HiPPO Functional Memory: Coordinate-Queryable Attention for Irregular Time Series

Anonymous authors
Paper under double-blind review

Abstract

Irregular time series require models to answer temporal queries at arbitrary coordinates, yet existing approaches typically retain observation-dependent representations that must be repeatedly resolved for each query. We introduce HiPPO Functional Memory (HiPPO-FM), a functional-memory attention operator that compiles an irregular history into a reusable coordinate-addressable state. Instead of storing observation-indexed tokens or latent trajectories, HiPPO-FM writes observations into a fixed-rank functional space and performs direct coordinate evaluation through matched functional geometry. Kernel-smoothed HiPPO coordinates and Gram-based readout define a shared write/read mechanism, while observed-support self-attention enables content-dependent representations before compression. We characterize the induced finite-dimensional query space and derive an error decomposition separating functional geometry approximation from learned-memory mismatch. Across continuous-coordinate reconstruction, irregular representation learning, real-world extrapolation, and event-sequence prediction, HiPPO-FM consistently improves strong irregular time-series baselines, achieving strong reconstruction accuracy on Spiral, competitive performance across 20 UEA datasets, improved real-world extrapolation, and superior results on several irregular event benchmarks. After encoding, the compact functional state can be reused across different query sets without reprocessing the observation history.

1 Introduction

An irregular time series is observed at times determined by data collection and queried at times specified by the task. Clinical measurements and environmental observations arrive at uneven intervals, while downstream analysis may require interpolation at missing times, extrapolation beyond the observed interval, or evaluation at a different temporal resolution (Rubanova et al., 2019; Schirmer et al., 2022). These requests operate on the same history at different query coordinates. The representation retained from that history therefore determines how new temporal coordinates can be evaluated.

Existing approaches handle irregularity through continuous latent dynamics, time-aware attention, or structured representations of the observed samples. Latent ODEs and Neural CDEs model continuous hidden trajectories (Rubanova et al., 2019; Kidger et al., 2020); mTAN and ContiFormer introduce continuous-time structure into attention (Shukla & Marlin, 2021; Chen et al., 2023); and patch- or graph-based methods reorganize irregular observations before prediction (Zhang et al., 2024; Liu et al., 2026a). These approaches substantially improve how irregular histories are processed. In observation-indexed readouts, however, a requested coordinate is still resolved through a representation tied to the observed history. The form of this retained representation determines both how temporal queries access the history and what information must remain available for subsequent prediction.

How can an irregular history be compressed into a function-valued memory that supports direct evaluation at arbitrary temporal coordinates? We address this question by giving the retained state an explicit continuous coordinate semantics. Observations are written into a shared functional coordinate system, and requested times are evaluated in the same system. Content interaction occurs

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.

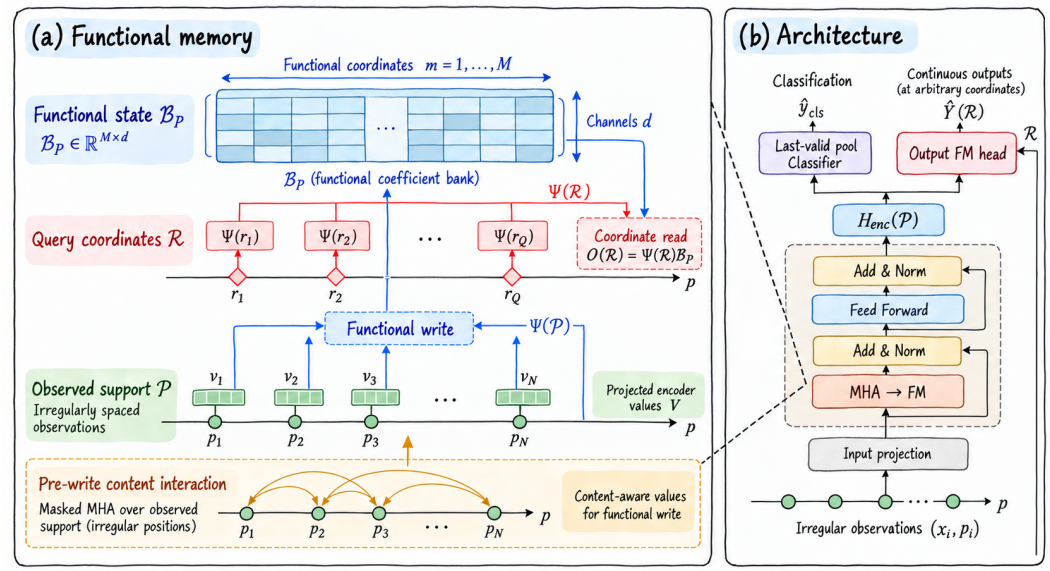

Figure 1: Figure 1: Overview of HiPPO Functional Memory (HiPPO-FM). **(a) Functional memory.** Irregular observations are compiled into a fixed-rank, coordinate-addressable functional state B_P using HiPPO-derived coordinate features and matched functional geometry. The same geometry is used for memory writing and coordinate readout, where query coordinates are evaluated through $O(R) = \Psi(R)B_P$, enabling reuse of the retained state across different query sets. **(b) Architecture.** Masked self-attention first performs content interaction on the observed support before functional compression. The resulting representation is either pooled for classification or passed to an output FM head for coordinate-valued prediction.

before compression, allowing the values written into memory to depend on relationships among the observations.

We introduce **HiPPO Functional Memory (HiPPO-FM)**, a functional-memory attention operator that compiles an irregular history into a fixed-rank, coordinate-queryable state:

$$\mathcal{D}_P \longmapsto B_P, \qquad O(R) = \Psi_\theta(R)B_P,$$

where P is the observed support, R is the requested support, and $B_P \in \mathbb{R}^{M \times d}$ contains the retained functional coefficients. Prediction at a new coordinate is obtained by evaluating these coefficients through its functional features rather than resolving a new query against the full observation-indexed history. After encoding, the history-specific state occupies $O(Md)$ storage, independent of the number of observed timestamps, and the same state can be evaluated on different query sets.

HiPPO-FM combines Legendre polynomial modes with interdomain kernel geometry to construct its temporal coordinate functions (Gu et al., 2020; Lazaro-Gredilla & Figueiras-Vidal, 2009). The same geometry governs the observation write and coordinate read, with a matched Gram solve accounting for correlations among the functional coordinates. Masked self-attention acts before the write, so content interaction determines what enters memory while the functional coordinates determine how that information is stored and queried. This **MHA-to-FM** ordering preserves the functional query space while allowing content-dependent representations to be formed before compression.

The resulting state admits an explicit functional interpretation. The raw memory consists of interdomain functionals of the kernel field induced by the projected observations, and the matched decoder restricts coordinate dependence to a finite-dimensional functional space. We characterize this space and derive a pointwise decomposition of approximation error into functional-geometry error and learned-memory mismatch. The coordinate-wise readout also makes the retained state reusable across different query sets without changing the prediction associated with a fixed requested coordinate.

We study these properties through continuous-coordinate reconstruction, irregular representation learning across 20 UEA datasets, real irregular extrapolation on PhysioNet, and controlled functional-memory diagnostics. Together, these settings evaluate both the predictive behavior of the representation and its ability to support repeated coordinate queries.

Our contributions are:

- **Coordinate-queryable functional attention.** We introduce an attention operator that compiles an irregular history into a fixed-rank functional state and performs prediction by direct coordinate evaluation. The retained state supports arbitrary query sets with history-length-independent post-encoding storage.
- **Matched functional memory with pre-write content interaction.** HiPPO-derived coordinates and interdomain kernel geometry define a common functional system for memory write and coordinate readout, while observed-support self-attention forms content-dependent representations before functional compression.
- **Functional analysis and reusable inference.** We establish the semantics of the retained coefficients and their finite query space, derive a geometry–memory error decomposition, and study the resulting representation across reconstruction, classification, extrapolation, and repeated-query evaluation.

2 Related Work

Irregular Time-Series Modeling. Continuous-time models handle irregular observations through latent dynamics or time-aware interactions, including Latent ODEs, Neural CDEs, CRU, mTAN, and ContiFormer (Rubanova et al., 2019; Kidger et al., 2020; Schirmer et al., 2022; Shukla & Marlin, 2021; Chen et al., 2023). Recent forecasting methods instead reorganize irregular observations through warping, patches, graphs, or point-based representations (Zhang et al., 2023; Yalavarthi et al., 2024; Zhang et al., 2024; Hu et al., 2025; Liu et al., 2026b;a). Frequency-based modeling provides another way to summarize irregular observations, as in TFMixer (Qiu et al., 2026). HiPPO-FM differs in the representation retained after encoding: it stores a coordinate-queryable functional state rather than an observation-indexed collection of tokens, patches, graph nodes, or latent trajectories.

Compact and Functional Memory. Linear attention and memory-efficient sequence architectures compress sequence interactions into compact statistics or latent states (Katharopoulos et al., 2020; Jaegle et al., 2022; Feng et al., 2024). HiPPO represents signal histories through structured orthogonal polynomial projections (Gu et al., 2020), while interdomain kernel methods define function-valued features through linear functionals (Lazaro-Gredilla & Figueiras-Vidal, 2009). Recent functional attention and memory-inspired operators explore basis-coefficient representations beyond token-level attention (Kiyohara et al., 2026; Xiao et al., 2026). HiPPO-FM differs in using these coordinate functions to define both a retained functional state and its direct temporal readout; Section 3 develops this operator and its representation properties.

3 Methodology

3.1 Problem Setup and Overview

Irregular observations and queries. An irregular observation history is

$$\mathcal{D}_P = \{(t_i, x_i, m_i)\}_{i=1}^N, \qquad P = \{t_i\}_{i=1}^N, \qquad R = \{r_a\}_{a=1}^Q, \tag{1}$$

where t_i is an observed coordinate, $x_i \in \mathbb{R}^{d_x}$ is the corresponding measurement, m_i optionally records channel validity, and R contains the requested coordinates. Coordinate-valued tasks produce $\widehat{Y}(R) \in \mathbb{R}^{Q \times d_y}$. Interpolation queries lie within the observed interval, whereas extrapolation queries extend beyond it. Forecasting encodes only the permitted history.

Shared temporal coordinates. Observations and queries use one observation-derived affine map,

$$a_P(t) = \frac{t - t_{\min}}{s_P}, \qquad t_{\min} = \min P, \qquad s_P = \max P - t_{\min}, \tag{2}$$

where the extrema are computed over valid observed coordinates for a nondegenerate history. The same transform is reused for every query set, and future coordinates may exceed one. We suppress a_P in the remaining feature-map notation.

Functional-memory architecture. HiPPO-FM separates history encoding from coordinate evaluation:

$$S_P = E(\mathcal{D}_P), \qquad \widehat{y}(r) = D(S_P, a_P(r)). \tag{3}$$

As shown in Figure 1, the observed-support encoder applies MHA-to-FM on $P \to P$. For coordinate-valued tasks, an independently parameterized output FM head writes $H_{\mathrm{enc}}(P)$ into the final functional state and evaluates that state on R. Classification instead pools the observed-support representation directly.

3.2 Coordinate-Queryable Functional Memory

An FM operator receives a representation $H \in \mathbb{R}^{N\times d}$ defined on P, writes its projected values into functional memory, and evaluates the retained coefficients on R. We use H_{head} heads of width d_v, with $d = H_{\mathrm{head}} d_v$.

Functional coordinate geometry. We use the Legendre polynomial family underlying HiPPO-LegS (Gu et al., 2020):

$$\phi_m(s) = \sqrt{2m+1}\, P_m(2s-1), \qquad m = 0, \dots, M-1, \tag{4}$$

where P_m is the degree-m Legendre polynomial on $[-1, 1]$. Let $S = \{s_g\}_{g=1}^{G} \subset [0, 1]$ be a fixed reference grid, and set $\Phi_{g,m} = \phi_m(s_g)$, with $\Phi \in \mathbb{R}^{G\times M}$. The polynomial modes provide the reference basis for the functional features evaluated at observed and requested coordinates. For a positive-semidefinite kernel k_θ, define

$$\Psi_\theta(X) = \frac{1}{G} K_\theta(X, S)\Phi, \qquad \psi_\theta(r) = \frac{1}{G}\Phi^\top K_\theta(S, r). \tag{5}$$

Each row of $\Psi_\theta(X)$ is $\psi_\theta(x)^\top$. The corresponding functional Gram matrix is

$$K_u = \frac{1}{G^2}\Phi^\top K_\theta(S, S)\Phi, \qquad \bar{K}_u = K_u + \epsilon I_M, \tag{6}$$

where $\epsilon > 0$ is numerical jitter. The same functional geometry is used for both memory writing and coordinate readout.

Functional write. For head h, an affine value projection produces $V^h \in \mathbb{R}^{N\times d_v}$, followed by aggregation over valid observed positions:

$$V^h = HW_V^h + \mathbf{1}_N (b_V^h)^\top, \qquad \widetilde{M}^h = \Psi_\theta(P)^\top V^h. \tag{7}$$

The contraction replaces the observation dimension N with the functional dimension M. We normalize the aggregated memory and solve for its retained coefficients:

$$M^h = \mathrm{LN}_{M\times d_v}(\widetilde{M}^h), \qquad \bar{K}_u B_P^h = M^h. \tag{8}$$

The Gram solve corrects for correlations and scaling among the kernel-smoothed coordinate functions and yields $B_P^h \in \mathbb{R}^{M\times d_v}$. We use a Cholesky factorization followed by triangular solves rather than forming an explicit inverse.

Coordinate read. Each head evaluates its retained coefficients at the requested coordinates through the same functional features:

$$O^h(R) = \Psi_\theta(R) B_P^h. \tag{9}$$

We denote the resulting multi-head operator by $\mathrm{FM}_R(H; P)$; head aggregation and output projection are specified in Appendix B.

3.3 Content-Conditioned Encoding and Output Query

Content interaction before compression. A finite-rank functional write may map different representations to the same memory state. To preserve content-dependent interactions before compression, HiPPO-FM applies self-attention on the observed support before functional projection. The resulting representation is then compressed into functional memory, while the coordinate geometry remains unchanged.

Our default MHA-to-FM encoder applies full masked self-attention over the valid observed support and writes the content-updated representation into the encoder FM:

$$\begin{aligned} L(H) &= \lambda\ \mathrm{MHA}_P(H), \\ A(H) &= L(H) + \mathrm{FM}_P^{\mathrm{enc}}\big(H + L(H); P\big). \end{aligned} \tag{10}$$

Here λ is a learned scalar. Self-attention forms content-dependent combinations of the observed representations, and the FM stores their projections in functional coordinates. The requested support R enters only through the output decoder. For forecasting, the encoder receives only the permitted observation history.

Observed-support encoder. We embed the inner operator in a single post-normalization Transformer block. Suppressing dropout notation,

$$\begin{aligned} Z(P) &= \mathrm{LN}\big(H(P) + A(H(P))\big), \\ H_{\mathrm{enc}}(P) &= \mathrm{LN}\big(Z(P) + \mathrm{FFN}(Z(P))\big). \end{aligned} \tag{11}$$

The input representation $H(P)$ is obtained by projecting observations with missing channels zero-filled. Validity masks remove invalid time positions from attention and functional contraction. The normalizations in Equation (11) act on token representations and are separate from the post-contraction memory normalization in Equation (8). Appendix B gives the architecture and masking details.

Task-specific readout. For coordinate-valued tasks, an independently parameterized output FM head maps the encoded history to the requested support:

$$\widehat{Y}(R) = \mathrm{Head}\big(\mathrm{FM}_R^{\mathrm{out}}\big(H_{\mathrm{enc}}(P); P\big)\big)\,. \tag{12}$$

Here Head is an affine map applied independently at each requested coordinate. For UEA classification, we instead apply last-valid pooling to $H_{\mathrm{enc}}(P)$, followed by a classification head.

3.4 Functional Semantics and Approximation

The write/read pair defines a finite-dimensional functional query space. Theorem 3.3 characterizes the resulting approximation behavior by separating functional geometry error from learned-memory mismatch.

The functional write and coordinate read are defined in the same kernel-smoothed basis. The Gram solve converts the stored functional values into synthesis coefficients, so the same functional geometry governs observation writing and coordinate evaluation.

Raw-memory semantics. For each mode m, define the discrete linear functional and its RKHS representer

$$\mathcal{L}_m^G f = \frac{1}{G}\sum_{g=1}^{G} \phi_m(s_g) f(s_g), \qquad h_m(\cdot) = \frac{1}{G}\sum_{g=1}^{G} \phi_m(s_g) k_\theta(\cdot, s_g). \tag{13}$$

Then $h_m(r) = \psi_{\theta,m}(r)$ and $[K_u]_{mn} = \langle h_m, h_n\rangle_{\mathcal{H}_{k_\theta}}$, so the coordinate features and Gram matrix refer to the same family of functions.

Proposition 3.1 (Raw-memory semantics)**.** *For scalar projected values $v \in \mathbb{R}^N$, let $g_{P,v}(\cdot) = \sum_{i=1}^N v_i k_\theta(\cdot, t_i)$. Every entry of the unnormalized memory satisfies*

$$\big[\Psi_\theta(P)^\top v\big]_m = \sum_{i=1}^{N} \psi_{\theta,m}(t_i) v_i = \langle g_{P,v}, h_m\rangle_{\mathcal{H}_{k_\theta}} = \mathcal{L}_m^G g_{P,v}. \tag{14}$$

The raw write therefore stores interdomain functional values of the empirical kernel field induced by the observations. Memory normalization changes these values before the Gram solve; the coordinate functions remain fixed.

Matched readout and finite query space. Write $A_\epsilon = (K_u + \epsilon I_M)^{-1}$ and $\mathcal{U}_M = \mathrm{span}\{h_0, \ldots, h_{M-1}\}$. The inverse is defined for $\epsilon > 0$, and also for $\epsilon = 0$ when K_u is nonsingular.

Proposition 3.2 (Matched readout and finite query space)**.** *For fixed kernel, basis, and reference grid, every scalar readout $r \mapsto \psi_\theta(r)^\top A_\epsilon m$, with $m \in \mathbb{R}^M$, belongs to $\mathcal{U}_M$. For exact functional values $u_m = \mathcal{L}_m^G f$, zero jitter, and nonsingular K_u,*

$$\psi_\theta(r)^\top K_u^{-1} u = \big(\Pi_{\mathcal{U}_M} f\big)(r), \qquad f \in \mathcal{H}_{k_\theta}. \tag{15}$$

The Gram inverse maps functional values to the coefficients of their RKHS projection. For an arbitrary learned memory, the coordinate readout remains in a space of dimension at most M. Shared geometry across heads and linear channel mixing preserve this span; affine biases add a constant function. Each scalar output function of the affine coordinate-prediction head therefore lies in $\mathcal{U}_M + \mathrm{span}\{1\}$. The observed-support encoder retains a separate nonlinear residual path.

Coverage and learned-memory mismatch. The functional geometry also separates approximation capacity from learned-memory quality. Define

$$c_\epsilon(r) = k_\theta(r, r) - \psi_\theta(r)^\top A_\epsilon \psi_\theta(r) \geq 0. \tag{16}$$

Theorem 3.3 (Pointwise functional approximation)**.** *Let $f \in \mathcal{H}_{k_\theta}$, let $u_m = \mathcal{L}_m^G f$, and let $\widehat{u} \in \mathbb{R}^M$ be any memory supplied to the decoder. For $\epsilon > 0$, or $\epsilon = 0$ with nonsingular K_u,*

$$\left|f(r) - \psi_\theta(r)^\top A_\epsilon \widehat{u}\right| \leq \|f\|_{\mathcal{H}_{k_\theta}} \sqrt{c_\epsilon(r)} + \left\|\psi_\theta(r)^\top A_\epsilon\right\|_2 \|\widehat{u} - u\|_2 . \tag{17}$$

The first term bounds approximation error from the functional geometry at r. The second measures mismatch between the learned and exact functional values, scaled by the readout norm. The theorem separates limitations caused by the functional coordinate system from those caused by imperfect memory estimation.

For fixed geometry, pre-write content interaction and memory normalization affect the learned-memory term while preserving the query space. The theorem applies to a fixed geometry; separately trained configurations may also differ through their learned kernels and optimization trajectories. Query-set consistency is verified in Appendix E.

3.5 Reusable Functional States

After encoding, HiPPO-FM retains a history-specific functional state $S_P = (B_P, t_{\min}, s_P)$, which is independent of the requested query set. A new query set can therefore be evaluated directly through the coordinate readout without recomputing the observation history. This separates history encoding from coordinate evaluation and allows the same functional state to support multiple downstream query resolutions.

The persistent storage of the functional state is $O(Md)$, independent of the number of observed timestamps N. For a new query set with Q points, the readout requires only functional feature construction and coefficient evaluation, while the encoder, functional write, and geometry computation are performed once and reused across queries. Detailed complexity accounting and implementation details are provided in Appendix B.

4 Experiments

We evaluate HiPPO-FM on continuous-coordinate reconstruction, irregular representation learning, and $P \neq R$ extrapolation. An auxiliary event-sequence benchmark from the ContiFormer reference setting is included in Section 4.3. Controlled analyses of the retained functional state are reported in the appendices. Unless stated otherwise, HiPPO-FM uses one observed-support encoder block, four heads, width 64, shifted-Legendre functional coordinates, and a learned squared-exponential kernel. Full preprocessing, optimization, model-selection rules, and detailed results are provided in Appendices A–D.

Table 1: Interpolation and extrapolation results of different models on 2-dimensional spirals. Values are mean ± sample SD over five seeds. Lower is better; bold marks the lowest available raw mean in each row. Full α results are in Appendix D.2.

α	Metric	Transformer	Neural ODE	Latent ODE	Neural CDE	ContiFormer	HiPPO-FM
Task = Interpolation ($\times 10^{-2}$)							
0.02	RMSE (↓)	0.95 ± 0.21	1.03 ± 0.19	1.16 ± 0.25	1.32 ± 0.95	0.69 ± 0.16	$\mathbf{0.42 \pm 0.08}$
	MAE (↓)	0.95 ± 0.17	1.01 ± 0.14	1.14 ± 0.20	1.31 ± 0.91	0.70 ± 0.13	$\mathbf{0.42 \pm 0.09}$
0.2	RMSE (↓)	2.09 ± 0.72	1.89 ± 0.25	2.50 ± 0.94	2.29 ± 1.59	$\mathbf{0.58 \pm 0.14}$	0.61 ± 0.22
	MAE (↓)	1.74 ± 0.29	1.65 ± 0.10	2.04 ± 0.47	1.89 ± 1.17	0.55 ± 0.11	$\mathbf{0.51 \pm 0.11}$
2	RMSE (↓)	2.49 ± 0.41	2.18 ± 0.14	2.32 ± 0.39	2.24 ± 0.94	2.06 ± 2.39	$\mathbf{0.69 \pm 0.22}$
	MAE (↓)	2.32 ± 0.25	2.10 ± 0.18	2.14 ± 0.25	1.83 ± 0.40	1.83 ± 1.98	$\mathbf{0.60 \pm 0.15}$
Task = Extrapolation ($\times 10^{-2}$)							
0.02	RMSE (↓)	1.29 ± 0.11	1.24 ± 0.07	1.18 ± 0.08	2.39 ± 2.73	0.88 ± 0.13	$\mathbf{0.62 \pm 0.20}$
	MAE (↓)	1.29 ± 0.10	1.16 ± 0.09	1.12 ± 0.06	2.32 ± 2.57	0.87 ± 0.10	$\mathbf{0.57 \pm 0.18}$
0.2	RMSE (↓)	3.47 ± 1.45	4.13 ± 0.94	4.72 ± 0.43	2.84 ± 2.50	0.93 ± 0.35	$\mathbf{0.67 \pm 0.23}$
	MAE (↓)	3.02 ± 1.10	3.28 ± 0.68	3.68 ± 0.23	2.46 ± 2.40	0.90 ± 0.35	$\mathbf{0.61 \pm 0.13}$
2	RMSE (↓)	5.23 ± 1.08	4.18 ± 1.03	4.94 ± 1.05	3.70 ± 2.03	3.11 ± 2.89	$\mathbf{0.80 \pm 0.17}$
	MAE (↓)	4.48 ± 0.86	3.61 ± 0.82	3.97 ± 0.74	3.19 ± 1.58	2.69 ± 2.39	$\mathbf{0.73 \pm 0.08}$

4.1 Continuous-Coordinate Reconstruction

We first study continuous-coordinate reconstruction on the two-dimensional Spiral benchmark used by ContiFormer (Chen et al., 2023). We compare the standard Transformer, Neural ODE, Latent ODE, and Neural CDE baselines with ContiFormer and HiPPO-FM, using the original references of the corresponding baselines (Vaswani et al., 2017; Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020). The benchmark contains noisy irregular observations of dense trajectories; we evaluate interpolation on the observed half and extrapolation on the withheld half using a fixed five-seed protocol. RMSE and MAE are reported in Table 1, with the complete settings in Appendix D.2.

Figure 2(a) shows the seed-42 trajectories from the best recorded checkpoints of the 5,000-iteration runs. Dashed boxes 1–3 align the inner turn, right endpoint, and lower-left outer turn across the three methods, making their local reconstruction differences easier to compare.

Table 1 reports the consolidated α sweep. HiPPO-FM has the lowest mean in most rows and remains competitive in the remaining conditions, while ContiFormer is strongest in one interpolation

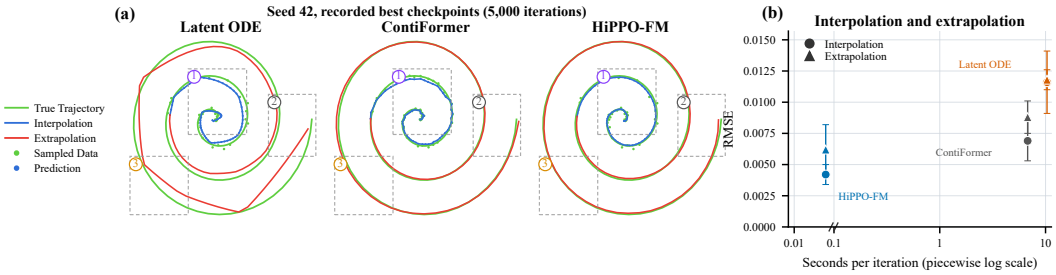

Figure 2: **Spiral 2D qualitative and runtime comparison.** **(a)** Interpolation and extrapolation on irregularly sampled spirals by Latent ODE, ContiFormer, and HiPPO-FM. **(b)** Reconstruction accuracy versus runtime for the same methods; lower-left is better.

Table 2: Irregular UEA classification results. Avg. ACC. stands for average accuracy over 20 datasets and Avg. Rank stands for average ranking over 20 datasets. The overall ACC and rank summarize the three observation-drop settings using equal-weight averages. ↑ (↓) indicates higher (lower) is better. Full results are in Appendix D.1.

Mask Ratio	30% Dropped		50% Dropped		70% Dropped		Overall (equal-drop)	
Metric	Avg. ACC. (↑)	Avg. Rank (↓)	Avg. ACC. (↑)	Avg. Rank (↓)	Avg. ACC. (↑)	Avg. Rank (↓)	Overall ACC (↑)	Overall Rank (↓)
GRU-D	0.7140	5.50	0.7080	5.88	0.6793	6.14	0.7004	9.00
GRU-Δt	0.7255	5.35	0.7124	5.68	0.6876	5.10	0.7085	7.00
ODE-RNN	0.7211	5.75	0.7150	5.23	0.6949	5.20	0.7104	6.00
Neural CDE	0.5156	9.18	0.5174	9.00	0.5028	8.90	0.5119	11.00
mTAN	0.5587	8.58	0.5447	9.23	0.5219	8.80	0.5418	10.00
CADN	0.7345	5.38	0.7251	4.85	0.6859	5.38	0.7152	4.00
S5	**0.7497**	4.70	0.7307	4.73	0.7012	5.08	0.7272	2.00
TST	0.7378	4.78	0.7130	5.28	0.6857	5.63	0.7122	5.00
ContiFormer	0.7257	5.18	0.7235	**4.43**	0.7010	**4.53**	0.7167	3.00
t-PatchGNN	0.7314	5.88	0.6977	6.47	0.6756	6.36	0.7011	8.00
HiPPO-FM	0.7496	**4.60**	**0.7364**	4.50	**0.7058**	4.55	**0.7306**	**1.00**

condition. The full generator, sampling procedure, and baseline implementation details are given in Appendix D.2.

4.2 Irregular Representation Learning

We evaluate irregular representation learning on the 20-dataset UEA suite used by ContiFormer (Chen et al., 2023). The protocol removes different proportions of observations while retaining their timestamps and compares recurrent, continuous-time, state-space, and attention-based baselines under the reproduced data split. The comparison includes GRU-D, GRU-Δt (Che et al., 2018), ODE-RNN, CADN, Neural CDE, mTAN, S5, TST, and ContiFormer (Che et al., 2018; Rubanova et al., 2019; Chien & Chen, 2023; Kidger et al., 2020; Shukla & Marlin, 2021; Smith et al., 2023; Zerveas et al., 2021; Chen et al., 2023). The additional t-PatchGNN accuracy values are included where the benchmark reports a cell, as an additional baseline (Zhang et al., 2024).

Table 2 summarizes the three observation-drop settings with t-PatchGNN included whenever it reports a cell. HiPPO-FM provides the strongest overall accuracy and rank under the equal-drop aggregation, with especially strong performance as the masking ratio increases. S5 remains competitive at light masking, while ContiFormer leads the per-setting rank at the two larger drop ratios. Full per-dataset values and reporting caveats are given in Appendix D.1.

UEA classification pools the observed-support representation and bypasses the final $P \to R$ functional readout. This experiment therefore measures the history encoder and retained representation. Appendix D.1 reports all per-dataset results.

4.3 Predicting Irregular Event Sequences

We further evaluate next-event prediction on irregular marked event sequences. Given an ordered history of event times and marks, a model predicts both the mark and occurrence time of the next event. We use one synthetic and five real-world datasets (Neonate, Traffic, MIMIC, StackOverflow, and BookOrder), with four-fold cross-validation for Synthetic, Neonate, and Traffic, and five-fold cross-validation for the remaining datasets. We report log-likelihood (LL), event-type accuracy, and next-event-time RMSE; additional protocol details are provided in Appendix A.3.

Table 3 compares classical Hawkes, recurrent marked point-process, neural Hawkes, and self-attentive or Transformer Hawkes models (Hawkes, 1971; Du et al., 2016; Mei & Eisner, 2017; Zhang et al., 2020; Zuo et al., 2020; Ma et al., 2022). HiPPO-FM is strongest on several displayed datasets, especially in event-type accuracy and next-event-time RMSE, showing that its functional state can also support marked event prediction.

Table 3: Prediction results for irregular event sequences. LL denotes log-likelihood and ACC denotes event-type accuracy. ↑ (↓) indicates that higher (lower) is better. Bold values mark the best displayed performance.

Model	Metric	Synthetic	Neonate	Traffic	MIMIC	StackOverflow	BookOrder
HP	LL (↑)	-3.084 ± 0.005	-4.618 ± 0.005	-1.482 ± 0.005	-4.618 ± 0.005	-5.794 ± 0.005	-1.036 ± 0.000
	Accuracy (↑)	0.756 ± 0.000	–	0.570 ± 0.000	0.795 ± 0.000	0.441 ± 0.000	0.604 ± 0.000
	RMSE (↓)	0.953 ± 0.000	10.957 ± 0.012	0.407 ± 0.000	1.021 ± 0.000	1.341 ± 0.000	3.781 ± 0.000
RMTPP	LL (↑)	-1.025 ± 0.030	-2.817 ± 0.023	-0.546 ± 0.012	-1.184 ± 0.023	-2.374 ± 0.001	-0.952 ± 0.007
	Accuracy (↑)	0.841 ± 0.000	–	0.805 ± 0.002	0.823 ± 0.004	0.461 ± 0.000	0.624 ± 0.000
	RMSE (↓)	0.369 ± 0.014	9.517 ± 0.023	0.337 ± 0.001	0.864 ± 0.017	0.955 ± 0.000	3.647 ± 0.003
NeuralHP	LL (↑)	-1.371 ± 0.004	-2.795 ± 0.012	-0.643 ± 0.004	-1.239 ± 0.027	-2.608 ± 0.000	-1.104 ± 0.005
	Accuracy (↑)	0.841 ± 0.000	–	0.759 ± 0.001	0.814 ± 0.001	0.450 ± 0.000	0.621 ± 0.000
	RMSE (↓)	0.631 ± 0.002	9.614 ± 0.013	0.358 ± 0.001	0.846 ± 0.007	1.022 ± 0.000	3.734 ± 0.003
SAHP	LL (↑)	-0.619 ± 0.063	-2.646 ± 0.057	$\mathbf{0.372 \pm 0.022}$	-1.110 ± 0.030	-2.404 ± 0.002	$\mathbf{-0.304 \pm 0.002}$
	Accuracy (↑)	0.841 ± 0.000	–	0.780 ± 0.001	0.830 ± 0.004	0.455 ± 0.000	0.622 ± 0.000
	RMSE (↓)	0.521 ± 0.055	9.403 ± 0.060	0.337 ± 0.000	0.851 ± 0.006	0.963 ± 0.000	3.676 ± 0.011
THP	LL (↑)	-0.589 ± 0.017	-2.702 ± 0.045	-0.569 ± 0.015	-1.137 ± 0.038	-2.354 ± 0.001	-1.102 ± 0.052
	Accuracy (↑)	0.841 ± 0.000	–	0.818 ± 0.001	0.834 ± 0.005	-0.468 ± 0.000	0.622 ± 0.004
	RMSE (↓)	0.205 ± 0.006	9.471 ± 0.045	0.332 ± 0.000	0.843 ± 0.017	0.951 ± 0.000	3.688 ± 0.004
NSTKA	LL (↑)	-1.001 ± 0.025	-2.747 ± 0.061	-0.667 ± 0.027	-1.188 ± 0.012	-2.406 ± 0.003	-0.919 ± 0.019
	Accuracy (↑)	0.842 ± 0.000	–	0.767 ± 0.001	0.833 ± 0.003	0.465 ± 0.000	0.621 ± 0.000
	RMSE (↓)	0.379 ± 0.015	9.502 ± 0.049	0.346 ± 0.001	$\mathbf{0.842 \pm 0.008}$	0.956 ± 0.000	3.731 ± 0.007
GRU-Δt	LL (↑)	-0.871 ± 0.050	-2.736 ± 0.031	-0.613 ± 0.062	-1.164 ± 0.026	-2.389 ± 0.002	-0.915 ± 0.006
	Accuracy (↑)	0.841 ± 0.000	–	0.800 ± 0.004	0.832 ± 0.007	0.466 ± 0.000	$\mathbf{0.627 \pm 0.000}$
	RMSE (↓)	0.249 ± 0.013	9.421 ± 0.050	0.335 ± 0.001	0.850 ± 0.010	0.950 ± 0.000	3.666 ± 0.016
ODE-RNN	LL (↑)	-1.032 ± 0.102	-2.732 ± 0.080	-0.491 ± 0.011	-1.183 ± 0.028	-2.395 ± 0.001	-0.988 ± 0.006
	Accuracy (↑)	0.841 ± 0.000	–	0.812 ± 0.000	0.827 ± 0.006	0.467 ± 0.000	0.624 ± 0.000
	RMSE (↓)	0.342 ± 0.034	$\mathbf{9.289 \pm 0.048}$	0.334 ± 0.000	0.865 ± 0.021	0.952 ± 0.000	$\mathbf{3.605 \pm 0.004}$
mTAN	LL (↑)	-0.920 ± 0.036	-2.722 ± 0.026	-0.548 ± 0.023	-1.149 ± 0.029	-2.391 ± 0.002	-0.980 ± 0.004
	Accuracy (↑)	0.842 ± 0.000	–	0.811 ± 0.002	0.832 ± 0.009	0.466 ± 0.000	0.620 ± 0.000
	RMSE (↓)	0.286 ± 0.008	9.363 ± 0.042	0.334 ± 0.001	0.848 ± 0.006	0.950 ± 0.000	3.680 ± 0.015
ContiFormer	LL (↑)	-0.617 ± 0.031	-2.700 ± 0.031	-0.628 ± 0.025	-1.162 ± 0.028	-2.450 ± 0.003	-0.310 ± 0.014
	Accuracy (↑)	0.841 ± 0.002	–	0.799 ± 0.004	0.820 ± 0.007	0.458 ± 0.002	0.613 ± 0.003
	RMSE (↓)	$\mathbf{0.188 \pm 0.006}$	9.472 ± 0.040	0.350 ± 0.004	0.972 ± 0.012	1.010 ± 0.002	3.750 ± 0.025
HiPPO-FM	LL (↑)	$\mathbf{-0.2239 \pm 0.0683}$	$\mathbf{-2.6214 \pm 0.3164}$	0.0844 ± 0.1064	$\mathbf{-0.8632 \pm 0.2641}$	$\mathbf{-1.1649 \pm 0.0093}$	-0.9205 ± 0.1028
	Accuracy (↑)	$\mathbf{0.9215 \pm 0.0080}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9532 \pm 0.0040}$	$\mathbf{0.8731 \pm 0.0314}$	$\mathbf{0.7481 \pm 0.0029}$	$\mathbf{0.6508 \pm 0.0071}$
	RMSE (↓)	0.2182 ± 0.0142	9.5208 ± 4.4988	0.2494 ± 0.0466	0.8868 ± 0.1281	$\mathbf{0.6577 \pm 0.0039}$	$\mathbf{3.5624 \pm 0.4127}$

5 Conclusion

We introduced HiPPO-FM, a functional-memory attention operator that maps irregular observations to a fixed-rank, coordinate-addressable state. HiPPO-derived functional coordinates and matched kernel geometry define both memory construction and arbitrary-coordinate readout, with content interaction applied on the observed support before compression. The resulting state has $O(Md)$ post-encoding storage and can be reused across query sets. Experiments evaluate this representation through Spiral reconstruction, 20 UEA classification datasets, PhysioNet extrapolation, and repeated-query diagnostics. The method therefore provides an explicit interface for encoding an irregular history once and evaluating the retained functional state at requested coordinates. Across these experiments, it performs strongly on controlled reconstruction and PhysioNet extrapolation, remains competitive on UEA, and supports consistent repeated querying.

Observed-support self-attention still has quadratic encoding cost, while the functional write can be sensitive to sampling density and finite-rank geometry. Future work should study lighter encoders and density-aware or adaptive functional bases.

AI use statement

In this work, we used generative AI tools, including ChatGPT and Codex, as auxiliary tools for literature organization, code debugging, mathematical expression refinement, manuscript language improvement, and proofreading. AI tools were not used to generate the core research ideas, design the proposed methodology, formulate the experimental protocols, create or modify experimental data, or determine scientific conclusions.

All methodological decisions, theoretical analyses, experimental implementations, result interpretations, and final manuscript contents were developed and verified by the authors. The authors

carefully reviewed all AI-assisted outputs, including citations, code, figures, and text, and take full responsibility for the accuracy and integrity of the submitted work.

REPRODUCIBILITY STATEMENT

Appendix A records the preprocessing, training, selection, and metric conventions; Appendix B specifies the encode/query contract and architecture. Appendix C gives the theoretical analysis, Appendix D contains detailed results, and Appendix E reports query-set reuse.

ETHICS STATEMENT

The clinical experiments concern retrospective benchmark prediction, not clinical decision-making or deployment. Lower reconstruction error does not establish medical utility, calibration, fairness across patient groups, or safety. The use of evaluation-split statistics in the inherited CRU preprocessing is disclosed and limits deployment-oriented interpretation. Any subsequent clinical use would require independent validation, appropriate data governance, and domain-specific review.

REFERENCES

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1), 2018. URL https://www.nature.com/articles/s41598-018-24271-9.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. ContiFormer: Continuous-time transformer for irregular time series modeling. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2402.10635.

Jen-Tzung Chien and Yi-Hsiang Chen. Learning continuous-time dynamics with attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1906–1918, 2023. doi: 10.1109/TPAMI.2022.3162711. URL https://doi.org/10.1109/TPAMI.2022.3162711.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564, 2016. doi: 10.1145/2939672.2939875. URL https://doi.org/10.1145/2939672.2939875.

Leo Feng, Frederick Tung, Hossein Hajimirsadeghi, Yoshua Bengio, and Mohamed Osama Ahmed. Memory efficient neural processes via constant memory attention block. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13365–13386, 2024. URL https://proceedings.mlr.press/v235/feng24i.html.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/102f0bb6efb3a6128a3c750dd16729be-Abstract.html.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. doi: 10.1093/biomet/58.1.83.

Zhangyi Hu, Jiemin Wu, Hua Xu, Mingqian Liao, Ninghui Feng, Bo Gao, Songning Lai, and Yutao Yue. IMTS is worth time $\times$ channel patches: Visual masked autoencoders for irregular multivariate time series prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24805–24822. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25y.html.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Henaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs and outputs. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.14795.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Naoki Kiyohara, Harrison Bo Hua Zhu, Riccardo El Hassanin, Zhuo Sun, Wenlong Chen, Samir Bhatt, and Yingzhen Li. Interdomain attention: Beyond token-level key-value memory. *arXiv preprint arXiv:2605.24330*, 2026. URL https://arxiv.org/abs/2605.24330.

Miguel Lazaro-Gredilla and Anibal R. Figueiras-Vidal. Inter-domain gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. ASTGI: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/28d38c036365420f61ce03300418e44a-Abstract-Conference.html.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23873–23881, 2026b. doi: 10.1609/aaai.v40i28.39563. URL https://doi.org/10.1609/aaai.v40i28.39563.

Yu Ma, Zhining Liu, Chenyi Zhuang, Yize Tan, Yi Dong, Wenliang Zhong, and Jinjie Gu. Non-stationary time-aware kernelized attention for temporal event prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1224–1232, 2022. doi: 10.1145/3534678.3539470. URL https://doi.org/10.1145/3534678.3539470.

Hongyuan Mei and Jason M. Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6463c88460bd63bbe256e495c63aa40b-Abstract.html.

Xiangfei Qiu, Kangjia Yan, Xvyuan Liu, Xingjian Wu, and Jilin Hu. Bridging time and frequency: A joint modeling framework for irregular multivariate time series forecasting. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.00582.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ODEs for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 19388–19405. PMLR, 2022. URL https://proceedings.mlr.press/v162/schirmer22a.html.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.10318.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.04933.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Jiefang Xiao, Maolin Gao, Simon Weber, Guandao Yang, and Daniel Cremers. Functional attention: From pairwise affinities to functional correspondences. *arXiv preprint arXiv:2605.31559*, 2026. URL https://arxiv.org/abs/2605.31559.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. GraFITi: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024. doi: 10.1609/aaai.v38i15.29560. URL https://doi.org/10.1609/aaai.v38i15.29560.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021. URL https://arxiv.org/abs/2010.02803.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3273–3285. ACM, 2023. doi: 10.1145/3580305.3599543. URL https://doi.org/10.1145/3580305.3599543.

Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-attentive hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11183–11193. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhang20q.html.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 60179–60196. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24bw.html.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11692–11702. PMLR, 2020. URL https://proceedings.mlr.press/v119/zuo20a.html.

A EXPERIMENTAL SETUP AND REPRODUCIBILITY

This appendix records the common data, training, selection, and evaluation conventions used by the main-text experiments. Unless a task-specific exception is stated, all comparisons use the same preprocessing and evaluation protocol for every method.

A.1 UEA IRREGULAR TIME-SERIES CLASSIFICATION

We use the official TRAIN/TEST partitions for the 20 UEA datasets in Table 4. For each dataset, 30%, 50%, or 70% of the observations are removed at random while retained values remain paired with their original timestamps. All reproduced methods use seeds $\{27, 42, 1024\}$ and the registered `train_es` protocol: the complete official TRAIN split is used for optimization, training lasts at most 1,000 epochs, and it stops when training accuracy has not improved for 100 consecutive epochs. Classification is performed directly from the observed-support representation and does not use the final coordinate readout. The same registered training, masking, and evaluation protocol is used for all three observation-drop settings. The TEST split is reserved for final reporting and is not used for checkpoint selection.

The final 50% MHA-to-FM archive is complete for all 20 datasets and three seeds. The baseline columns follow the established full 20-dataset `train_es` report. Cells marked † retain that report's declared exception: TST on LSST uses its registered reporting substitution, and Neural CDE on HandMovementDirection at 30% and 50% uses the finite checkpoint accuracy from runs whose recorded loss was nonfinite. No such exception is used for HiPPO-FM.

A.1.1 DATASET INFORMATION

Table 4: Dataset information for the 20-dataset UEA classification suite.

Dataset	Code	# Train	# Test	Length	# Classes	# Variables
ArticularyWordRecognition	AWR	275	300	144	25	9
BasicMotions	BM	40	40	100	4	6
CharacterTrajectories	CT	1422	1436	182	20	3
DuckDuckGeese	DDG	60	40	270	5	1345
Epilepsy	EP	137	138	207	4	3
ERing	ER	30	270	65	6	4
FingerMovements	FM	316	100	50	2	28
HandMovementDirection	HMD	160	74	400	4	10
Handwriting	HW	150	850	152	26	3
Heartbeat	HB	204	205	405	2	61
JapaneseVowels	JV	270	370	29	9	12
Libras	LB	180	180	45	15	2
LSST	LSST	2459	2466	36	14	6
NATOPS	NATOPS	180	180	51	6	24
PEMS-SF	PEMS	267	173	144	7	963
PenDigits	PD	7494	3498	8	10	2
RacketSports	RS	151	152	30	4	6
SelfRegulationSCP1	SR	268	293	896	2	6
SpokenArabicDigits	SA	6599	2199	93	10	13
UWaveGestureLibrary	UGL	2238	2241	315	8	3

A.2 SPIRAL RECONSTRUCTION

The diagnostic uses 300 trajectories, a 200/100 training/testing split, 150 equally spaced time positions, and 30 observations sampled from the first half. The registered noise scale is 0.1 and the main setting uses $\alpha = 0.02$. Data scaling is computed from the training trajectories. Adam uses learning rate 10^{-3}, dropout 0, clipping norm 1, a maximum of 1,000 epochs, and validation-MSE patience 10. Five seeds $\{13, 27, 42, 100, 2024\}$ are used. The complete α sweep and qualitative visualization are reported in Appendix D.2.

A.3 IRREGULAR EVENT-SEQUENCE PREDICTION

Each example is an ordered marked sequence. The model receives the history up to the current event and predicts the mark and time of the next event. We use one synthetic and five real-world

Algorithm 1. Encode once, query many times

Input: observed history (x_P, t_P, m_P) and a query set R.
Output: reusable state $\mathcal{S}_P = (B_P, t_{\min}, s_P)$.
Determine valid observed positions and the observed-support coordinate transform.
Encode the observed support with the MHA-to-FM encoder.
Write the encoded values into the output FM and solve for B_P.
Discard observation-token activations after returning $\mathcal{S}_P$.
For each query set R **do**
  Apply the cached coordinate transform to R and evaluate $\Psi(R)$.
  Compute $O(R) = \Psi(R)B_P$ and apply the output projection.
  Return predictions in the caller's query order.

Table 5: Encode/query decomposition for reusable-state evaluation. The observation encoder, memory write, and linear solve are performed once and are not repeated for subsequent query sets.

datasets: Neonate, Traffic, MIMIC, StackOverflow, and BookOrder. Synthetic, Neonate, and Traffic use four-fold cross-validation; MIMIC, StackOverflow, and BookOrder use five folds. The reported metrics are next-event log-likelihood, event-type accuracy, and next-event-time RMSE. Dataset sizes, event cardinalities, and split details are given in Appendix D.3.

A.4 PHYSIONET EXTRAPOLATION

We follow the released Continuous-Recurrent-Units preprocessing and evaluation setup of Schirmer et al. (2022). PhysioNet Challenge 2012 v1.0.0, sets A and B, contains 8,000 patient records. Static attributes are removed, leaving 37 time-varying features. Times are quantized to 0.1 hour. Patient splitting with random state 0 yields 4,800 training, 1,600 validation, and 1,600 test records; training and validation are combined for model fitting. Each split is separately min–max normalized and missing values are reset to zero. Observations at or after 24 hours are hidden, and the task evaluates available targets over the withheld part of the first 48 hours.

Both methods use Adam with learning rate 10^{-3}, gradient clipping at norm 1, a batch size of 100, and a fixed 100-epoch budget. Evaluation uses seeds $1, \ldots, 5$. The CRU baseline retains its native masked Gaussian NLL objective, whereas HiPPO-FM uses a deterministic coordinate decoder. The hidden-region metric is the arithmetic mean of valid per-batch MSE values, not a global ratio accumulated across test samples.

A.5 PREPROCESSING AND METRIC CONVENTIONS

Spiral reconstruction uses RMSE and MAE on clean target trajectories. UEA classification reports accuracy and per-dataset rank, with the aggregate quantities formed from the three registered observation-drop settings. Event prediction reports LL, event-type accuracy, and next-event-time RMSE. PhysioNet extrapolation reports hidden-region MSE. Coordinate normalization is always computed from the observed support rather than from the requested query grid.

B IMPLEMENTATION DETAILS AND COMPLEXITY

Encoder and functional head. The natural-data configuration uses $d = 64$, one encoder block, four attention heads, and a two-layer FFN. Each FM operator uses a Legendre polynomial basis with $G = 128$ and rank $M = 16$. The output FM head follows the same functional construction with separate parameters. Missing values are zero-filled and validity masks are applied to attention and functional contraction; classification uses last-valid pooling.

Coordinate transformation. The coordinate origin and scale are computed from the observed support, never from the query grid. The same affine map is applied to observations and queries, and future coordinates may exceed one. This preserves the meaning of an extrapolation query independently of the requested query set.

Encode/query pseudocode.

Table 6: Persistent history-state storage after encoding, excluding model parameters and temporary solver workspace.

Representation	Persistent state	Query access
Observation-indexed attention	$O(Nd)$	resolve against retained observations
Latent ODE/CDE	$O(d)$ latent state	query-dependent integration
HiPPO-FM	$O(Md)$ plus metadata	direct functional readout

Storage. After encoding, persistent functional-state storage is $O(Md)$ plus coordinate metadata, independent of the number of observed timestamps. A new query set requires only functional feature construction and coordinate readout; the history encoder and functional write are reused.

This is a symbolic comparison of the persistent history representation, not a runtime benchmark. The encoder and functional write are paid once for a fixed history; subsequent query sets only construct functional features and evaluate the retained coefficients.

C Functional Geometry and Proofs

The results in this appendix concern the functional-memory operator. The finite-grid formulation is exact for the implemented feature map. A continuous-functional formulation is then obtained by replacing the reference-grid sum with an integral. Neither formulation identifies the learned normalized state with a Bayesian posterior.

For notational compactness, this appendix writes $k = k_\theta$, $\mathcal{H}_k = \mathcal{H}_{k_\theta}$, and $\psi = \psi_\theta$ unless a distinction is stated explicitly.

C.1 Finite-grid representers and covariance consistency

Fix a symmetric positive-semidefinite kernel k with RKHS $\mathcal{H}_k$, a reference grid S, and finite values $\phi_m(s_g)$. Define bounded linear functionals

$$\mathcal{L}_m^G f = G^{-1} \sum_{g=1}^{G} \phi_m(s_g) f(s_g), \qquad h_m = G^{-1} \sum_{g=1}^{G} \phi_m(s_g) k(\cdot, s_g). \tag{18}$$

Point evaluation is bounded in an RKHS; finite linear combinations are therefore bounded. The reproducing property gives $\mathcal{L}_m^G f = \langle f, h_m \rangle_{\mathcal{H}_k}$. Define the synthesis operator $\mathsf{H} : \mathbb{R}^M \to \mathcal{H}_k$ by $\mathsf{H}a = \sum_m a_m h_m$. Then

$$K_u = \mathsf{H}^*\mathsf{H} = G^{-2} \Phi^\top K(S,S) \Phi, \qquad \psi(r) = \mathsf{H}^* k(r, \cdot) = G^{-1} \Phi^\top K(S, r). \tag{19}$$

Thus the write features and the Gram solve refer to *the same* functional coordinates. For every a, $a^\top K_u a = \|\mathsf{H}a\|_{\mathcal{H}_k}^2 \geq 0$, and $\bar{K}_u = K_u + \epsilon I$ is positive definite for $\epsilon > 0$.

The same matrices can be derived from a conceptual prior $f \sim \mathcal{GP}(0, k)$. For $u_m = \mathcal{L}_m^G f$, covariance bilinearity yields

$$\mathrm{Cov}(f(r), u) = \psi(r)^\top, \qquad \mathrm{Cov}(u, u) = K_u. \tag{20}$$

Consequently, $\psi(r)^\top K_u^{-1} u$ would be the conditional mean under noiseless exact functional observations, when K_u is invertible. HiPPO-FM uses this matched linear algebra with a neural state in place of u; it does not perform this conditioning operation or retain a posterior covariance.

C.2 Raw-memory semantics

Proof of Proposition 3.1. Let $v \in \mathbb{R}^N$ and $g_{P,v} = \sum_i v_i k(\cdot, t_i)$. For every functional coordinate, the reproducing property gives

$$[\Psi(P)^\top v]_m = \sum_i v_i h_m(t_i) = \left\langle \sum_i v_i k(\cdot, t_i), h_m \right\rangle_{\mathcal{H}_k} = \mathcal{L}_m^G g_{P,v}. \tag{21}$$

This is the functional value of the empirical kernel field induced by v.

C.3 Matched readout and finite query space

Write $A_\epsilon = (K_u + \epsilon I_M)^{-1}$ and $\mathcal{U}_M = \mathrm{span}\{h_0, \dots, h_{M-1}\}$. The inverse is defined for $\epsilon > 0$, and also for $\epsilon = 0$ when K_u is nonsingular.

Proof of Proposition 3.2: matched projection. Assume K_u is nonsingular, let $u = \mathsf{H}^* f$, and set $s = \mathsf{H} K_u^{-1} u$. For each n,

$$\langle f - s, h_n \rangle = u_n - [K_u K_u^{-1} u]_n = 0. \tag{22}$$

Since $s \in \mathcal{U}_M$ and $f - s$ is orthogonal to every element of $\mathcal{U}_M$, $s = \Pi_{\mathcal{U}_M} f$. Evaluation gives $s(r) = \psi(r)^\top K_u^{-1} u$. The inverse corrects the Gram geometry of the representers; replacing it with the identity is equivalent only in the special case where those representers are orthonormal.

Arbitrary normalized memory. For any m, let $A_\epsilon = (K_u + \epsilon I)^{-1}$. Then

$$o_m(r) = \psi(r)^\top A_\epsilon m = \sum_j [A_\epsilon m]_j h_j(r) \in \mathcal{U}_M. \tag{23}$$

LayerNorm changes m, not the functions h_j. With shared geometry, concatenation and linear channel mixing preserve this functional span. An affine output bias adds a constant function, yielding the upper bound $\mathcal{U}_M + \mathrm{span}\{1\}$. A nonlinear task head would not generally preserve this span; the natural coordinate-prediction head considered here is affine. The observed-support encoder has its own nonlinear residual path and is not itself asserted to have rank M.

C.4 Induced raw kernel and finite-rank compression

Proposition C.1 (Raw induced kernel)**.** *The raw, unnormalized coordinate operator has kernel*

$$k_{M,\epsilon}(r, t) = \psi(r)^\top A_\epsilon \psi(t), \qquad o_{\mathrm{raw}}(r) = \sum_i k_{M,\epsilon}(r, t_i) v_i. \tag{24}$$

It is positive semidefinite and has rank at most M on any finite coordinate set.

Proof. For a matrix Ψ_R of query features, $K_M(R, R) = \Psi_R A_\epsilon \Psi_R^\top$. For any vector z, $z^\top K_M z = \left\| A_\epsilon^{1/2} \Psi_R^\top z \right\|_2^2 \geq 0$. The product factors through an M-dimensional space, so its rank is at most M. Substitution of the raw write gives the displayed prediction. □

Proposition C.2 (Collisions of the raw writer)**.** *If $\delta \in \mathbb{R}^{N \times d_v}$ satisfies $\Psi(P)^\top \delta = 0$, then V and $V + \delta$ have identical raw memories. Applying the same deterministic memory normalization and decoder preserves this equality in exact arithmetic.*

Proof. $\Psi(P)^\top (V + \delta) = \Psi(P)^\top V$. Equal inputs remain equal under deterministic functions. □

The condition applies at the input of the *linear contraction*, not automatically to raw sensor inputs before arbitrary nonlinear encoders. For a pre-write transform F, the relevant difference becomes $\Psi(P)^\top [F(H + \delta) - F(H)]$. It need not vanish, but it also need not be nonzero. For example, the identity retains the collision; a nonlinear MLP or MHA can change it. None of these statements orders their downstream risks or proves injectivity of the compressed representation.

C.5 Proof of Theorem 3.3

Proof. Write the error as $[f(r) - \psi(r)^\top A_\epsilon u] + \psi(r)^\top A_\epsilon (u - \widehat{u})$. The second term is bounded by Euclidean Cauchy–Schwarz. Define $q_r = k(r, \cdot) - \mathsf{H} A_\epsilon \psi(r)$. The first term is $\langle f, q_r \rangle_{\mathcal{H}_k}$, and

$$\|q_r\|_{\mathcal{H}_k}^2 = k(r, r) - 2\psi(r)^\top A_\epsilon \psi(r) + \psi(r)^\top A_\epsilon K_u A_\epsilon \psi(r) \tag{25}$$

$$= c_\epsilon(r) - \epsilon \psi(r)^\top A_\epsilon^2 \psi(r) \leq c_\epsilon(r), \tag{26}$$

using $A_\epsilon K_u A_\epsilon = A_\epsilon - \epsilon A_\epsilon^2$. This also proves $c_\epsilon(r) \geq 0$. RKHS Cauchy–Schwarz completes the bound. □

The first term measures functional geometry; the second measures coefficient mismatch amplified by the readout. The network's normalized memory is one possible $\widehat{u}$. The theorem bounds representation error for fixed geometry; the effects of attention and learned rank on optimization and test error lie outside its scope. When $\widehat{u} = \mathsf{H}^* g_{P,v}$, it describes approximation of the empirical field; that field need not equal the target trajectory.

For a perturbed decoder $\widehat{D}_r$ in place of $D_r = \psi(r)^\top A_\epsilon$, triangle inequality adds $\left\|\widehat{D}_r - D_r\right\|_2 \|\widehat{u}\|_2$ to Equation (17). This separates a numerical decoder discrepancy from representational error without assuming a particular floating-point implementation.

C.6 Query consistency and reuse

Proposition C.3 (Query-set consistency)**.** *Fix the encoded state S_P. In evaluation mode, the prediction at coordinate r is unchanged by the order, grouping, or chunking of other query coordinates, provided every query uses the same observed-support coordinate map and the post-state decoder acts pointwise in r.*

For the frozen state B_P, a row of the output has the form

$$\widehat{y}(r) = W_y \left(\mathrm{Concat}_h[\psi(a_P(r))^\top B_P^h] W_O + b_O\right) + b_y. \tag{27}$$

No quantity on the right depends on any other query in R. Permutation, grouping, and chunking therefore preserve the row corresponding to r. This proves Proposition C.3. The assumptions exclude batch statistics across queries, query self-attention, query-dependent coordinate normalization, and stochastic dropout at evaluation.

Caching B_P, $t_{\min}$, s_P, and validity changes only the execution schedule. The kernel parameters, reference grid, normalization parameters, and output projections are fixed model parameters at inference. Exact equivalence therefore holds algebraically. Changing any such parameter invalidates a cache, as does changing P or its observed values. The numerical tests evaluate these statements for the implemented path; they do not establish invariance under changes to the observation set.

D Additional Experimental Results

This appendix provides detailed results and qualitative evidence that would make the main text unnecessarily long. The UEA tables preserve the complete registered comparison, including t-PatchGNN cells where they are available.

D.1 Full UEA results

Tables 7–9 report mean test accuracy over the three seeds for all 20 datasets. Row-wise boldface marks the highest reported accuracy. The ACC row is the macro average over the reported dataset cells at that drop ratio. The Rank row is computed per dataset over the displayed methods with reported cells, including t-PatchGNN. Each method's mean rank uses its available cells; t-PatchGNN therefore contributes 17, 17, and 18 datasets at drop ratios 0.3, 0.5, and 0.7, respectively. The ACC row is likewise the macro average over each method's available dataset cells. Only the selected HiPPO-FM model is shown. Earlier MHA-to-FM settings, Parallel, V4, pooling alternatives, and external adaptations are excluded from the rank calculation, while t-PatchGNN is included whenever its reported cell is available. The Overall columns in the main text average the three macro-ACC columns with equal weight and rank those resulting means; they do not replace the per-dataset Rank rows reported below.

Table 7: UEA test accuracy with 30% of observations dropped under the `train_es` protocol. Entries are three-seed means; † marks the declared baseline-reporting exception described in the text.

Dataset	GRU-D	GRU-Δt	ODE-RNN	CADN	N. CDE	S5	mTAN	TST	ContiFormer	t-PatchGNN	HiPPO-FM
AWR	0.8467	0.8878	0.8878	0.9378	0.7056	0.8933	0.7322	0.9244	0.9322	0.7789	**0.9500**
BM	0.8417	0.8917	0.8417	0.8500	0.3500	0.8167	0.7500	**0.9583**	0.9250	0.9500	0.9000
CT	0.9740	0.9705	0.9817	0.9612	0.8572	**0.9749**	0.4053	0.9640	0.9687	0.9615	0.9626
DDG	0.5067	0.5600	0.5600	0.5267	0.1933	0.5733	0.6067	0.6600	0.6600	–	**0.6667**
EP	0.9203	0.9251	0.9155	0.9010	0.4662	0.8696	0.5894	0.9469	0.9058	0.9251	**0.9565**
ER	0.7679	0.7568	0.7309	0.8259	0.7259	0.8173	0.7136	0.8321	**0.8358**	0.7889	0.7469
FM	0.5500	0.5167	0.5167	0.5233	0.5233	0.5467	**0.5600**	0.5200	0.4500	0.5367	0.5433
HMD	0.2973	0.3153	0.2928	0.2432	0.2027†	**0.4324**	0.2838	0.3288	0.3018	0.2973	0.3694
HW	0.2055	0.2545	0.1580	**0.2588**	0.0886	0.2561	0.0749	0.1976	0.1549	0.1576	0.1447
HB	0.6683	0.6634	0.6504	**0.6959**	0.5447	0.6098	0.6829	0.6553	0.6715	0.6911	0.6537
JV	0.9685	0.9694	0.9667	0.9613	0.9685	0.9595	0.9550	**0.9784**	0.9766	0.8766	0.9631
LB	0.7685	0.7870	0.7926	0.8593	**0.8704**	0.7778	0.2889	0.8056	0.8093	0.5870	0.7833
LSST	0.4692	0.4695	0.5088	0.4640	0.0933	0.5377	0.4501	0.4938†	**0.5969**	–	0.5603
NATOPS	0.9278	**0.9444**	0.9389	0.9259	0.7204	0.9296	0.7204	0.8574	0.9019	0.7852	0.9130
PEMS	0.6956	0.6513	0.6821	0.7534	0.6994	0.7380	0.7572	0.7726	0.7842	–	**0.8054**
PD	0.8526	0.8095	0.8433	0.8435	0.3878	0.9301	0.4528	0.7925	0.6518	0.9066	**0.9302**
RS	0.7390	0.7500	0.7763	**0.7785**	0.3311	0.7018	0.4890	0.7259	0.6711	0.7215	0.7193
SR	0.6018	0.6553	0.6268	0.6098	0.4881	**0.8578**	0.6644	0.7452	0.6985	0.7861	0.7418
SA	0.9615	0.9554	0.9662	0.9375	0.3232	**0.9864**	0.4869	0.9422	0.9507	0.9565	0.9536
UGL	0.7167	0.7760	0.7854	**0.8333**	0.7729	0.7844	0.5104	0.6552	0.6667	0.7271	0.7281
ACC ↑	0.7140	0.7255	0.7211	0.7345	0.5156	**0.7497**	0.5587	0.7378	0.7257	0.7314	0.7496
Rank ↓	5.88	5.35	5.75	5.38	9.18	<u>4.70</u>	8.58	4.78	5.18	5.88	**4.60**

Table 8: UEA test accuracy with 50% of observations dropped under the `train_es` protocol. Entries are three-seed means; † marks the declared baseline-reporting exception described in the text.

Dataset	GRU-D	GRU-Δt	ODE-RNN	CADN	N. CDE	S5	mTAN	TST	ContiFormer	t-PatchGNN	HiPPO-FM
AWR	0.8611	0.8844	0.8900	0.9367	0.6933	0.8911	0.7189	0.9233	0.9367	0.7756	**0.9544**
BM	0.9000	0.9333	0.9250	0.9083	0.3083	0.7750	0.8500	0.9417	0.9417	**0.9750**	0.9667
CT	0.9656	0.9754	0.9779	0.9181	0.8647	**0.9659**	0.3942	0.9582	0.9663	0.9568	0.9582
DDG	0.5667	0.5533	0.5600	0.5467	0.2267	0.5467	0.5867	0.6200	0.6600	–	0.6467
EP	0.9348	0.9493	0.9275	0.9251	0.4541	0.8575	0.5628	0.9517	0.9130	0.9155	**0.9614**
ER	0.7827	0.7728	0.7469	0.8284	0.7346	0.7988	0.7123	0.8321	**0.8420**	0.7556	0.7247
FM	0.5200	0.5000	0.4867	**0.5300**	0.5067	0.5167	0.5033	0.5200	0.5067	0.4967	0.5267
HMD	0.2838	0.2838	0.3108	0.3333	0.2297†	**0.4234**	0.2793	0.3063	0.3198	0.3514	0.3423
HW	0.1863	**0.2361**	0.1596	0.2161	0.0878	0.2063	0.0690	0.1616	0.1408	0.1231	0.1224
HB	0.6488	0.6748	0.6309	0.6813	0.6098	**0.7154**	0.6537	0.6585	0.6683	0.6634	0.6537
JV	0.9532	0.9550	0.9622	0.9423	0.9604	**0.9658**	0.9432	0.9577	0.9559	0.6306	0.9523
LB	0.7352	0.7556	0.7556	0.8167	**0.8630**	0.7463	0.2704	0.7926	0.8111	0.5407	0.7944
LSST	0.4566	0.4669	0.4669	0.4393	0.1160	0.5188	0.4324	0.4620†	**0.5673**	–	0.5270
NATOPS	**0.9370**	0.9167	**0.9370**	0.9204	0.8093	0.8926	0.6778	0.8333	0.9296	0.7778	0.9185
PEMS	0.6782	0.6455	0.7013	0.7399	0.6744	0.6532	0.7206	0.7726	**0.8112**	–	0.7861
PD	0.7707	0.6662	0.7216	0.6275	0.2249	0.7616	0.3630	0.5171	0.5587	**0.7985**	0.7906
RS	0.7346	0.7346	0.7456	**0.7632**	0.2785	0.7127	0.5000	0.7259	0.6294	0.6447	0.7325
SR	0.5404	0.6291	0.6200	0.6837	0.4972	**0.8794**	0.6519	0.7486	0.6849	0.7907	0.7315
SA	0.9488	0.9451	0.9586	0.9322	0.4390	**0.9760**	0.4846	0.9280	0.9579	0.9530	0.9468
UGL	0.7552	0.7698	**0.8167**	0.8125	0.7698	0.8115	0.5198	0.6479	0.6677	0.7115	0.6917
ACC ↑	0.7080	0.7124	0.7150	0.7251	0.5174	0.7307	0.5447	0.7130	0.7235	0.6977	**0.7364**
Rank ↓	5.95	5.68	5.23	4.85	9.00	4.73	9.23	5.28	**4.43**	6.47	<u>4.50</u>

Table 9: UEA test accuracy with 70% of observations dropped under the `train_es` protocol. Entries are three-seed means; † marks the declared baseline-reporting exception described in the text.

Dataset	GRU-D	GRU-Δt	ODE-RNN	CADN	N. CDE	S5	mTAN	TST	ContiFormer	t-PatchGNN	HiPPO-FM
AWR	0.8611	0.8944	0.9067	**0.9356**	0.7778	0.8956	0.7600	0.9133	0.9144	0.7600	0.9322
BM	0.8833	0.8833	0.9333	0.8750	0.2667	0.7917	0.8250	0.9333	0.9333	**0.9917**	0.9500
CT	0.9615	0.9731	0.9659	0.8728	0.8635	**0.9547**	0.3679	0.9420	0.9633	0.9422	0.9325
DDG	0.5933	0.5467	0.5200	0.5400	0.2467	0.5667	0.5733	0.5800	0.5733	–	**0.6867**
EP	0.9058	0.9348	0.9251	0.8865	0.5531	0.8623	0.5362	0.9275	0.9082	0.9034	**0.9565**
ER	0.7556	0.7457	0.7111	0.7691	0.7765	0.7654	0.6543	0.7741	**0.8333**	0.7395	0.6210
FM	0.4900	**0.5533**	0.5200	0.5200	0.5033	**0.5533**	0.4533	0.4600	0.4867	0.5333	0.4767
HMD	0.2838	0.2658	0.3514	0.2928	0.2658	**0.3919**	0.3018	0.2658	0.2793	0.3108	0.3559
HW	0.1725	**0.1867**	0.1502	0.1376	0.0796	0.1427	0.0537	0.1302	0.1220	0.0996	0.1024
HB	0.6472	0.6846	0.6439	0.6520	0.6163	0.6325	**0.6959**	0.6179	0.6911	0.6569	0.6634
JV	0.9243	0.9333	0.9297	0.9423	0.9288	0.9414	0.9324	**0.9495**	**0.9495**	0.7297	0.9468
LB	0.5537	0.6444	0.6611	0.7407	**0.8037**	0.6426	0.2222	0.7796	0.7519	0.6741	0.7056
LSST	0.4303	0.4565	0.4244	0.4347	0.1245	0.4808	0.4207	0.4556†	**0.5242**	0.4109	0.4861
NATOPS	0.9074	**0.9222**	0.8981	0.9111	0.7481	0.9056	0.6167	0.8611	0.8981	0.7222	0.9148
PEMS	0.6455	0.6185	0.6724	0.6686	0.4297	0.6474	0.6647	**0.7553**	0.7476	–	0.7148
PD	0.6142	0.5006	0.5880	0.4982	0.1454	0.6118	0.2933	0.4018	0.4751	**0.6809**	0.6561
RS	0.6711	0.6798	0.7171	**0.7346**	0.2851	0.6075	0.4430	0.6908	0.6162	0.5636	0.6316
SR	0.6303	0.6246	0.6735	0.5825	0.5358	**0.8635**	0.6553	0.7406	0.7099	0.8100	0.7804
SA	0.9218	0.9113	0.9335	0.9110	0.3352	**0.9573**	0.4705	0.8739	0.9297	0.9163	0.9053
UGL	0.7323	0.7917	0.7729	**0.8125**	0.7698	0.8094	0.4979	0.6625	0.7125	0.7156	0.6969
ACC ↑	0.6793	0.6876	0.6949	0.6859	0.5028	0.7012	0.5219	0.6857	0.7010	0.6756	**0.7058**
Rank ↓	6.03	5.10	5.20	5.38	8.90	5.08	8.80	5.63	**4.53**	6.36	<u>4.55</u>

D.2 Full Spiral settings and visualization

This subsection records the complete protocol underlying the main-text Spiral comparison. It uses 300 trajectories, a 200/100 training/testing split, 150 equally spaced time positions, 30 observations from the first half, five seeds $\{13, 27, 42, 100, 2024\}$, and 5,000 optimization iterations. The registered noise scale is $\beta = 0.1$ and the reported main setting uses $\alpha = 0.02$; the additional α values are used only as a robustness check and are not repeated in a separate table here.

For normalized time $t \in [s, e]$, the counter-clockwise family is

$$x_{\mathrm{cc}}(t) = (a + bt)\cos(t) + 5, \tag{28}$$
$$y_{\mathrm{cc}}(t) = (a + bt)\sin(t) + 5. \tag{29}$$

The clockwise family reverses the radial-angular coordinate. With $z(t) = e + 1 - t$ and $r(t) = a + 50b/z(t)$, we use

$$x_{\mathrm{cw}}(t) = r(t)\cos(z(t)) - 5, \tag{30}$$
$$y_{\mathrm{cw}}(t) = r(t)\sin(z(t)) - 5. \tag{31}$$

Clean trajectories are generated first, followed by Gaussian observation noise on the sparse input coordinates; clean trajectories remain the evaluation targets. Interpolation is measured on the observed half and extrapolation on the withheld half. The complete numerical comparison is kept in the main table to avoid duplicating the same results in the appendix.

For the qualitative comparison, the baselines use their respective temporal or continuous-time readouts, while HiPPO-FM uses the functional readout described in Section 3. All rows use the same five seeds and plotting routine.

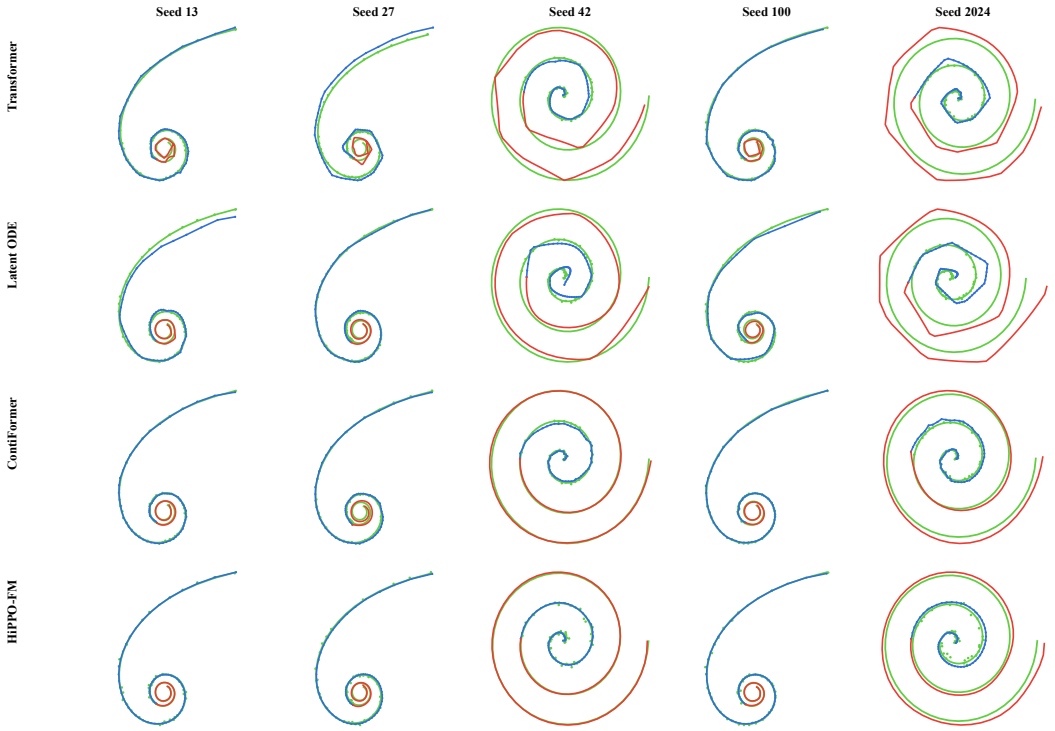

Figure 3: **Five-seed qualitative Spiral visualizations under the protocol-I setting.** Columns are seeds $\{13, 27, 42, 100, 2024\}$ and all rows use the same plotting routine with 50 observed support points. Rows show Transformer, Latent ODE, ContiFormer, and HiPPO-FM. This visualization is baseline-aligned and qualitative; the main-text aggregate table reports the registered $N_{\text{obs}} = 30$ setting.

D.3 Event benchmark details

The event benchmark contains one synthetic and five real-world marked sequences. The archived statistics are: Synthetic ($K = 10$, 8,618 events, average length 17, 375/125 train/test sequences), Neonate ($K = 1$, 534 events, average length 11, 38/10 sequences), Traffic ($K = 2$, 154,747 events, average length 60, 1,938/647 sequences), MIMIC ($K = 75$, 2,419 events, average length 4, 527/65 sequences), StackOverflow ($K = 22$, 480,413 events, average length 72, 4,777/1,326 sequences), and BookOrder ($K = 2$, 414,800 events, average length 2,074, 90/100 sequences). Here K denotes the number of event types and the two sequence counts are train/test counts.

The synthetic benchmark uses ten event types and 500 generated sequences with four-fold cross-validation. Neonate records seizure-start events from 79 patients; only 48 patients have at least two seizures. Traffic contains two event types derived from large changes in hourly sensor measurements, with events close in time deduplicated. MIMIC treats each ICU visit as a sequence. BookOrder contains buy and sell events with sequences clipped to length 500 and a maximum inter-event gap of 70, while StackOverflow records user badge-reward histories over two years. The real-data folds follow the partition counts above; MIMIC, StackOverflow, and BookOrder use five folds, whereas Neonate and Traffic use four folds.

The event objective evaluates the next-event mark and time from each observed prefix. We report log-likelihood, event-type accuracy, and next-event-time RMSE, with mean and sample standard deviation over the registered folds or seeds. Neonate has $K = 1$, so its event-type accuracy is structurally non-discriminative; its timing metrics provide the meaningful comparison. Per-type frequency tables are not present in the archived benchmark summary, so we do not make a class-balance claim.

D.4 PHYSIONET EXTRAPOLATION RESULTS

This appendix gives the complete result record for irregular extrapolation on PhysioNet (Schirmer et al., 2022). The task hides observations at or after 24 hours and evaluates predictions on the withheld coordinates. The matched comparison uses the released preprocessing, masks, batch size, and a fixed 100-epoch budget, following the protocol in Appendix A.4; values are mean $\pm$ sample standard deviation over five seeds.

We compare HiPPO-FM with CRU, a continuous-time recurrent baseline designed for irregular observations. The comparison uses hidden-region MSE and also records epoch time and peak allocated memory, separating predictive behavior on real irregular data from the controlled coordinate-reconstruction experiments. HiPPO-FM improves PhysioNet hidden-region MSE relative to CRU while using less epoch time and peak allocated memory.

Table 10: **PhysioNet extrapolation, five seeds.** MSE and seconds per epoch are mean $\pm$ sample SD. Memory is peak allocated memory.

Dataset	Model	MSE ↓	s/epoch ↓	GiB ↓	Parameters
PhysioNet	CRU	0.006465 ± 0.000409	45.84 ± 12.13	4.121	28,654
	HiPPO-FM	$\mathbf{0.005099} \pm 0.000100$	$\mathbf{2.01 \pm 0.14}$	**0.196**	55,978

The relative difference uses CRU as denominator: $100(E_{\mathrm{HiPPO}} - E_{\mathrm{CRU}})/E_{\mathrm{CRU}}$. HiPPO-FM improves by 21.14% relative to CRU on PhysioNet. The comparison is descriptive and uses five optimization seeds on the same dataset.

HiPPO-FM reduces mean epoch time from 45.84 s to 2.01 s on PhysioNet, corresponding to approximately a $22.8\times$ ratio, and peak allocated memory from 4.121 to 0.196 GiB. The CRU comparison is an implementation-matched descriptive benchmark: CRU retains its native masked Gaussian NLL objective, whereas HiPPO-FM uses a deterministic coordinate decoder. It therefore does not evaluate posterior uncertainty or calibrated clinical risk.

E QUERY REUSE AND CONSISTENCY

For a fixed observation history, we compare direct evaluation with cached-state evaluation under different query orderings, groupings, and chunk sizes. The maximum coordinate-wise difference is 2.611×10^{-5}. This check supports reuse of one encoded state without reprocessing the observation history; it does not test changes to the observed history or model parameters.