a)} = a\, P_{v^{(a)}} \int_0^x e^{-a(x-\xi)} H(x, \xi; v^{(a)}(x), v^{(a)}(\xi))\, d\xi.$$

The identity

$$a\int_0^x e^{-a(x-\xi)}g(\xi)\,d\xi - g(x) = a\int_0^x e^{-a(x-\xi)}(g(\xi)-g(x))\,d\xi - e^{-ax}g(x)$$

shows that Lipschitz g gives an $O(1/a)$ interior error. Under the compatibility condition $\mathcal{G}(0,u_0(0)) = 0$, the boundary term also satisfies $e^{-ax}\|\mathcal{G}(x,v(x,\tau))\| \le L_{\mathrm{loc}}\Lambda(\tau)/(ae)$; without it, the same estimate holds on each strip $x \ge x_0 > 0$. ∎

The mean lag under the exponential kernel is

$$L_{\mathrm{eff}}(a) := \frac{\int_0^1\int_0^x (x-\xi)\,e^{-a(x-\xi)}\,d\xi\,dx}{\int_0^1\int_0^x e^{-a(x-\xi)}\,d\xi\,dx} = \frac{a-2+(a+2)e^{-a}}{a\big(a-1+e^{-a}\big)},$$

so $L_{\mathrm{eff}}(0) = \frac{1}{3}$ and $L_{\mathrm{eff}}(a) \sim 1/a$ as $a \to \infty$.

Appendix E. Supplementary Experimental Figures and Readout

This appendix expands the numerical evidence from Section 4. Each figure is meant to answer one diagnostic question rather than serve as an image dump: whether the convergence is genuinely first order in token spacing, whether recomputing the gates changes the qualitative finite-depth behavior, and whether the memory-rate limits show up in finite-dimensional dynamics.

Protocols. All particle systems are integrated with fourth-order Runge–Kutta time stepping and projected back to $\mathcal{S}^{d-1}_{\mathrm{rms}}$ after each step. Volterra reference solutions use the same time integrator together with trapezoidal quadrature on a fine spatial grid, and particle solutions are compared to the reference by piecewise-constant interpolation. The convergence diagnostic uses synthetic $d = 2$ fixed-clock systems with $t_f = 3$, a 4096-point reference grid, 300 time steps, and $T \in \{32, 64, 128, 256, 512, 1024\}$. The mean-to-token-1 diagnostic uses 40 random seeds with $T = 8$, $t_f = 20$, and 800 time steps. It reports

$$D_1(t_f) := d_{\mathcal{S}^{d-1}_{\mathrm{rms}}}(m(t_f), u_1(0))\,, \qquad m(t) := \sqrt{d}\,\frac{T^{-1}\sum_{i=1}^T u_i(t)}{\left\|T^{-1}\sum_{i=1}^T u_i(t)\right\|},$$

where $m(t)$ is the normalized Euclidean centroid on $\mathcal{S}^{d-1}_{\mathrm{rms}}$, $d_{\mathcal{S}^{d-1}_{\mathrm{rms}}}(u,v) := \sqrt{d}\,\arccos(\langle u, v\rangle/d)$ is the geodesic distance on the RMS sphere, and the denominator is nonzero in all reported trials. Figure 2 uses the same convergence protocol and displays spatially binned rescaled profiles $T\,E_T(x,t_f)$. For Figure 3, each of 20 random seeds is evaluated for both Mamba-2 and SISO Mamba-3 ($\theta = 3$) by integrating matched fixed-gate and dynamic-gate trajectories from the same initial tokens and weights. For Figure 4, the memory-rate experiment sweeps $a \in [10^{-1}, 10^2]$ over 30 logarithmically spaced values and 10 random seeds at $T = 32$; the mean-to-token-1 panels use 100 seeds, $T = 8$, $t_f = 30$, 1500 time steps, and a causal-transformer baseline with $\beta = 2$.

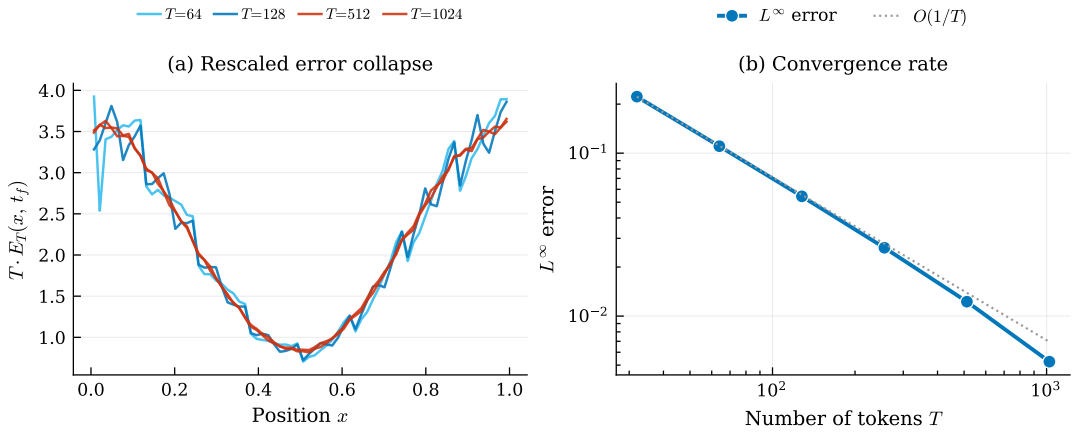

Figure 2: Convergence details (fixed clock, $d = 2$, 4096-point reference grid). (a) Spatially binned rescaled error profiles $T \cdot E_T(x, t_f)$ collapse across token counts, confirming structural $O(1/T)$ rather than a fitted final-time slope alone. (b) L^∞ error vs. T (log-log); dashed = $O(1/T)$. Together the panels isolate the quadrature mechanism in Theorem 1: after multiplying by the token count, the spatial error profile stabilizes, and the final-time sup-norm slope follows the predicted first-order rate.

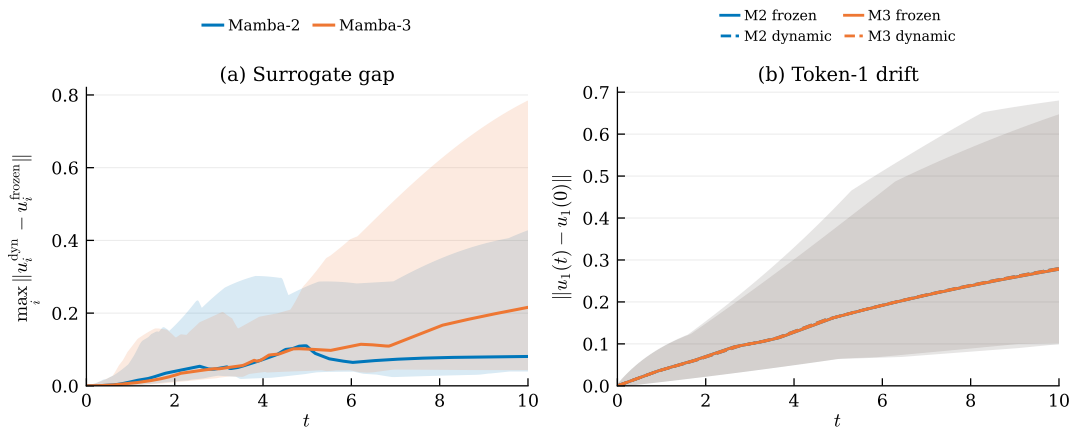

Figure 3: Fixed- versus dynamic-gate sanity check over 20 synthetic SISO trials per architecture ($T = 8$, $t_f = 10$). Panel (a) reports the maximum tokenwise separation between dynamic and fixed-gate trajectories from matched tokens and weights; panel (b) compares token-1 drift, with solid curves for fixed gates, dashed curves for dynamic gates, and shaded interquartile ranges. Median final max-token gaps are 0.081 for Mamba-2 and 0.216 for SISO Mamba-3 with rotation frequency $\theta = 3$, while median final token-1 drift is essentially unchanged (0.278 fixed, 0.280 dynamic). Thus the diagnostic supports the fixed-gate approximation as a local probe, but it does not extend the theorem beyond the fixed-clock particle system.

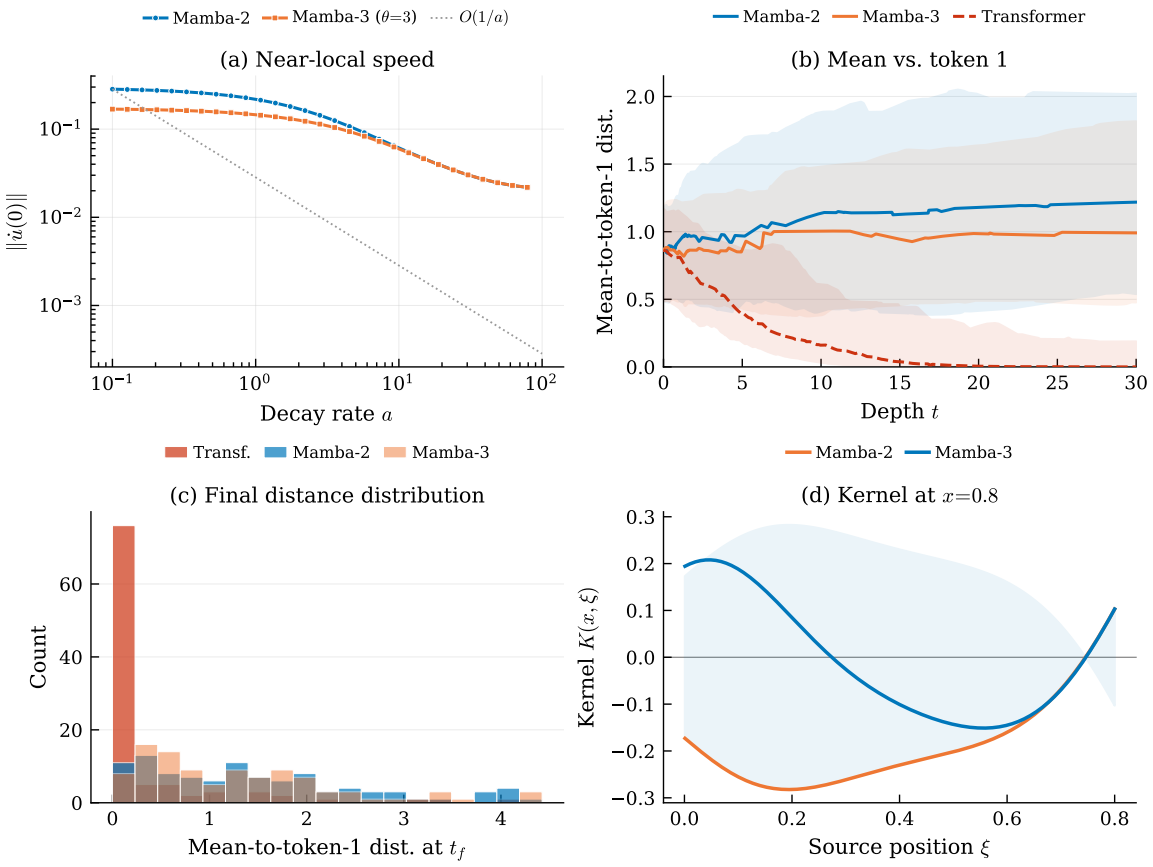

Figure 4: Additional dynamical consequences ($d = 2$, synthetic parameters). (a) Initial speed vs. a: both Mamba-2 and Mamba-3 ($\theta = 3$) confirm the $O(1/a)$ near-local scaling. (b) Mean-to-token-1 distance uses the same normalized-centroid-to-initial-token metric as the main text; solid curves are medians and shaded bands are interquartile ranges, with the dashed transformer baseline collapsing toward zero while both SSM variants stay large. (c) Final distribution of this mean-to-token-1 distance over seeds. (d) Content-modulated kernel slice for Mamba-2 and SISO Mamba-3; the zero line highlights the sign change introduced by the rotation. The readout is that the Volterra memory parameter has visible finite-dimensional consequences: large a suppresses motion, the causal-attention anchor mechanism is absent for the SSM flows, and the Mamba-3 rotation turns the non-oscillatory Mamba-2 slice into an oscillatory signed interaction.

Appendix F. Empirical Memory-Horizon Distribution

We motivate the memory-scale discussion by analyzing all 576 heads (24 layers $\times$ 24 heads) of pretrained Mamba-2-130M [7].

Setup. For each head h in layer ℓ, the architectural decay rate is $a_h = \exp(A_{\log,h})$ and the data-dependent step size is $\Delta_t = \mathrm{softplus}(S_\Delta u_t + b_h)$. As a bias-only architectural proxy, we estimate the typical step size as $\bar{\Delta}_h = \mathrm{softplus}(b_h)$ (since $S_\Delta u$ averages to ≈ 0 for random inputs), and define the effective per-step decay $\gamma_h = a_h\,\bar{\Delta}_h$. The *memory horizon* in tokens is $1/\gamma_h$: the number of past tokens whose influence remains above $1/e$. This is not an activation-conditioned measurement of Δ_t on real sequences; it is a lightweight proxy for the learned decay scale.

Results. The memory horizon spans four orders of magnitude across heads:

Memory horizon	Heads	Fraction
<1 token (ultra-local)	66	11%
1–5 tokens	173	30%
5–20 tokens	105	18%
20–100 tokens	91	16%
100–500 tokens	68	12%
>500 tokens (near-global)	73	13%

Quantiles: P25 = 2.3, median = 7.0, P75 = 98, P95 = 1233 tokens. Because these horizons are fixed in token count by the learned parameters a_h, b_h, they are not evidence that fixed learned parameters remain constant-horizon as T grows: at $T = 2048$ the median span is only 0.34% of the context, although the longest-horizon heads still cover an $O(1)$ fraction. At $T = 128$, 60% of heads span 1–50% of the sequence, placing many heads in a constant-horizon operating regime where the Volterra approximation $e^{-a(x-\xi)}$ with $a = O(1)$ is relevant; 33% remain in that range at $T = 2048$. Thus the pretrained model is multi-scale at operating lengths; beyond the training distribution, fixed-rate dynamics become increasingly local, consistent with the known difficulty of length generalization in SSMs. The model employs a *multi-scale strategy*: different heads specialize in different temporal scales, from sub-token (high-frequency features) to hundreds of tokens (long-range dependencies).

Per-layer structure. The median horizon varies by layer: early layers (0, 8, 17) tend toward short horizons (5–40 tokens), while certain middle and late layers (4, 19, 20, 21) have median horizons exceeding 100 tokens. This is consistent with the hierarchical composition observed in deep transformers, where lower layers capture local syntax and upper layers integrate longer-range context.