DECOUPLING TIME AND SPACE IN SELECTIVE CDES USING TENSOR TRAIN DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective state-space models (SSMs) and linear controlled differential equations (LNCDEs) have emerged as effective approaches for long-sequence modeling. However, existing models face a trade-off: diagonal state transitions offer computational efficiency but struggle to capture cross-channel interactions, whereas dense transitions successfully capture these interactions at a computational cost that structured alternatives have since aimed to reduce. To address this, we propose the Tensor-Decoupled CDE (TD-CDE), an architecture that decouples temporal updating from spatial mixing. By restricting time evolution to diagonal, data-dependent decays, TD-CDE maintains efficient parallel scans. Cross-channel interactions are modeled along the depth dimension using Tensor Train (TT) decomposition, extracting high-order signature-type features without exponential parameter growth. Furthermore, the model employs gating via raw inputs, avoiding normalizations to preserve causality and exact, additive interpretability. Theoretically, we prove that models with commuting generators—including diagonal SSMs—realise only the shuffle subalgebra and cannot represent complex interactions like Lévy area at any width, and that TD-CDE's graded state decomposes predictions exactly by interaction order and feature words. Additionally, we formalize an order contract for selective discretisation. Empirically, TD-CDE captures cross-channel dynamics in path-dependent synthetic tasks that challenge existing models like Mamba. Evaluated on the UEA archive, TD-CDE achieves state-of-the-art accuracy with up to two orders of magnitude fewer parameters specifically on the datasets our order profiler flags as high-order, while showing no advantage on low-order ones—a separation predicted prior to training. Crucially, we demonstrate that a high-order discretisation scheme is essential for resolution invariance: TD-CDE preserves robust predictions and stable exact attributions even under severe sequence subsampling, where standard endpoint approximations fail.

1 INTRODUCTION

Linear state-space models (SSMs) have emerged to overcome the quadratic cost of Transformers (Vaswani et al., 2017), evolving from continuous signal compression (Gu et al., 2020; 2021; Smith et al., 2022) to input-dependent selective mechanisms (Gu & Dao, 2023). Despite their success, limitations in state-tracking capabilities persist (Merrill et al., 2024; Jelassi et al., 2024), driving rapid architectural refinements including duality frameworks (Dao & Gu, 2024), delta updates for associative recall (Schlag et al., 2021; Yang et al., 2024b;a), and vectorized gating (Peng et al., 2025). For continuous-valued time series, these advances raise a complementary question: do the predictions and the temporal interactions supporting them remain stable when the observation grid changes?

Muca Cirone et al. (2024) unified selective SSMs as continuous-time linear controlled differential equations (CDEs), showing that input-dependent gating drives expressivity. This motivated structured state-transition matrices that balance efficiency and expressivity (Walker et al., 2026; Siems et al., 2025). However, most approaches structure a single time-evolution matrix spatially. Since structures such as Diagonal Plus Low-Rank (DPLR) are not closed under multiplication, associative scans over them incur an $\mathcal{O}(d_h^3)$ composition cost, requiring complex approximations (Siems et al., 2025; Walker et al., 2026). These models are also typically optimized for key-value associative memory: effective for exact retrieval, but limited for tasks requiring non-Markovian deep histories, such as extracting

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.

continuous multilinear interactions via path signatures for high-frequency financial modeling or physics simulations.

To process such interactions efficiently, tensor networks such as Tensor-Train (TT) decompositions (Oseledets, 2011) have been applied to recurrent models (Yang et al., 2017; Onu et al., 2020; Khrulkov et al., 2019) for parameter compression and interpretability (Hamreras et al., 2025). Rough path theory offers the path signature to capture geometric time-series features (Chevyrev & Kormilitzin, 2025; Kidger et al., 2019), though integrating high-order signatures into deep models requires careful scaling due to their factorial decay at higher orders.

Building on this, we propose the Tensorized Decoupled Selective CDE (TD-CDE), which structurally decouples time evolution from channel interactions. Restricting time updates to a rank-wise diagonal decay enables efficient chunked parallel scans without the $\mathcal{O}(d_h^3)$ cost, while channel mixing is delegated to a graded, depth-wise bilinear tensor coupling that uses TT cores to prevent dimensional explosion during high-order signature extraction.

This decoupling also changes what discretization means. In a selective CDE, holding the input-dependent gates at the step endpoint—as in the standard zero-order-hold view of selective SSMs—caps the global numerical order at one, regardless of how the diagonal decay is integrated. We separate the step approximation into five error axes and state an *order contract* whose prediction is the minimum over these axes, using exponential-integrator weights for within-level coupling (Hochbruck & Ostermann, 2010) and barycentric quadrature on the Chen simplices (Morrill et al., 2021; Walker et al., 2024). This numerical order is nearly invisible in task accuracy at the training resolution, but governs behavior off the training grid: a fully corrected scheme preserves its predictions under subsampling, whereas baseline endpoint schemes suffer severe performance drops.

Beyond post-hoc analyses of selective SSMs (Rezaei Jafari et al., 2024; Ali et al., 2025), we introduce an interpretable mode that preserves the algebraic structure of signatures: dimension-wise scaling calibrated before training replaces opaque inter-layer mixing such as LayerNorm, and the selection gate is driven directly from raw inputs. Because each level is multilinear in the cores, masking yields exact additive decompositions of the prediction by level, word, and space-time stratum in a fixed number of forward passes, without surrogate samplers such as SHAP (Lundberg & Lee, 2017).

Our contributions are fourfold:

1. **Graded TT-CDE architecture:** We propose TD-CDE, a tensorized bilinear selective CDE decoupling time updates and channel interactions. Diagonal updates maintain parallel scans, while depth-wise TT coupling extracts high-order signature spaces efficiently.
2. **Order contract for selective discretization:** We identify five error axes of a selective CDE step—including the evaluation points of the gates—and give a predictive order rule, verified on held-out order measurements.
3. **Order buys invariance, not accuracy:** We show empirically that numerical discretization order is largely undetectable in native-grid task accuracy, yet it is the primary factor determining whether predictions and attributions survive a change of sampling resolution.
4. **Exact conditional attribution:** Level-wise scaling and raw-input gating make level and word decompositions exact identities of the trained model at a deterministic cost, enabling the quantification of explanation stability across observation grids.

2 Related Work

Selective State-Space Models and Linear CDEs Linear-time sequence models evolved from orthogonal memory compression (Gu et al., 2020; 2021; Smith et al., 2022) to selective mechanisms (Gu & Dao, 2023) and structural duality (Dao & Gu, 2024). Muca Cirone et al. (2024) established a continuous-time linear-CDE perspective on selective SSMs, relating gating and state mixing to path functionals. TD-CDE imposes a graded structure in which within-level transitions are diagonal while adjacent-level TT couplings carry rank mixing; its core distinction lies in this algebraic organization of interactions and the resulting strict decomposition, rather than in pursuing dense-equivalent expressivity or parallel evaluation alone.

Structured State-Transition Matrices To optimize the trade-off between expressivity and complexity, models such as DeltaNet (Schlag et al., 2021; Yang et al., 2024b;a; Siems et al., 2025) extend associative memory using generalized Householder matrices, and SLiCEs (Walker et al., 2026) shows that specific structured matrices achieve dense-equivalent expressivity. However, packing complex structures into a single time step creates algebraic composition issues during associative scans; architectures that fully decouple time updates from channel interactions remain an underexplored design space.

Tensor Decompositions and Path Signatures Tensor factorization has long been used to reduce computational load in sequence modeling (Oseledets, 2011; Novikov et al., 2015; Dao et al., 2022). Path signatures provide ordered iterated-integral features for sequential learning (Chevyrev & Kormilitzin, 2025; Kidger et al., 2019), and Random Fourier Signature Features approximate signature kernels using random features and tensor projections (Toth et al., 2025), avoiding explicit high-dimensional spaces. TD-CDE goes beyond low-rank signature compression: its learned TT cores dynamically organize selective, decaying dynamics while preserving exact core-word decompositions without input-dependent normalization of the states.

Discretization of Continuous-Time Sequence Models Selective SSMs are typically discretized by holding the input-dependent coefficients constant over a step: S4 uses a bilinear map (Gu et al., 2021), S5 and Mamba use a zero-order hold (Smith et al., 2022; Gu & Dao, 2023), and Mamba-3 adopts an exponential-trapezoidal rule (Lahoti et al., 2026). Treating gates as step-wise constants is equivalent to a right-endpoint quadrature of an input-dependent integrand; the corresponding numerical-analysis tools are exponential integrators (Hochbruck & Ostermann, 2010) and log-ODE methods for iterated integrals (Morrill et al., 2021; Walker et al., 2024). We adopt exact ψ_1 weights for adjacent-level coupling and barycentric quadrature on Chen simplices, treating the resulting numerical order as a design variable rather than an implementation detail.

Exact Decomposition and Interpretability Post-hoc analyses of SSMs examine how internal computations support predictions (Rezaei Jafari et al., 2024; Ali et al., 2025). TD-CDE instead leverages its multilinear core structure to obtain additive decompositions directly through masked forward passes: with the input, gates, and calibration constants held fixed, these passes isolate levels, words, and sets of letters deterministically, even under selective decay. The resulting attributions describe algebraic computational pathways rather than feature-removal effects, and we distinguish algebraic completeness at a fixed grid from the numerical stability of the same decomposition after resampling.

3 METHOD

Scope. All attribution results concern the single-layer *interpretable* configuration: one TD-CDE layer, identity input projection, no normalization, zero initial state, and calibration constants fixed before training. Multi-layer and latent variants are used only where stated and are outside the attribution claims. The cubic path reconstruction described below is an offline method; the order-contract validation assumes constant decay rate. All contribution statements are conditional on the observed input and the resulting gates being fixed.

3.1 CONTINUOUS-TIME FORMULATION

Let $\tilde{Z}_t \in \mathbb{R}^D$ denote the gated driving path. The state at level $m \in \{0, \dots, M-1\}$ is $H_t^{(m)} \in \mathbb{C}^{H \times R}$ with $H_{t_0}^{(m)} = 0$:

$$dH_t^{(m)} = \left(-\gamma_t^{(m)} \lambda_t^{(m)} + i\Omega\right) \odot H_t^{(m)}\, dt + \gamma_t^{(m)}\, \mathcal{C}^{(m)}[H_t^{(m-1)}, d\tilde{Z}_t], \qquad H^{(-1)} \equiv 1. \quad (1)$$

The TT contractions are

$$\mathcal{C}^{(m)}[u,v]_{h,s} = \sum_{r=1}^{R}\sum_{k=1}^{D} u_{h,r}v_k\mathcal{C}^{(m)}_{h,k,r,s}, \qquad m \geq 1, \tag{2}$$

$$\mathcal{C}^{(0)}[1,v]_{h,s} = \sum_{k=1}^{D} v_k\mathcal{C}^{(0)}_{h,k,s}. \tag{3}$$

There are $H(DR + (M-1)DR^2)$ core parameters. Two structural properties drive the analysis: *gradedness* makes the level coupling strictly lower-triangular, while *core independence* means that gates, decay and rotation depend on the input but not on the TT cores. With zero initialization, $H^{(m)}$ is therefore multilinear of degree one in $\mathcal{C}^{(0)}, \ldots, \mathcal{C}^{(m)}$ for fixed input. Input-dependent gates can still mix all raw channels; the restriction is only that this mixing does not depend on the cores.

3.2 Driving path and gates

The spatial feature z_l is either the raw input or an optional causal filter bank. For time constant τ, the filter is the exact solution of $\dot{e} = (x-e)/\tau$ under the piecewise-linear input interpolant; the full EMA update, optional lag, and Log-ODE preprocessing are given in Appendix C.2. The gated augmented path is

$$d\tilde{Z}_t = \Big[s(x_t) \odot c_{\text{static}} \odot dz_t,\ dt\Big], \qquad s(x_t) = \sigma(W_{\text{sel}}x_t + b_{\text{sel}}). \tag{4}$$

The selection gate acts on raw channels and is shared across filter taps; $c_{\text{static}} > 0$ is calibrated once before training to equalize the mean total variation of spatial channels. Three discrete objects are distinguished: the endpoint increment, node values of the within-step integrand, and zero increments on padded steps. The time channel belongs to the alphabet and can be disabled from the cores by construction.

Clock and decay are

$$\gamma_t^{(m)} = \sigma(W^{(m)}_{\text{clock}}u_t + b^{(m)}_{\text{clock}}), \qquad \lambda_t^{(m)} = \text{softplus}(W^{(m)}_{\text{decay}}u_t + b^{(m)}_{\text{decay}}), \tag{5}$$

with u_t equal to the spatial input, time, or nothing. Time-only and constant modes are resolution-preserving; the whole model has this property only when the selection gate is frozen or itself order preserving. Rotation depends on dt only and is used only in synthetic probes ($\Omega = 0$ in real-data experiments); its initialization and exact real-valued implementation are in Appendix C.1.

3.3 Discretization and order contract

For $\Delta t_l = \text{ReLU}(t_l - t_{l-1})$, the step exponent is

$$Z_l^{(m)} = \text{clip}_{[0,x_{\max}]}\Big(\Delta t_l\, Q_l[\gamma^{(m)}\lambda^{(m)}]\Big) - i\Omega\Delta t_l, \qquad x_{\max} = 690/64. \tag{6}$$

The homogeneous factor $e^{-Z_l^{(m)}}$ is exact. Hence all numerical error comes from the quadrature of the selective coefficient and from the injection term.

Endpoint scheme (A0).

$$H_l^{(m)} = e^{-Z_l^{(m)}} \odot H_{l-1}^{(m)} + \gamma_l^{(m)}\mathcal{C}^{(m)}[H_{l-1}^{(m-1)}, \Delta\tilde{Z}_l]. \tag{7}$$

A0 omits within-step simplex terms, so for a single increment it produces no higher-order self-interaction and has global order one.

Corrected scheme (A7h). A7h restores the Chen cascade

$$b_l^{(m)} = \sum_{j=1}^{m+1}\frac{1}{j!}T^{(m)}_{j,l},$$

Table 1: Five error axes of the selective step. The level-one order is the minimum over these axes; levels ≥ 2 inherit a common upper bound.

Axis	error source	choice	order
A	j=1 weight	barycentre / exact ψ_1	2 / —
B	quadrature of $\gamma\lambda$	right / trapezoid / Simpson	1 / 2 / 4
C	within-step $\gamma\lambda$ variation	off / first-order / GL	2 / 3–4 / 4
D	gate evaluation points	endpoint / averaged / nodes	1 / 2 / —
E	path reconstruction	linear / cubic	2 / 4

where $T^{(m)}_{j,l}$ approximates the ordered j-fold within-step term. The leading $j = 1$ term uses the exact exponential-integrator kernel

$$\psi_1(u, w) = \frac{e^{-w} - e^{-u}}{u - w}, \tag{8}$$

with Gauss–Legendre quadrature replacing the closed form when the gates vary within the step. Higher-j terms use barycentric decay, node-evaluated gates, fractional increments, and the cubic reconstruction when enabled. The full cascade, interpolation, tilt correction, and stability analysis are in Appendix D.

The *order contract* is the minimum of the five axes, with two structural dependencies: cubic reconstruction helps only together with barycentric increments (D and E), while tilt correction requires a non-endpoint decay rule (B and C). The barycentric level-2 decay approximation gives a structural upper bound of order three. A0 therefore has contract order one, whereas A7h has level-one order four and level-two order three. The contract was derived on seven calibration configurations and then tested blindly on nine additional configurations, three gate regimes, and depth three: all 144 held-out and all 56 calibration measurements match within ± 0.45. The validation holds λ constant; input-dependent decay is outside its certified range. On smooth synthetic paths, the measured orders are 0.89 (A0) and 2.97 (A7h).

Cost and causality. A0 uses M core contractions per step, while the Chen cascade uses $M(M + 1)(M + 2)/6$ contractions; with input-dependent gates the $j = 1$ term is repeated at n_q quadrature nodes. All gates and nodes depend only on the input, so every scheme remains multilinear in the cores. A0, the Chen cascade, and linear reconstruction are causal; cubic reconstruction uses future samples and is therefore offline-only. Further stencil and error details are in Appendix D.2.

3.4 Parallel scan

Each level is the affine recurrence $H_l = e^{-Z_l} \odot H_{l-1} + b_l$. We split the sequence into chunks of 64 steps, perform the magnitude scan in log space in FP64, and propagate only the $L/64$ chunk boundaries sequentially, giving sequential depth $O(ML/64)$. A global phase alignment reduces complex rotation to a real non-expansive scan; the exponent budget and complete derivation are in Appendix E.

3.5 Level calibration and read-out

Because higher levels have smaller magnitude, each level is multiplied by a constant s_m calibrated once on the training set to unit RMS. Calibration equalizes average block scales but does not whiten within- or cross-level correlations. It is fixed after discretization and before training, and held fixed across resolutions.

The read-out state is

$$Y_T = \Big[\bigoplus_{m=0}^{M-1} \mathrm{vec}(\tilde{H}^{(m)}_T); z_T\Big], \qquad \hat{y} = W_{\mathrm{lin}} Y_T + \sum_{i,k} \delta_i \Theta_{\cdot ik} (P Y_T)_k + W_{\mathrm{disp}} \delta + b, \quad \delta = x_T - x_0. \tag{9}$$

Thus $\hat{y} = \langle \theta(x), Y_T \rangle + \beta(x)$ with $\theta(x) = \theta^{\mathrm{lin}} + \theta^{\mathrm{bil}}(x)$. In the interpretable configuration there is no normalization, convolution, or output gate; variance is controlled by c_{static} and the fixed level scales.

3.6 EXACT CONDITIONAL ATTRIBUTION

All decompositions below are algebraic identities of the implemented discrete model, up to floating-point error. They are computed by masking TT cores while holding the input, gates, and calibration constants fixed, so they describe conditional computational pathways rather than input-removal effects.

Level m contributes

$$\hat{y} = \sum_{m=0}^{M-1} \left(\langle \theta_m^{\mathrm{lin}}, \tilde{H}_T^{(m)} \rangle + \langle \theta_m^{\mathrm{bil}}(x), \tilde{H}_T^{(m)} \rangle \right) + \langle \theta_z(x), z_T \rangle + \beta(x). \tag{10}$$

Masking each core to one letter isolates a word contribution $c_I(x)$, and its covariance share is $\mathrm{Cov}(c_I, \hat{y})/\mathrm{Var}(\hat{y})$. A word keeps the full-input gates and decays fixed; it is not the effect of deleting a channel. For the experiments we additionally group words into space/time strata $p \in \{\mathtt{S}, \mathtt{T}\}^M$. All 2^M stratum shares are obtained in exactly 2^M forward passes and sum to the level share.

To resolve dominant words without enumerating all D^M words, we retain a set of strata $\mathcal{P}$. The deterministic cost is

$$2^M + \sum_{p \in \mathcal{P}} (D-1)^{\#\mathrm{spatial}(p)},$$

and the retained aggregate share is omitted by at most the sum of absolute shares of the discarded strata. For $D = 129, M = 3$, keeping strata with at most two spatial positions costs $49{,}545$ passes rather than $2{,}146{,}689$. Marginal tables and functional ANOVA are not used in the main experiments and are given in Appendix G. The full stratified-search analysis, including the loose branch-and-bound bound, is in Appendix G.4.

The interpretation of a level is conditional. Without the time letter, the linear part of level m has raw spatial degree $m + 1$ only when the gates and decay are order preserving; the bilinear read-out can add one degree. A space/time stratum is a decay-weighted temporal kernel in the trained model, not a literal iterated integral unless decay and gating are removed.

3.7 ORDER PROFILER

Independently of TD-CDE, we estimate whether a dataset benefits from temporal order. For truncation level K, a regularized linear classifier is fit on the time-augmented signature and again after jointly permuting spatial increments within each series while keeping time increments fixed. Endpoint displacement is preserved, whereas Lévy area and higher order-order-sensitive terms change. The accuracy difference therefore serves as a ranking statistic for order-dependent structure, not a certificate; finite-sample fitting and regularization also contribute. The full permutation protocol is in Appendix H.

4 THEORETICAL ANALYSIS

We analyze the post-gate path as a fixed driving path Y, following the path-transformation view of selective CDEs (Muca Cirone et al., 2024). Under bounded variation, a bounded true vector field, and polynomial singular-value decay $\sigma_j \lesssim j^{-\alpha}$ with $\alpha > 1/2$, TT-SVD truncation at rank R gives

$$\sup_{t \in [0,1]} \| Z_t^* - \widehat{Z}_t^{(R)} \|_2 \leq C_{\mathrm{tot}} R^{-\alpha+1/2},$$

with C_{tot} independent of R; the complete assumptions and proof are in Appendix I.1.

4.1 CP WIDTH VERSUS TT RANK

For interaction order K, identify a pure-signature functional with the coefficient tensor

$$F_{\mathcal{W}}(\omega) = \left\langle \mathcal{W}, \mathrm{Sig}(\omega)_{0,1}^{(K)} \right\rangle. \tag{11}$$

Channel-wise independent diagonal stacks realize CP-format coefficient tensors, while a single TD-CDE module with depth K and bond dimension R realizes TT-format tensors. This distinction yields the following structural separation.

Table 2: UEA accuracy (%). "Net" is the order profiler's net signal. "Baseline" is the better of SLiCEs and Mamba under our protocol. "Gap" is TD-CDE minus Baseline.

Dataset	Net	TD-CDE	Params	Baseline	Gap	Published
SelfRegulationSCP1	0.263	**91.1**	2,120	85.6	+5.5	90
EigenWorms	0.168	**88.0**	6,797	79.0	+9.0	88
JapaneseVowels	0.089	93.1	11,449	96.9	−3.8	–
Epilepsy	0.077	88.9	1,552	92.9	−4.0	–
SelfRegulationSCP2	0.037	47.4	2,366	48.8	−1.4	54
EthanolConcentration	0.019	24.7	2,338	22.5	+2.2	34
MotorImagery	0.010	47.6	19,694	48.3	−0.7	54
Heartbeat	0.003	73.5	46,876	75.5	−2.0	78
BasicMotions	0.000	94.4	6,050	100.0	−5.6	–

Theorem 4.1 (Spatial separation: CP width versus TT rank)**.** *Let $K \geq 2$, $d \geq 2$, $R \geq 2$, and $q = \min(d, R)$. There exists a coefficient tensor $\mathcal{W}^\star$ of TT rank at most R such that*

$$F_{\mathcal{W}^\star} \in \mathcal{M}_{\mathrm{TT}}(K, R), \qquad F_{\mathcal{W}^\star} \in \mathcal{M}_{\mathrm{Diag}}(H) \Longrightarrow H \geq q^{\lfloor K/2 \rfloor}.$$

Thus exact representation of a generic TT-rank-R interaction can require exponentially many independent diagonal channels, whereas TD-CDE needs $\mathcal{O}(KdR^2)$ core parameters. This is an exact-representation statement; distribution-specific approximation and deep diagonal stacks are different comparison classes. Proof in Appendix I.6.

4.2 Implicit low-rank signature projection

In the pure-integration case, TD-CDE computes an exact linear functional of the order-M signature with a TT-constrained coefficient tensor.

Proposition 4.2 (Implicit low-rank signature projection)**.** *Let $H_t^{(m)} = \int_0^t \sum_i (H_s^{(m-1)} G_i^{(m)})\, d\omega_s^i$ with $H^{(0)} \equiv 1$. Then*

$$H_t^{(M)} = \sum_{I=(i_1,\ldots,i_M)} G_{i_1}^{(1)} \cdots G_{i_M}^{(M)}\ \mathrm{Sig}(\omega)_{0,t}^I.$$

A fixed linear read-out closes the final bond and therefore yields a scalar signature functional whose coefficient tensor has TT rank bounded by the chosen bond dimension. The induction proof is in Appendix I.9.

This result explains why the architecture's depth dimension corresponds to ordered interaction order, while TT rank controls cross-channel width.

5 Experiments

We ask four questions: where the graded model helps, whether the discretization delivers the predicted order, what order buys under resolution transfer, and what the exact decomposition reveals.

5.1 Setup

We use UEA multivariate classification tasks with a 70/15/15 split resampled per seed; hyperparameters are selected on validation data only. The interpretable TD-CDE uses one layer with $M = 3$, $H = 4$, $R = 4$ and keeps the time letter. EigenWorms uses a depth-2 Log-ODE path with interval 700 and $L = 26$. Baselines are SLiCEs (Walker et al., 2026) and Mamba (Gu & Dao, 2023) trained under the same search budget; Mamba receives Δt as an input channel, while $\mathrm{Mamba}_{\mathrm{step}}$ receives no time information. Published SLiCEs numbers use different splits and are therefore indicative only.

5.2 Where the graded model helps

TD-CDE has its largest gains on SelfRegulationSCP1 and EigenWorms, the same two datasets with the largest profiler signal: $+5.5$ and $+9.0$ points with 40 and 13 times fewer parameters than the corresponding baselines. On the other seven datasets the gap ranges from -5.6 to $+2.2$ points.

Table 3: SelfRegulationSCP1, five seeds. Accuracy loss (points) under test-path subsampling with timestamps preserved. The $k = 2$ accuracy loss is 0.0 for all arms and is omitted for space. "Worst" is the maximum over seeds and $k \in \{2, 4, 8\}$. "Flips" counts changed predictions among the corresponding 1,275 seed–stride cells; $D(2)$ is relative logit drift.

Arm	Params	Native	$k = 4$	$k = 8$	$k = 64$	Worst	$D(2)$
A7h	2,120	90.1 ± 4.4	0.0 ± 0.0	-0.2 ± 0.5	2.8 ± 2.7	0.0	2.3×10^{-3}
A7h_{ns}	2,108	89.2 ± 5.6	-0.2 ± 0.5	0.0 ± 0.8	0.2 ± 2.6	1.2	2.0×10^{-3}
A0	2,120	90.8 ± 4.0	5.4 ± 9.7	7.8 ± 10.9	2.6 ± 3.9	24.7	1.3×10^{-1}
A0_{ns}	2,108	88.7 ± 7.6	1.9 ± 7.5	4.0 ± 7.9	-0.2 ± 4.4	17.6	1.5×10^{-1}
$\text{Mamba}_{\text{step}}$	68,258	80.7 ± 2.4	17.2 ± 10.3	28.7 ± 8.8	28.9 ± 5.2	37.6	4.7×10^{-1}
Mamba	68,322	81.9 ± 3.5	23.3 ± 8.6	37.2 ± 13.9	44.7 ± 15.0	60.0	5.2×10^{-1}

Flips: A7h 1, A0 90, Mamba 433.

The alignment between the two largest profiler scores and the two largest gains is suggestive but exploratory because the ranking criterion was not fixed in advance. TD-CDE is therefore not a general-purpose replacement for diagonal SSMs; its empirical advantage is concentrated on datasets with strong order-dependent structure. Results in this table use A0, and A7h matches A0 at the training resolution.

5.3 Does the discretisation deliver its order?

Predictions of the order contract were logged before measurement. All 144 held-out measurements and all 56 calibration measurements agree with the contract within ± 0.45, including depth-three level-3 measurements. On smooth paths, A0 and A7h attain orders 0.89 and 2.97; A7h costs 5.2 times the runtime of A0 for a 9.3×10^4 smaller error, while adding the cubic reconstruction to an otherwise second-order scheme costs 2.7% for a 145-fold error reduction. The detailed synthetic table is in Appendix K.1.

Native-grid accuracy does not reveal these numerical differences: the four combinations of axes D and E all give task R^2 between 0.99993 and 0.99996, although the two corrections together reduce the integration error by a factor of 359. A live input-dependent gate supplies one effective Lie level but not two: it closes both the level-2 Lévy-area target and a level-3 Lie target when the internal depth is raised by one, while remaining rank restricted. The full effective-depth table is in Appendix J.

5.4 What order buys on real data: resolution invariance

At the training resolution, A7h minus A0 is -0.7 points (95% CI $[-2.0, +0.6]$, five seeds). The measured step exponent is tiny (median 7×10^{-4}, 99th percentile 8.9×10^{-3}, no clipping), so the difference between first- and third-order numerical error is below the native accuracy quantum. The constant-decay contract is used here only as a predictor of this null because real-data decay is input dependent.

Off the training grid, A7h never loses accuracy for $k \in \{2, 4, 8\}$, whereas A0 loses up to 24.7 points and Mamba up to 60.0. A7h changes one of 1,275 predictions, compared with 90 for A0 and 433 for Mamba; the exact two-sided sign tests give $p = 3.9 \times 10^{-3}$ versus A0 and $p = 6.1 \times 10^{-5}$ versus Mamba. At $k = 64$, only 14 of 896 samples remain: A7h retains 87.3% accuracy, compared with 37.2% for Mamba and 62.4% for an endpoint-only ridge probe. A0 is not uniformly worse; its loss is bimodal across seeds.

Mamba's explicit Δt channel worsens under subsampling because the feature itself is multiplied by the stride, whereas TD-CDE absorbs Δt inside the transition operator. The invariance is also not an endpoint shortcut: the endpoint-only TD-CDE reaches 63.3%, matching the 62.4% ridge ceiling, permutation of increments reduces it to 56.5%, and removing the shortcut leaves A7h invariant while A0 remains grid-sensitive. Under extreme coarsening of EigenWorms to two endpoints, Chen-cascade arms retain 82.5–85.0% while A0 drops to 43.5%, consistent with the fact that A0 annihilates all higher levels for a single zero-initialized increment. The transfer figure and the full controls are in Appendix K.2.

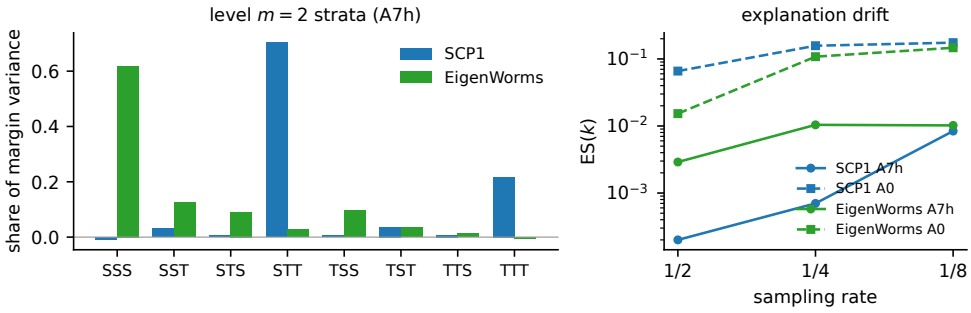


Figure 1: Exact level-2 space/time attribution and explanation drift under subsampling. Left: A7h stratum shares. Right: explanation drift $\mathrm{ES}(k)$ on a log scale.

5.5 What the exact decomposition reveals

All decompositions reconstruct the prediction to numerical precision and sum to one within 10^{-4}. Level $m = 2$ carries 66–91% of the margin variance on the two tasks, while endpoint terms contribute 5.6–8.8% on SCP1 and essentially nothing on EigenWorms. Thus the level most affected by low-order within-step integration is also the dominant predictive level.

The dominant structure differs across tasks. SelfRegulationSCP1 is dominated by the `STT` stratum, a spatial increment followed by two time letters, corresponding without decay to the kernel $\int_0^T \frac{1}{2}(T-u)^2 dx(u)$. EigenWorms is dominated by the purely spatial `SSS` stratum. In two of three seeds, more than 98% of its mass lies in words formed from three Lévy-area letters, which become degree six in the raw path after Log-ODE preprocessing; the remaining seed instead relies on increment letters. Detailed per-seed and channel-level analyses are in Appendix K.3.

Order stabilises the explanation, not its content. Under halved sampling, A7h's stratum shares move by $\mathrm{ES} = 0.0002$, 330 times less than A0; at $k = 8$ A0 moves roughly one sixth of its attribution mass and reduces the rank correlation to 0.614, versus 0.986 for A7h. Gate activations at matched timestamps are identical across the arms ($r = 1.000$), so the drift arises in the accumulation of iterated integrals rather than in the gates. The content learned by the model is therefore distinct from the numerical stability with which that content is expressed across grids.

6 Conclusion

We propose TD-CDE to resolve the trade-off between computational efficiency and cross-channel expressivity in continuous-time sequence modeling. By structurally decoupling diagonal time updates from spatial mixing via Tensor Train decomposition, TD-CDE efficiently extracts high-order path signatures without the $\mathcal{O}(d_h^3)$ composition cost of dense CDEs. It theoretically guarantees superior spatio-temporal expressivity beyond standard diagonal SSMs, and empirically outperforms state-of-the-art models like Mamba on high-order, path-dependent tasks with significantly fewer parameters. Crucially, we demonstrated that a high-order discretization scheme is essential for robust resolution invariance, allowing both the model's predictions and our exact, additive attributions to remain stable even under severe sequence subsampling. While future work will explore integrating non-diagonal transitions to enhance long-term associative memory, TD-CDE establishes a transparent, data-efficient foundational architecture for modeling complex physical systems.

References

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)*, pp. 1516–1534, 2025.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. In *Signature Methods in Finance: An Introduction with Computational Applications*, pp. 3–64. Springer, 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Tri Dao, Beidi Chen, Nimit S Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, and Christopher Ré. Monarch: Expressive structured matrices for efficient and accurate training. In *International Conference on Machine Learning*, pp. 4690–4721. PMLR, 2022.

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487, 2020.

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

Safa Hamreras, Sukhbinder Singh, and Román Orús. Tensorization is a powerful but underexplored tool for compression and interpretability of neural networks. *arXiv preprint arXiv:2505.20132*, 2025.

Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. *arXiv preprint arXiv:1711.00811*, 2017.

Valentin Khrulkov, Oleksii Hrinchuk, and Ivan Oseledets. Generalized tensor models for recurrent neural networks. *arXiv preprint arXiv:1901.10801*, 2019.

Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. *Advances in neural information processing systems*, 32, 2019.

Aakash Sunil Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, volume 2026, pp. 86173–86202, 2026.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *International Conference on Machine Learning*, pp. 7829–7838. PMLR, 2021.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. *Advances in Neural Information Processing Systems*, 37:127226–127272, 2024.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015.

Charles C Onu, Jacob E Miller, and Doina Precup. A fully tensorized recurrent neural network. *arXiv preprint arXiv:2010.04196*, 2020.

Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5): 2295–2317, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, et al. Rwkv-7" goose" with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*, 2025.

Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. Mambalrp: Explaining selective state space sequence models. *Advances in neural information processing systems*, 37:118540–118570, 2024.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pp. 9355–9366. PMLR, 2021.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. Deltaproduct: Improving state-tracking in linear rnns via householder products. *arXiv preprint arXiv:2502.10297*, 2025.

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

Csaba Toth, Harald Oberhauser, and Zoltan Szabo. Random fourier signature features. *SIAM Journal on Mathematics of Data Science*, 7(1):329–354, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Benjamin Walker, Andrew D McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The lie brackets make a difference. *arXiv preprint arXiv:2402.18512*, 2024.

Benjamin Walker, Lingyi Yang, Nicola Muca Cirone, Cristopher Salvi, and Terry Lyons. Structured linear cdes: Maximally expressive and parallel-in-time sequence models. *Advances in Neural Information Processing Systems*, 38:170358–170396, 2026.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37: 115491–115522, 2024b.

Yinchong Yang, Denis Krompass, and Volker Tresp. Tensor-train recurrent neural networks for video classification. In *International conference on machine learning*, pp. 3891–3900. PMLR, 2017.

A Background: Linear CDEs, Signatures, and Tensor Trains

The following material collects the background definitions used by the main text. It is included for completeness so that the main paper can introduce only the notation needed by the TD-CDE construction.

B PRELIMINARIES

B.1 LINEAR CONTROLLED DIFFERENTIAL EQUATIONS (LINEAR CDES)

Linear CDEs provide a framework for describing hidden state dynamics driven by continuous-time input paths, serving as a continuous-time foundation for modern selective state-space models (SSMs). Affine controlled vector fields can be represented in this homogeneous form by augmenting the state with a constant coordinate and the path with time, $X_t = (t, x_t)$; we adopt this convention throughout. The time evolution of a hidden state $h_t \in \mathbb{R}^{d_h}$, driven by a continuous path of bounded variation $X : [0, T] \to \mathbb{R}^{d_x}$, is formulated as follows:

$$h_t = h_0 + \int_0^t \sum_{i=1}^{d_x} A_i h_s \, dX_s^i \tag{12}$$

where $A_i \in \mathbb{R}^{d_h \times d_h}$ is the linear vector field corresponding to the i-th input channel. In practical settings with discrete observations $0 = t_0 < t_1 < \cdots < t_N = T$, the state update is evaluated over small intervals $[t_j, t_{j+1}]$. If the transition matrices A_i are diagonal or mutually commutative, the state update is given exactly by $h_{t_{j+1}} = \exp\left(\sum_i A_i (X^i_{t_{j+1}} - X^i_{t_j})\right) h_{t_j}$. In the diagonal case, this enables efficient parallel evaluation of the entire sequence using an associative scan algorithm. Dense matrices for A_i capture complex interactions among channels, but require $\mathcal{O}(d_x d_h^2)$ parameters and increase the state composition cost in scan operations to $\mathcal{O}(d_h^3)$, creating a bottleneck for practical parallelization.

B.2 PATH SIGNATURES AND FACTORIAL DECAY

The expressivity of Linear CDEs is related to their ability to extract geometric features of the input path. By applying Picard iteration to the integral equation in Eq. (12), the solution of the CDE is expanded as a series involving the iterated integrals of the input path. In rough path theory, this infinite series of iterated integrals is called the path signature. For a multi-index $I = (i_1, \ldots, i_k) \in \{1, \ldots, d_x\}^k$, the k-th order iterated integral of path X over the interval $[s, t]$ is defined as:

$$S(X)^I_{s,t} = \int_{s<u_1<\cdots<u_k<t} dX^{i_1}_{u_1} \cdots dX^{i_k}_{u_k} \tag{13}$$

On compact sets of time-augmented paths of bounded variation with a common initial point, equipped with the 1-variation topology, linear functionals of the full signature are dense in the space of continuous real-valued functions. Directly optimizing this in deep learning is nonetheless challenging. Let V_1 be the 1-variation (length) of the path over $[s, t]$. Under a compatible tensor norm, the k-th order signature tensor has an upper bound of $\|S(X)^{(k)}_{s,t}\| \leq \frac{V_1^k}{k!}$. While this factorial bound guarantees the convergence of the series, the upper bound peaks near $k \approx V_1$ and then decays rapidly, causing scale differences across levels rather than uniform numerical underflow. This motivates a level-wise scaling mechanism that compensates for the $1/k!$ decay, combined with a normalization that controls V_1, to stabilize the learning of high-order temporal interactions.

B.3 TENSOR TRAIN (TT) DECOMPOSITION

To parameterize high-order signature interactions while overcoming the parameter and composition cost issues of dense vector fields in Linear CDEs, we introduce the concept of Tensor Train (TT) decomposition. Any high-dimensional tensor $\mathcal{W} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ admits an exact TT decomposition, and truncating the ranks to a maximum rank R reduces the parameter count from $\mathcal{O}(n^D)$ to $\mathcal{O}(DnR^2)$, generally at the cost of approximation. An element of $\mathcal{W}$ at index $(i_1, \ldots, i_D)$ is expressed as the following matrix product:

$$\mathcal{W}_{i_1,\ldots,i_D} = G^{(1)}_{i_1} G^{(2)}_{i_2} \cdots G^{(D)}_{i_D} \tag{14}$$

where each $G^{(k)} \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$ is a TT core, and r_k denotes the TT rank. Standard boundary conditions use $r_0 = r_D = 1$, but here we leave the final rank index open to act as the state dimension, closing it later with a linear readout. The chain topology of the TT format mirrors the temporal ordering of an iterated integral: contracting a TT with the level-k signature can be evaluated by a

left-to-right recursion of width R. We integrate the TT format into the CDE as a structural prior for computing high-order signatures, rather than using it only for static parameter compression, yielding efficient continuous-time dynamics.

C Continuous-Time Implementation Details

C.1 Complex rotation

Complex rotation (optional). Rotation is used only in the synthetic probes; all real-data experiments use $\Omega = 0$. With $\Omega = 0$ the state is real. With $\Omega \neq 0$ the level-zero kernel becomes a damped oscillation $e^{-(\gamma\lambda)(T-t)}e^{i\Omega(T-t)}$ rather than a pure decay, which gives the model Fourier–Laplace-type kernels directly; representing an oscillatory kernel by combinations of real decaying exponentials is possible but ill-conditioned. Half of the rank channels are initialised with $\Omega = 0$ and the other half on a log-spaced grid from 2π to $L_{\mathrm{ref}}/4$ radians per unit of normalised time; all entries are learned and unconstrained. The cores remain real: a complex state is contracted by applying $\mathcal{C}^{(m)}$ to its real and imaginary parts separately. The rotation depends on time only, so it is order preserving (Definition C.3).

C.2 Driving path: filter bank and Log-ODE preprocessing

The raw input is $x_l \in \mathbb{R}^d$ at increasing times t_l. The spatial features $z_l \in \mathbb{R}^{d_s}$ are either the raw input ($d_s = d$) or the output of an optional causal filter bank that concatenates the identity with exponential moving averages. For time constant τ, with $e_0 = x_0$, $h_l = \Delta t_l/\tau$, $a_l = e^{-h_l}$ and $\varphi_1(h) = (1 - e^{-h})/h$,

$$e_l = a_l e_{l-1} + \big(\varphi_1(h_l) - a_l\big)x_{l-1} + \big(1 - \varphi_1(h_l)\big)x_l, \tag{15}$$

which is the exact solution of $\dot{e} = (x - e)/\tau$ for the piecewise-linear interpolant of x; its coefficients are non-negative and sum to one for $\Delta t_l \geq 0$. A one-step lag can also be appended; since it is not a fixed physical delay, its meaning changes with resolution, and it is disabled in all resolution experiments.

For very long series the raw path may be replaced by the cumulative depth-N log-signature over coarse intervals (Log-ODE preprocessing), implemented in the data pipeline. Its area coordinates are quadratic in the raw path, so one core application may then raise the raw spatial degree by more than one.

The gated driving path is defined as

$$d\tilde{Z}_t = \Big[\, s(x_t) \odot c_{\mathrm{static}} \odot dz_t \;,\; dt \,\Big], \qquad s(x_t) = \sigma\big(W_{\mathrm{sel}}x_t + b_{\mathrm{sel}}\big) \in (0,1)^d, \tag{16}$$

where the gate acts on the d raw channels and is replicated across the filters, and $c_{\mathrm{static}} \in \mathbb{R}^{d_s}_{>0}$ is a fixed constant setting the mean total variation of every spatial channel, measured before gating, to a common value on the training set. The selection gate is initialised at $W_{\mathrm{sel}} = 0$, $b_{\mathrm{sel}} = 4$. Three discrete versions of (16) are used and must be distinguished: the endpoint increment $\Delta\tilde{Z}_l = [s(x_l) \odot c_{\mathrm{static}} \odot (z_l - z_{l-1}), \Delta t_l]$; the node values $v_l(\sigma_q)$ of the integrand on the normalised step $\sigma \in [0,1]$ (Section 3.3); and zero on padded steps.

Because the time channel belongs to the alphabet, the cores also contract the time letter. Optionally the time slice of every core is fixed to zero from construction; this is condition (i) of Proposition G.5.

C.3 Clock, decay, and resolution-preserving modes

The main text records the gate parameterizations and their role in resolution transfer. Here we collect the initialization and mode-specific details: $W_{\mathrm{clock}} = W_{\mathrm{decay}} = 0$ at initialization, so every clock and decay head starts from a time-invariant gate. In time-only and constant modes the gates are unchanged by spatial scaling of the path. The selection gate is raw-input driven and therefore must itself be frozen or order preserving when a resolution-invariant interpretation is claimed. Rotation depends only on dt and is therefore independent of the clock.

D Discretisation Details and the Order Contract

D.1 Full step construction

Let $\Delta t_l = \mathrm{ReLU}(t_l - t_{l-1})$. For each level the step exponent is

$$Z_l^{(m)} = \mathrm{clip}_{[0,\, x_{\max}]}\Big(\Delta t_l\, Q_l\big[\gamma^{(m)}\lambda^{(m)}\big]\Big) - i\,\Omega\,\Delta t_l, \qquad x_{\max} = 690/64, \tag{17}$$

where Q_l is the right-endpoint, trapezoidal or Simpson rule on the step (axis B below). The homogeneous update $H \mapsto e^{-Z_l^{(m)}} \odot H$ is then applied in closed form. Thus the rotation is integrated exactly, while the decay carries the quadrature error of Q_l; the remaining error comes from the injection term. All convergence statements assume that the clip in (17) is inactive; we report its activation rate.

Endpoint scheme (A0). Gates are held at the step endpoint:

$$H_l^{(m)} = e^{-Z_l^{(m)}} \odot H_{l-1}^{(m)} + b_l^{(m)}, \qquad b_l^{(m)} = \gamma_l^{(m)}\, \mathcal{C}^{(m)}\big[H_{l-1}^{(m-1)},\, \Delta\tilde{Z}_l\big]. \tag{18}$$

This combines an exact decay step with an endpoint injection without the φ_1 weight, and is called ZOH in the implementation although it is not the exact zero-order-hold discretisation. It accumulates products of increments at strictly increasing indices (iterated *sums*); for a single step with increment v the level-two term is 0 rather than $\frac{1}{2}v^{\otimes 2}$. Its global order is one.

Chen cascade. Within-step terms are restored by expanding each step over the simplices Δ_j, $j = 1, \dots, m+1$, where j is the number of increments that fall inside step l:

$$b_l^{(m)} = \sum_{j=1}^{m+1} \frac{1}{j!}\, T_{j,l}^{(m)}, \qquad T_{j,l}^{(m)} = E_j^{(m)}\, \gamma^{(m)} \mathcal{C}^{(m)}\Big[\cdots E_j^{(m-j+1)}\, \gamma^{(m-j+1)} \mathcal{C}^{(m-j+1)}\big[E_j^{(m-j)} H_{l-1}^{(m-j)},\, \Delta\tilde{Z}_l\big] \cdots,\, \Delta\tilde{Z}_l\Big], \tag{19}$$

with barycentric decay $E_j^{(k)} = e^{-Z_l^{(k)}/(j+1)}$ and $E_j^{(-1)} H^{(-1)} := 1$. With zero decay, zero rotation, unit clock and a constant increment within the step, (19) reproduces the TT contraction of $\frac{1}{j!}v^{\otimes j}$, i.e. of the signature of the piecewise-linear path. The leading $j = 1$ term is integrated exactly for frozen coefficients:

$$T_{1,l,h,s}^{(m)} = \gamma_{l,h}^{(m)} \sum_{r=1}^{R} \sum_{k=1}^{D} H_{l-1,h,r}^{(m-1)}\, \Delta\tilde{Z}_{l,k}\, \mathcal{C}_{h,k,r,s}^{(m)}\, \psi_1\big(Z_{l,h,s}^{(m)},\, Z_{l,h,r}^{(m-1)}\big), \qquad \psi_1(u,w) = \frac{e^{-w} - e^{-u}}{u - w}, \tag{20}$$

so the weight depends on the pair of input and output rank channels, and $\psi_1(u, 0) = \varphi_1(u)$ at level zero. We evaluate $\psi_1 = e^{-\min(u,w)}\varphi_1(|u - w|)$, a product of two factors bounded by one; for complex arguments we use the equivalent $\sinh$ form, which is bounded because the real part of (17) is clipped.

With input-dependent gates, the $j = 1$ term is instead computed by an n_q-point Gauss–Legendre rule in σ. At each node the clock, the selection gate and the path derivative are evaluated, and the within-step decay is reconstructed from the step exponent and its change across the step (the "tilt"), with the node exponents clipped to $[0, x_{\max}]$. This reconstruction is affine in time, which limits the term to global order four when $\gamma\lambda$ varies within a step; the rule's own error, of degree $2n_q$, is attained only when $\gamma\lambda$ is constant over the step. For $j \geq 2$ the clock is evaluated at the barycentres $i/(j+1)$, and the increments there are obtained by Lagrange interpolation from the node values. The path within a step is either linear or the cubic polynomial through four consecutive samples ("Hermite" in the implementation), whose derivative integrates exactly to the step increment. The cubic reconstruction assumes a uniform grid and equal-length sequences; its higher-order accuracy is not claimed for irregular sampling or padded batches.

Order contract. The error decomposes into the five axes of Table 4. The level-one order is the minimum over the axes, and all levels ≥ 2 share a common bound that cannot exceed the level-one order. Two dependencies are enforced: at levels ≥ 2 the cubic reconstruction helps only together with barycentric increments (axes D and E), and the tilt corrections require a non-endpoint rule Q_l (axes B and C). The barycentric approximation of the decay in (19) leaves a local $O(\Delta^4)$ error at

Table 4: Error axes and their controls (level one; levels ≥ 2 in Appendix D.3). "—" means the axis imposes no algebraic limit. The Gauss–Legendre entry of axis C is set by the affine reconstruction of the within-step decay; the rule itself has degree $2n_q$, and $n_q = 8$ satisfies the implementation's requirement $n_q \geq \lceil e\, x_{\max}/4 \rceil$.

Axis	error source	choice	order
A	weight of the j=1 term	barycentre / exact ψ_1	2 / —
B	quadrature Q_l of $\gamma\lambda$	right / trapezoid / Simpson	1 / 2 / 4
C	within-step variation of $\gamma\lambda$	off / first-order / Gauss–Legendre	2 / 3–4 / 4
D	evaluation points of γ and s	endpoint s / nodes, averaged γ / nodes	1 / 2 / —
E	path reconstruction	linear / cubic (uniform grid)	2 / 4

level two, so order three is the structural upper bound there. The corrections that make the contract predictive were derived on a calibration set of seven configurations and then tested blind on nine further configurations under three gate regimes and at depth three: all 144 held-out measurements (72 configuration, regime and level slots in two measurement modes) agree with the prediction within ± 0.45, as do all 56 calibration measurements (Appendix D.3). The validation holds the decay rate λ constant and varies the clock γ and the selection gate s between constant and input dependent; input-dependent decay is not covered.

We report two schemes. **A0** is the endpoint scheme (18). **A7h** is the Chen cascade in which the $j = 1$ term is computed by the $n_q = 8$ Gauss–Legendre rule in place of the closed-form ψ_1, with Simpson quadrature for Q_l, node-evaluated clock and selection gate, barycentric clock and increments at levels ≥ 2, and cubic reconstruction; its level-one order is therefore four, limited jointly by axes B, C and E. The contract predicts order one for A0 and order three at level two for A7h; on smooth synthetic paths sampled on a uniform grid of up to 2048 steps, with constant decay and input-dependent gates, the measured orders are 0.89 and 2.97. In the real-data experiments the decay is also input dependent, which lies outside the validated regime; there we use A7h for its empirical behaviour rather than for a certified order.

Causality. The endpoint scheme, the Chen cascade and the linear reconstruction use only past and present observations. The cubic reconstruction uses future observations, including additional look-ahead in the boundary stencils; it is suitable for offline classification but not for streaming.

Multilinearity and cost. In every scheme the gates, exponents, nodes and weights depend only on the input. Expanding the source state $H^{(m-j)}$ in (19) recursively, every resulting term contains each of $\mathcal{C}^{(0)}, \ldots, \mathcal{C}^{(m)}$ exactly once, so each level remains multilinear of degree one in the cores for every scheme. Per step, A0 performs M core contractions and the Chen cascade $M(M+1)(M+2)/6$ (ten for $M = 3$), with the $j = 1$ terms repeated at n_q nodes when input-dependent gates are used.

D.2 Causality of reconstruction stencils

The endpoint scheme, the Chen cascade, and linear interpolation use only past and present observations. The cubic reconstruction uses four consecutive samples and therefore uses future observations at the interior and boundary stencils. It is an offline reconstruction and is not suitable for streaming. The higher-order accuracy of the cubic stencil is claimed only on uniform grids with equal-length sequences; irregular sampling and padded boundary segments are excluded from that claim.

D.3 Complete validation of the order contract

The order contract in the main text was derived from seven calibration configurations and then frozen before the blind measurements. We varied the clock and selection-gate modes at fixed decay rate and evaluated both grid and refinement measurements. The complete held-out counts are 54 grid slots, 54 refinement slots, and an additional $18 + 18$ depth-three slots, for 144 measurements in total; all agree with the predicted order within ± 0.45. The 56 calibration measurements also agree.

Table 5: Full synthetic validation and effective-depth experiment. (a) The contract prediction error is measured as $|\Delta p| \leq 0.45$. (b) $1 - R^2$ for the Lévy-area target and a level-3 Lie target.

(a) Order predictor Set	Slots	Grid	Refine
Calibration	28	28/28	28/28
Held-out	54	54/54	54/54
Held-out, depth 3	18	18/18	18/18

(b) Effective depth Route	Lévy area	Lie, level 3
$M = 1$, constant	7.6×10^{-1}	9.8×10^{-1}
$M = 1$, live	2.9×10^{-5}	8.5×10^{-1}
$M = 2$, constant	3.1×10^{-5}	7.6×10^{-1}
$M = 2$, live	6.6×10^{-5}	3.2×10^{-4}
$M = 3$, constant	1.3×10^{-4}	1.6×10^{-4}

D.4 Why axes D and E form a conjunction

The barycentric increment correction and cubic reconstruction are each insufficient to improve the level-two error on their own: relative error remains at factor 1.00 when either is enabled alone, whereas the two together reduce it by a factor of 359. This interaction is why the contract is expressed as a minimum with explicit dependencies rather than as a sum of independent truncation orders.

The $j = 1$ branch is computed by $n_q = 8$ Gauss–Legendre nodes in A7h. The rule has degree $2n_q$, but the within-step decay is reconstructed affinely, which limits the full term to global order four when $\gamma\lambda$ varies inside a step. For constant $\gamma\lambda$, the quadrature rule itself is the limiting accuracy.

E Numerically Stable Chunked Parallel Scan

The main text gives the computational summary. This section gives the exact FP64 exponent budget, phase alignment, chunk recurrence, and reconstruction.

We describe the specific algorithm for the chunked parallel scan used to propagate the discretized TD-CDE dynamics introduced by the discretized TD-CDE recurrence.

Once the intra-step increment $b_{m,l}$ has been assembled (as described in Appendix L), every tensor level of TD-CDE reduces to one and the same first-order linear recurrence:

$$H_l^{(m)} \;=\; e^{-Z_{m,l}} \odot H_{l-1}^{(m)} \;+\; b_{m,l}, \qquad Z_{m,l} \;=\; \xi_{m,l} - i\,\Omega_{m,l}, \qquad H_0^{(m)} = 0, \tag{21}$$

with $\xi_{m,l} \in [0, x_{\max}]^{N_H \times R}$ and $b_{m,l} \in \mathbb{C}^{N_H \times R}$ (which is real-valued when rotation is disabled). This section specifies the kernel that solves (21). It is deliberately *agnostic to how $b_{m,l}$ was produced*: the entire high-order machinery (e.g., Chen-expansion and quadrature) lives inside $b_{m,l}$, so the discretization scheme can be moved along all accuracy axes without modifying the parallel kernel. Note that ξ carries *both* a head index and a rank index; the scan therefore consists of $N_H R$ independent scalar recurrences per level, all executed in parallel. We write N_H for the number of heads to avoid a clash with the hidden state $H^{(m)}$.

Why Chunking, and Where $x_{\max}$ Comes From. A single global log-space scan would require $\exp\big(\sum_{j \leq l} \xi_j\big)$, whose exponent grows like $L\,x_{\max}$ and is unrepresentable for any realistic sequence length L. Confining the log-space accumulation to chunks of length C bounds the exponent by $C\,x_{\max}$. Requiring that this term remains representable in FP64,

$$C\,x_{\max} \;\leq\; \ln(\texttt{DBL_MAX}) \;\approx\; 709.8, \tag{22}$$

is exactly what determines the clipping bound imposed on ξ. We adopt the conservative budget

$$x_{\max} \;=\; \min\big(30,\; 690/C\big), \qquad \text{so } x_{\max} \approx 10.78 \text{ for } C = 64. \tag{23}$$

Chunk size and admissible per-step decay are therefore two sides of one FP64 exponent budget rather than independent hyperparameters: a larger C buys parallelism but pays in dynamic range. The

constant 30 acts as a saturation cap (since $e^{-30} \approx 10^{-13}$ effectively annihilates the state), relevant only for small C.

Indexing and Padding. The sequence is zero-padded from L to $L' = C\lceil L/C \rceil$, giving $K = L'/C$ chunks. The sequence index is remapped as $l \leftrightarrow (k, c)$ with $l = (k-1)C + c$, where $k \in [1, K]$ and $c \in [1, C]$. Padded positions receive $\xi \leftarrow 0$ (i.e., no decay) and $b \leftarrow 0$, so they propagate the state unchanged and can be safely discarded after the scan.

Step 1: Global Phase Alignment. The coefficient in (21) is complex, which precludes a real-valued log-space scan directly. Let $\Omega_{m,l}$ be the phase increment and

$$\Phi_{m,l} \;=\; \sum_{j \le l} \Omega_{m,j} \tag{24}$$

be the *global* accumulated phase, computed as a two-level prefix sum in FP64 (an intra-chunk cumulative sum plus an exclusive cumulative sum of the chunk totals, both fully parallel). Substituting $\hat{H}_l^{(m)} = e^{-i\Phi_{m,l}} H_l^{(m)}$ and $\hat{b}_{m,l} = e^{-i\Phi_{m,l}} b_{m,l}$ into (21) yields:

$$\hat{H}_l^{(m)} \;=\; e^{-\xi_{m,l}} \odot \hat{H}_{l-1}^{(m)} \;+\; \hat{b}_{m,l}, \tag{25}$$

which is a recurrence with a *purely real, non-expansive* coefficient, allowing the log-space scan to apply verbatim. Because Φ is global rather than chunk-local, the boundary state carried across chunks in Step 3 lives in a single common frame and requires no phase-twist correction, decoupling the complex rotation from the magnitude decay at no additional cost. Phases are accumulated in FP64 so that the absolute error of $\cos\Phi$ and $\sin\Phi$ stays near the machine epsilon even when Φ grows linearly with L.

Step 2: Intra-Chunk Scan in Log Space (FP64). Let $\Sigma_{k,c} \in \mathbb{R}^{N_H \times R}$ be the local cumulative log-magnitude and ν_k be the stabilizing shift:

$$\Sigma_{k,c} \;=\; -\sum_{j=1}^{c} \xi_{k,j} \;\; (\le 0), \qquad \nu_k \;=\; \max_{1 \le c \le C} \big(-\Sigma_{k,c}\big) \;=\; -\Sigma_{k,C}, \tag{26}$$

where the closed form holds because $\xi \ge 0$ makes $\Sigma_{k,\cdot}$ monotonically non-increasing. ν_k is *constant within a chunk*: a c-dependent shift would factor directly out of the inner sum, leaving the accumulator unprotected from overflow. Defining

$$p_{k,c} = \exp\big(\mathrm{clip}(-\Sigma_{k,c} - \nu_k, \pm 700)\big) \in (0, 1], \tag{27}$$

$$q_{k,c} = \exp\big(\mathrm{clip}(\Sigma_{k,c} + \nu_k, \pm 700)\big) \in [1, e^{C x_{\max}}], \tag{28}$$

the stabilized partial sums within the chunk are obtained by a single cumulative sum:

$$\tilde{b}_{k,c} \;=\; q_{k,c} \odot \sum_{j=1}^{c} p_{k,j} \odot \hat{b}_{k,j}. \tag{29}$$

The shift cancels identically in the product $p_{k,j} q_{k,c}$, so (29) is algebraically exact; its role is only to bound the summands ($\|p_{k,j} \odot \hat{b}_{k,j}\| \le \|\hat{b}_{k,j}\|$), capping the accumulator at $\mathcal{O}(C \max_j \|b_{k,j}\|)$ and moving the only exponentially large factor into the scalar multiplier $q_{k,c}$. The residual requirement that $q_{k,c}$ itself remains representable is exactly the budget defined in (22). Both (29) and the cumulative sums from Step 1 are evaluated in FP64.

Step 3: Sequential Boundary Pass (K Steps). Only the chunk boundaries are computed sequentially. With $h_0 = 0$:

$$h_k \;=\; \exp(\Sigma_{k,C}) \odot h_{k-1} \;+\; \tilde{b}_{k,C}, \qquad k = 1, \ldots, K, \tag{30}$$

where $h_k = \hat{H}_{k,C}^{(m)}$ is the aligned state at the end of chunk k (equivalently, the state entering chunk $k{+}1$). This loop and everything downstream run in the standard working precision of the model (FP32 by default); $\exp(\Sigma_{k,C})$ flushes to zero once the accumulated decay of a chunk exceeds approximately 88 nats, which is numerically harmless because the carried state is strictly negligible by construction.

Step 4: Parallel Reconstruction and De-Rotation. All L' states are reconstructed in parallel:

$$\hat{H}_{k,c}^{(m)} \;=\; \exp(\Sigma_{k,c}) \odot h_{k-1} \;+\; \tilde{b}_{k,c}, \qquad H_{k,c}^{(m)} \;=\; e^{+i\Phi_{m,k,c}} \odot \hat{H}_{k,c}^{(m)}, \tag{31}$$

after which the padding is truncated. Real and imaginary components are maintained as two separate real-valued tensors throughout this process, so no custom complex-valued kernels are required.

Cost. The scan performs $\mathcal{O}(L\,N_H R)$ work per level with a sequential depth of only $K = \lceil L/C \rceil$ steps, and stores $\mathcal{O}(L\,N_H R)$ activations per level. Because this kernel only consumes (ξ, Ω, b), the choice of discretization scheme alters neither its arithmetic cost nor its parallel depth: the computational cost of a higher-order scheme is paid entirely inside the construction of $b_{m,l}$ (Appendix L).

F Level Calibration and Read-Out Details

F.1 Level calibration and read-out

The magnitude of $H^{(m)}$ decreases rapidly with m. Each level is rescaled by a constant, $\tilde{H}_l^{(m)} = s_m H_l^{(m)}$, calibrated once on the training set so that every level has unit root-mean-square. The calibrated value is the reciprocal of the unscaled level RMS and does not depend on the initial value of s_m. Calibration equalises the average block scales of the read-out Gram matrix; it does not whiten it, and within-level anisotropy and cross-level correlation remain (Proposition F). Calibration is performed after the discretisation is fixed and before training, and the constants s_m and c_{static} are never changed afterwards; in particular they are held fixed across resolutions in convergence and transfer experiments.

The read-out state and prediction are

$$Y_T = \Big[\bigoplus_{m=0}^{M-1} \mathrm{vec}\big(\tilde{H}_T^{(m)}\big) \;;\; z_T \Big], \qquad \hat{y} = W_{\text{lin}} Y_T + \sum_{i,k} \delta_i\, \Theta_{\cdot ik} (P Y_T)_k + W_{\text{disp}}\delta + b, \qquad \delta = x_T - x_0, \tag{32}$$

where complex levels contribute their real and imaginary parts. For each sample and output, $\hat{y} = \langle \theta(x), Y_T \rangle + \beta(x)$ with $\theta(x) = \theta^{\text{lin}} + \theta^{\text{bil}}(x)$, $\theta^{\text{lin}} = W_{\text{lin}}$, $\theta^{\text{bil}}(x) = \delta^\top \Theta P$ and $\beta(x) = W_{\text{disp}}\delta + b$. In the interpretable configuration no normalisation, convolution or output gate is used; variance is controlled only by c_{static} and s_m. Pseudocode for the forward pass is given in Appendix L.

The RMS calibration has a simple fixed-point interpretation: if an unscaled level has training RMS ρ_m, the calibrated value is $s_m = 1/\rho_m$. Changing the initialization of s_m therefore changes only the transient of the calibration procedure and not the final value. Calibration equalizes average block scales in the read-out Gram matrix; it does not remove within-level anisotropy or cross-level correlations.

G Exact Attribution: Marginals, ANOVA, and Search

G.1 Exact attribution by forward passes

Every identity below is an algebraic decomposition of the implemented discrete model, exact up to floating-point error; it does not claim exact integration of the continuous system (1). Each is obtained by forward passes of the trained model with some cores masked, with the input, gates and calibration constants held fixed.

Level decomposition.

$$\hat{y} = \sum_{m=0}^{M-1} \Big(\underbrace{\langle \theta_m^{\text{lin}}, \tilde{H}_T^{(m)} \rangle}_{\text{linear part}} + \underbrace{\langle \theta_m^{\text{bil}}(x), \tilde{H}_T^{(m)} \rangle}_{\text{bilinear part}} \Big) + \langle \theta_z(x), z_T \rangle + \beta(x) \tag{33}$$

holds exactly (Theorem G); separating the linear and bilinear parts costs nothing, since θ is additive. We check at run time that the components reconstruct the prediction, and report the sum of shares with every decomposition.

Word contributions and shares. Masking core k to a single letter i_k yields the contribution $c_I(x)$ of the word $I = (i_1, \ldots, i_{m+1})$, and $\sum_{|I|=m+1} c_I = \langle \theta_m, \tilde{H}_T^{(m)} \rangle$ exactly. A word contribution is the part of the computation in which the cores select those letters, with the gates and decays still determined by the full input; it is not the effect of removing an input channel. We report the covariance share $\mathrm{sh}_I = \mathrm{Cov}(c_I, \hat{y}) / \mathrm{Var}(\hat{y})$, which is additive but signed and may exceed one; shares sum to one only when the z_T and $\beta(x)$ terms are included (Remark G).

Sums over sets of words. Because each level is multilinear, masking a core to a *set* of letters rather than a single letter returns the sum of the corresponding word contributions in one pass. For positions $S \subseteq [M]$ this gives the marginal $\sum_{I:\, I|_S = \alpha} c_I$ (Theorem G.2), and Möbius inversion of such marginals gives a functional ANOVA over word positions (Theorem G.2; an ANOVA over positions, not a Sobol decomposition over input features, Remark G.2). Masking every position to the same letter set A gives the sum $g(A)$ over words whose letters all lie in A; Möbius inversion $\varphi(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} g(B)$ then yields the exact share of words whose set of distinct letters is exactly A. We use this support decomposition, and per-letter splits of a stratum, in Section 5.5; the marginal tables and ANOVA are not used in our experiments.

Strata and stratified search. A stratum $p \in \{\mathtt{S}, \mathtt{T}\}^M$ collects the words whose positions carry a spatial or the time letter as prescribed by p. Its total share Σ_p is one forward pass with each core masked to the spatial letters or to the time letter, so all 2^M strata cost 2^M passes; these are the strata reported in our experiments, and they sum to the level share exactly. To locate individual words when D^M is large, one enumerates only a retained set $\mathcal{P}$. This costs $2^M + \sum_{p \in \mathcal{P}} (D-1)^{\#\mathrm{spatial}(p)}$ passes, and the error in the retained aggregate share is at most $\sum_{p \notin \mathcal{P}} |\Sigma_p|$ (Corollary G.4). The bound does not certify individual words or top-k recall. Retaining the strata with at most two spatial positions for $D = 129$, $M = 3$ costs 49,545 passes against 2,146,689 (43× fewer). The strata are informative only when the time letter is kept; when it is excluded, only the all-spatial stratum is non-zero, which is why our experiments keep it.

We use this stratified search rather than an admissible branch-and-bound procedure: a bound on $|\mathrm{sh}_I|$ built from operator norms and Cauchy–Schwarz returns the exact top-k set, but prunes almost nothing in practice, being loose by roughly 73× per level (Remark G.4), whereas the stratified search has a deterministic cost.

What the levels mean. Level m is a level of the gated driving path. It coincides with raw spatial degree $m + 1$ (homogeneous degree under spatial scaling of the input) only when the time letter is excluded, the clock, decay and selection gate are all order preserving, and the linear part of (33) is used; the bilinear part adds one degree (Proposition G.5). Degree counts letters with repetition, so it is not the number of distinct channels involved; a stratum such as SSS contains same-channel words as well as cross-channel ones. Likewise the reading of a stratum as an explicit iterated integral, e.g. STT as $\int_0^T \frac{1}{2}(T-u)^2\, dx(u)$, holds only without decay and gating; in the trained model the time letters carry the corresponding decay-weighted kernels.

G.2 Functional ANOVA over word positions

For positions $S \subseteq [M]$ and a fixed letter assignment α, masking the corresponding cores to α gives the marginal

$$g_S(\alpha; x) = \sum_{I:\, I|_S = \alpha} c_I(x).$$

Möbius inversion over nested position subsets recovers the associated interaction components. These components form an ANOVA over *word positions*; they are not a Sobol decomposition over the original input channels. The marginal and ANOVA constructions are exact because masking is performed on the multilinear core factors while all gates remain fixed.

G.3 Support decomposition by distinct letters

For a letter set A, define $g(A)$ as the sum of all word contributions whose letters are drawn from A. The exact contribution of words whose set of distinct letters is *exactly* A is

$$\varphi(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} g(B).$$

This is the decomposition used to analyze the purely spatial SSS stratum in EigenWorms. It distinguishes repeated-letter words from genuine multi-channel interactions without enumerating every word.

G.4 Stratified search and the branch-and-bound comparison

All 2^M space/time strata cost 2^M masked passes. To recover dominant individual words, we enumerate only a retained stratum set $\mathcal{P}$ and then search within each retained spatial stratum. The deterministic cost is

$$2^M + \sum_{p \in \mathcal{P}} (D-1)^{\#\mathrm{spatial}(p)}.$$

For $D = 129$ and $M = 3$, keeping strata with at most two spatial positions costs $49{,}545$ passes, compared with $2{,}146{,}689$ for complete enumeration.

We compared this strategy with an admissible operator-norm/Cauchy–Schwarz bound for word contributions. The bound returns the exact top-k set when used as a pruning rule, but its numerical value is sufficiently loose that it prunes little; across levels it is roughly $73\times$ looser than the observed magnitude scale. We therefore use the deterministic stratified search rather than the older branch-and-bound implementation. The aggregate error bound reported in the main text applies to omitted *stratum mass*, not to individual-word recall.

G.5 Conditional meaning of levels and strata

Level m corresponds to $m+1$ letters in the gated path. It equals raw spatial degree $m+1$ only when the time letter is excluded, the clock, decay, and selection gate are order preserving, and the linear read-out component is used. The bilinear read-out contributes one additional spatial degree. Repeated letters count separately, so a stratum such as SSS also contains same-channel words. A notation such as STT corresponds to $\int_0^T \frac{1}{2}(T-u)^2\, dx(u)$ only in the zero-decay, ungated limit; the trained model instead uses decay-weighted kernels.

H Order Profiler Details

Independently of TD-CDE, we probe whether a dataset carries order-dependent structure. For each truncation level K we fit a regularised linear classifier on the truncated signature of the time-augmented path, and repeat the fit after permuting, within each series, the spatial increments jointly across channels (time increments are unchanged). Permutation leaves the endpoint displacement and every polynomial in it invariant but alters the Lévy area and other order-dependent terms (Lemma H), so a fixed predictor that depends only on the endpoint displacement is unaffected (Proposition H). The accuracy difference between the original and permuted fits, assessed over repeated permutations with a held-out split, is therefore evidence that the chosen estimator benefits from temporal order up to truncation K. It is a ranking statistic, not a certificate: finite-sample, optimisation and regularisation effects also contribute.

The profiler is intended as a dataset-level exploratory statistic. Joint permutation of spatial increments preserves the endpoint displacement and polynomials of it, while changing order-sensitive iterated integrals such as Lévy area. Therefore a performance difference between original and permuted paths indicates reliance of the fitted signature probe on order-sensitive coordinates, but does not by itself separate finite-sample estimation effects, optimization, or regularization.

I PROOFS FOR THE THEORETICAL RESULTS

I.1 TRAJECTORY APPROXIMATION BOUND

Theorem I.1 (Trajectory approximation bound)**.** *Assume: (i) the singular values of the mode-wise unfoldings of the target tensor satisfy $\sigma_{\ell,j} \le C_\ell j^{-\alpha}$ with $\alpha > 1/2$; (ii) the driving path has bounded variation $V_1 < \infty$; and (iii) $\|\mathcal{W}^*\|_{\mathrm{op}} \le M_0$. Let $\mathcal{W}_{\mathrm{TT}}^{(R)}$ be its TT-SVD approximation with maximum rank R. Then the corresponding trajectories satisfy*

$$\sup_{t\in[0,1]} \|Z_t^* - \widehat{Z}_t^{(R)}\|_2 \le C_{\mathrm{tot}} R^{-\alpha+1/2},$$

where $C_{\mathrm{tot}} = \|Z_0\|_2 V_1 e^{2MV_1} K_{\mathrm{TT}}$ is independent of R and $M = M_0 + K_{\mathrm{TT}}$.

This appendix provides the formulation and proof of the trajectory approximation theorem regarding the expressivity of the TD-CDE core and the propagation of approximation error to the trajectory presented in the theoretical approximation result.

I.2 SETUP AND ASSUMPTIONS

Consider a linear CDE core driven by a multivariate input path $Y \in C_{1,0}([0,1];\mathbb{R}^{d_{\mathrm{aug}}})$. Let $\mathcal{W}^* \in \mathbb{R}^{d_{\mathrm{aug}}\times\cdots\times d_{\mathrm{aug}}}$ be the order-k tensor defining the linear vector field of the true dense CDE, and let $\mathcal{W}_{TT}^{(R)}$ be its TT-SVD approximation with a maximum TT-rank R. To establish the bounds, we construct both the true and approximated dynamics on a fixed, common state space $\mathbb{R}^N$ with a fixed linear mapping scheme. We formulate the true and approximated dynamics as follows:

$$dZ_t^* = \mathcal{L}_{\mathcal{W}^*}(Z_t^*)dY_t, \qquad d\widehat{Z}_t^{(R)} = \mathcal{L}_{\mathcal{W}_{TT}^{(R)}}(\widehat{Z}_t^{(R)})dY_t$$

To align with the implementation where the recursion starts with a non-zero constant ($h^{(-1)} \equiv 1$), we take the state to be the depth-truncated augmented vector $Z_t = (1, h_t^{(0)}, \ldots, h_t^{(k-1)}) \in \mathbb{R}^N$. The initial condition is $Z_0 = (1, 0, \ldots, 0)$, so $\|Z_0\|_2 = 1$. Here, $\mathcal{L}_{\mathcal{W}}$ is the linear map induced by the tensor $\mathcal{W}$, which is linear in $\mathcal{W}$ (i.e., $\mathcal{L}_{\mathcal{W}_1} - \mathcal{L}_{\mathcal{W}_2} = \mathcal{L}_{\mathcal{W}_1-\mathcal{W}_2}$). We define the operator norm jointly in the state and input arguments,

$$\|\mathcal{W}\|_{\mathrm{op}} := \sup_{\|u\|_2=1,\ \|y\|_2=1} \|\mathcal{L}_{\mathcal{W}}(u)\, y\|_2,$$

so that the bound $\|\mathcal{L}_{\mathcal{W}}(u)dY\|_2 \le \|\mathcal{W}\|_{\mathrm{op}}\|u\|_2\|dY\|_2$ holds, and naturally $\|\mathcal{W}\|_{\mathrm{op}} \le \|\mathcal{W}\|_F$.

For the analysis, we make the following three assumptions.

Assumption I.1 (Target Tensor and Singular Value Decay)**.** For the true tensor $\mathcal{W}^*$ responsible for high-order channel interactions, let $\sigma_{\ell,j}$ be the singular values of the mode-ℓ unfolding matrix $\mathcal{W}_*^{\langle\ell\rangle}$. We assume there exist constants $C_\ell > 0$ and $\alpha > 1/2$ such that the following polynomial decay holds for each $\ell = 1, \ldots, k-1$:

$$\sigma_{\ell,j} \le C_\ell j^{-\alpha}$$

Assumption I.2 (Bounded-Variation Input)**.** The driving path Y is continuous and of bounded variation (i.e., $Y \in C_{1,0}$), so that the total variation function $V_t := \|Y\|_{1\text{-var},[0,t]}$ is continuous, non-decreasing, and $V_1 < \infty$.

Assumption I.3 (Uniform Operator Bound)**.** The true vector field is bounded by a constant $M_0 > 0$. That is,

$$\|\mathcal{W}^*\|_{\mathrm{op}} \le M_0$$

Remark I.1. By Lemma 1 and the triangle inequality, $\|\mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \le M_0 + K_{\mathrm{TT}}R^{-\alpha+1/2} \le M_0 + K_{\mathrm{TT}} =: M$ uniformly for $R \ge 1$. We use this M throughout Lemma 2.

I.3 LEMMA 1 (TT-SVD TRUNCATION ERROR) AND PROOF

Lemma I.2. *Under Assumption 1, let $\mathcal{W}_{TT}^{(R)}$ be the TT-SVD approximation of $\mathcal{W}^*$ truncated at a maximum TT-rank R. The error in the operator norm is bounded by:*

$$\|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \le K_{\mathrm{TT}} R^{-\alpha+1/2}$$

where $K_{\mathrm{TT}} := \left(\sum_{\ell=1}^{k-1} \frac{C_\ell^2}{2\alpha-1}\right)^{1/2}$.

Proof of Lemma 1. Following the error analysis of TT-SVD by Oseledets (2011), the squared Frobenius norm error of the entire tensor is bounded by the sum of squared truncated singular values across all unfolding matrices:

$$\|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_F^2 \le \sum_{\ell=1}^{k-1} \sum_{j=R+1}^{\mathrm{rank}_\ell} \sigma_{\ell,j}^2 \le \sum_{\ell=1}^{k-1} \sum_{j=R+1}^{\infty} \sigma_{\ell,j}^2$$

Substituting the polynomial decay $\sigma_{\ell,j} \le C_\ell j^{-\alpha}$ from Assumption 1 yields:

$$\sum_{j=R+1}^{\infty} \sigma_{\ell,j}^2 \le C_\ell^2 \sum_{j=R+1}^{\infty} j^{-2\alpha}$$

Since the function $f(x) = x^{-2\alpha}$ is monotonically decreasing for $x > 0$, this infinite series is bounded above by the corresponding integral:

$$\sum_{j=R+1}^{\infty} j^{-2\alpha} \le \int_R^{\infty} x^{-2\alpha} dx = \left[\frac{x^{-2\alpha+1}}{-2\alpha+1}\right]_R^{\infty}$$

By the assumption $\alpha > 1/2$, which implies $-2\alpha + 1 < 0$, the upper limit converges to 0. Thus,

$$\left[\frac{x^{-2\alpha+1}}{-2\alpha+1}\right]_R^{\infty} = 0 - \frac{R^{-2\alpha+1}}{-2\alpha+1} = \frac{1}{2\alpha-1} R^{-2\alpha+1}$$

Returning this to the summation over ℓ, we obtain the following bound:

$$\|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_F^2 \le \sum_{\ell=1}^{k-1} C_\ell^2 \left(\frac{1}{2\alpha-1} R^{-2\alpha+1}\right) = \left(\sum_{\ell=1}^{k-1} \frac{C_\ell^2}{2\alpha-1}\right) R^{-2\alpha+1}$$

Using the property that the operator norm is bounded by the Frobenius norm for any tensor, where $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$, and taking the square root of both sides gives:

$$\|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \le \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_F \le \left(\sum_{\ell=1}^{k-1} \frac{C_\ell^2}{2\alpha-1}\right)^{1/2} R^{-\alpha+1/2} = K_{\mathrm{TT}} R^{-\alpha+1/2}$$

This concludes the proof. □

I.4 Lemma 2 (Lipschitz Stability of the Linear CDE Core) and Proof

Lemma I.3. *Under Assumptions 2 and 3, the trajectory error for any $t \in [0,1]$ is bounded by:*

$$\|Z_t^* - \widehat{Z}_t^{(R)}\|_2 \le \|Z_0\|_2 V_t \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} e^{2MV_t}$$

Proof of Lemma 2. Let $e_t := \|Z_t^* - \widehat{Z}_t^{(R)}\|_2$ denote the norm of the difference. Expressing both sides in integral form and adding $\pm \int_0^t \mathcal{L}_{\mathcal{W}^*}(\widehat{Z}_s^{(R)}) dY_s$ to the right hand side, we rearrange the terms as:

$$\begin{aligned} Z_t^* - \widehat{Z}_t^{(R)} &= \int_0^t \mathcal{L}_{\mathcal{W}^*}(Z_s^*) dY_s - \int_0^t \mathcal{L}_{\mathcal{W}_{TT}^{(R)}}(\widehat{Z}_s^{(R)}) dY_s \\ &= \int_0^t \mathcal{L}_{\mathcal{W}^*}(Z_s^* - \widehat{Z}_s^{(R)}) dY_s + \int_0^t \left(\mathcal{L}_{\mathcal{W}^*} - \mathcal{L}_{\mathcal{W}_{TT}^{(R)}}\right)(\widehat{Z}_s^{(R)}) dY_s \end{aligned}$$

Taking the norm of both sides and applying the triangle inequality for integrals along with the operator norm property $\|\mathcal{L}_{\mathcal{W}}(u) dY_s\|_2 \le \|\mathcal{W}\|_{\mathrm{op}} \|u\|_2 dV_s$, we replace dY_s with the total variation measure dV_s via the Lebesgue-Stieltjes integral to obtain:

$$e_t \le \int_0^t \|\mathcal{W}^*\|_{\mathrm{op}} e_s dV_s + \int_0^t \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \|\widehat{Z}_s^{(R)}\|_2 dV_s$$

We now bound the norm of the approximated trajectory $\|\widehat{Z}_t^{(R)}\|_2$ using Grönwall's inequality.

$$\|\widehat{Z}_t^{(R)}\|_2 \leq \|Z_0\|_2 + \int_0^t \|\mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \|\widehat{Z}_s^{(R)}\|_2 dV_s$$

Applying Assumption 3 where $\|\mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \leq M$ yields:

$$\|\widehat{Z}_t^{(R)}\|_2 \leq \|Z_0\|_2 \exp\left(\int_0^t M dV_s\right) = \|Z_0\|_2 e^{MV_t}$$

Substituting this into the second term of the previous inequality and applying Assumption 3 where $\|\mathcal{W}^*\|_{\mathrm{op}} \leq M$ to the first term gives:

$$e_t \leq \int_0^t M e_s dV_s + \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \|Z_0\|_2 \int_0^t e^{MV_s} dV_s$$

We evaluate the integral in the second term $\int_0^t e^{MV_s} dV_s$. Using substitution for the total variation V_s with $u = MV_s$ and $du = M dV_s$, and noting $V_0 = 0$, we obtain:

$$\int_0^t e^{MV_s} dV_s = \frac{1}{M}\left[e^{MV_s}\right]_0^t = \frac{e^{MV_t} - 1}{M} \leq V_t e^{MV_t}$$

Therefore, the inequality is rearranged as follows:

$$e_t \leq \underbrace{\left(\|Z_0\|_2 \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} V_t e^{MV_t}\right)}_{:=A(t)} + \int_0^t M e_s dV_s$$

Since V_t is monotonically increasing, the function $A(t)$ on the right hand side is non-negative and monotonically increasing. Thus, we can apply the standard Grönwall's inequality $u(t) \leq A(t) + \int_0^t B(s)u(s)dV_s \implies u(t) \leq A(t)\exp(\int_0^t B(s)dV_s)$:

$$e_t \leq A(t) \exp\left(\int_0^t M dV_s\right) = \left(\|Z_0\|_2 \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} V_t e^{MV_t}\right) e^{MV_t}$$

By the laws of exponents,

$$e_t \leq \|Z_0\|_2 V_t \|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} e^{2MV_t}$$

This concludes the proof. □

I.5 Proof of the trajectory approximation theorem

Proof of the trajectory approximation theorem. Substituting the operator norm bound $\|\mathcal{W}^* - \mathcal{W}_{TT}^{(R)}\|_{\mathrm{op}} \leq K_{\mathrm{TT}} R^{-\alpha+1/2}$ derived in Lemma 1 into the result of Lemma 2 gives:

$$\|Z_t^* - \widehat{Z}_t^{(R)}\|_2 \leq \|Z_0\|_2 V_t \left(K_{\mathrm{TT}} R^{-\alpha+1/2}\right) e^{2MV_t}$$

For the time interval $t \in [0, 1]$, the definition of bounded variation from Assumption 2 ensures that $V_t \leq V_1$ always holds. Therefore, the exponential term on the right hand side is bounded above by e^{2MV_1}.

$$\sup_{t\in[0,1]} \|Z_t^* - \widehat{Z}_t^{(R)}\|_2 \leq \left(\|Z_0\|_2 V_1 e^{2MV_1} K_{\mathrm{TT}}\right) R^{-\alpha+1/2} = C_{\mathrm{tot}} R^{-\alpha+1/2}$$

Setting $C_{\mathrm{tot}} := \|Z_0\|_2 V_1 e^{2MV_1} K_{\mathrm{TT}}$ completes the proof. □

I.6 CP–TT spatial separation and temporal separation

Theorem I.4 (Temporal separation by internal depth)**.** *Let $f^*_{\mathrm{temp}}(\omega) = \int_0^1 (\int_0^t x_s^1 \, dx_s^2) \, dx_t^3$. It is realised by a single TD-CDE module with internal depth 3 and bond dimension 1, but it is not in the uniform closure of the class $\mathcal{M}^{(1)}_{\mathrm{Diag}}$ of single-layer selective diagonal functionals with contemporaneous input-dependent gates.*

This appendix provides the proof of the space-time separation theorem, which establishes the finite-depth expressivity gap between a single-layer TD-CDE and a diagonal CDE, as presented in the space-time separation result.

I.7 Proof of Theorem 4.1 (Spatial Separation)

The proof has four steps: (i) the coefficient tensor is *identifiable* from the path functional, so lower bounds on tensor rank transfer to lower bounds on architecture width; (ii) the two architectures realise the CP and TT formats respectively; (iii) any matricization lower-bounds the CP-rank; (iv) an explicit TT-rank-q witness whose *interleaved* matricization has rank $q^{\lfloor K/2 \rfloor}$.

Step (i): identifiability of the coefficient tensor.

Lemma I.5 (Identifiability)**.** *The map $\mathcal{W} \mapsto F_{\mathcal{W}}$ defined in* (11) *is injective on $(\mathbb{R}^d)^{\otimes K}$.*

Proof. Let $w^{(1)}, \dots, w^{(K)} \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}^K$, and let ω_ε be the piecewise-linear path obtained by concatenating the K straight segments with increments $\varepsilon_1 w^{(1)}, \dots, \varepsilon_K w^{(K)}$. By Chen's identity,

$$\mathrm{Sig}(\omega_\varepsilon) = \exp^{\otimes}\big(\varepsilon_1 w^{(1)}\big) \cdots \exp^{\otimes}\big(\varepsilon_K w^{(K)}\big), \qquad \exp^{\otimes}(w) = \sum_{k \geq 0} \tfrac{1}{k!} w^{\otimes k},$$

so that the order-K component is

$$\mathrm{Sig}(\omega_\varepsilon)^{(K)} = \sum_{\substack{k_1 + \cdots + k_K = K \\ k_j \geq 0}} \frac{\varepsilon_1^{k_1} \cdots \varepsilon_K^{k_K}}{k_1! \cdots k_K!} \, \big(w^{(1)}\big)^{\otimes k_1} \otimes \cdots \otimes \big(w^{(K)}\big)^{\otimes k_K}.$$

Hence $\varepsilon \mapsto F_{\mathcal{W}}(\omega_\varepsilon)$ is a polynomial, and the unique term contributing the multilinear monomial $\varepsilon_1 \varepsilon_2 \cdots \varepsilon_K$ is the one with $k_1 = \cdots = k_K = 1$. Its coefficient equals $\langle \mathcal{W}, w^{(1)} \otimes \cdots \otimes w^{(K)} \rangle$. If $F_{\mathcal{W}} \equiv 0$, all polynomial coefficients vanish, so $\langle \mathcal{W}, w^{(1)} \otimes \cdots \otimes w^{(K)} \rangle = 0$ for all $w^{(1)}, \dots, w^{(K)}$. Since simple tensors span $(\mathbb{R}^d)^{\otimes K}$, $\mathcal{W} = 0$. □

Remark I.2 (Why the order-K signature and not $x^{\otimes K}$)**.** Lemma I.5 is what makes a tensor-rank lower bound meaningful, and it is specific to sequential features. If one instead evaluates a coefficient tensor against the symmetric power $x^{\otimes K}$ of a *single* vector x, the map $\mathcal{W} \mapsto \langle \mathcal{W}, x^{\otimes K} \rangle$ factors through the symmetrisation $\mathrm{Sym}(\mathcal{W})$ and is therefore far from injective, so no rank lower bound on $\mathcal{W}$ can be transferred. The witness $\mathcal{W}^\star$ of Lemma I.8 with $q = 2$ illustrates this concretely: it satisfies $\mathrm{rank}_{\mathrm{CP}}(\mathcal{W}^\star) \geq 2^{K/2}$, yet $\langle \mathcal{W}^\star, x^{\otimes K} \rangle = (x_1^2 + x_2^2)^{K/2}$ is a binary form of degree K and can be written as a sum of $\mathcal{O}(K)$ products of linear forms. The exponential gap exists only when the K tensor slots are filled by *ordered, distinct* time increments — which is precisely what the iterated integrals of a path provide.

Step (ii): architectures realise CP and TT formats.

Lemma I.6 (Architecture ⇒ tensor format)**.** *Consider pure-integration dynamics as in Proposition 5.3.*

- *(a) Let a selective diagonal model consist of H scalar channels that evolve independently and are mixed only by the final linear readout, i.e., $H_{t,h}^{(0)} \equiv 1$ and $H_{t,h}^{(k)} = \int_0^t H_{s,h}^{(k-1)} \, d\langle \mathbf{u}_k^{(h)}, \omega_s \rangle$ for $k = 1, \dots, K$, with readout $y_t = \sum_{h=1}^H c_h H_{t,h}^{(K)}$. Then $y_1 = F_{\mathcal{W}}$ with $\mathcal{W} = \sum_{h=1}^H c_h \, \mathbf{u}_1^{(h)} \otimes \cdots \otimes \mathbf{u}_K^{(h)}$, so $\mathrm{rank}_{\mathrm{CP}}(\mathcal{W}) \leq H$; conversely every CP-rank-$\leq H$ tensor arises this way.*
- *(b) A single TD-CDE module of internal depth K, bond dimension R and one head realises $\mathcal{W}_I = G_{i_1}^{(1)} \cdots G_{i_K}^{(K)}$, i.e., exactly the tensors of TT-rank $\leq R$; with H heads, TT-rank $\leq HR$.*

In particular $\mathcal{M}_{\mathrm{Diag}}(H)$ coincides with TD-CDE restricted to bond dimension $R = 1$ and H heads: the diagonal model is the rank-one specialisation of our architecture.

Proof. Part (b) is Proposition 5.3. Part (a) is Proposition 5.3 applied with scalar TT-cores $G_{h,i}^{(k)} = (\mathbf{u}_k^{(h)})_i \in \mathbb{R}^{1 \times d \times 1}$ for each channel h, followed by the linear readout; the converse follows by reading a CP decomposition off as the H channel parameters. □

Step (iii): matricization bounds CP-rank.

Lemma I.7 (**?**, Lemma 1)**.** *For any tensor $\mathcal{X}$ and any bipartition (s,t) of its modes,* $\operatorname{rank}\mathcal{X}^{(s,t)} \leq \operatorname{rank}_{\mathrm{CP}}(\mathcal{X})$.

Proof. A CP-rank-one tensor matricizes to a rank-one matrix under *any* bipartition; matricization is linear and matrix rank is subadditive. □

Lemma I.7 holds for arbitrary bipartitions, including *interleaved* ones. This is essential: for the *sequential* unfolding $\mathcal{W}^{\langle m\rangle}$ (first m modes versus the rest) one has $\operatorname{rank}\mathcal{W}^{\langle m\rangle} = r_m \leq R$ exactly (**?**, Thm. 2.1), so sequential unfoldings can never certify an exponential lower bound. The exponential growth is visible only in interleaved matricizations.

Step (iv): explicit witness.

Lemma I.8 (Witness tensor)**.** *Let $q = \min(d, R)$ and assume K is even (for odd K leave the last mode free). Define $\mathcal{W}^\star \in (\mathbb{R}^d)^{\otimes K}$ by*

$$\mathcal{W}^\star_{i_1 i_2 \cdots i_K} := \prod_{j=1}^{K/2} \delta^{(q)}_{i_{2j-1}\, i_{2j}}, \qquad \delta^{(q)}_{ab} := \mathbf{1}\{a = b \leq q\}.$$

Then (a) $\operatorname{rank}_{\mathrm{TT}}(\mathcal{W}^\star) = (q, 1, q, 1, \ldots, q) \leq R$ *entrywise, and (b) the interleaved matricization $s = \{1, 3, \ldots, K-1\}$, $t = \{2, 4, \ldots, K\}$ satisfies* $\operatorname{rank}(\mathcal{W}^\star)^{(s,t)} = q^{K/2}$.

Proof. (a) Take TT-cores $G^{(2j-1)}_{i,\alpha} = \delta^{(q)}_{i\alpha} \in \mathbb{R}^{1\times d\times q}$ and $G^{(2j)}_{\alpha,i} = \delta^{(q)}_{i\alpha} \in \mathbb{R}^{q\times d\times 1}$; contracting the bond index α between modes $2j-1$ and $2j$ gives $\sum_{\alpha\leq q} \delta^{(q)}_{i_{2j-1}\alpha}\delta^{(q)}_{i_{2j}\alpha} = \delta^{(q)}_{i_{2j-1}i_{2j}}$, which reproduces $\mathcal{W}^\star$; the bonds between modes $2j$ and $2j+1$ are trivial. (b) Restricted to indices in $\{1, \ldots, q\}$ the matrix $(\mathcal{W}^\star)^{(s,t)}$ is exactly the identity of size $q^{K/2} \times q^{K/2}$, and it vanishes elsewhere. □

Proof of Theorem 4.1.

Suppose $F_{\mathcal{W}^\star} \in \mathcal{M}_{\mathrm{Diag}}(H)$, i.e., $F_{\mathcal{W}^\star} = F_{\mathcal{V}}$ for some $\mathcal{V}$ with $\operatorname{rank}_{\mathrm{CP}}(\mathcal{V}) \leq H$. By Lemma I.5, $\mathcal{V} = \mathcal{W}^\star$. Combining Lemmas I.7 and I.8,

$$H \;\geq\; \operatorname{rank}_{\mathrm{CP}}(\mathcal{W}^\star) \;\geq\; \operatorname{rank}(\mathcal{W}^\star)^{(s,t)} \;=\; q^{K/2}.$$

Meanwhile $\mathcal{W}^\star$ has TT-rank $\leq q \leq R$, so by Lemma I.6(b) a single TD-CDE module of bond dimension R realises $F_{\mathcal{W}^\star}$ with $\mathcal{O}(KdR^2)$ parameters. The genericity statement follows from Theorem 1 of Khrulkov et al. (2017): the set $\{\mathcal{W} : \operatorname{rank}_{\mathrm{TT}}\mathcal{W} \leq R,\ \operatorname{rank}_{\mathrm{CP}}\mathcal{W} < q^{K/2}\}$ has Lebesgue measure zero in the variety of TT-rank-$\leq R$ tensors. □

Remark I.3 (Interpretation of the witness)**.** $F_{\mathcal{W}^\star}(\omega) = \sum_{i_1,\ldots,i_{K/2}\leq q} \mathrm{Sig}(\omega)^{i_1 i_1 i_2 i_2 \cdots i_{K/2} i_{K/2}}_{0,1}$ is a sum of *matched-pair* iterated integrals: it measures $K/2$ successive self-reinforcing excursions of individual channels. This gives a directly testable synthetic target, used in Section 5.

Remark I.4 (Scope of the separation)**.** Three caveats delimit the statement. (i) The bound concerns *exact* representation; functional approximation measured under specific path distributions may require far fewer channels, and Khrulkov et al. (2017) report that the empirical gap between CP and TT architectures is often only polynomial. (ii) Lemma I.6(a) models a diagonal recurrence whose selection gates act linearly on the driving path; models with nonlinear input-dependent input matrices can realise functionals outside this class (see Remark I.6). (iii) Stacking K diagonal layers with inter-layer projections changes the comparison class. The separation of Theorem 4.1 is a statement about a channel-wise independent architecture; we do not establish strict width lower bounds against densely-coupled deep diagonal stacks.

I.8 Proof of the space-time separation theorem, Part 2 (Temporal Separation)

Definition I.1 (Depth-one selective diagonal class)**.** Let $\mathcal{M}^{(1)}_{\mathrm{Diag}}$ denote the class of functionals realised by a *single* selective diagonal layer, i.e., those of the form $\Phi(\omega) = \sum_{j=1}^{d} \int_0^1 f_j(x_s)\, dx^j_s$ with

continuous f_j, where x_s is the time-augmented instantaneous input. $\mathcal{M}^{(1)}_{\mathrm{Diag}}$ neither contains nor is contained in the class $\mathcal{M}_{\mathrm{Diag}}(H)$ of Lemma I.6(a): the latter has internal depth K with *constant* gates, while the former has depth one with *input-dependent* gates. In particular, $\int(\int x^1 dx^2)dx^3$ has CP-rank one and therefore *does* belong to $\mathcal{M}_{\mathrm{Diag}}(1)$; Part 2 is a statement about $\mathcal{M}^{(1)}_{\mathrm{Diag}}$ only, and the two parts of the space-time separation theorem separate along orthogonal axes (external depth vs. width).

Proof of the temporal separation claim. Let the target signature $f^*_{\text{temp}} = \int(\int x^1 dx^2)dx^3$ be an iterated integral composed of non-commutative cross-terms of length $K = 3$, which depends on the temporal order of events.

We first establish the limitation of $\mathcal{M}^{(1)}_{\mathrm{Diag}}$. Because the state transition is diagonal, the channels do not mix during the continuous-time evolution; mixing only occurs via the input-dependent gates, which rely solely on the contemporaneous input x_s. The closure of the function space for a single-layer selective diagonal model with pointwise gating therefore consists of functionals of the form $\Phi(\omega) = \sum_{j=1}^{d} \int f_j(x_s) dx^j_s$.

We show that this closure cannot express f^*_{temp}. For $n \in \mathbb{Z}_{\geq 0}$ and $T > 0$, let $\omega_{n,T}$ be the concatenation of (i) n identical traversals of a fixed closed loop γ in the (x^1, x^2)-plane based at p, with $x^3 \equiv 0$ and signed area $a := \oint_\gamma x^1\, dx^2 \neq 0$, and (ii) a straight segment of length T in the x^3-direction with $(x^1, x^2) \equiv p$. We fix the time parametrisation once and for all: segment (i) occupies $[0, \frac{1}{2}]$ (the n traversals are simply run n times faster) and segment (ii) occupies $[\frac{1}{2}, 1]$, independently of n and T. Iterated integrals are invariant under reparametrisation, so this entails no loss of generality.

Any Φ in the closure is additive over concatenation and its integrand depends only on the current (time-augmented) state, with no dependence on past accumulations; hence $\Phi(\omega_{n,T}) = \Phi_{(\mathrm{i})} + \Phi_{(\mathrm{ii})}$. By the fixed parametrisation, segment (i) as a parametrised path is determined by n alone, so $\Phi_{(\mathrm{i})} = C_1(n)$; and segment (ii) is determined by T alone — it occupies $[\frac{1}{2}, 1]$ and starts at $(p_1, p_2, 0)$ regardless of n — so $\Phi_{(\mathrm{ii})} = C_2(T)$. Hence $\Phi(\omega_{n,T}) = C_1(n) + C_2(T)$, which is *additively separable* in (n, T). For the target, $dx^3 = 0$ on (i), while on (ii) the inner integral is frozen at $A_s \equiv na$, so $f^*_{\text{temp}}(\omega_{n,T}) = n\, a\, T$, which is *bilinear*. The mixed second difference $\Delta\phi := \phi(1,2) - \phi(1,1) - \phi(0,2) + \phi(0,1)$ annihilates every additively separable function, whereas $\Delta f^*_{\text{temp}} = a \neq 0$. Since $|\Delta\psi| \leq 4 \sup|\psi|$, we conclude $\sup_{n\leq 1,\, T\leq 2} \left| f^*_{\text{temp}} - \Phi \right| \;\geq\; |a|/4$ for *every* Φ in the class, so f^*_{temp} is not even in the uniform closure.

For the function class $\mathcal{M}_{\mathrm{TT}}(K, R)$, by contrast, the temporal update is directly executed as a tensor contraction $dH^{(m)} = H^{(m-1)} \times_1 \mathcal{C}^{(m)} \times_2 dx$ embedded within the solver step inside a single module ($D = 1$). Isolating a specific word with a rank-1 core $G^{(m)}_i = \delta_{i,i_m}$, TD-CDE maintains and extracts non-commutative signatures as internal states: a single TD-CDE module computes order-K interactions within its depth-K integration hierarchy at bond dimension $R = 1$, with no external layer stacking required. □

I.9 Implicit signature projection

The pure-integration proposition in the main text follows by induction on the internal depth. The implementation-specific decay, rotation, and calibrated scales are omitted in this proposition; they act after the same multilinear TT topology has been formed.

Proof of Proposition 4.2. For $m = 1$,

$$H^{(1)}_t = \sum_{i_1=1}^{d_\omega} G^{(1)}_{i_1} \operatorname{Sig}(\omega)^{i_1}_{0,t}.$$

Assume

$$H^{(m-1)}_t = \sum_{(i_1,\ldots,i_{m-1})} G^{(1)}_{i_1} \cdots G^{(m-1)}_{i_{m-1}} \operatorname{Sig}(\omega)^{(i_1,\ldots,i_{m-1})}_{0,t}.$$

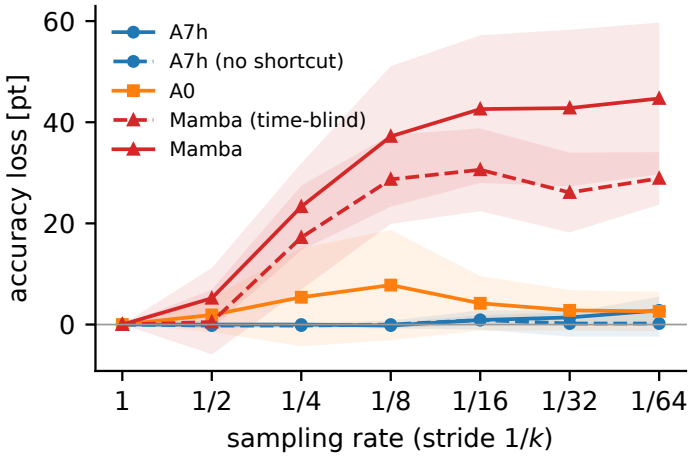


Figure 2: SelfRegulationSCP1: accuracy loss under stride-k subsampling (mean $\pm$ sd over five seeds). At $k = 64$ only 14 of 896 samples remain.

Substituting into the update for level m and using

$$\int_0^t \mathrm{Sig}(\omega)_{0,s}^{J}\, d\omega_s^{i_m} = \mathrm{Sig}(\omega)_{0,t}^{J\cdot i_m}$$

gives

$$H_t^{(m)} = \sum_{(i_1,\dots,i_m)} G_{i_1}^{(1)} \cdots G_{i_m}^{(m)}\, \mathrm{Sig}(\omega)_{0,t}^{(i_1,\dots,i_m)}.$$

Thus the statement holds for all $m \le M$. □

J Effective Depth of a Live Gate

The live-gate experiment isolates the contribution of an input-dependent multiplicative gate from additional internal levels. With constant gates, internal depth M reaches signature interaction order M in the pure integration probe. Enabling a live gate at the input supplies one additional multiplicative level, but the resulting contribution remains rank restricted. For the symmetric level-3 target, the best live-gate route reaches 5.2×10^{-2} error compared with 1.7×10^{-3} for an additional real level. It therefore behaves as one effective level rather than two.

K Extended Experimental Results

K.1 Synthetic discretisation results

The complete synthetic results are reproduced in Table 5. On the 2048-step smooth paths, A7h required $5.2\times$ the runtime of A0 but reduced the integration error by 9.3×10^4. Adding the cubic reconstruction to the otherwise second-order scheme cost only 2.7% and reduced error by $145\times$. Task R^2 remained between 0.99993 and 0.99996 for all four D/E combinations, demonstrating why native-grid task accuracy is too coarse to diagnose integration order.

K.2 Resolution-transfer figure and controls

At the training resolution, the A7h–A0 accuracy difference is -0.7 points with a 95% interval of $[-2.0, +0.6]$. The median step exponent is 7×10^{-4} and its 99th percentile is 8.9×10^{-3}; no clipping occurs. Under $k \in \{2, 4, 8\}$, A7h changes one prediction among $1{,}275$ seed–stride cells, A0 changes 90, and Mamba changes 433. At $k = 64$, the A7h accuracy remains 87.3%, compared with 37.2%

Table 6: Margin-variance shares and explanation stability (mean over seeds; up to 64 test samples). "Endpt" is the explicit endpoint-displacement share. $\mathrm{ES}(k) = \frac{1}{2}\|s^{(k)} - s^{(1)}\|_1$ over the level-2 strata and τ is their Kendall correlation at $k = 1$ and $k = 8$. EigenWorms uses a Log-ODE path, so these degree statements are indicative.

Dataset	Arm	Endpt	$m = 0$	$m = 1$	$m = 2$	Stratum	ES(2)	ES(8)	τ
SCP1	A7h	0.056	−0.027	0.082	0.889	STT 0.703	0.0002	0.0084	0.986
SCP1	A0	0.088	−0.052	0.051	0.912	STT 0.506	0.0659	0.1750	0.614
EW	A7h	−0.001	0.002	0.239	0.760	SSS 0.619	0.0029	0.0102	0.976
EW	A0	0.001	0.024	0.311	0.664	SSS 0.425	0.0153	0.1468	0.786

for Mamba and 62.4% for the endpoint-only ridge probe. The corresponding relative logit drift at $k = 2$ is 2.3×10^{-3} for A7h and 1.3×10^{-1} for A0.

The endpoint controls are: an endpoint-only A7h model reaches 63.3%, increment permutation with endpoints fixed gives 56.5%, and removing the explicit displacement shortcut leaves A7h resolution-stable while A0 remains grid-sensitive. In EigenWorms reduced to two endpoints, Chen-cascade arms retain 82.5–85.0% while A0 drops to 43.5%; a ridge probe using the same aggregate increment reaches 79.2%.

K.3 Detailed interpretation table

All decompositions sum to one within 10^{-4}, and each finer split reconstructs its parent within 10^{-13}. On SCP1, the STT stratum dominates and its mass is distributed across all six channels; none exceeds a quarter after per-channel scale normalization. The channel ordering is consistent across seeds (Kendall's $W = 0.70$, $p = 0.004$), with channels 0 and 2 in the top three of every seed. The purely spatial stratum is below 0.03 in absolute value in all five seeds.

On EigenWorms, more than 98% of the SSS stratum lies in words built from three Lévy-area letters in two of three seeds for both arms; those letters correspond to degree six in the raw path after the depth-2 Log-ODE preprocessing. The remaining seed relies instead on increment letters, dominated by the pair (dx_3, dx_5). Across the seeds, these pathway regimes can differ even when accuracy is similar. Within a seed, A0 and A7h share 8, 4, and 5 of their top 12 pathways in the three seeds, respectively.

K.4 Explanation drift

The A7h stratum shares move by $\mathrm{ES} = 0.0002$ under halved sampling, $330\times$ less than A0. At $k = 8$, A7h has Kendall correlation 0.986 with the native stratum ordering, while A0 falls to 0.614. Gate activations at matched timestamps are identical across the arms ($r = 1.000$), isolating the difference to numerical accumulation of the iterated interactions.

L Forward Pass and Implementation

In this section, we present the pseudocode for the forward pass process of TD-CDE. This algorithm reflects the practical computational details, and in particular makes explicit the objects that a standard zero-order-hold (ZOH) linear CDE hides: (i) the intra-step Chen expansion with the $1/j!$ simplex weights and the barycentric decay factors $e^{-Z/(j+1)}$ (Algorithm 2), (ii) the exponential-integrator kernels (φ_1, Ψ_1, K_1) that integrate the decay over the step exactly instead of freezing it at the right endpoint, and (iii) the evaluation points of the *input-dependent* gates γ and s — Gauss–Legendre nodes for $j = 1$ and ordered-simplex barycentres $i/(j + 1)$ for $j \geq 2$.

Algorithm 1 acts as the main driver. Algorithm 2 constructs the intra-step increments (the exponential-integrator / Chen-expansion core of the method). Algorithm 3 performs the chunked parallel scan, where the complex rotation is decoupled from the magnitude decay via a global phase alignment so that the scan remains a real-valued log-space prefix sum. Table 7 fixes the notation and Table 8 lists the intra-step kernels selected by the correction mode configuration.

Throughout, $\odot$ denotes elementwise product with broadcasting over the head and rank axes, $\oplus$ denotes concatenation, and all quantities carry an implicit head index $h \in [H]$ and rank index $r, s \in [R]$. The one-step propagator of level m is $e^{-Z_{m,l}}$ with $Z_{m,l} = \xi_{m,l} - i\Omega_{m,l}$.

Table 7: Notation. d_x is the number of raw spatial channels; the input tensor has $d_x + 1$ columns, column 0 being absolute time.

Symbol	Meaning
$x_l^{\text{raw}} \in \mathbb{R}^{d_x}, t_l$	raw spatial channels, absolute time
$M_l \in \{0, 1\}$	validity mask from sequence lengths; ℓ = last valid index
$z_l \in \mathbb{R}^{d_s}$	spatial state; $d_s = K d_x$ in interpretable mode
$\Delta Z_l \in \mathbb{R}^{d_a}$	gated augmented increment, $d_a = d_s + 1$
$\gamma_{m,l} \in (0, 1)^{H\times 1}$	clock gate at the right endpoint
$\Gamma_{m,l}$	step-averaged clock gate $\mathcal{Q}[\gamma_{m,\cdot}]_l$
$\lambda_{m,l} \in \mathbb{R}_+^{H\times R}$	decay rate
$\Lambda_{m,l}$	$\gamma_{m,l} \odot \lambda_{m,l}$ if decay is clock-gated, else $\lambda_{m,l}$
$\xi_{m,l} \in [0, x_{\max}]$	real decay exponent
$\tau_{m,l}$	intra-step tilt $(\Lambda_{m,l} - \Lambda_{m,l-1})\Delta t_l$
$\Omega_{m,l}, \Phi_{m,l}$	phase increment / global phase
$Z_{m,l}$	$\xi_{m,l} - i\Omega_{m,l}$
$\{(\sigma_q, w_q)\}_{q=1}^{Q}$	Gauss–Legendre nodes and weights on $[0, 1]$
$x_{\max}$	maximum float64 exponent (e.g., $\ln(\text{DBL_MAX})/C$)

Multilinear Core Operator. For the TT cores $\mathcal{C}^{(0)} \in \mathbb{R}^{H\times d_a\times R}$ and $\mathcal{C}^{(k)} \in \mathbb{R}^{H\times d_a\times R\times R}$ ($k \geq 1$), we define the contractions:

$$\mathcal{A}_0[\,\cdot\,, \Delta Z; g]_{h,r} = g_h \textstyle\sum_d \Delta Z_d \, \mathcal{C}^{(0)}_{h,d,r}$$
$$\mathcal{A}_k[u, \Delta Z; g]_{h,s} = g_h \textstyle\sum_{d,r} u_{h,r} \, \Delta Z_d \, \mathcal{C}^{(k)}_{h,d,r,s}$$

Table 8: Intra-step kernels for $j = 1$ branch by correction mode. We define $\delta = (\tau_{m,l} - \tau_{m-1,l})/12$. $\varphi_1(x) = \frac{1-e^{-x}}{x}$, $\Psi_1(u, w) = \frac{e^{-w}-e^{-u}}{u-w}$, and $K_1(x) = \frac{x-2+(x+2)e^{-x}}{x^3}$.

Correction Mode	$\kappa^{(1)}$ (level $m = 0$)	$\kappa^{(2)}_{r,s}$ (level $m \geq 1$)
None	$\varphi_1(Z_{0,l})$	$\Psi_1(Z_{m,l,s}, Z_{m-1,l,r})$
Level-1 Shift	$\varphi_1(Z_{0,l} + \tau_{0,l}/6)$	$\Psi_1(Z_{m,l,s}, Z_{m-1,l,r})$
Global Shift	$\varphi_1(Z_{0,l} + \tau_{0,l}/6)$	$\Psi_1(Z_{m,l,s}+\delta, Z_{m-1,l,r}+\delta)$
Exact Level-1	$\varphi_1(Z_{0,l}) - \frac{\tau_{0,l}}{2} K_1(Z_{0,l})$	$\Psi_1(Z_{m,l,s}+\delta, Z_{m-1,l,r}+\delta)$
Quadrature	Evaluated via Gauss–Legendre quadrature (lines 6–12 of Alg. 2)	

M Additional Implementation Information

Experiments were run on an Nvidia A100 (40GB) GPU using PyTorch (Paszke et al., 2019). The interpretable UEA configuration uses the single-layer architecture described in the main text. Full hyperparameter selection uses the 70/15/15 train/validation/test split per seed; validation data are used for hyperparameter selection and test data are used only for the reported evaluation. Published baseline values are retained only as indicative references when the split or training protocol differs.

Algorithm 1 Forward Pass of TD-CDE (Main Driver)

Input: raw spatial channels $x^{\text{raw}}_{1:L}$, times $t_{1:L}$, valid length ℓ
Parameters: TT cores $\{\mathcal{C}^{(m)}\}_{m=0}^{D-1}$, clock/decay heads, frequencies $\omega \in \mathbb{R}^{H\times R}$, selection gate $(W_{\text{sel}}, b_{\text{sel}})$, static scale S_{static}, depth scales $\{c_m\}$, readout $(W_{\text{lin}}, b_{\text{lin}}, P, \mathcal{T}, W_{\text{disp}})$
Scheme: chunk size C, depth D, heads H, rank R, discretization flags, $x_{\max}$
Output: interpretable sequence features $Y_{1:L}$ or prediction $\hat{y}$

 # 1. Masking, Spatial Lifting, and Driving Increments
$M_l \leftarrow \mathbf{1}[l \le \ell]$; $\quad \Delta t_l \leftarrow \text{ReLU}(t_l - t_{l-1})$ with $\Delta t_1 \leftarrow 0$
$z_{1:L} \leftarrow \mathcal{F}(x^{\text{raw}}_{1:L}; \Delta t_{1:L}) \odot M$ {ramp-invariant EMA taps $\oplus$ identity $\oplus$ lag-1}
$\Delta z_l \leftarrow (z_l - z_{l-1}) \odot M_l$ with $\Delta z_1 \leftarrow 0$
$s_l \leftarrow \sigma(W_{\text{sel}} x^{\text{raw}}_l + b_{\text{sel}}) \in (0,1)^{d_x}$; $\quad \tilde{s}_l \leftarrow \mathbf{1}_K \otimes s_l$ {gate shared across K taps}
$\Delta Z_l \leftarrow [\, \Delta z_l \odot S_{\text{static}} \odot \tilde{s}_l \,,\, \Delta t_l \,]$ {time channel is neither scaled nor gated}
 # 2. Step Coefficients (Clock, Decay, Tilt, Phase)
for $m = 0$ **to** $D-1$ **do**
  $\gamma_{m,l} \leftarrow \sigma(\text{head}^{\text{clk}}_m(z_l, t_l))$, $\quad \lambda_{m,l} \leftarrow \text{softplus}(\text{head}^{\text{dec}}_m(z_l, t_l))$
  $\Gamma_{m,l} \leftarrow \mathcal{Q}[\gamma_{m,\cdot}]_l$ {step-averaged clock via quadrature}
  $\Lambda_{m,l} \leftarrow \gamma_{m,l} \odot \lambda_{m,l}$ if decay is clock-gated else $\lambda_{m,l}$
  $\xi_{m,l} \leftarrow \text{clip}\,(\mathcal{Q}[\Lambda_{m,\cdot}]_l \cdot \Delta t_l, 0, x_{\max})$ {log-magnitude decay}
  $\tau_{m,l} \leftarrow (\Lambda_{m,l} - \Lambda_{m,l-1}) \cdot \Delta t_l$ {intra-step tilt}
  $\Omega_{m,l} \leftarrow \omega \cdot \Delta t_l$ ($\odot \Gamma_{m,l}$ if rotation is clock-gated) else 0
  $Z_{m,l} \leftarrow \xi_{m,l} - i\Omega_{m,l}$
end for
 # 3. Depth Recursion (Tensor Levels)
$H^{(-1)} \equiv \mathbf{1}$
for $m = 0$ **to** $D-1$ **do**
  $b_{m,1:L} \leftarrow$ BUILDINCREMENT(m) {Alg. 2; consumes causally shifted $\widehat{H}^{(p)}_l := H^{(p)}_{l-1}$}
  $H^{(m)}_{1:L} \leftarrow$ PHASEALIGNEDCHUNKSCAN$(\xi_{m,\cdot}, \Omega_{m,\cdot}, b_{m,\cdot})$ {Alg. 3}
  $\widehat{H}^{(m)}_l \leftarrow H^{(m)}_{l-1}$ (with $\widehat{H}^{(m)}_1 = 0$) {strict causality}
end for
 # 4. Rescaling and Interpretable Readout
$Y^{\text{CDE}}_l \leftarrow \bigoplus_{m=0}^{D-1} c_m \cdot \text{vec}\left([\Re(H^{(m)}_l), \Im(H^{(m)}_l)]\right)$ {c_m initialized to $(m+1)!$}
$Y_l \leftarrow (Y^{\text{CDE}}_l \oplus z_l) \odot M_l$
if interpretable mode (sequence-to-one) **then**
  $disp \leftarrow x^{\text{raw}}_\ell - x^{\text{raw}}_1$ {global displacement of raw path}
  $\hat{y}_o \leftarrow (W_{\text{lin}} Y_\ell + b_{\text{lin}})_o + \sum_{i,k} disp_i \mathcal{T}_{o,i,k} (P Y_\ell)_k + (W_{\text{disp}} disp)_o$
  return $\hat{y}$
else
  return $Y_{1:L}$
end if

Algorithm 2 BUILDINCREMENT(m): Intra-step increment $b_{m,l}$ (all l in parallel)

if using Zero-Order Hold (ZOH) scheme **then**
  return $\mathcal{A}_m\left[\widehat{H}_l^{(m-1)}, \Delta Z_l; \Gamma_{m,l}\right]$
end if
$b \leftarrow 0$
for $j = 1$ **to** $m+1$ **do**
  $p \leftarrow m - j$ {source level; $p = -1$ starts from $H^{(-1)} = \mathbf{1}$}
  if $j = 1$ **and** using Gauss-Legendre quadrature **then**
    $\rho_q \leftarrow \frac{1}{2}\sigma_q(1-\sigma_q)$
    if $m = 0$ **then**
      $u \leftarrow \sum_q w_q e^{-\text{clip}[(1-\sigma_q)Z_{0,l}+\rho_q\tau_{0,l}]} \odot \mathcal{A}_0\left[\cdot, \Delta Z_l^{(\sigma_q)}; \gamma_{0,l}^{(\sigma_q)}\right]$
    else
      $F_q \leftarrow e^{-\text{clip}[(1-\sigma_q)Z_{m,l}+\rho_q\tau_{m,l}]}, \quad G_q \leftarrow e^{-\text{clip}[\sigma_q Z_{m-1,l}-\rho_q\tau_{m-1,l}]}$
      $u \leftarrow \sum_q w_q F_q \odot \mathcal{A}_m\left[\widehat{H}_l^{(m-1)} \odot G_q, \Delta Z_l^{(\sigma_q)}; \gamma_{m,l}^{(\sigma_q)}\right]$
    end if
  else if $j = 1$ **and** using exact analytic kernels **then**
    if $m = 0$ **then**
      $u \leftarrow \Gamma_{0,l}\left(\sum_d \Delta Z_{l,d}\mathcal{C}_{h,d,r}^{(0)}\right) \odot \kappa^{(1)}$
    else
      $u_{h,s} \leftarrow \Gamma_{m,l,h}\sum_{r,d}\widehat{H}_{l,h,r}^{(m-1)}\Delta Z_{l,d}\mathcal{C}_{h,d,r,s}^{(m)}\kappa_{r,s}^{(2)}$ {$R \times R$ kernel from Table 8}
    end if
  else
    $u \leftarrow \widehat{H}_l^{(p)} \odot e^{-Z_{p,l}/(j+1)}$ if $p \geq 0$, else $u \leftarrow \mathbf{1}$
    for $k = p+1$ **to** m **do**
      $i \leftarrow k - p$ {ordinal position of this increment, $1 \leq i \leq j$}
      $g \leftarrow \gamma_{k,l}^{(i/(j+1))}$ if using simplex-centroid gates else $\Gamma_{k,l}$
      $dZ \leftarrow \Delta Z_l^{(i/(j+1))}$ if using fractional increments else ΔZ_l
      $u \leftarrow \mathcal{A}_k[u, dZ; g] \odot e^{-Z_{k,l}/(j+1)}$ {barycentric approximation}
    end for
  end if
  $b \leftarrow b + u/j!$ {$1/j!$ = volume of ordered simplex}
end for
return b

Algorithm 3 PHASEALIGNEDCHUNKSCAN(ξ, Ω, b)

Input: $\xi_{1:L} \in [0, x_{\max}]$, $\Omega_{1:L}$, $b_{1:L}$ **Output:** $H_{1:L}$ solving $H_l = e^{-\xi_l + i\Omega_l}H_{l-1} + b_l$, $H_0 = 0$

Zero-pad to $L' = C\lceil L/C\rceil$; reindex $l \leftrightarrow (k, c)$ where chunk $k \in [K']$, pos $c \in [C]$
 # Phase: Two-Level Prefix Sum (Fully Parallel)
$\Omega_k^{\text{tot}} \leftarrow \sum_c \Omega_{k,c}; \quad \Phi_k^{\text{start}} \leftarrow \sum_{k'<k}\Omega_{k'}^{\text{tot}}; \quad \Phi_{k,c} \leftarrow \Phi_k^{\text{start}} + \sum_{j\leq c}\Omega_{k,j}$
$\hat{b}_{k,c} \leftarrow b_{k,c}e^{-i\Phi_{k,c}}$ {align to global phase frame}
 # Magnitude: Log-Space Scan in FP64
$\Sigma_{k,c} \leftarrow -\sum_{j\leq c}\xi_{k,j}\ (\leq 0); \quad \nu_k \leftarrow \max_c(-\Sigma_{k,c}) = -\Sigma_{k,C}$
$p_{k,c} \leftarrow e^{\text{clip}(-\Sigma_{k,c}-\nu_k, \pm 700)} \in (0, 1]; \quad q_{k,c} \leftarrow e^{\text{clip}(\Sigma_{k,c}+\nu_k, \pm 700)} \geq 1$
$\theta_{k,c} \leftarrow q_{k,c} \odot \sum_{j\leq c} p_{k,j} \odot \hat{b}_{k,j}$ {ν_k shift cancels out; forces $x_{\max} \leq \ln(\text{DBL_MAX})/C$}
 # Chunk Boundaries: Sequential Steps
$h_0 \leftarrow 0; \quad$ **for** $k = 1..K'$ **do** $h_k \leftarrow e^{\Sigma_{k,C}} \odot h_{k-1} + \theta_{k,C}$
$\hat{H}_{k,c} \leftarrow e^{\Sigma_{k,c}} \odot h_{k-1} + \theta_{k,c}$ {h_{k-1} is state entering chunk k}
$H_{k,c} \leftarrow \hat{H}_{k,c}e^{+i\Phi_{k,c}}$ {rotate back to local frame}
return $H_{1:L}$ (truncate the padding)