UPDATING BELIEFS OVER EVENTUAL OUTCOMES AS EVIDENCE ARRIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world decisions often require updating beliefs about the same eventual outcome as evidence arrives. Learning these belief paths is challenging because the true outcome probabilities at intermediate stages are not directly observable, even after outcomes are known. We introduce **BELIEF**, a generative framework for learning belief paths from event streams. Rather than directly predicting outcomes, BELIEF learns how incoming evidence is distributed under each candidate outcome, given the full preceding history. Observed evidence then updates the relative support for these candidate outcomes through accumulated log likelihoods and Bayesian inference, yielding entire belief paths without direct supervision of intermediate beliefs. It supports binary, multiclass, and continuous outcomes. We provide quantitative guarantees for belief-path recovery and prove exact recovery under correct prior and evidence models. Experiments show accurate belief-path recovery on synthetic evidence streams and the closest alignment with reference probability paths on real data, while preserving downstream task performance. Each belief update decomposes into per-event evidence contributions, providing a basis for risk monitoring and timely decisions.

1 INTRODUCTION

Forecasts often need to be revised before the event of interest is resolved. Such revisions arise across medicine (Klug et al., 2024; Thorsen-Meyer et al., 2020), finance (Medina-Olivares et al., 2023), weather forecasting (Regnier & Harr, 2006), and sports (Brill et al., 2026), where probabilities are updated as new information accumulates. Consider a World Cup tournament: goals, match results, and performance measurements arrive over time and continually change the evidence about the eventual champion. Although the available evidence changes, each update concerns the same eventual outcome. Such updates can inform clinical monitoring, financial risk management, and weather preparedness.

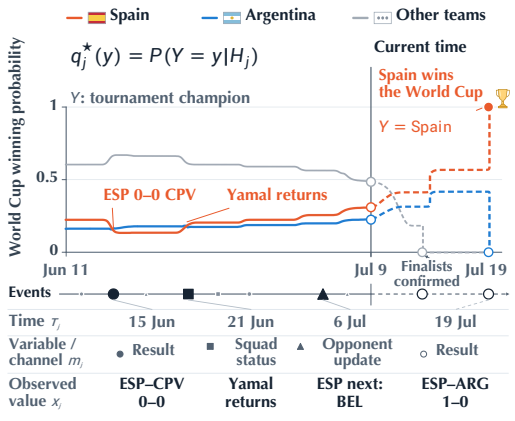

Figure 1: An example of belief-path learning. Consider a World Cup tournament: as the schedule unfolds, its events arrive as a stream of observations, each revising every team's probability of winning the title. The curves schematically show these belief paths, solid before the current time and dashed after. Only the observed events and the realized champion are recorded; the intermediate beliefs cannot be observed.

We call this problem *belief-path learning*: recovering how probabilistic beliefs about the same eventual outcome evolve as evidence arrives. Formally, the target belief after observation j is $q_j^\star(\cdot) = P(Y \in \cdot \mid H_j)$, where Y denotes the eventual outcome and $H_j = (o_1, \ldots, o_j)$ contains the evidence observed after the first j observations. We represent incoming evidence as an event stream, with each observation $o_j = (\tau_j, m_j, x_j)$ recording its time, variable, and value; observations may arrive irregularly and asynchronously (Park et al., 2025). Training data consist of completed historical event

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.

streams and their realized outcomes. Thus, the eventual outcome is known retrospectively, while the intermediate conditional probabilities remain unobserved. The learning problem is therefore to recover how beliefs should change as evidence accumulates, without intermediate probability labels. Unlike models trained on recorded probability paths (Lin et al., 2021), our setting provides event streams and realized outcomes rather than intermediate forecasts. Figure 1 illustrates belief-path learning through a World Cup example.

A direct way to approach this task is to train an outcome predictor on successive event prefixes (Teinemaa et al., 2019); target replication similarly applies a sequence-level outcome label to intermediate prediction steps (Lipton et al., 2016). Under a strictly proper scoring rule, unrestricted prefix prediction has the true posterior as its population optimum (Gneiting & Raftery, 2007). However, each new observation's evidential contribution remains implicit, and supervising weakly informative early prefixes with final outcome labels can yield noisy training signals (Maystre et al., 2025).

We introduce **BELIEF**, a generative framework that treats candidate eventual outcomes as alternative explanations of incoming evidence. Building on generative Bayesian inference, BELIEF extends the focus from a single outcome prediction to learning entire belief paths with explicit evidence contributions. For each candidate outcome, it learns the next-observation distribution conditioned on the full preceding history. Following an unexpected group-stage defeat, for example, it compares how likely the observed result is under competing eventual-champion hypotheses. The conditional log-likelihood ratio gives the observation's explicit contribution to the change in log posterior odds. Thus, BELIEF connects event-specific evidence learning to both the construction and the interpretation of the belief path.

BELIEF learns these evidence distributions through outcome-conditioned next-observation prediction. For each completed trajectory, the realized outcome provides the conditioning outcome, while each observed value serves as a likelihood target given the preceding history, observation time, and variable. At inference, the conditional likelihoods accumulated along the observed history are combined with the prior and normalized over candidate outcomes to obtain the posterior at each prefix. The same construction applies to binary, multiclass, and continuous outcomes, and associates each belief revision with the likelihood contribution of the newly observed evidence.

This construction also admits direct guarantees for belief-path recovery. At each observed prefix, the expected posterior KL error is bounded by the prior KL error plus the accumulated expected conditional evidence-model KL errors; when the prior and evidence models are correct, the Bayesian construction exactly recovers the posterior path.

Our contributions are threefold:

1. We formalize belief-path learning and introduce BELIEF, a generative framework that learns outcome-conditioned evidence distributions from the full preceding history and converts them into posterior paths through Bayesian inference. The framework supports binary, multiclass, and continuous outcomes without intermediate probability labels, with explicit per-observation likelihood contributions.
2. We establish theoretical guarantees for belief-path recovery, quantifying error propagation under model misspecification and proving exact recovery under correct prior and evidence models.
3. Experiments on linear and nonlinear synthetic streams demonstrate accurate belief-path recovery across binary, multiclass, and continuous outcomes. On financial, sports, and weather event streams, BELIEF aligns most closely with reference probability paths while maintaining competitive downstream predictive performance.

2 Related Work

2.1 Belief-Path Learning and Bayesian Inference

A central question in dynamic prediction is how beliefs about the same eventual outcome evolve as evidence accumulates. Some work studies probability trajectories directly. For example, GLIM learns distributions over recorded forecast paths (Lin et al., 2021). When intermediate probability paths are not observed, a common alternative is to repeatedly predict the eventual outcome from

growing partial histories. Predictive process monitoring applies outcome predictors to event prefixes (Teinemaa et al., 2019), target replication supervises intermediate predictions using sequence-level labels (Lipton et al., 2016), and Dynamic-DeepHit updates event-risk predictions as longitudinal observations accumulate (Lee et al., 2020). Because these prefix-wise predictions are repeatedly produced along the same evolving process, subsequent work has also introduced explicit temporal-consistency objectives to couple predictions across observation times (Maystre et al., 2025; Fah et al., 2026).

A complementary line derives sequential outcome beliefs from probabilistic evidence rather than predicting them directly at each prefix. Class-conditional generative classifiers obtain outcome probabilities from learned input likelihoods and priors through Bayes' rule (Ng & Jordan, 2001). Sequential density-ratio methods learn relative evidence for competing hypotheses directly: SPRT-TANDEM estimates sequential density ratios for early classification (Ebihara et al., 2021), while its multiclass extension estimates density-ratio matrices (Miyagawa & Ebihara, 2021). FIRMBOUND uses accumulated evidence to determine when to stop observing and make a decision (Ebihara et al., 2025). Bayesian generative approaches instead connect observations and outcome risk through latent structures, updating latent disease-state probabilities from incoming measurements (Alaa et al., 2017) or linking longitudinal measurements and event risk through shared random effects (Rizopoulos, 2011). Unlike direct ratio estimation, BELIEF learns normalized next-observation distributions under each candidate outcome and derives likelihood-ratio evidence from these distributions; Bayesian accumulation then yields the posterior path without intermediate probability supervision.

2.2 Time Series and Event Stream Modeling

Time series modeling has advanced rapidly in recent years (Wang et al., 2026). For regularly sampled data, forecasting models learn temporal dependencies from aligned historical sequences, while probabilistic approaches represent future uncertainty through predictive distributions (Salinas et al., 2020). Real-world event streams, however, may involve multiple variables observed at irregular times, making irregular multivariate time-series modeling a closely related line of research. Irregular multivariate observations involve uneven time intervals and variable-specific sampling rates. GRU-D incorporates missingness and elapsed time into recurrent representations, while mTAND aggregates observations through time-aware attention (Che et al., 2018; Shukla & Marlin, 2021). Continuous-time approaches such as GRU-ODE-Bayes and ACSSM represent latent dynamics between irregular measurements and incorporate observations into state inference (De Brouwer et al., 2019; Park et al., 2025). Viewing irregular measurements as time–variable–value events makes their asynchronous structure explicit (Horn et al., 2020; Luo et al., 2025). Set-based methods aggregate individual measurements, while graph-based approaches such as Raindrop and Hi-Patch organize temporal and inter-variable dependencies (Horn et al., 2020; Zhang et al., 2022; Luo et al., 2025). BELIEF adopts this irregular event stream representation, allowing it to accommodate flexible observation patterns across times and variables.

3 Methodology

3.1 Task Definition

Belief-path learning aims to recover how the conditional distribution over the same eventual outcome evolves as evidence accumulates. Let Y denote the eventual outcome, taking values in an outcome space $\mathcal{Y}$. Evidence arrives as a sequence of observations $o_j = (\tau_j, m_j, x_j)$, where τ_j, m_j, and x_j denote the observation time, variable (or channel), and observed value, respectively. Let $H_j = (o_1, \ldots, o_j)$ denote the evidence history after the first j observations, with $H_0 = \varnothing$. We define the belief at prefix j as the conditional distribution of the eventual outcome given the evidence observed so far:

$$q_j^\star(\cdot) = P(Y \in \cdot \mid H_j), \qquad j = 0, \ldots, J. \tag{1}$$

The sequence $(q_0^\star, \ldots, q_J^\star)$ forms the *belief path*. The goal of belief-path learning is to recover the belief at every observed prefix from the accumulating event stream, without labels for these intermediate beliefs.

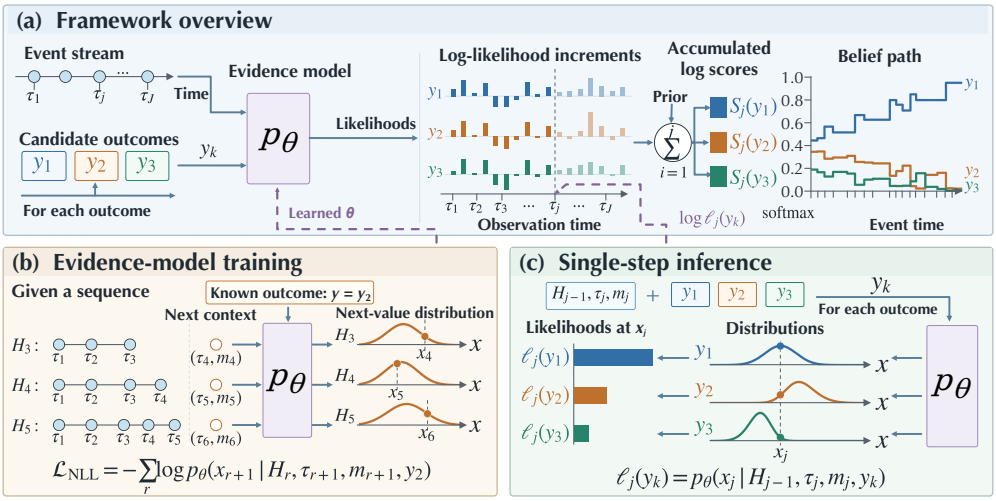

Figure 2: BELIEF framework. (a) **Framework overview.** Given an event stream $H_J = (o_1, \ldots, o_J)$, BELIEF processes its observations sequentially. Starting from the prior $q_{\theta,0} = \pi$, each o_j is evaluated under the candidate outcomes, producing likelihoods $\ell_j(y)$. Their log-likelihood contributions are accumulated with the log prior; exponentiation and normalization at each prefix yield $q_{\theta,j}$. Repeating this update for $j = 1, \ldots, J$ traces the full belief path. (b) **Evidence-model training.** With the realized outcome y fixed, each prefix H_{j-1} together with the next observation time and variable defines a distribution over x_j. The actual x_j is scored by NLL before entering the next prefix, yielding multiple training positions from one trajectory. (c) **Single-step inference.** For one arriving observation, H_{j-1}, τ_j, m_j, and x_j are fixed while y varies. Evaluating the corresponding conditional distributions at x_j gives candidate-wise likelihoods used for the current update in (a).

3.2 Framework Overview

The central challenge in belief-path learning is that intermediate beliefs are not available as training labels. BELIEF addresses this by shifting the learning target from beliefs to evidence. Using Bayes' rule, it learns the conditional distribution of the next observed value under each candidate eventual outcome, $p_\theta(x_j \mid H_{j-1}, \tau_j, m_j, y)$, rather than directly supervising the corresponding belief. A single historical trajectory provides supervision at every observation position, yielding multiple next-observation likelihood targets without intermediate belief labels. At inference, the same arriving value is evaluated under candidate outcomes, and the resulting likelihoods are accumulated with the prior to recover the belief path. Figure 2 summarizes the framework.

3.3 Outcome-Conditioned Evidence Modeling and Learning

Evidence model. The evidence model outputs a normalized distribution over possible observed values rather than a point prediction. Given the preceding history H_{j-1}, observation time τ_j, variable m_j, and a candidate eventual outcome y, it models $p_\theta(x \mid H_{j-1}, \tau_j, m_j, y)$. Evaluating these normalized distributions at the same observed value provides candidate-wise likelihoods; their ratios quantify the observation's relative support for competing outcomes. In our experiments, we use a simple instantiation consisting of a causal encoder and a probabilistic decoder. The encoder maps the observed history to $h_{j-1} = \mathrm{Enc}_\theta(H_{j-1})$, and its output is combined with embeddings of the current observation time and variable to form a context representation u_j. The decoder parameterizes

$$p_\theta(x \mid H_{j-1}, \tau_j, m_j, y) = \sum_{k=1}^{K} w_{\theta,k}(u_j, y)\, f_{\theta,k}(x \mid u_j). \tag{2}$$

Here K is the number of mixture components rather than the number of outcome classes. Each $f_{\theta,k}$ is a normalized component distribution, while the weights are nonnegative and sum to one. The component distributions are shared across candidate outcomes, whereas the outcome-conditioned weights allow different outcome hypotheses to assign different likelihoods to the same observed value. Conditioning the shared mixture on the candidate outcome allows the same construction to extend naturally to continuous outcomes. With positive weights and fixed context, each pairwise likelihood ratio lies between the smallest and largest corresponding weight ratios; the derivation and implementation details are in Appendix B.1.

Autoregressive likelihood training. Training uses N historical event streams with known eventual outcomes. For trajectory n, the realized outcome $y^{(n)}$ conditions every prediction, while each observed value provides a likelihood target. The evidence model is trained by conditional maximum likelihood,

$$\mathcal{L}_{\mathrm{NLL}}(\theta) = -\sum_{n=1}^{N}\sum_{j=1}^{J_n} \log p_\theta\Big(x_j^{(n)} \mid H_{j-1}^{(n)}, \tau_j^{(n)}, m_j^{(n)}, y^{(n)}\Big). \tag{3}$$

A trajectory of length J_n therefore contributes J_n likelihood terms, one at each observation position, analogous to position-wise next-token likelihood training. Each value is scored before entering the history used for subsequent predictions, so predicting x_j uses neither x_j itself nor future observations. Conditioning on the full preceding history captures temporal and cross-variable dependencies.

3.4 Bayesian Belief Updating

We use the evidence model learned in Section 3.3 to construct a belief distribution at each observed prefix. At inference, the eventual outcome is unknown, so the same observed value is evaluated under each candidate outcome, $\ell_j(y) = p_\theta(x_j \mid H_{j-1}, \tau_j, m_j, y)$. Starting from the prior $q_{\theta,0} = \pi$, BELIEF reweights the preceding belief by the new value's likelihood and normalizes over outcomes. Expanding the recursion gives

$$q_{\theta,j}(y) = \frac{q_{\theta,j-1}(y)\ell_j(y)}{\int_{\mathcal{Y}} q_{\theta,j-1}(u)\ell_j(u)\,\mathrm{d}\mu(u)} = \frac{\pi(y)\prod_{i=1}^{j}\ell_i(y)}{\int_{\mathcal{Y}} \pi(u)\prod_{i=1}^{j}\ell_i(u)\,\mathrm{d}\mu(u)}, \qquad q_{\theta,0} = \pi. \tag{4}$$

Here μ denotes a reference measure over the outcome space. Repeating the update produces the path $(q_{\theta,0}, \ldots, q_{\theta,J})$. Section 3.5 shows that, under the observation-context assumption, this model-based path exactly recovers the target belief path when the prior and conditional evidence models are correct, and further bounds the posterior error under model misspecification.

Taking the logarithm of the unnormalized numerator converts the likelihood product into an accumulated log score:

$$S_j(y) = \log\left[\pi(y)\prod_{i=1}^{j}\ell_i(y)\right] = \log\pi(y) + \sum_{i=1}^{j}\log\ell_i(y), \qquad q_{\theta,j}(y) \propto e^{S_j(y)}. \tag{5}$$

Hence $\Delta S_j(y) = \log \ell_j(y)$ is the score increment contributed by the newly observed value. For two candidates a, b with positive prior and likelihoods,

$$\log\frac{q_{\theta,j}(a)}{q_{\theta,j}(b)} - \log\frac{q_{\theta,j-1}(a)}{q_{\theta,j-1}(b)} = \log\frac{\ell_j(a)}{\ell_j(b)}. \tag{6}$$

Thus, each observation contributes an explicit pairwise evidence increment, quantified by the conditional log-likelihood ratio. Detailed readouts and evidence-ratio forms for binary classification, multiclass classification, and continuous regression are provided in Appendix B.3.

3.5 BELIEF-PATH RECOVERY GUARANTEES

The following results establish conditions for exact belief-path recovery and bound expected posterior KL error under model misspecification. The guarantees apply to the Bayesian readout in Section 3.4 and do not depend on the particular parameterization of the evidence model. Technical details and proofs are provided in Appendix A. Let P denote the true data-generating distribution, P_Y its outcome marginal, and $p^\star(\cdot \mid H_{i-1}, \tau_i, m_i, Y)$ the true conditional value distribution. Fix a finite horizon J, and assume throughout that the relevant conditional distributions and posterior quantities are well defined.

Observation-context assumption. For every $i \le J$,

$$P\big((\tau_i, m_i) \in \cdot \mid H_{i-1}, Y\big) = P\big((\tau_i, m_i) \in \cdot \mid H_{i-1}\big) \quad P\text{-a.s.} \tag{7}$$

Conditioned on the observed history, the mechanism determining the next observation time and variable is independent of the eventual outcome.

Theorem 3.1 (Exact belief-path recovery)**.** *Suppose the observation-context assumption holds. If $\pi = P_Y$ and, for every $i \le J$,*

$$p_\theta(\cdot \mid H_{i-1}, \tau_i, m_i, Y) = p^\star(\cdot \mid H_{i-1}, \tau_i, m_i, Y) \quad P\textit{-a.s.},$$

then the recovered outcome distributions satisfy

$$P\big(q_{\theta,j} = q_j^\star \textit{ for all } j = 0, \dots, J\big) = 1.$$

Correct prior and evidence models therefore recover the entire path. For misspecified models, define the average conditional evidence error

$$\varepsilon_i(\theta) = \mathbb{E}_P[D_{\mathrm{KL}}(p^\star(\cdot \mid H_{i-1}, \tau_i, m_i, Y) \,\|\, p_\theta(\cdot \mid H_{i-1}, \tau_i, m_i, Y))]\,. \tag{8}$$

This averages the conditional value-distribution KL over the outcome, observed history, observation time, and channel drawn from P.

Theorem 3.2 (Posterior error under misspecification)**.** *Under the observation-context assumption, for every fixed prefix $j \le J$ with finite right-hand side,*

$$\mathbb{E}_{H_j \sim P} D_{\mathrm{KL}}(q_j^\star \,\|\, q_{\theta,j}) \le D_{\mathrm{KL}}(P_Y \,\|\, \pi) + \sum_{i=1}^{j} \varepsilon_i(\theta). \tag{9}$$

The bound links posterior error to the conditional distributions fitted during training. At the population level, the expected next-value NLL differs from the corresponding conditional KL by a model-independent term, whenever both are finite. Vanishing prior error and accumulated conditional evidence error therefore imply vanishing expected posterior KL.

For a fixed $1 \le j \le J$, let L_j, R_j, and M_j denote population prefix evidence NLL, expected posterior KL, and history-marginal KL, respectively.

Relative marginal-stability condition. The compared models share the prior π and observation-context kernels and, for some $0 \le \kappa < 1$, satisfy $|M_j(\theta_B) - M_j(\theta_A)| \le \kappa |L_j(\theta_B) - L_j(\theta_A)|$.

Theorem 3.3 (Posterior improvement under relative marginal stability)**.** *Under the observation-context and relative marginal-stability conditions, any two models θ_A, θ_B with finite relevant losses and KL divergences satisfy*

$$L_j(\theta_B) < L_j(\theta_A) \implies R_j(\theta_A) - R_j(\theta_B) \ge (1 - \kappa)\big[L_j(\theta_A) - L_j(\theta_B)\big] > 0. \tag{10}$$

Under these conditions, the decrease in expected posterior KL is at least $(1 - \kappa)$ times the decrease in population prefix evidence NLL at the same prefix.

Table 1: Belief-path recovery of BELIEF, PREFIX, and TERMINAL across six synthetic datasets and six encoders. Metrics are averaged over applicable datasets; mv measures total squared path variation relative to the analytic posterior (ideal: 1). Best results within each encoder are bold. Per-setting results are in Appendix D.1.

		All settings			*Classification*		*Continuous*	
Encoder	Variant	KL ↓	MAE ↓	mv → 1	AUROC ↑	AUPRC ↑	corr ↑	CRPS ↓
Causal Transformer	**BELIEF**	**0.0154** ± 0.0017	**0.1033** ± 0.0057	**0.9505** ± 0.0132	**0.9711** ± 0.0079	**0.9457** ± 0.0116	**0.7873** ± 0.0055	**0.3406** ± 0.0043
	PREFIX	0.1519 ± 0.0250	0.3478 ± 0.0401	1.5959 ± 0.1859	0.8996 ± 0.0211	0.8288 ± 0.0290	0.6353 ± 0.0447	0.4190 ± 0.0163
	TERMINAL	0.3490 ± 0.0168	0.5045 ± 0.0290	13.8902 ± 1.6766	0.8879 ± 0.0221	0.8110 ± 0.0286	0.6051 ± 0.0532	0.4288 ± 0.0171
GRU	**BELIEF**	**0.0114** ± 0.0008	**0.0871** ± 0.0026	**0.9861** ± 0.0055	**0.9725** ± 0.0077	**0.9478** ± 0.0122	**0.7879** ± 0.0055	**0.3397** ± 0.0041
	PREFIX	0.0561 ± 0.0074	0.1941 ± 0.0122	0.5994 ± 0.0210	0.9617 ± 0.0072	0.9281 ± 0.0105	0.7567 ± 0.0071	0.3617 ± 0.0058
	TERMINAL	0.2360 ± 0.0308	0.3766 ± 0.0251	2.1174 ± 0.7652	0.9535 ± 0.0073	0.9137 ± 0.0101	0.7327 ± 0.0243	0.3783 ± 0.0158
GRU-ODE-Bayes	**BELIEF**	**0.0142** ± 0.0026	**0.0965** ± 0.0084	**0.9714** ± 0.0112	**0.9722** ± 0.0076	**0.9471** ± 0.0116	**0.7832** ± 0.0078	**0.3434** ± 0.0062
	PREFIX	0.0537 ± 0.0063	0.1961 ± 0.0130	1.2130 ± 0.0334	0.9620 ± 0.0066	0.9282 ± 0.0094	0.7558 ± 0.0109	0.3638 ± 0.0073
	TERMINAL	0.2636 ± 0.0441	0.4632 ± 0.0959	6.1352 ± 2.2470	0.9130 ± 0.0559	0.8753 ± 0.0551	0.3623 ± 0.3360	0.4744 ± 0.0759
mTAND	**BELIEF**	**0.0117** ± 0.0012	**0.0859** ± 0.0053	**0.9746** ± 0.0118	**0.9715** ± 0.0070	**0.9456** ± 0.0096	**0.7918** ± 0.0063	**0.3369** ± 0.0062
	PREFIX	0.1420 ± 0.0285	0.3618 ± 0.0524	0.7948 ± 0.1991	0.8799 ± 0.0392	0.8363 ± 0.0344	0.5279 ± 0.1363	0.4406 ± 0.0293
	TERMINAL	1.0939 ± 1.6168	0.6296 ± 0.0759	3.1656 ± 1.2798	0.7451 ± 0.1065	0.6694 ± 0.1170	0.2590 ± 0.1590	0.5162 ± 0.0316
ASTGI	**BELIEF**	**0.0188** ± 0.0041	**0.1082** ± 0.0120	**0.9444** ± 0.0314	**0.9708** ± 0.0078	**0.9452** ± 0.0125	**0.7838** ± 0.0141	**0.3429** ± 0.0105
	PREFIX	0.2179 ± 0.0173	0.4613 ± 0.0258	14.0238 ± 0.9769	0.8132 ± 0.0208	0.7043 ± 0.0215	0.4631 ± 0.0476	0.4801 ± 0.0116
	TERMINAL	0.4321 ± 0.0136	0.6506 ± 0.0077	23.5963 ± 4.8331	0.7609 ± 0.0229	0.6499 ± 0.0191	0.3176 ± 0.0710	0.5025 ± 0.0151
HSTU	**BELIEF**	**0.0182** ± 0.0030	**0.1097** ± 0.0088	**0.9722** ± 0.0146	**0.9703** ± 0.0069	**0.9439** ± 0.0105	**0.7848** ± 0.0055	**0.3426** ± 0.0040
	PREFIX	0.1550 ± 0.0271	0.3529 ± 0.0443	2.1579 ± 0.1845	0.9012 ± 0.0198	0.8363 ± 0.0283	0.6264 ± 0.0534	0.4202 ± 0.0171
	TERMINAL	26.6362 ± 58.3273	0.5762 ± 0.0328	33.3027 ± 5.1257	0.8695 ± 0.0461	0.7893 ± 0.0538	0.5170 ± 0.0951	0.4546 ± 0.0189

4 EXPERIMENTS

The experiments address four questions: **(RQ1)** Can BELIEF accurately recover posterior belief paths? **(RQ2)** Can BELIEF learn belief paths on real-world event streams that align with reference probability paths? **(RQ3)** Does BELIEF affect downstream task performance? **(RQ4)** Which design choices contribute to BELIEF's belief-path quality?

4.1 EXPERIMENTAL SETUP

Data. We evaluate BELIEF on three groups of tasks. (i) Synthetic recovery uses six linear/nonlinear event-stream settings with binary, multiclass, or continuous outcomes; each contains 8,571 trajectories and has an analytically tractable posterior at every prefix. (ii) Real-world reference-path evaluation covers Weather (12-hour precipitation; 37,499 station-windows; NBM references), Steel-RB (RB1905–RB2606 Chinese rebar futures; option-implied references), and Soccer (six seasons, 21 leagues; FiveThirtyEight SPI references). (iii) Downstream evaluation includes binary prediction on P12 and MIMIC-III, multiclass transport-mode recognition on Geolife, and continuous length-of-stay prediction on P12-LoS. Further details are in Appendix C.1.

Baselines. BELIEF is encoder-independent. Synthetic experiments instantiate it with the six backbones below; real-domain and downstream experiments use a causal Transformer with a probabilistic decoder, denoted BELIEF* in Tables 2–4. To test whether BELIEF improves path recovery beyond direct outcome supervision, we use two matched controls. PREFIX predicts the eventual outcome at every prefix using the same realized label; TERMINAL applies the outcome loss only to the complete trajectory. At test time, both are evaluated on successive prefixes to form belief paths.

To compare across encoder styles, the synthetic experiments use six backbones: Causal Transformer (self-attention; Vaswani et al., 2017), GRU (recurrent modeling; Cho et al., 2014), GRU-ODE-Bayes (continuous-time dynamics; De Brouwer et al., 2019), mTAND (time-aware attention; Shukla & Marlin, 2021), ASTGI (graph modeling; Liu et al., 2026), and HSTU (recommendation-oriented sequence modeling; Zhai et al., 2024).

Real-domain and downstream comparisons include GRU-D (Che et al., 2018), mTAND (Shukla & Marlin, 2021), Raindrop (Zhang et al., 2022), Hi-Patch (Luo et al., 2025), and ATENet (Lee et al., 2025); iTransformer (Liu et al., 2024) adapted to irregular inputs; Dynamic-DeepHit (Lee et al., 2020); an adapted SPRT-TANDEM baseline for sequential density-ratio estimation (Ebihara et al., 2021; Miyagawa & Ebihara, 2021); and same-backbone PREFIX and TERMINAL. Implementation details are in Appendix B; baseline sources, adaptations, and training settings are in Appendices C.2–C.3.

Table 2: Belief-path comparison of BELIEF* and baselines on Weather, Steel-RB, and Soccer. PACE-MAE measures reference-path error; AUROC uses each trajectory's last valid support point. Best results are bold; second-best are underlined.

	WEATHER		STEEL-RB		SOCCER	
Method	PACE-MAE ↓	AUROC ↑	PACE-MAE ↓	AUROC ↑	PACE-MAE ↓	AUROC ↑
GRU-D	<u>0.2681</u> ± 0.0019	<u>0.7824</u> ± 0.0074	0.3182 ± 0.0611	0.7871 ± 0.1131	0.1035 ± 0.0102	0.9591 ± 0.0074
mTAND	0.2757 ± 0.0066	0.7249 ± 0.0119	0.3290 ± 0.0583	0.8340 ± 0.0660	0.1097 ± 0.0038	0.9470 ± 0.0085
Raindrop	0.2697 ± 0.0014	0.7490 ± 0.0080	0.3808 ± 0.0171	0.6153 ± 0.0593	0.3250 ± 0.0990	0.8101 ± 0.1070
Hi-Patch	0.2711 ± 0.0028	0.7750 ± 0.0107	0.3910 ± 0.0357	0.7023 ± 0.0743	<u>0.0933</u> ± 0.0054	<u>0.9629</u> ± 0.0038
ATENet	0.2770 ± 0.0041	0.7093 ± 0.0120	0.3475 ± 0.0248	<u>0.8538</u> ± 0.0143	0.1250 ± 0.0065	0.9385 ± 0.0064
iTransformer	0.2715 ± 0.0015	0.7620 ± 0.0054	0.3365 ± 0.0644	0.7690 ± 0.0617	0.0949 ± 0.0042	0.9447 ± 0.0108
Dynamic-DeepHit	0.2758 ± 0.0008	0.6607 ± 0.0046	0.3305 ± 0.0147	0.5812 ± 0.0258	0.1132 ± 0.0067	0.9475 ± 0.0017
SPRT-TANDEM	0.3007 ± 0.0090	0.7418 ± 0.0164	0.5903 ± 0.1624	0.4714 ± 0.0435	0.1102 ± 0.0141	0.9614 ± 0.0058
BELIEF*	**0.2637** ± 0.0035	0.7762 ± 0.0088	**0.2524** ± 0.0340	**0.8573** ± 0.0447	**0.0896** ± 0.0048	**0.9667** ± 0.0027
PREFIX	0.2718 ± 0.0045	**0.8139** ± 0.0047	0.3993 ± 0.0377	0.6952 ± 0.0171	0.0934 ± 0.0117	0.9586 ± 0.0049
TERMINAL	0.2860 ± 0.0071	0.7265 ± 0.0142	<u>0.3089</u> ± 0.0851	0.7685 ± 0.0871	0.1076 ± 0.0108	0.9538 ± 0.0103

4.2 SYNTHETIC BELIEF-PATH RECOVERY (RQ1)

On synthetic streams, where the exact posterior is available at every prefix, BELIEF achieves lower path KL and MAE than PREFIX and TERMINAL across all six encoder architectures (Table 1), with average KL reductions of 85.7% and 96.6%, respectively. These gains may partly reflect the event-specific likelihood targets used by BELIEF. When early prefixes contain little outcome information, the realized outcome can provide noisy supervision for intermediate probabilities. PREFIX reuses that outcome label across positions, whereas BELIEF learns from the value observed at each position.

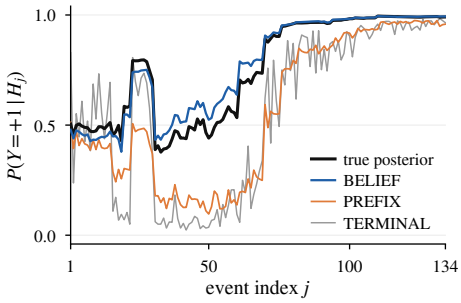

Figure 3: Synthetic belief paths on one nonlinear-binary test trajectory. Black denotes the analytic posterior; blue, orange, and gray denote BELIEF, PREFIX, and TERMINAL, respectively. The horizontal axis is the event index.

BELIEF also keeps mv close to one, indicating total squared variation comparable to the analytic posterior. In the synthetic example in Figure 3, all methods approach similar final probabilities, but PREFIX and TERMINAL deviate at intermediate events while BELIEF tracks the posterior more closely. Thus, endpoint agreement alone does not establish accurate belief-path recovery.

4.3 BELIEF PATHS ON REAL EVIDENCE STREAMS (RQ2)

On real event streams, the reference probability paths need not equal the true posterior, so we evaluate agreement with these references rather than exact recovery. BELIEF achieves the lowest mean PACE-MAE in all three domains (Table 2), with lower mean errors than matched PREFIX and TERMINAL and the adapted SPRT-TANDEM baseline. These comparisons support closer reference-path alignment in the evaluated settings. BELIEF also achieves the highest mean last-support AUROC on Steel-RB and Soccer. On Weather, PREFIX has higher mean AUROC while BELIEF has lower mean PACE-MAE, showing that path alignment and endpoint discrimination need not coincide.

4.4 DOWNSTREAM TASK PERFORMANCE (RQ3)

The downstream benchmarks test whether learning belief paths comes at the cost of predictive performance. BELIEF achieves the best mean results on Geolife and P12-LoS and remains among the top-performing methods on P12 and MIMIC-III (Table 3). It also achieves higher mean AUROC on the three classification tasks and lower mean MAE on P12-LoS than matched PREFIX and TERMINAL. These results support competitive downstream performance on the evaluated tasks.

Table 3: Downstream comparison of BELIEF* and baselines on classification and length-of-stay prediction. P12-LoS reports MAE. Best results are bold; second-best are underlined.

	P12	MIMIC-III	GEOLIFE	P12-LOS
Method	AUROC ↑	AUROC ↑	AUROC ↑	MAE ↓
GRU-D	0.7960 ± 0.0060	0.8220 ± 0.0180	0.7567 ± 0.0470	0.6177 ± 0.0023
mTAND	0.8530 ± 0.0030	0.8380 ± 0.0030	0.7064 ± 0.0121	0.6206 ± 0.0024
Raindrop	0.8200 ± 0.0060	0.7980 ± 0.0130	0.5343 ± 0.0851	0.6448 ± 0.0190
Hi-Patch	**0.8690** ± 0.0070	**0.8480** ± 0.0020	0.6920 ± 0.0296	<u>0.6099</u> ± 0.0021
ATENet	<u>0.8682</u> ± 0.0014	0.8013 ± 0.0229	0.6033 ± 0.0569	0.6106 ± 0.0009
iTransformer	0.8356 ± 0.0057	0.8159 ± 0.0032	0.6878 ± 0.0068	0.6472 ± 0.0034
Dynamic-DeepHit	0.8517 ± 0.0013	0.8406 ± 0.0036	0.6440 ± 0.0229	0.6644 ± 0.0073
SPRT-TANDEM	0.8059 ± 0.0063	0.7990 ± 0.0046	0.6199 ± 0.0199	0.6989 ± 0.0239
BELIEF*	0.8608 ± 0.0032	<u>0.8453</u> ± 0.0054	**0.7834** ± 0.0213	**0.6062** ± 0.0035
PREFIX	0.8065 ± 0.0070	0.8094 ± 0.0031	0.7724 ± 0.0100	0.6525 ± 0.0041
TERMINAL	0.8126 ± 0.0039	0.8087 ± 0.0017	<u>0.7743</u> ± 0.0122	0.6461 ± 0.0030

4.5 COMPONENT ABLATIONS (RQ4)

Ablation settings. To assess each component, we construct six one-component variants of BELIEF. *rp. free evidence increment* replaces likelihood-derived evidence with unconstrained increments while retaining accumulation. *rp. direct read-out* replaces Bayesian readout with a direct head on frozen BELIEF representations. *w/o learned prior* uses a fixed uniform prior instead of a prior estimated from training data. *w/o variable identity* removes variable/channel identity. *w/o shared atoms* uses outcome-specific rather than shared evidence components. *w/o inter-arrival encoding* removes inter-arrival encoding while retaining other time information. Detailed definitions are provided in Appendix C.5.

Table 4: Component ablations of BELIEF* on four synthetic classification datasets and three real-world domains. Synthetic results report mean path MAE; real-domain results report PACE-MAE. Best results are bold; second-best are underlined.

	SYNTHETIC	WEATHER	STEEL-RB	SOCCER
Method	MAE ↓	PACE-MAE ↓	PACE-MAE ↓	PACE-MAE ↓
BELIEF*	**0.0940**	**0.2637**	**0.2524**	**0.0896**
rp. free evidence increment	0.3506	0.2902	0.6874	0.1385
rp. direct read-out	0.6239	<u>0.2676</u>	0.4752	<u>0.1282</u>
w/o learned prior	<u>0.0953</u>	0.4310	0.3444	0.2735
w/o variable identity	0.7875	0.3284	0.3738	0.2696
w/o shared atoms	0.1078	0.4009	0.4120	0.4276
w/o inter-arrival encoding	0.1457	0.4267	<u>0.3120</u>	0.2332

Results. The full model achieves the lowest error in all four columns of Table 4. Both replacement variants increase path error, showing that likelihood-derived evidence with Bayesian accumulation preserves belief paths more accurately than free evidence increments or direct readout. Removing the learned prior or inter-arrival encoding also degrades performance, confirming the importance of the initial prior and temporal spacing. Removing variable identity or shared evidence components likewise worsens the results. Together, the ablations support all six design choices.

5 CONCLUSION

We formalized belief-path learning and introduced BELIEF, a generative framework for recovering evolving beliefs about the same eventual outcome from event streams without intermediate belief labels. BELIEF models incoming evidence conditioned on candidate outcomes and the full preceding history, then constructs belief paths through accumulated likelihoods and Bayesian updating. The framework supports binary, multiclass, and continuous outcomes with explicit evidence contributions. Under the observation-context assumption, our analysis bounds expected posterior KL error by prior and accumulated evidence-model errors and establishes exact recovery under correct prior and evidence models. Experiments show more accurate synthetic path recovery than prefix and terminal-only supervision across diverse encoder architectures, closer alignment with real-world reference probability paths, and competitive downstream performance. Together, these findings support conditional evidence modeling for recovering belief evolution, beyond predicting the final outcome alone.

AI USE STATEMENT

The original text of this paper was written by the authors, and all code was written by the authors. We used generative AI tools for two purposes. First, we used them to polish the language of the manuscript: making the expression read naturally for native English readers, correcting grammar, and improving the coherence between sentences and paragraphs. Second, we used them to help schedule and dispatch the experiment scripts across computing resources. The authors reviewed all AI-assisted edits and take full responsibility for the content of this paper, including its text, claims, and code.

REPRODUCIBILITY STATEMENT

Data. The synthetic benchmark is fully specified in Appendix C.1.1, including its value dynamics, exact belief paths, and splits. The three real-domain datasets are built from the sources whose links are given in Appendix C.1.2. The downstream benchmarks P12, MIMIC-III, and Geolife are publicly available, and their preprocessing and splits are described in Appendix C.1.3.

Code. The source code of BELIEF, the synthetic data generator, and the training and evaluation scripts are available in an anonymous GitHub repository at `https://anonymous.4open.science/r/belief-public-7346`.

REFERENCES

Ahmed M. Alaa, Scott Hu, and Mihaela van der Schaar. Learning from clinical judgments: Semi-Markov-modulated marked Hawkes processes for risk prognosis. In *International Conference on Machine Learning*, 2017.

Douglas T. Breeden and Robert H. Litzenberger. Prices of state-contingent claims implicit in option prices. *The Journal of Business*, 1978.

Ryan S. Brill, Ronald Yurko, and Abraham J. Wyner. Exploring the difficulty of estimating win probability: A simulation study. *Journal of Quantitative Analysis in Sports*, 2026.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, 2019.

Akinori F. Ebihara, Taiki Miyagawa, Kazuyuki Sakurai, and Hitoshi Imaoka. Sequential density ratio estimation for simultaneous optimization of speed and accuracy. In *International Conference on Learning Representations*, 2021.

Akinori F. Ebihara, Taiki Miyagawa, Kazuyuki Sakurai, and Hitoshi Imaoka. Learning the optimal stopping for early classification within finite horizons via sequential probability ratio test. In *International Conference on Learning Representations*, 2025.

Harrison Fah, Russell Greiner, and Roger A. Dixon. Temporally consistent survival prediction for non-uniform longitudinal data. *Journal of Biomedical Informatics*, 2026.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 2000.

Thomas M. Hamill, Eric Engle, David Myrick, Matthew Peroutka, Christina Finan, and Michael Scheuerer. The U.S. National Blend of Models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Monthly Weather Review*, 2017.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, 2020.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016.

Julian Klug, Guillaume Leclerc, Elisabeth Dirren, and Emmanuel Carrera. Machine learning for early dynamic prediction of functional outcome after stroke. *Communications Medicine*, 2024.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 2020.

Sangho Lee, Kyeongseo Min, Youngdoo Son, and Hyungrok Do. Adaptive time encoding for irregular multivariate time-series classification. In *Advances in Neural Information Processing Systems*, 2025.

Zhiyuan Jerry Lin, Hao Sheng, and Sharad Goel. Probability paths and the structure of predictions over time. In *Advances in Neural Information Processing Systems*, 2021.

Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with LSTM recurrent neural networks. In *International Conference on Learning Representations*, 2016.

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*, 2026.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-Patch: Hierarchical patch GNN for irregular multivariate time series. In *International Conference on Machine Learning*, 2025.

Lucas Maystre, Gabriel Barello, Tudor Berariu, Aleix Cambray, Rares Dolga, Alvaro Ortega Gonzalez, Andrei Nica, and David Barber. Incremental sequence classification with temporal consistency. In *Advances in Neural Information Processing Systems*, 2025.

Victor Medina-Olivares, Raffaella Calabrese, Jonathan Crook, and Finn Lindgren. Joint models for longitudinal and discrete survival data in credit scoring. *European Journal of Operational Research*, 2023.

Taiki Miyagawa and Akinori F. Ebihara. The power of log-sum-exp: Sequential density ratio matrix estimation for speed-accuracy optimization. In *International Conference on Machine Learning*, 2021.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*, 2001.

Byoungwoo Park, Hyungi Lee, and Juho Lee. Amortized control of continuous state space Feynman-Kac model for irregular time series. In *International Conference on Learning Representations*, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallègue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 2024.

Eva Regnier and Patrick A. Harr. A dynamic decision model applied to hurricane landfall. *Weather and Forecasting*, 2006.

Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 2011.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 2020.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Ikaro Silva, George B. Moody, Daniel J. Scott, Leo Anthony Celi, and Roger G. Mark. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology Challenge 2012. In *Computing in Cardiology*, 2012.

Francesco Spinnato and Cristiano Landi. PYRREGULAR: A unified framework for irregular time series, with classification benchmarks. In *International Conference on Learning Representations*, 2026.

Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Transactions on Knowledge Discovery from Data*, 2019.

Hans-Christian Thorsen-Meyer, Annelaura B. Nielsen, Anna P. Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J. Chmura, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Kirstine Belling, Søren Brunak, and Anders Perner. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *International Conference on Machine Learning*, 2024.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *International Conference on Machine Learning*, 2024.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022.

Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on GPS data. In *International Conference on Ubiquitous Computing*, 2008.

Yu Zheng, Xing Xie, and Wei-Ying Ma. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin*, 2010.

Appendix

A Technical Details and Proofs

Notation. Throughout this appendix, write $Z_i = (\tau_i, m_i)$. For each history and observation context, conditional value densities use a common reference measure across candidate outcomes and models. We write $p_\theta(\mathrm{d}x \mid \cdot)$ for the corresponding probability kernel and $q_{\theta,j}(\mathrm{d}y) = q_{\theta,j}(y)\,\mathrm{d}\mu(y)$ for the belief measure, with the same convention for $p^\star$ and $q_j^\star$.

A.1 Proof of Theorem 3.1

Theorem 3.1 (Exact belief-path recovery)**.** *Suppose the observation-context assumption holds. If $\pi = P_Y$ and, for every $i \le J$,*

$$p_\theta(\cdot \mid H_{i-1}, \tau_i, m_i, Y) = p^\star(\cdot \mid H_{i-1}, \tau_i, m_i, Y) \qquad P\text{-a.s.},$$

then the recovered outcome distributions satisfy

$$P\big(q_{\theta,j} = q_j^\star \text{ for all } j = 0, \dots, J\big) = 1.$$

Proof. The claim follows by induction on j. At $j = 0$, $q_{\theta,0} = \pi = P_Y = q_0^\star$. Suppose that $q_{\theta,j-1} = q_{j-1}^\star$ P-almost surely. The observation-context assumption gives

$$P(Y \in \cdot \mid H_{j-1}, Z_j) = q_{j-1}^\star \qquad P\text{-a.s.}$$

At P_{H_{j-1},Z_j}-almost every context, the conditional value kernels agree for $q_{j-1}^\star$-almost every outcome. Bayes' rule and the induction hypothesis therefore yield

$$\begin{aligned} q_j^\star(\mathrm{d}y) &= \frac{p^\star(x_j \mid H_{j-1}, Z_j, y)\, q_{j-1}^\star(\mathrm{d}y)}{\int_{\mathcal{Y}} p^\star(x_j \mid H_{j-1}, Z_j, u)\, q_{j-1}^\star(\mathrm{d}u)} \\ &= \frac{p_\theta(x_j \mid H_{j-1}, Z_j, y)\, q_{\theta,j-1}(\mathrm{d}y)}{\int_{\mathcal{Y}} p_\theta(x_j \mid H_{j-1}, Z_j, u)\, q_{\theta,j-1}(\mathrm{d}u)} = q_{\theta,j}(\mathrm{d}y) \qquad P\text{-a.s.} \end{aligned}$$

Induction and the finite intersection over $j = 0, \dots, J$ complete the proof. □

A.2 Proof of Theorem 3.2

Theorem 3.2 (Posterior error under misspecification)**.** *Under the observation-context assumption, for every fixed prefix $j \le J$ with finite right-hand side,*

$$\mathbb{E}_{H_j \sim P} D_{\mathrm{KL}}(q_j^\star \,\|\, q_{\theta,j}) \le D_{\mathrm{KL}}(P_Y \,\|\, \pi) + \sum_{i=1}^{j} \varepsilon_i(\theta).$$

Proof. Fix $j \le J$ with finite right-hand side, and fix versions of the kernels $g_i(\mathrm{d}z \mid h) = P(Z_i \in \mathrm{d}z \mid H_{i-1} = h)$. The observation-context assumption gives the first factorization below; the second defines the comparison law Q_θ:

$$P_{Y,H_j}(\mathrm{d}y, \mathrm{d}z_{1:j}, \mathrm{d}x_{1:j}) = P_Y(\mathrm{d}y) \prod_{i=1}^{j} g_i(\mathrm{d}z_i \mid H_{i-1}) p^\star(\mathrm{d}x_i \mid H_{i-1}, Z_i, y),$$

$$Q_{\theta,Y,H_j}(\mathrm{d}y, \mathrm{d}z_{1:j}, \mathrm{d}x_{1:j}) = \pi(\mathrm{d}y) \prod_{i=1}^{j} g_i(\mathrm{d}z_i \mid H_{i-1}) p_\theta(\mathrm{d}x_i \mid H_{i-1}, Z_i, y).$$

Applying the KL chain rule in the order $Y, Z_1, X_1, \dots, Z_j, X_j$ yields

$$\begin{aligned} D_j(\theta) &:= D_{\mathrm{KL}}(P_{Y,H_j} \,\|\, Q_{\theta,Y,H_j}) \\ &= D_{\mathrm{KL}}(P_Y \,\|\, \pi) + \sum_{i=1}^{j} \varepsilon_i(\theta) < \infty, \end{aligned} \tag{11}$$

since the shared context kernels contribute zero conditional KL. In particular, $P_{H_j} \ll Q_{\theta,H_j}$. For a fixed history, the context factors are independent of the candidate outcome and cancel in Bayes' rule. Hence

$$Q_\theta(\mathrm{d}y \mid H_j) = \frac{\prod_{i=1}^{j} p_\theta(x_i \mid H_{i-1}, Z_i, y)\,\pi(\mathrm{d}y)}{\int_{\mathcal{Y}} \prod_{i=1}^{j} p_\theta(x_i \mid H_{i-1}, Z_i, u)\,\pi(\mathrm{d}u)} = q_{\theta,j}(\mathrm{d}y), \tag{12}$$

first Q_{θ,H_j}-almost surely and, by absolute continuity, also P_{H_j}-almost surely. Applying the chain rule instead in the order H_j, Y gives

$$D_j(\theta) = D_{\mathrm{KL}}(P_{H_j} \,\|\, Q_{\theta,H_j}) + \mathbb{E}_{H_j\sim P} D_{\mathrm{KL}}(q_j^\star \,\|\, q_{\theta,j}). \tag{13}$$

The claimed bound follows from Equation 11 and nonnegativity of the marginal KL term. □

A.3 Proof of Theorem 3.3

Theorem 3.3 (Posterior improvement under relative marginal stability)**.** *Under the observation-context and relative marginal-stability conditions, any two models θ_A, θ_B with finite relevant losses and KL divergences satisfy*

$$\begin{aligned} L_j(\theta_B) < L_j(\theta_A) \quad &\Longrightarrow \\ R_j(\theta_A) - R_j(\theta_B) \geq (1-\kappa)\big[L_j(\theta_A) - L_j(\theta_B)\big] &> 0. \end{aligned}$$

Proof. Fix $1 \leq j \leq J$ and use the comparison laws from Appendix A.2, with a common prior π and observation-context kernels. Define

$$\begin{aligned} L_j(\theta) &= \mathbb{E}_P\left[-\sum_{i=1}^{j} \log p_\theta(X_i \mid H_{i-1}, Z_i, Y)\right], \\ R_j(\theta) &= \mathbb{E}_{H_j\sim P} D_{\mathrm{KL}}(q_j^\star \,\|\, q_{\theta,j}), \\ M_j(\theta) &= D_{\mathrm{KL}}(P_{H_j} \,\|\, Q_{\theta,H_j}). \end{aligned} \tag{14}$$

For both compared models, assume that the likelihood ratios below are well-defined P-almost surely, the prefix losses are integrable, and $D_j(\theta) < \infty$. Here L_j is a population loss rather than an empirical training loss, and the common prior need not equal P_Y.

By Equation 13,

$$D_j(\theta) = M_j(\theta) + R_j(\theta), \qquad \theta \in \{\theta_A, \theta_B\}.$$

Iterated expectation gives

$$\varepsilon_i(\theta) = \mathbb{E}_P\left[\log \frac{p^\star(X_i \mid H_{i-1}, Z_i, Y)}{p_\theta(X_i \mid H_{i-1}, Z_i, Y)}\right].$$

Because the prior and observation-context kernels are shared, Equation 11 therefore yields

$$\begin{aligned} D_j(\theta_A) - D_j(\theta_B) &= \sum_{i=1}^{j} \big[\varepsilon_i(\theta_A) - \varepsilon_i(\theta_B)\big] \\ &= \mathbb{E}_P\left[\sum_{i=1}^{j} \log \frac{p_{\theta_B}(X_i \mid H_{i-1}, Z_i, Y)}{p_{\theta_A}(X_i \mid H_{i-1}, Z_i, Y)}\right] \\ &= L_j(\theta_A) - L_j(\theta_B). \end{aligned}$$

The true conditional likelihoods cancel in the second equality. All quantities are finite, so subtraction yields the exact identity

$$\begin{aligned} R_j(\theta_A) - R_j(\theta_B) &= D_j(\theta_A) - D_j(\theta_B) - \big[M_j(\theta_A) - M_j(\theta_B)\big] \\ &= L_j(\theta_A) - L_j(\theta_B) - \big[M_j(\theta_A) - M_j(\theta_B)\big]. \end{aligned} \tag{15}$$

Let $d_j = L_j(\theta_A) - L_j(\theta_B) > 0$. Relative marginal stability gives $|M_j(\theta_A) - M_j(\theta_B)| \leq \kappa d_j$. Hence,

$$\begin{aligned} R_j(\theta_A) - R_j(\theta_B) &= d_j - \big[M_j(\theta_A) - M_j(\theta_B)\big] \\ &\geq d_j - |M_j(\theta_A) - M_j(\theta_B)| \\ &\geq (1-\kappa) d_j > 0, \end{aligned}$$

where the last inequality uses $0 \leq \kappa < 1$. □

The following lemma verifies the relative marginal-stability condition of Theorem 3.3 for a Gaussian evidence model family under spherical additive noise.

Lemma A.1 (Relative marginal stability in a Gaussian model family)**.** *Fix a finite J and let the observation contexts Z_i be deterministic, so that H_j reduces to $X_{1:j}$. Suppose $X_i = Y\mu^\star + \zeta_i$, where $Y \in \{-1, +1\}$ and the ζ_i are i.i.d., spherically symmetric, independent of Y, and have finite second moment. Consider the evidence models*

$$p_w(x \mid H_{i-1}, Z_i, y) = \mathcal{N}(x; yw, I_d), \qquad \|w\| = r > 0,$$

with a shared prior $\pi(\pm 1) > 0$. Under the preceding integrability and finite-KL conditions throughout this family, there exists $0 \le \kappa < 1$, uniform over model pairs and $1 \le j \le J$, such that, with L_j and M_j as in Equation 14,

$$|M_j(w_A) - M_j(w_B)| \le \kappa |L_j(w_A) - L_j(w_B)|.$$

The noise need not be Gaussian, and r need not equal $\|\mu^\star\|$.

Proof. Set $b = \frac{1}{2}\log[\pi(+1)/\pi(-1)]$, $c_w = w^\top \mu^\star$, and $\mathcal{C} = [-r\|\mu^\star\|, r\|\mu^\star\|]$. Spherical symmetry gives $w^\top \sum_{i \le j} \zeta_i$ the same law for all $\|w\| = r$; let ξ_j have this law independently of Y. Define $F_j(c) = \mathbb{E}\log\cosh(b + jYc + \xi_j)$. Expanding the Gaussian loss and marginal density gives, with $c_A = c_{w_A}$ and $c_B = c_{w_B}$,

$$\begin{aligned} L_j(w_A) - L_j(w_B) &= j(c_B - c_A), \\ M_j(w_A) - M_j(w_B) &= F_j(c_B) - F_j(c_A). \end{aligned}$$

The 1-Lipschitz property of $\log\cosh$ and bounded convergence yield

$$F_j'(c) = j\,\mathbb{E}[Y \tanh(b + jYc + \xi_j)],$$

a continuous derivative with $|F_j'(c)| < j$. Compactness of $\mathcal{C}$ therefore gives

$$\kappa_j = \max_{c \in \mathcal{C}} |F_j'(c)|/j < 1.$$

The mean value theorem yields the claimed bound with κ_j; when $c_A = c_B$, both differences vanish. Taking $\kappa = \max_{1 \le j \le J} \kappa_j < 1$ completes the proof. □

B Implementation Details

This appendix specifies the instantiation of BELIEF used in the real-domain and downstream experiments: its architecture (Appendix B.1), training and inference procedures (Appendix B.2), and the readouts for each outcome type (Appendix B.3). Training and hyperparameter settings are given in Appendix C.3.

B.1 A Simple Instantiation of BELIEF

The instantiation pairs a causal Transformer history encoder with an outcome-conditioned mixture decoder (Figure 4). The encoder summarizes the events that precede the current one; the decoder turns this summary, the current observation context, and a candidate outcome into a normalized distribution over the value about to be recorded.

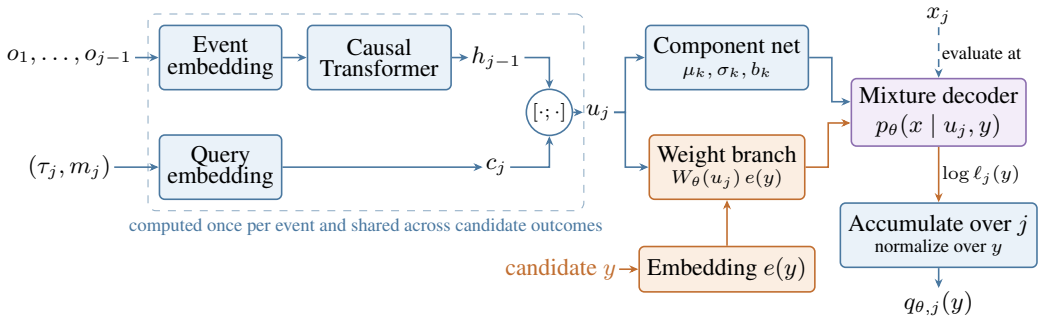

Figure 4: Architecture of the simple instantiation. Blue modules do not depend on the candidate outcome and are computed once per event; orange modules are evaluated for each candidate y. The recorded value x_j enters only as the point at which the decoder density is evaluated and joins the history after it has been scored.

Event embedding. Each observation $o_i = (\tau_i, m_i, x_i)$ is mapped to a vector

$$z_i = \mathrm{LN}\big(v(x_i, m_i) + E_{\mathrm{var}}[m_i] + \phi(\tau_i, \Delta\tau_i)\big), \qquad \Delta\tau_i = \tau_i - \tau_{i-1},$$

where LN is layer normalization and E_{var} is a learned variable embedding. For numeric variables, v adds a linear projection of the standardized value to an embedding of its quantized level, so the token retains the exact value while attention can also match on coarse levels; categorical variables use a per-variable lookup table instead. The time encoding ϕ applies sinusoidal features at geometrically spaced frequencies to the normalized time τ_i/T_{H}, where T_{H} is the task horizon, and to the log inter-arrival time $\log(1 + \Delta\tau_i)$, followed by a linear map. Irregular sampling therefore enters the model through $\Delta\tau_i$ rather than through a fixed grid.

Causal history encoder. The embedded events are shifted by one position behind a learned start token, $(z_{\mathrm{s}}, z_1, \ldots, z_{J-1})$, and processed by a causal Transformer with pre-layer normalization. Each block applies masked multi-head self-attention and a position-wise feed-forward network, both with residual connections. The output at position j is $h_{j-1} = \mathrm{Enc}_\theta(H_{j-1})$, which depends on $o_1, \ldots, o_{j-1}$ only. The current observation context is embedded in the same way as an event without its value, $c_j = E_{\mathrm{var}}[m_j] + \phi(\tau_j, \Delta\tau_j)$, and the decoder input is $u_j = [h_{j-1}; c_j]$. When a task provides covariates known at the start of the stream, their projection is added to the start token and to c_j. The encoder never receives a candidate outcome, so one encoder pass serves all candidates.

Outcome-conditioned mixture decoder. The decoder has a component branch that does not see the outcome and a weight branch that does. A multilayer perceptron maps u_j to base mixture logits $b_\theta(u_j) \in \mathbb{R}^K$ and to the location $\mu_{\theta,k}(u_j)$ and scale $\sigma_{\theta,k}(u_j) > 0$ of K Student-t components with shared degrees of freedom $\nu > 2$. A linear map produces $W_\theta(u_j) \in \mathbb{R}^{K \times E}$, which acts on an outcome embedding $e(y) \in \mathbb{R}^E$:

$$w_\theta(u_j, y) = \mathrm{softmax}\big(b_\theta(u_j) + W_\theta(u_j)\, e(y)\big), \qquad p_\theta(x \mid u_j, y) = \sum_{k=1}^{K} w_{\theta,k}(u_j, y)\, f_{\theta,k}(x \mid u_j), \tag{16}$$

where $f_{\theta,k}(\cdot \mid u_j)$ is the Student-t density with parameters $(\mu_{\theta,k}, \sigma_{\theta,k}, \nu)$. This is the shared-mixture construction of Equation 2. The embedding e is a learned table for discrete outcomes and a small network of the outcome value for continuous outcomes; one reference outcome y_{ref} has $e(y_{\mathrm{ref}}) = 0$, and W_θ is initialized at zero, so training starts from outcome-independent evidence and a belief path that stays at the prior. Evaluating Equation 16 at the recorded value gives the evidence likelihood $\ell_j(y) = p_\theta(x_j \mid u_j, y)$.

Bounded evidence. Because the components are shared across outcomes, the likelihood ratio between two candidates is a ratio of two mixtures of the same densities and lies between the smallest and largest ratio of their weights. For every recorded value x,

$$\left|\log \frac{p_\theta(x \mid u_j, y)}{p_\theta(x \mid u_j, y')}\right| \le \max_k \left|\log \frac{w_{\theta,k}(u_j, y)}{w_{\theta,k}(u_j, y')}\right|.$$

A single observation, however extreme its value, can therefore move the log-odds between two outcomes by no more than the largest log-ratio of their mixture weights at that position. Appendix C.5 removes the sharing in the *w/o shared atoms* ablation.

B.2 Training and Inference Procedures

Algorithms 1 and 2 separate conditional evidence learning from sequential belief updating. Training fixes each stream's realized outcome and scores every recorded value under the preceding history and the current observation context. Inference evaluates the same value under every candidate outcome and normalizes the accumulated log scores. In both algorithms, $p_\theta(x \mid u_j, y)$ is the decoder of Equation 16. Algorithm 1 optimizes the likelihood objective of Equation 3.

Algorithm 1 BELIEF: evidence-model training

Require: Training streams $\{(H^{(n)}_{J_n}, y^{(n)})\}_{n=1}^{N}$; evidence model p_θ; optimization settings
Ensure: Trained evidence-model parameters θ

Initialize θ
repeat
  Sample a minibatch $\mathcal{B}$ of streams; $\mathcal{L}_{\mathrm{NLL}} \leftarrow 0$
  for each $(H_J, y) \in \mathcal{B}$ **do**
    $H_0 \leftarrow \varnothing$
    for $j = 1, \ldots, J$ **do**
      Read $o_j = (\tau_j, m_j, x_j)$
      $h_{j-1} \leftarrow \mathrm{Enc}_\theta(H_{j-1})$ ▷ past events only
      $u_j \leftarrow [h_{j-1}; c_j]$ with c_j built from (τ_j, m_j)
      $\mathcal{L}_{\mathrm{NLL}} \leftarrow \mathcal{L}_{\mathrm{NLL}} - \log p_\theta(x_j \mid u_j, y)$ ▷ realized outcome; x_j is the target
      $H_j \leftarrow (H_{j-1}, o_j)$
    end for
  end for
  Take an optimizer step on $\mathcal{L}_{\mathrm{NLL}}$ with respect to θ
until the stopping criterion is met
return θ

Algorithm 2 BELIEF: belief-path inference

Require: Trained parameters θ; event stream $H_J = (o_1, \ldots, o_J)$; prior π on $\mathcal{Y}$
Ensure: Belief path $(q_{\theta,0}, \ldots, q_{\theta,J})$

$H_0 \leftarrow \varnothing$; $q_{\theta,0} \leftarrow \pi$; $S_0(y) \leftarrow \log \pi(y)$ for all $y \in \mathcal{Y}$
for $j = 1, \ldots, J$ **do**
  Receive $o_j = (\tau_j, m_j, x_j)$
  $h_{j-1} \leftarrow \mathrm{Enc}_\theta(H_{j-1})$; $u_j \leftarrow [h_{j-1}; c_j]$ ▷ once per event
  $\log \ell_j(y) \leftarrow \log p_\theta(x_j \mid u_j, y)$ for all $y \in \mathcal{Y}$ ▷ same x_j, different y
  $S_j(y) \leftarrow S_{j-1}(y) + \log \ell_j(y)$ for all $y \in \mathcal{Y}$
  $q_{\theta,j}(y) \leftarrow e^{S_j(y)} \big/ \int_{\mathcal{Y}} e^{S_j(u)} \, \mathrm{d}\mu(u)$
  $H_j \leftarrow (H_{j-1}, o_j)$
end for
return $(q_{\theta,0}, \ldots, q_{\theta,J})$

The loops state information dependencies rather than execution order: each observation joins the history only after its value has been scored. With shifted inputs and a causal mask, all positions of a stream are evaluated in one parallel pass during training, and the full belief path of a stream is obtained in one encoder pass followed by $|\mathcal{Y}|$ decoder evaluations per event at inference. For continuous outcomes, $\mathcal{Y}$ is represented by a finite grid of candidate values.

B.3 Readouts for Binary, Multiclass, and Continuous Outcomes

All readouts normalize the accumulated score of Equation 5,

$$S_j(y) = \log \pi(y) + \sum_{i=1}^{j} \log \ell_i(y),$$

over the outcome space. For discrete outcomes, the prior π is a learned categorical distribution fitted by its own likelihood on the training outcomes; for continuous outcomes, it is a fixed density on the standardized outcome scale.

Binary outcomes. For $\mathcal{Y} = \{0, 1\}$,

$$q_{\theta,j}(1) = \frac{e^{S_j(1)}}{e^{S_j(0)} + e^{S_j(1)}} = \sigma\big(S_j(1) - S_j(0)\big), \qquad S_j(1) - S_j(0) = \log \frac{\pi(1)}{\pi(0)} + \sum_{i=1}^{j} \log \frac{\ell_i(1)}{\ell_i(0)}, \tag{17}$$

so the log-odds is the prior log-odds plus one log-likelihood ratio per event.

Multiclass outcomes. For $\mathcal{Y} = \{1, \ldots, C\}$, softmax normalization gives

$$q_{\theta,j}(c) = \frac{e^{S_j(c)}}{\sum_{c'=1}^{C} e^{S_j(c')}}, \qquad c = 1, \ldots, C. \tag{18}$$

The per-event evidence is a vector of $C - 1$ log-likelihood ratios against the reference class; any other reference yields the same beliefs.

Continuous outcomes. For a continuous outcome, normalization yields a posterior density, and the point prediction is its mean:

$$q_{\theta,j}(y) = \frac{e^{S_j(y)}}{\int_{\mathcal{Y}} e^{S_j(u)}\, \mathrm{d}u}, \qquad \widehat{y}_j = \int_{\mathcal{Y}} y\, q_{\theta,j}(y)\, \mathrm{d}y. \tag{19}$$

The normalizing integral is computed over a finite grid of candidate values.

Evidence ratios. For any two candidates with positive prior and likelihoods, all three readouts share the decomposition

$$\log \frac{q_{\theta,j}(y)}{q_{\theta,j}(y')} = \log \frac{\pi(y)}{\pi(y')} + \sum_{i=1}^{j} \log \frac{\ell_i(y)}{\ell_i(y')}, \tag{20}$$

where the ratio is a density ratio in the continuous case.

C Experimental Details

This appendix describes the datasets and tasks (Appendix C.1), the compared methods (Appendix C.2), the implementation and training configurations (Appendix C.3), the evaluation metrics (Appendix C.4), and the ablation variants (Appendix C.5).

C.1 Datasets and Tasks

C.1.1 Synthetic event streams

The synthetic benchmark evaluates belief-path recovery where the target posterior is known exactly. Each trajectory has an eventual outcome Y and an irregular sequence of observations $o_j = (\tau_j, m_j, x_j)$. Trajectories are generated in three steps: the outcome is drawn first, the observation times and variables are drawn independently of it, and each value is drawn given the preceding values, its observation context, and the outcome. The observation context is therefore uninformative about Y, as assumed in Section 3.5, while the values carry the evidence.

Outcomes and observation times. Binary outcomes are uniform on two classes with codes $c(Y) \in \{-1, +1\}$; four-class outcomes are uniform with codes $c(Y) \in \{-3, -1, +1, +3\}$; continuous outcomes follow $Y \sim \mathcal{N}(0, 1)$ with $c(Y) = Y$. Observation times follow a Gamma renewal process on $[0, 48]$ with mean gap 0.32 and coefficient of variation 0.82, which matches the inter-arrival variability of P12 and gives about 150 events per stream. Each event's variable is drawn uniformly from $M = 8$ variables.

Value dynamics. With $\varepsilon_j \sim \mathcal{N}(0,1)$, the linear mechanism is

$$x_j = e^{-0.7\,\Delta\tau_j}\, x_{j-1} + \mu_{m_j} c(Y) + \sigma_{m_j}\varepsilon_j, \qquad x_0 = 0,$$

and the nonlinear mechanism is

$$x_j = 0.8\tanh(0.7\,\bar{x}_j) + g_j\,\mu_{m_j} c(Y) + \sigma_{m_j}\varepsilon_j, \qquad g_j = 1 + 0.5\tanh(\bar{x}_j),$$

where $\bar{x}_j$ is the average of earlier values weighted by $e^{-0.7(\tau_j-\tau_i)}$ ($\bar{x}_1 = 0$). In both mechanisms, the history term and the gain g_j depend on past values only, so the outcome enters through the loading $\mu_{m_j} c(Y)$ alone. For each data seed, noise scales σ_m and loading magnitudes u_m are drawn log-uniformly from $[0.5, 2]$, and $\mu_m = s\,\mathrm{sign}_m u_m$ with four positive and four negative signs. The common scale s is calibrated on 400 pilot trajectories so that the expected accumulated information $\mathbb{E}[A_J]$ defined below equals 1.894, which corresponds to an expected terminal log-odds magnitude of 4 in the balanced binary case. All six settings, the three outcome types under each mechanism, therefore carry comparable total evidence.

Exact belief paths. Let η_j denote the history term, $e_j = x_j - \eta_j$, $a_j = g_j\mu_{m_j}$ (with $g_j \equiv 1$ in the linear mechanism), $b_j = \sum_{i\le j} a_i e_i/\sigma^2_{m_i}$, and $A_j = \sum_{i\le j} a_i^2/\sigma^2_{m_i}$. Because the history term is identical under every outcome and the observation context is independent of Y, the posterior after each prefix has a closed form:

$$\operatorname{logit} P(Y{=}1 \mid H_j) = 2b_j, \qquad q_j^\star(c) \propto \exp\!\big(c\,b_j - \tfrac{1}{2}c^2 A_j\big), \qquad Y \mid H_j \sim \mathcal{N}\Big(\frac{b_j}{1+A_j}, \frac{1}{1+A_j}\Big)$$

for the binary, four-class, and continuous settings. Exact paths are used only for evaluation; no method receives them during training.

Scale and splits. Each setting contains 8,571 trajectories, split by trajectory into 5,999 training, 1,286 validation, and 1,286 test trajectories. One data seed fixes the variable parameters, the trajectories, and the split; Table 1 uses seeds 3407–3411, and Appendix D.1 reports the six settings separately.

C.1.2 Real-domain reference-path tasks

Weather. The evidence consists of METAR and SPECI surface reports from US Automated Surface Observing System (ASOS) stations, distributed by NOAA's Meteorological Assimilation Data Ingest System[1] and obtained from the WeatherBench 2 archive (Rasp et al., 2024). Each trajectory is a station and a 12-hour verification window ending at 00 or 12 UTC. The label is positive if any report in the window records at least 0.01 in (0.254 mm) of one-hour precipitation, the event forecast by the 12-hour probability of precipitation of the National Blend of Models (Hamill et al., 2017). The evidence stream contains the reports of the preceding 84 hours, time-stamped at reception, with nine variables: temperature, dew point, wind speed, visibility, altimeter setting, sea-level pressure, one-hour precipitation, present-weather code, and report type (METAR or SPECI). The data cover 131 hourly ASOS stations with NBM coverage. The reference path is the NBM version 4.0 probability for the same window, taken from the text bulletins in NOAA's open-data archive at leads of 12–84 hours on a 6-hour grid. The data cover the December–February rainy seasons of 2020–2021 and 2021–2022, 37,499 windows in total. Within each season, windows are split chronologically into training (December 4–January 31), validation (February 1–14), and test (February 15–28) periods separated by 84-hour gaps, which yields 27,669, 4,991, and 4,839 windows with no shared reports.

Steel-RB. The task asks whether the settlement price of a rebar futures contract on the Shanghai Futures Exchange at its option-expiry date T reaches a strike K. Each trajectory is a contract and a relative offset $\delta \in \{-8, -6, -4, -2, 0, 2, 4\}\%$, with $K = S_{T-30}(1+\delta)$ for the spot price S 30 trading days before expiry. The evidence consists of 48 industrial series, covering prices, costs and margins, social and mill inventories, output, capacity utilization, and demand at daily, weekly, and monthly frequency; each observation is time-stamped at its publication date, and the stream covers trading days $T-60$ to $T-1$. Two events at the anchor date carry δ and the product variety.

[1] https://madis.ncep.noaa.gov/.

Futures prices and quantities derived from them are excluded from the evidence. The reference path is the option-implied probability of expiring above K, obtained from daily closing prices of the exchange's rebar call options as the central difference of call prices across adjacent strikes (Breeden & Litzenberger, 1978), interpolated linearly to K, and made non-increasing in strike by isotonic regression. The references are used only for evaluation, at reference dates within 20 trading days of T. Contracts are split by expiry into RB1905–RB2308 for training (364 trajectories), RB2402–RB2408 for validation (49), and RB2502–RB2606 for testing (119); contracts expiring between these ranges are excluded.

Soccer. The data are FiveThirtyEight's club soccer match records and season forecasts[2], released under CC BY 4.0. The files are retrieved from the Internet Archive snapshot of March 6, 2025.[3] Each trajectory follows one team through a league season, with end-of-season relegation as the outcome. Each match contributes up to nine variables: home or away status, goals scored and conceded, goal difference, result, the team's and the opponent's points per game over earlier matches of the season, and expected goals for and against where available. Time is measured in hours from the league's first match day. The reference path is the daily relegation probability published by FiveThirtyEight's Soccer Power Index (SPI) model; the SPI ratings, match probabilities, projected scores, and match-importance fields are excluded from the evidence. Keeping leagues with relegation yields 1,482 team-seasons in 21 leagues over 2016–2021, split forward in time into 2016–2019 for training (983), 2020 for validation (234), and 2021 for testing (265).

C.1.3 Downstream benchmarks

P12. P12 is the PhysioNet/Computing in Cardiology Challenge 2012 dataset (Silva et al., 2012; Goldberger et al., 2000): 11,988 intensive-care stays, the first 48 hours of each, and in-hospital mortality as the outcome. We use the preprocessing and the five splits released with Raindrop (Zhang et al., 2022), each with 9,590 training, 1,199 validation, and 1,199 test stays. Inputs are 36 time-varying variables and static fields (age, gender, height, ICU type, and weight), which enter as events at $\tau = 0$.

MIMIC-III. MIMIC-III (Johnson et al., 2016) is accessed under the PhysioNet credentialed data-use agreement. We use the in-hospital mortality cohort and split released with Hi-Patch (Luo et al., 2025), which follows the preprocessing of t-PatchGNN (Zhang et al., 2024): 21,107 stays, 16 time-varying variables, and a 48-hour observation window, split into 14,662 training, 3,217 validation, and 3,228 test stays.

Geolife. Geolife (Zheng et al., 2008; 2010) contains GPS trajectories with user-provided transport-mode labels. We use the 5,977 labeled segments with 11 transport modes distributed by PYRREGULAR (Spinnato & Landi, 2026) and split them by stratified sampling on the label, with random seed 42, into 3,587 training, 1,195 validation, and 1,195 test segments. Each GPS fix produces simultaneous latitude and longitude events, and at most the last 2,048 events of a segment are retained.

P12-LoS. P12-LoS uses the P12 event streams and splits to predict the length of stay of the 10,006 survivors with a recorded stay longer than two days. The target is transformed by $\log(1 + x)$ and then by a rank-based inverse normal transformation fitted on the training partition, so that the training targets are approximately standard normal.

C.1.4 Event representation and preprocessing

All datasets are represented as time–variable–value events. Numeric variables in the synthetic, Weather, Steel-RB, Soccer, and Geolife data are standardized with training-partition statistics, and categorical variables are represented by identifiers. Events are ordered by time and then by variable index, with static events first at equal times. Time is measured in hours, except for Steel-RB, which uses trading days.

[2]Forecast methodology: `https://web.archive.org/web/20240107152717/https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/`.

[3]Match records: `https://web.archive.org/web/20250306125411/https://projects.fivethirtyeight.com/soccer-api/club/spi_matches.csv`.

C.2 Baselines and Task Adaptations

C.2.1 Outcome-supervision strategies

BELIEF, PREFIX, and TERMINAL learn from completed event streams and their realized outcomes but differ in the training objective and in how the belief is read out. BELIEF fits outcome-conditioned observation distributions and obtains beliefs by accumulating their likelihoods. PREFIX predicts the outcome after every observed prefix, whereas TERMINAL applies the outcome loss only after the final observation. Both direct predictors output class probabilities for discrete outcomes and a Gaussian mean and scale for continuous outcomes, and both are applied to successive prefixes at evaluation to produce a belief path. Within each backbone, the three strategies share the event representation and the encoder architecture. BELIEF predicts x_j from H_{j-1} and the current time and variable, whereas the direct predictors use H_j to predict the outcome after observing o_j, so both have seen the same evidence when their belief at j is read. In the real-domain and downstream experiments, PREFIX and TERMINAL use the causal Transformer.

C.2.2 Backbone comparison

The synthetic comparison instantiates each strategy with six encoders. Each encoder maps the available history to a representation for the likelihood decoder or the outcome head, and every representation used to score a value is computed from earlier observations only.

Causal Transformer. The causal Transformer (Vaswani et al., 2017) applies masked self-attention over event tokens whose encoding includes observation times and inter-arrival intervals.

GRU. The GRU (Cho et al., 2014) processes event tokens sequentially and uses its hidden state at each position as the history representation.

GRU-ODE-Bayes. The encoder derived from GRU-ODE-Bayes (De Brouwer et al., 2019) evolves its hidden state in continuous time between observations and updates it at each arrival.

mTAND. The time-attention encoder of mTAND (Shukla & Marlin, 2021) attends from the history's own observation times to the available observations and passes the result through a recurrent layer.

ASTGI. ASTGI (Liu et al., 2026) builds spatio-temporal neighborhoods among the available observations to model temporal and cross-variable relations.

HSTU. HSTU (Zhai et al., 2024) encodes the event history with causal attention and relative-time information.

C.2.3 Real-world and downstream baselines

Every baseline receives only the observations available at the evaluation time. Models that require a regular grid receive a binned representation of the prefix, with observation masks and elapsed-time channels where their inputs include them; observations are truncated at the evaluation time before binning.

GRU-D. GRU-D (Che et al., 2018) incorporates values, missingness masks, and elapsed times through learned decay in a recurrent network.

mTAND. mTAND (Shukla & Marlin, 2021) aggregates the available observations onto a set of reference time points through time attention before classification or regression.

Raindrop. Raindrop (Zhang et al., 2022) models dependencies among variables through a learned sensor graph over binned values and observation masks.

Hi-Patch. Hi-Patch (Luo et al., 2025) keeps the observed timestamps and groups simultaneous values into a hierarchy of patches.

ATENet. ATENet (Lee et al., 2025) aggregates values, masks, and normalized observation times through learnable reference points.

iTransformer. iTransformer (Liu et al., 2024) represents the binned prefix as one token per variable containing its values, mask, and time since the last observation.

Dynamic-DeepHit. Dynamic-DeepHit (Lee et al., 2020) predicts event-time distributions from longitudinal observations. It receives the prefix-restricted observations with missingness indicators, and its event-time distribution is mapped to the outcome space of each task.

SPRT-TANDEM. We adapt SPRT-TANDEM (Ebihara et al., 2021; Miyagawa & Ebihara, 2021) to the event-stream setting. The model uses the shared tokenizer and encoder with capacity matched to PREFIX, and a head directly emits a log-ratio increment for each event. Accumulated scores are normalized into prefix-wise outcome probabilities. For multiclass outcomes, we use the LSEL formulation from the density-ratio-matrix extension (Miyagawa & Ebihara, 2021); evaluation is performed at every prefix rather than through an early-stopping rule. The continuous target uses a regression adaptation of the same accumulation. Unlike BELIEF, these evidence increments are learned directly rather than derived from normalized outcome-conditioned value distributions.

C.3 Implementation and Training Details

Computing environment. GPU experiments run on NVIDIA RTX A6000 GPUs with 48 GB of memory, one GPU per run, under Ubuntu 22.04.5 with Python 3.13.12, PyTorch 2.12.1, and CUDA 13.0. Dynamic-DeepHit runs on CPU in a separate environment with Python 3.7.12 and TensorFlow 1.15.5. BELIEF, PREFIX, TERMINAL, and SPRT-TANDEM are implemented in one PyTorch codebase and share the event tokenizer and encoders. All external baselines use publicly available implementations of the original methods.

Training and hyperparameter search. All models are trained with AdamW, a fixed batch size of 32, weight decay of 0.01, and gradient clipping at norm 1.0. Training uses a fixed epoch budget with early stopping on the validation split, retaining the checkpoint with the best validation performance.

The hyperparameters of BELIEF and all comparison methods are tuned on the validation split, including the learning rate, hidden width, and number of layers. In the synthetic comparison, the learning rate of every backbone and strategy is chosen from eight values between 10^{-5} and 3×10^{-2}. For BELIEF, the architecture search covers widths $\{64, 128, 256, 512\}$, numbers of layers $\{1, 2, 4, 6\}$, and dropout rates $\{0, 0.1, 0.2, 0.3\}$. Two BELIEF-specific hyperparameters are also tuned: the decoder mixture size $K \in \{1, 3, 6, 12\}$ and the component family (Student-t or Gaussian).

The selected BELIEF configurations use width 128, $K = 12$, and a backbone-specific learning rate in the synthetic comparison; width 256, $K = 12$, and learning rate 3×10^{-4} on the three real domains; width 128, $K = 12$, and learning rate 10^{-4} on P12 and MIMIC-III; and width 256, $K = 3$, and learning rate 3×10^{-4} on Geolife and P12-LoS. The synthetic ablations use the corresponding synthetic full-model configuration with width 128 and $K = 12$, and the real-domain ablations use the corresponding real-domain full-model configuration with width 256, $K = 12$, and learning rate 3×10^{-4}. All other parameters are inherited from the corresponding full-model configuration.

C.4 Evaluation Metrics

The metrics follow the order of Tables 1–4. Let $\widehat{q}_{n,j}$ and $q^\star_{n,j}$ denote the predicted and exact beliefs after prefix j of test trajectory n. Path metrics average over all observed event positions of the test set, $j = 1, \ldots, J_n$, excluding the prior at $j = 0$.

Path KL. For discrete outcomes, the error at a position is $D_{\mathrm{KL}}(q^\star_{n,j} \,\|\, \widehat{q}_{n,j})$. For continuous outcomes, the predicted belief is summarized by a Gaussian with its mean $\widehat{\mu}_{n,j}$ and standard deviation $\widehat{\sigma}_{n,j}$, and the error is $D_{\mathrm{KL}}\big(q^\star_{n,j} \,\|\, \mathcal{N}(\widehat{\mu}_{n,j}, \widehat{\sigma}^2_{n,j})\big)$, which has a closed form because $q^\star_{n,j}$ is Gaussian.

Path MAE. The error at a position is

$$\mathrm{MAE}_{n,j} = \begin{cases} \sum_{y\in\mathcal{Y}} \big|\widehat{q}_{n,j}(y) - q^\star_{n,j}(y)\big|, & \text{discrete outcomes,} \\ |\widehat{\mu}_{n,j} - \mu^\star_{n,j}| + |\widehat{\sigma}_{n,j} - \sigma^\star_{n,j}|, & \text{continuous outcomes.} \end{cases}$$

Relative movement (mv)**.** With $v_{n,j}$ the class-probability vector for discrete outcomes and the posterior mean for continuous outcomes,

$$\mathrm{mv} = \frac{\sum_n \sum_{j=2}^{J_n} \|\widehat{v}_{n,j} - \widehat{v}_{n,j-1}\|_2^2}{\sum_n \sum_{j=2}^{J_n} \|v^\star_{n,j} - v^\star_{n,j-1}\|_2^2}.$$

The ideal value is 1; larger values indicate more total squared movement than the exact path, and smaller values indicate less.

AUROC. For scores s_i^+ of n_+ positive and s_k^- of n_- negative examples,

$$\mathrm{AUROC} = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{k=1}^{n_-} \Big[\mathbf{1}\{s_i^+ > s_k^-\} + \tfrac{1}{2}\mathbf{1}\{s_i^+ = s_k^-\}\Big],$$

the probability that a positive example outranks a negative one. Multiclass AUROC is the unweighted mean of one-vs-rest AUROC over the classes present in the evaluation partition. On the synthetic and downstream tasks, the score is the belief at the last observed event. On the real-domain tasks, it is the log-odds at the last reference time of each trajectory; trajectories without reference coverage use the latest evaluation time (a lead of 12 hours for Weather and trading day $T-1$ for Steel-RB).

AUPRC. AUPRC is computed as average precision. With P_k and R_k the precision and recall at the kth of L distinct score thresholds in decreasing order and $R_0 = 0$,

$$\mathrm{AUPRC} = \sum_{k=1}^{L} (R_k - R_{k-1})\, P_k.$$

It uses the belief at the last observed event, with one-vs-rest macro averaging for multiclass outcomes.

Correlation and CRPS. For continuous synthetic outcomes, correlation is the Pearson correlation between the posterior mean at the last observed event and the realized outcome. The continuous ranked probability score of a Gaussian belief $\mathcal{N}(\mu, \sigma^2)$ at outcome y has the closed form

$$\mathrm{CRPS} = \sigma\Big[z\big(2\Phi(z) - 1\big) + 2\varphi(z) - \tfrac{1}{\sqrt{\pi}}\Big], \qquad z = \frac{y-\mu}{\sigma},$$

where Φ and φ are the standard normal distribution and density functions; it is averaged over test trajectories at the last observed event.

PACE-MAE. PACE-MAE measures agreement with a reference probability path $r_n(t)$ at the times $\mathcal{T}_n$ at which the reference is issued: NBM leads of 12–84 hours for Weather, trading days with option quotes within 20 trading days of expiry for Steel-RB, and the daily forecasts for Soccer. The model value $\widehat{q}_n(t)$ is the belief after the last event at or before t, or the training-partition outcome frequency before the first event. Over the N_{ref} test trajectories with reference coverage,

$$\mathrm{PACE\text{-}MAE} = \frac{1}{N_{\mathrm{ref}}} \sum_{n=1}^{N_{\mathrm{ref}}} \frac{1}{|\mathcal{T}_n|} \sum_{t\in\mathcal{T}_n} \big|\widehat{q}_n(t) - r_n(t)\big|.$$

The reference paths never enter training as targets or inputs.

Regression MAE. For P12-LoS, the prediction is the posterior mean at the last observed event, and MAE is computed on the transformed outcome scale of Appendix C.1.3.

ECE. The supplementary synthetic tables report top-label expected calibration error over all observed prefixes. The confidence is the largest predicted class probability, binned into 15 equal-width bins on $[1/C, 1]$ for C classes.

C.5 Ablation Settings

Table 4 reports path MAE averaged over the four synthetic classification settings and PACE-MAE on the three real domains. Each variant changes one component of the corresponding full-model configuration specified in Appendix C.3, with all other settings held fixed, and is run with seeds 3407–3411.

rp. free evidence increment. A score head replaces the likelihood decoder and emits an unconstrained increment after each observation. The increments are added to a learnable log prior and normalized into outcome probabilities, and the model is trained from scratch with the outcome cross-entropy at every observed prefix instead of the evidence likelihood. The score head starts from zero output.

rp. direct read-out. The event embedding and history encoder of a trained BELIEF model are frozen, with dropout disabled, and a new classification head is trained on their prefix representations with the outcome cross-entropy. Only the head is updated, whereas PREFIX learns the representation and the head jointly from scratch.

w/o learned prior. The outcome prior is fixed to the uniform distribution instead of being fitted to the training outcomes.

w/o variable identity. Variable-identity information is removed from both the event tokens and the query context, while value and time information are retained.

w/o shared atoms. The component locations and scales may depend on the candidate outcome instead of being shared across outcomes, which removes the bound of Appendix B.1. The history encoder remains shared.

w/o inter-arrival encoding. The encoding of the inter-arrival time $\Delta\tau_j$ is removed from the event representation and the query context; the other time features are unchanged.

D SUPPLEMENTARY EXPERIMENTS

D.1 DETAILED SYNTHETIC RESULTS

Tables 5–7 report the six synthetic settings in three outcome-specific tables. Each table stacks linear and nonlinear generation rather than averaging the mechanisms together. Every block retains the six backbones and three variants from Table 1. Every cell is the mean ± sample standard deviation over five seeds (3407–3411), each seed regenerating the data, the split, and the model initialization.

Table 5: Binary outcomes, with linear generation above nonlinear generation. Each block reports all six backbones and three variants from Table 1. Aggregation and metric definitions follow Appendix C.4. Within each backbone and mechanism, the best value is bold and the second-best underlined; mv is compared by proximity to one.

Encoder	Variant	KL ↓	MAE ↓	ECE ↓	mv → 1	AUROC ↑	AUPRC ↑
Linear							
Causal Transformer	**BELIEF**	**0.0076** ± 0.0016	**0.0634** ± 0.0056	**0.0101** ± 0.0043	**0.9786** ± 0.0132	**0.9703** ± 0.0103	**0.9711** ± 0.0093
	PREFIX	0.1985 ± 0.0337	0.4241 ± 0.0594	0.0274 ± 0.0067	1.3078 ± 0.2686	0.8107 ± 0.0478	0.8029 ± 0.0548
	TERMINAL	0.2809 ± 0.0187	0.4965 ± 0.0454	0.1115 ± 0.0241	12.9771 ± 3.2573	0.7890 ± 0.0559	0.7830 ± 0.0601
GRU	**BELIEF**	**0.0062** ± 0.0014	**0.0568** ± 0.0068	**0.0090** ± 0.0060	**0.9882** ± 0.0156	**0.9707** ± 0.0111	**0.9715** ± 0.0097
	PREFIX	0.0627 ± 0.0162	0.1924 ± 0.0248	0.0272 ± 0.0123	0.6947 ± 0.0786	0.9512 ± 0.0074	0.9524 ± 0.0032
	TERMINAL	0.1598 ± 0.0484	0.3022 ± 0.0390	0.0911 ± 0.0248	2.2591 ± 0.8933	0.9348 ± 0.0152	0.9375 ± 0.0108
GRU-ODE-Bayes	**BELIEF**	**0.0080** ± 0.0021	**0.0658** ± 0.0080	**0.0093** ± 0.0033	**0.9512** ± 0.0197	**0.9702** ± 0.0105	**0.9709** ± 0.0093
	PREFIX	0.0573 ± 0.0103	0.1911 ± 0.0198	0.0225 ± 0.0096	1.2288 ± 0.0545	0.9526 ± 0.0087	0.9534 ± 0.0078
	TERMINAL	0.2040 ± 0.0945	0.4149 ± 0.1873	0.0623 ± 0.0296	10.3951 ± 9.8693	0.7682 ± 0.2260	0.7738 ± 0.2157
mTAND	**BELIEF**	**0.0103** ± 0.0023	**0.0756** ± 0.0079	**0.0085** ± 0.0035	**0.9607** ± 0.0294	**0.9698** ± 0.0088	**0.9710** ± 0.0070
	PREFIX	0.2503 ± 0.0871	0.5215 ± 0.1571	0.0307 ± 0.0139	0.3036 ± 0.4731	0.6554 ± 0.1650	0.6556 ± 0.1563
	TERMINAL	0.3033 ± 0.0099	0.6105 ± 0.0152	0.0373 ± 0.0445	0.1672 ± 0.3248	0.5421 ± 0.0888	0.5465 ± 0.0727
ASTGI	**BELIEF**	**0.0056** ± 0.0010	**0.0545** ± 0.0048	**0.0096** ± 0.0057	**0.9901** ± 0.0138	**0.9706** ± 0.0101	**0.9716** ± 0.0088
	PREFIX	0.2486 ± 0.0253	0.5207 ± 0.0437	0.0186 ± 0.0039	7.3892 ± 2.5402	0.6803 ± 0.0558	0.6642 ± 0.0527
	TERMINAL	0.3052 ± 0.0082	0.6107 ± 0.0152	0.0514 ± 0.0430	6.5949 ± 5.0069	0.5513 ± 0.0678	0.5509 ± 0.0513
HSTU	**BELIEF**	**0.0102** ± 0.0025	**0.0744** ± 0.0086	**0.0099** ± 0.0051	**0.9545** ± 0.0178	**0.9694** ± 0.0098	**0.9704** ± 0.0084
	PREFIX	0.2099 ± 0.0427	0.4347 ± 0.0856	0.0352 ± 0.0176	1.7186 ± 0.2206	0.8018 ± 0.0581	0.8028 ± 0.0580
	TERMINAL	0.3256 ± 0.0813	0.5670 ± 0.0526	0.0963 ± 0.0531	33.1478 ± 28.0478	0.7359 ± 0.1330	0.7312 ± 0.1323
Nonlinear							
Causal Transformer	**BELIEF**	**0.0155** ± 0.0044	**0.0987** ± 0.0174	**0.0164** ± 0.0032	**0.8552** ± 0.0956	**0.9666** ± 0.0114	**0.9679** ± 0.0105
	PREFIX	0.0407 ± 0.0106	0.1559 ± 0.0195	0.0205 ± 0.0084	1.9435 ± 0.2367	0.9568 ± 0.0082	0.9588 ± 0.0064
	TERMINAL	0.1973 ± 0.0135	0.2980 ± 0.0123	0.1245 ± 0.0152	14.5661 ± 2.1857	0.9507 ± 0.0084	0.9528 ± 0.0068
GRU	**BELIEF**	**0.0067** ± 0.0009	**0.0611** ± 0.0040	**0.0093** ± 0.0039	**0.9812** ± 0.0271	**0.9699** ± 0.0089	**0.9709** ± 0.0081
	PREFIX	0.0280 ± 0.0033	0.1304 ± 0.0087	0.0191 ± 0.0099	0.5752 ± 0.0315	0.9630 ± 0.0102	0.9643 ± 0.0083
	TERMINAL	0.0786 ± 0.0218	0.2067 ± 0.0201	0.0743 ± 0.0199	1.3716 ± 0.6388	0.9588 ± 0.0131	0.9600 ± 0.0115
GRU-ODE-Bayes	**BELIEF**	**0.0065** ± 0.0006	**0.0599** ± 0.0026	**0.0106** ± 0.0042	**0.9757** ± 0.0175	**0.9698** ± 0.0095	**0.9708** ± 0.0083
	PREFIX	0.0292 ± 0.0034	0.1310 ± 0.0087	0.0134 ± 0.0032	1.0965 ± 0.0979	0.9628 ± 0.0095	0.9635 ± 0.0079
	TERMINAL	0.0664 ± 0.0075	0.1944 ± 0.0095	0.0646 ± 0.0135	4.3921 ± 0.9534	0.9600 ± 0.0101	0.9608 ± 0.0091
mTAND	**BELIEF**	**0.0066** ± 0.0014	**0.0604** ± 0.0066	**0.0104** ± 0.0052	0.9794 ± 0.0362	**0.9694** ± 0.0100	**0.9707** ± 0.0088
	PREFIX	0.0575 ± 0.0052	0.1951 ± 0.0099	0.0143 ± 0.0061	**1.0005** ± 0.0634	0.9510 ± 0.0072	0.9520 ± 0.0059
	TERMINAL	0.2720 ± 0.0602	0.4822 ± 0.1383	0.1078 ± 0.0478	3.1524 ± 3.2358	0.7550 ± 0.1766	0.7574 ± 0.1727
ASTGI	**BELIEF**	**0.0312** ± 0.0187	**0.1363** ± 0.0421	0.0200 ± 0.0118	**0.8128** ± 0.1212	**0.9652** ± 0.0096	**0.9660** ± 0.0089
	PREFIX	0.1131 ± 0.0089	0.2778 ± 0.0125	**0.0195** ± 0.0076	21.9035 ± 0.9355	0.8766 ± 0.0103	0.8767 ± 0.0077
	TERMINAL	0.3238 ± 0.0540	0.4545 ± 0.0354	0.1465 ± 0.0200	42.4433 ± 12.9596	0.8602 ± 0.0119	0.8598 ± 0.0109
HSTU	**BELIEF**	**0.0166** ± 0.0049	**0.0978** ± 0.0150	**0.0087** ± 0.0022	**0.9417** ± 0.0643	**0.9660** ± 0.0079	**0.9675** ± 0.0067
	PREFIX	0.0487 ± 0.0137	0.1695 ± 0.0287	0.0235 ± 0.0107	2.6868 ± 0.3047	0.9546 ± 0.0085	0.9568 ± 0.0078
	TERMINAL	0.1945 ± 0.0407	0.3333 ± 0.0408	0.0971 ± 0.0113	28.6359 ± 14.0649	0.9492 ± 0.0098	0.9508 ± 0.0078

Table 6: Multiclass outcomes, with linear generation above nonlinear generation. Each block reports all six backbones and three variants from Table 1. Aggregation and metric definitions follow Appendix C.4. Within each backbone and mechanism, the best value is bold and the second-best underlined; mv is compared by proximity to one.

Encoder	Variant	KL ↓	MAE ↓	ECE ↓	mv → 1	macro-AUROC ↑	macro-AUPRC ↑
Linear							
Causal Transformer	**BELIEF**	**0.0132** ± 0.0029	**0.0929** ± 0.0100	**0.0121** ± 0.0055	**0.9696** ± 0.0117	**0.9748** ± 0.0060	**0.9238** ± 0.0164
	PREFIX	0.3405 ± 0.0634	0.5539 ± 0.0764	0.0487 ± 0.0217	1.1252 ± 0.1829	0.8671 ± 0.0303	0.6631 ± 0.0515
	TERMINAL	0.5571 ± 0.0682	0.6800 ± 0.0631	0.1605 ± 0.0130	10.1649 ± 1.4352	0.8521 ± 0.0283	0.6315 ± 0.0440
GRU	**BELIEF**	**0.0095** ± 0.0017	**0.0792** ± 0.0067	**0.0114** ± 0.0062	**0.9810** ± 0.0158	**0.9753** ± 0.0060	**0.9251** ± 0.0170
	PREFIX	0.0849 ± 0.0144	0.2417 ± 0.0200	0.0313 ± 0.0130	0.6525 ± 0.0560	0.9634 ± 0.0065	0.8883 ± 0.0182
	TERMINAL	0.3484 ± 0.0971	0.4483 ± 0.0445	0.1578 ± 0.0416	2.5597 ± 0.9473	0.9542 ± 0.0036	0.8628 ± 0.0109
GRU-ODE-Bayes	**BELIEF**	**0.0144** ± 0.0034	**0.0983** ± 0.0122	**0.0112** ± 0.0052	**0.9449** ± 0.0234	**0.9748** ± 0.0056	**0.9232** ± 0.0156
	PREFIX	0.0787 ± 0.0140	0.2419 ± 0.0256	0.0212 ± 0.0059	1.2931 ± 0.1057	0.9641 ± 0.0042	0.8915 ± 0.0098
	TERMINAL	0.3658 ± 0.0480	0.4903 ± 0.0370	0.1701 ± 0.0135	8.1466 ± 0.8711	0.9573 ± 0.0056	0.8696 ± 0.0182
mTAND	**BELIEF**	**0.0151** ± 0.0068	**0.0987** ± 0.0226	**0.0120** ± 0.0049	0.9868 ± 0.0287	**0.9746** ± 0.0053	**0.9223** ± 0.0130
	PREFIX	0.1308 ± 0.0182	0.3144 ± 0.0280	0.0238 ± 0.0086	**0.9906** ± 0.1016	0.9509 ± 0.0072	0.8498 ± 0.0223
	TERMINAL	0.6236 ± 0.1798	0.8073 ± 0.2571	0.0906 ± 0.0695	3.4392 ± 4.3748	0.7310 ± 0.1907	0.5181 ± 0.2593
ASTGI	**BELIEF**	**0.0091** ± 0.0023	**0.0774** ± 0.0101	**0.0115** ± 0.0051	**0.9816** ± 0.0166	**0.9754** ± 0.0057	**0.9250** ± 0.0158
	PREFIX	0.3737 ± 0.0375	0.5982 ± 0.0511	0.0202 ± 0.0079	7.5453 ± 1.1436	0.7955 ± 0.0132	0.5469 ± 0.0194
	TERMINAL	0.6825 ± 0.0348	0.8586 ± 0.0331	0.1200 ± 0.0276	13.6755 ± 2.4075	0.7318 ± 0.0331	0.4627 ± 0.0382
HSTU	**BELIEF**	**0.0197** ± 0.0065	**0.1136** ± 0.0191	**0.0115** ± 0.0045	**0.9646** ± 0.0176	**0.9737** ± 0.0058	**0.9193** ± 0.0155
	PREFIX	0.3136 ± 0.0943	0.5023 ± 0.0994	0.0662 ± 0.0132	1.8257 ± 0.5214	0.8870 ± 0.0350	0.7006 ± 0.0730
	TERMINAL	0.7995 ± 0.1021	0.8202 ± 0.0759	0.1813 ± 0.0268	27.5749 ± 14.2842	0.8408 ± 0.0518	0.6189 ± 0.0891
Nonlinear							
Causal Transformer	**BELIEF**	**0.0215** ± 0.0052	**0.1212** ± 0.0156	**0.0134** ± 0.0062	**0.9714** ± 0.0337	**0.9729** ± 0.0045	**0.9200** ± 0.0123
	PREFIX	0.0635 ± 0.0136	0.2083 ± 0.0200	0.0290 ± 0.0134	1.7196 ± 0.1816	0.9637 ± 0.0056	0.8905 ± 0.0151
	TERMINAL	0.3894 ± 0.0386	0.4465 ± 0.0209	0.1785 ± 0.0128	14.7021 ± 1.3766	0.9597 ± 0.0051	0.8769 ± 0.0153
GRU	**BELIEF**	**0.0108** ± 0.0020	**0.0846** ± 0.0075	**0.0110** ± 0.0064	**0.9785** ± 0.0106	**0.9742** ± 0.0054	**0.9238** ± 0.0149
	PREFIX	0.0493 ± 0.0035	0.1867 ± 0.0069	0.0281 ± 0.0082	0.5624 ± 0.0605	0.9691 ± 0.0059	0.9073 ± 0.0163
	TERMINAL	0.4326 ± 0.0117	0.5205 ± 0.0189	0.1981 ± 0.0129	0.7337 ± 0.2244	0.9662 ± 0.0055	0.8945 ± 0.0174
GRU-ODE-Bayes	**BELIEF**	**0.0104** ± 0.0029	**0.0830** ± 0.0105	**0.0123** ± 0.0055	**0.9706** ± 0.0194	**0.9741** ± 0.0054	**0.9237** ± 0.0146
	PREFIX	0.0486 ± 0.0046	0.1852 ± 0.0111	0.0194 ± 0.0068	1.1051 ± 0.0835	0.9685 ± 0.0053	0.9045 ± 0.0147
	TERMINAL	0.4433 ± 0.0456	0.5372 ± 0.0262	0.1908 ± 0.0188	2.6973 ± 0.4649	0.9667 ± 0.0057	0.8970 ± 0.0154
mTAND	**BELIEF**	**0.0208** ± 0.0064	**0.1177** ± 0.0191	**0.0126** ± 0.0052	**0.9689** ± 0.0325	**0.9722** ± 0.0046	**0.9184** ± 0.0117
	PREFIX	0.0808 ± 0.0082	0.2453 ± 0.0163	0.0173 ± 0.0064	1.0877 ± 0.1177	0.9623 ± 0.0050	0.8877 ± 0.0130
	TERMINAL	0.3959 ± 0.0405	0.4556 ± 0.0266	0.1667 ± 0.0059	6.5184 ± 0.3321	0.9523 ± 0.0079	0.8556 ± 0.0235
ASTGI	**BELIEF**	**0.0235** ± 0.0127	**0.1255** ± 0.0370	**0.0122** ± 0.0054	**0.9310** ± 0.0340	**0.9719** ± 0.0067	**0.9182** ± 0.0179
	PREFIX	0.1797 ± 0.0296	0.3712 ± 0.0326	0.0209 ± 0.0059	17.0242 ± 1.3171	0.9004 ± 0.0111	0.7292 ± 0.0200
	TERMINAL	0.5765 ± 0.0774	0.6144 ± 0.0064	0.1878 ± 0.0388	30.0744 ± 5.7004	0.9004 ± 0.0089	0.7261 ± 0.0219
HSTU	**BELIEF**	**0.0205** ± 0.0051	**0.1155** ± 0.0132	**0.0176** ± 0.0046	**1.0194** ± 0.0358	**0.9719** ± 0.0053	**0.9183** ± 0.0136
	PREFIX	0.0822 ± 0.0173	0.2361 ± 0.0269	0.0374 ± 0.0090	2.4709 ± 0.5483	0.9615 ± 0.0043	0.8850 ± 0.0122
	TERMINAL	0.3305 ± 0.0652	0.4840 ± 0.0634	0.1082 ± 0.0139	23.3934 ± 9.6498	0.9521 ± 0.0053	0.8561 ± 0.0147

Table 7: Continuous outcomes, with linear generation above nonlinear generation. Each block reports all six backbones and three variants from Table 1. Aggregation and metric definitions follow Appendix C.4. Within each backbone and mechanism, the best value is bold and the second-best underlined; mv is compared by proximity to one.

Encoder	Variant	KL ↓	MAE ↓	mv → 1	corr ↑	CRPS ↓
Linear						
Causal Transformer	**BELIEF**	**0.0199** ± 0.0030	**0.1250** ± 0.0082	**0.9990** ± 0.0116	**0.7869** ± 0.0067	**0.3406** ± 0.0053
	PREFIX	0.2176 ± 0.0268	0.5242 ± 0.0622	1.4614 ± 0.3444	0.5134 ± 0.0794	0.4743 ± 0.0268
	TERMINAL	0.3309 ± 0.0212	0.6248 ± 0.0402	14.3409 ± 3.9559	0.4656 ± 0.0913	0.4883 ± 0.0251
GRU	**BELIEF**	**0.0169** ± 0.0017	**0.1177** ± 0.0067	**0.9816** ± 0.0275	**0.7886** ± 0.0040	**0.3394** ± 0.0042
	PREFIX	0.0758 ± 0.0129	0.2390 ± 0.0232	0.5881 ± 0.0638	0.7422 ± 0.0132	0.3718 ± 0.0107
	TERMINAL	0.2325 ± 0.0579	0.4407 ± 0.0567	4.3525 ± 3.9962	0.7069 ± 0.0453	0.3954 ± 0.0305
GRU-ODE-Bayes	**BELIEF**	**0.0273** ± 0.0061	**0.1476** ± 0.0171	**0.9924** ± 0.0130	**0.7825** ± 0.0088	**0.3442** ± 0.0061
	PREFIX	0.0671 ± 0.0107	0.2392 ± 0.0298	1.3387 ± 0.1459	0.7458 ± 0.0203	0.3700 ± 0.0148
	TERMINAL	0.2832 ± 0.0430	0.6451 ± 0.1612	6.9983 ± 6.9138	0.2721 ± 0.3391	0.5057 ± 0.0688
mTAND	**BELIEF**	**0.0115** ± 0.0034	**0.0962** ± 0.0219	**0.9758** ± 0.0460	**0.7921** ± 0.0071	**0.3367** ± 0.0072
	PREFIX	0.2579 ± 0.0649	0.6259 ± 0.1493	0.3653 ± 0.4670	0.3173 ± 0.2621	0.5055 ± 0.0545
	TERMINAL	0.3100 ± 0.0072	0.7500 ± 0.0388	0.1992 ± 0.3721	0.1149 ± 0.1445	0.5438 ± 0.0228
ASTGI	**BELIEF**	**0.0154** ± 0.0016	**0.1112** ± 0.0055	**1.0007** ± 0.0204	**0.7918** ± 0.0048	**0.3372** ± 0.0054
	PREFIX	0.2644 ± 0.0103	0.6485 ± 0.0284	8.9580 ± 3.9437	0.3155 ± 0.0670	0.5234 ± 0.0113
	TERMINAL	0.3251 ± 0.0256	0.7490 ± 0.0199	5.7220 ± 7.2429	0.0607 ± 0.1411	0.5493 ± 0.0185
HSTU	**BELIEF**	**0.0257** ± 0.0025	**0.1408** ± 0.0064	**1.0067** ± 0.0394	**0.7842** ± 0.0072	**0.3433** ± 0.0052
	PREFIX	0.2210 ± 0.0293	0.5638 ± 0.0782	1.8250 ± 0.4562	0.4996 ± 0.0911	0.4773 ± 0.0248
	TERMINAL	157.8439 ± 350.0165	0.7597 ± 0.0677	54.6992 ± 40.5781	0.3254 ± 0.1976	0.5163 ± 0.0361
Nonlinear						
Causal Transformer	**BELIEF**	**0.0149** ± 0.0025	**0.1186** ± 0.0098	**0.9292** ± 0.0161	**0.7877** ± 0.0051	**0.3407** ± 0.0040
	PREFIX	0.0505 ± 0.0058	0.2203 ± 0.0109	2.0176 ± 0.4264	0.7573 ± 0.0115	0.3638 ± 0.0085
	TERMINAL	0.3382 ± 0.0565	0.4809 ± 0.0227	16.5900 ± 4.0860	0.7445 ± 0.0172	0.3692 ± 0.0101
GRU	**BELIEF**	**0.0183** ± 0.0010	**0.1232** ± 0.0033	**1.0060** ± 0.0214	**0.7871** ± 0.0082	**0.3400** ± 0.0056
	PREFIX	0.0358 ± 0.0026	0.1746 ± 0.0058	0.5236 ± 0.0572	0.7712 ± 0.0089	0.3516 ± 0.0044
	TERMINAL	0.1639 ± 0.0203	0.3414 ± 0.0197	1.4276 ± 0.1550	0.7585 ± 0.0101	0.3612 ± 0.0054
GRU-ODE-Bayes	**BELIEF**	**0.0186** ± 0.0020	**0.1244** ± 0.0067	**0.9937** ± 0.0170	**0.7839** ± 0.0080	**0.3427** ± 0.0072
	PREFIX	0.0412 ± 0.0064	0.1884 ± 0.0125	1.2155 ± 0.1237	0.7658 ± 0.0075	0.3576 ± 0.0027
	TERMINAL	0.2190 ± 0.0764	0.4976 ± 0.2184	4.1818 ± 2.0148	0.4524 ± 0.4013	0.4431 ± 0.0981
mTAND	**BELIEF**	**0.0056** ± 0.0007	**0.0670** ± 0.0039	0.9759 ± 0.0192	**0.7915** ± 0.0066	**0.3370** ± 0.0059
	PREFIX	0.0746 ± 0.0040	0.2686 ± 0.0159	**1.0213** ± 0.2385	0.7385 ± 0.0128	0.3757 ± 0.0056
	TERMINAL	4.6584 ± 9.5955	0.6722 ± 0.1167	5.5174 ± 5.2007	0.4032 ± 0.2588	0.4886 ± 0.0631
ASTGI	**BELIEF**	**0.0282** ± 0.0273	**0.1444** ± 0.0648	**0.9505** ± 0.0662	**0.7758** ± 0.0242	**0.3486** ± 0.0165
	PREFIX	0.1277 ± 0.0091	0.3511 ± 0.0215	21.3226 ± 1.4345	0.6106 ± 0.0285	0.4368 ± 0.0121
	TERMINAL	0.3796 ± 0.0480	0.6165 ± 0.0177	43.0680 ± 8.3947	0.5745 ± 0.0289	0.4557 ± 0.0170
HSTU	**BELIEF**	**0.0162** ± 0.0041	**0.1160** ± 0.0175	**0.9461** ± 0.0342	**0.7855** ± 0.0049	**0.3419** ± 0.0039
	PREFIX	0.0548 ± 0.0106	0.2110 ± 0.0206	2.4201 ± 0.3066	0.7532 ± 0.0168	0.3631 ± 0.0103
	TERMINAL	0.3233 ± 0.1082	0.4928 ± 0.0824	32.3648 ± 18.2540	0.7086 ± 0.0390	0.3929 ± 0.0164

BELIEF obtains lower path error for binary and multiclass outcomes under both mechanisms, with favorable discrimination metrics. The continuous-outcome comparison shows the same pattern in path KL and MAE together with strong correlation and CRPS. The outcome-specific tables therefore support the aggregate recovery result without conflating metrics defined on different outcome spaces.

D.2 Belief Paths on Real Event Streams

Figure 5 shows belief paths on twelve test trajectories, four per domain, together with the reference paths used by PACE-MAE. Each model's belief is drawn as its value after the most recent event, which is the quantity that PACE-MAE compares with the reference at each issue time. Across the twelve trajectories, BELIEF follows the direction and timing of the main revisions of the reference, including the late reversals on Weather and Soccer and the decline towards expiry on Steel-RB. TERMINAL stays near a constant level on Weather and Steel-RB, and on Steel-RB, PREFIX and the strongest external baseline revise their beliefs much less than the reference.

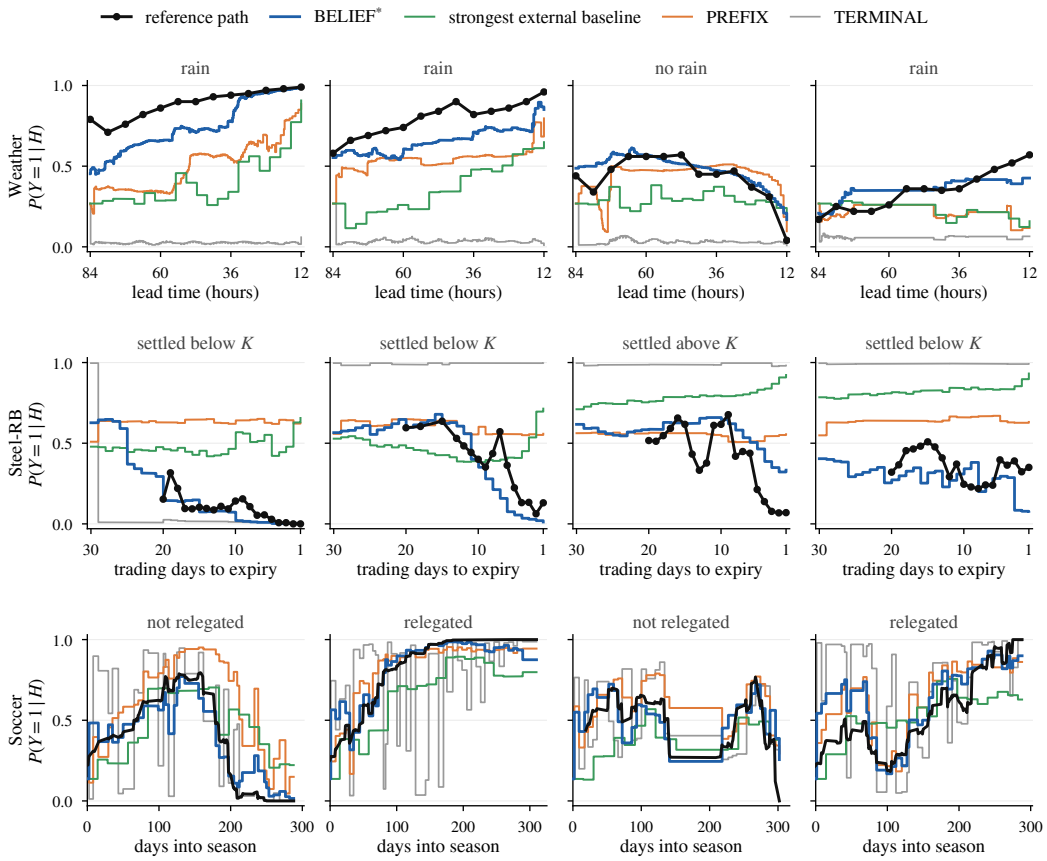

Figure 5: Belief paths on real-domain test trajectories. Rows show Weather, Steel-RB, and Soccer; panel titles give the realized outcome. Black: reference path (NBM, option-implied, and FiveThirtyEight SPI probabilities); blue: BELIEF*; green: the strongest external baseline of Table 2 in each domain (GRU-D for Weather and Steel-RB, Hi-Patch for Soccer); orange: PREFIX; gray: TERMINAL.

D.3 Sensitivity Analysis

Figure 6 evaluates sensitivity to mixture size, learning rate, hidden dimension, and encoder depth on nonlinear-binary synthetic streams, varying one parameter at a time.

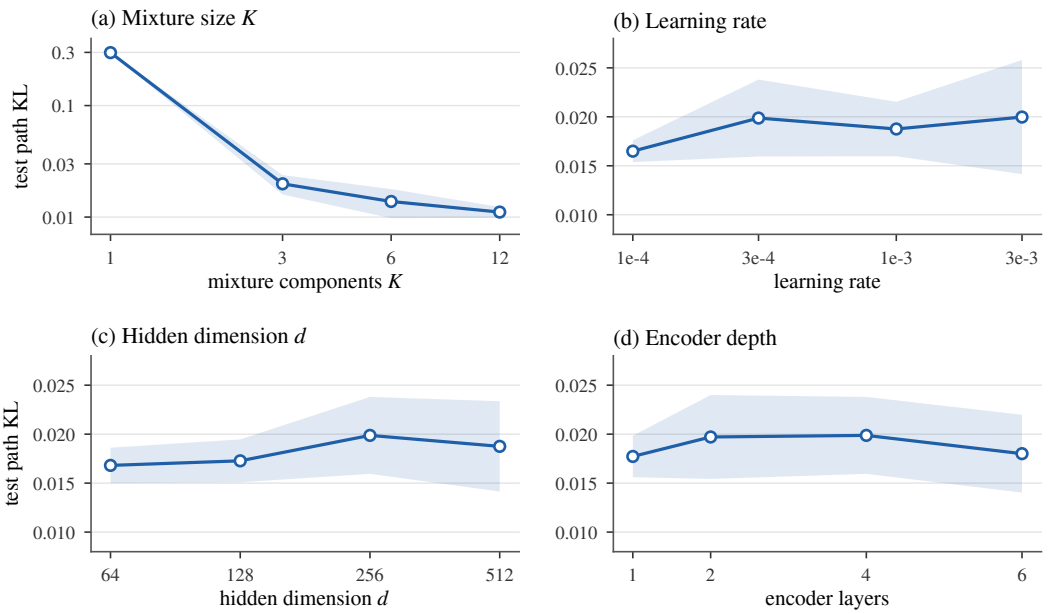

Figure 6: Hyperparameter sensitivity of BELIEF on nonlinear-binary synthetic streams. Test path KL when varying (a) mixture size K, (b) learning rate, (c) hidden dimension d, and (d) encoder depth, with other settings fixed. Lines show means over five seeds; shaded bands indicate ± 1 standard deviation. Panel (a) uses a logarithmic vertical axis.

Increasing K from 3 to 12 reduces path KL, whereas the other three parameters have a smaller effect over the tested ranges. With $K = 1$, the shared-mixture decoder assigns the same likelihood to every candidate outcome, so log-likelihood ratios vanish and the belief remains at the prior. This illustrates how outcome-conditioned mixture weights enable belief updates.

D.4 Scaling Analysis

Figure 7 examines how belief-path quality changes with event-stream length, training-set size, and variable count on nonlinear-binary synthetic streams.

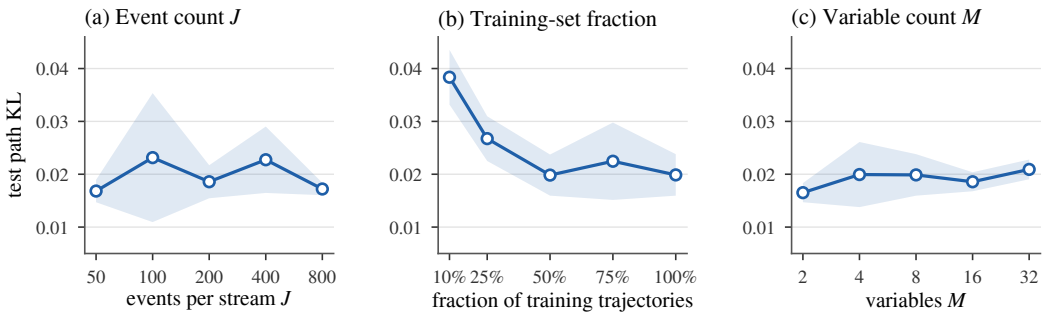

Figure 7: Belief-path quality across data scales on nonlinear-binary synthetic streams. Test path KL when varying (a) event count J, (b) training-set fraction, and (c) variable count M, one factor at a time. Lines show means over five seeds; shaded bands indicate ± 1 standard deviation. Validation and test sets remain fixed in (b).

Path KL remains comparable across $J = 50$–800 events and $M = 2$–32 variables. Increasing the training fraction from 10% to 50% reduces path error, with half of the training trajectories already yielding performance close to the full-data result.

Computational cost. Figure 8 shows that training time and peak training memory increase with stream length, while computing the complete belief path of one trajectory takes approximately 3 ms over the tested range. All prefix beliefs are obtained using a single causal encoder pass followed by likelihood accumulation, without separately re-encoding each prefix.

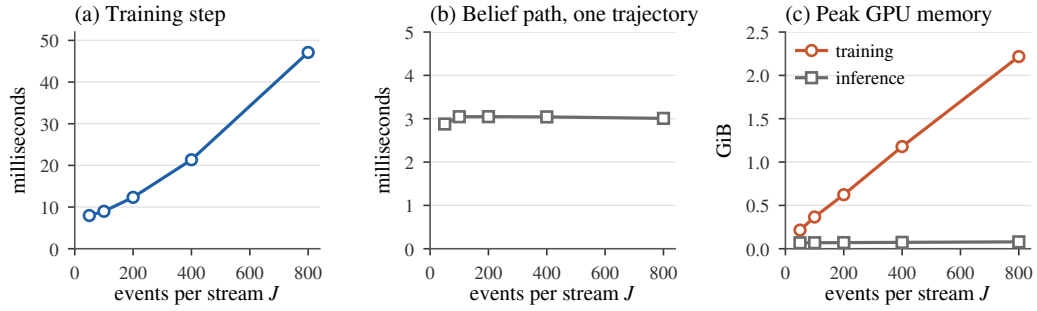

Figure 8: Computational cost of BELIEF across event-stream lengths. On nonlinear-binary synthetic streams, (a) training-step time at batch size 32, (b) inference time for the complete belief path of one trajectory, and (c) peak GPU memory during training and inference, as J increases from 50 to 800. Points report medians over 30 repetitions.

E Limitations

The recovery guarantees assume that observation times and variable identities carry no additional outcome information beyond the preceding history. In settings where the observation process is informative, jointly modeling when and what is measured would allow this information to contribute to the belief update. The posterior-improvement result also requires relative marginal stability be-

tween the compared models. Characterizing broader model classes and training procedures that satisfy this condition is a direction for further theoretical work.

The current framework uses standard likelihood accumulation with an evidence model fitted on historical streams. Adapting this model and calibrating belief updates under distribution shift remain directions for extension. Variable-wise likelihood tempering and time-aware evidence weighting could be investigated for this purpose; these modifications change the accumulation rule and would require their own analysis of belief-path recovery. Real-domain experiments evaluate agreement with external reference probabilities. Further deployment studies could examine how the resulting updates support decisions under task-specific costs and utilities. A changing risk estimate can inform when to reassess a situation. Connecting these beliefs to intervention policies would additionally require modeling actions and evaluating their consequences.