Dual Path Neural CDE for Probabilistic Stepwise Forecasting

Shengnan Di, Zhidong Li, Fang Chen, Bin Liang
Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia
shengnan.di@student.uts.edu.au

***Abstract*—Probabilistic forecasting for multivariate stepwise time series is difficult because observations are dominated by long stable regimes, while predictive changes occur sparsely and asynchronously across variables. We propose a dual path neural controlled differential equation framework that disentangles persistent trend structure from event driven updates in continuous time. A trend pathway summarizes regime level dynamics via smoothing, whereas an event pathway sparsifies the primary channel by retaining endpoints and statistically significant change windows to form an irregular event sequence. The two representations are fused by a learned time varying gate and mapped to predictive distribution parameters. We study Gaussian, Laplace, and Student t likelihoods and evaluate forecast quality using the continuous ranked probability score together with calibration and sharpness metrics. Experiments show consistent improvements over classical and neural baselines, and indicate that heavy tailed likelihoods improve distributional accuracy under regime shifts, while calibrated interval reliability remains subject to a sharpness trade-off.**

***Index Terms*—probabilistic forecasting, uncertainty quantification, neural controlled differential equation, stepwise time series, event driven modeling, heavy tailed likelihood, calibration**

I. Introduction

Accurate forecasting under uncertainty is essential for decision-making in many real-world systems. Across domains such as weather-related risk management [1], public health surveillance [2]–[4], healthcare operations [5], energy systems [6], finance [7], and large-scale retail [8], decisions depend not only on point estimates but on understanding the range of plausible future outcomes. In these settings, uncertainty-aware forecasts directly inform risk-sensitive actions.

Probabilistic forecasting has therefore emerged as a principled alternative to point prediction by explicitly modeling predictive distributions rather than single-valued estimates. Prior work has shown that probabilistic forecasts provide decision-relevant information unavailable from point predictions alone [1], [8]. This shift has motivated the use of proper scoring rules such as the continuous ranked probability score (CRPS), which jointly evaluates calibration and sharpness and serves as a standard criterion for distributional forecast quality [2], [9].

A first challenge arises from the mismatch between the temporal structure of many real-world signals and the assumptions underlying mainstream forecasting models. Recurrent, convolutional, and Transformer-based approaches typically represent temporal evolution through a single continuous latent stream, implicitly assuming smooth variation or approximate stationarity [10], [11]. While effective for densely evolving sequences, such representations are ill-suited to stepwise dynamics characterized by long periods of persistence punctuated by abrupt regime changes.

Empirical evidence across domains suggests that many real-world series are better described as piecewise regimes separated by sudden transitions rather than continuously evolving processes [12], [13]. Under such conditions, a single latent representation tends to conflate stable trends with event-driven changes, leading to inefficient use of temporal information. Although regime-switching models explicitly target abrupt transitions, they often rely on strong assumptions (e.g., homogeneous Markov dynamics) and sufficient data for stable estimation, which are frequently violated in practice [14].

Moreover, stepwise time series are often information-sparse: most observations convey nearly identical values, while informative signals concentrate on rare change events. This imbalance hampers effective signal propagation during training and increases the risk of overfitting to isolated jumps, further challenging standard learning-based forecasting pipelines [15], [16]. Fig. 1 summarizes the key difficulties of multivariate stepwise forecasting. As shown in Fig. 1(a–b), stepwise signals are dominated by long stable regimes, making sliding-window training highly redundant while leaving transition-containing windows sparse but decisive. Fig. 1(c) motivates our dual-path design that separates continuous trend dynamics from sparse event-driven updates.

In practice, this mismatch is further exacerbated in multivariate settings, where variables evolve at heterogeneous temporal scales. Some channels change smoothly, while others remain constant for long periods and update only in response to sparse external events. Naïvely aligning such asynchronous signals on a shared grid can blur decisive changes, introduce spurious delays, and obscure which components drive future outcomes. This pattern appears in housing markets, where rental prices are fixed between contract renewals while related covariates (e.g., commuting costs) shift only after major infrastructure updates, and in finance, where lending rates stay constant for long intervals before abrupt policy-driven changes. More broadly, heterogeneous update frequencies are known to induce aliasing and temporal misalignment when forced onto uniform sampling grids [17], [18].

A second challenge lies in probabilistic modeling, where distributional assumptions used to quantify uncertainty are frequently misaligned with jump-driven dynamics. Likelihood-

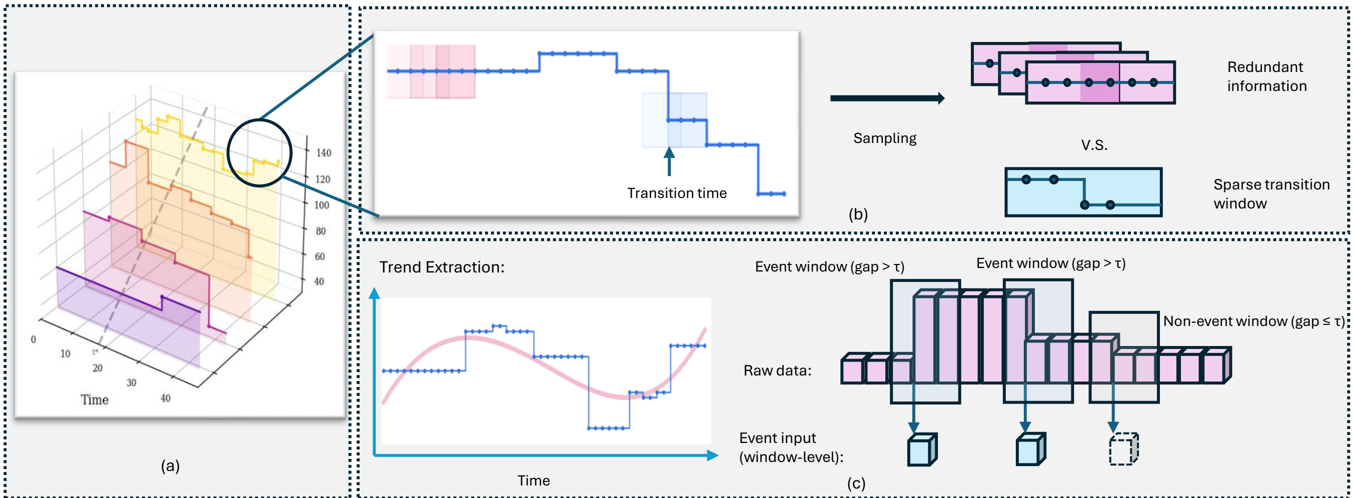

Fig. 1. Illustration of multivariate stepwise dynamics and the motivation for dual-path modeling. (a) A multivariate stepwise signal where each feature remains piecewise constant and updates occur asynchronously across variables. (b) Zoom-in on one feature illustrating the sampling imbalance induced by sliding-window training: windows extracted from stable regimes are highly redundant (pink), whereas windows containing transition points are sparse but decisive (blue). (c) Dual-path representation with explicit event abstraction. The trend pathway encodes a smooth continuous surrogate trajectory, while the event pathway sparsifies the raw series by retaining only windows that contain significant changes exceeding a threshold, yielding an irregular sequence of event occurrences (blue) interspersed with non-events (gray).

based approaches commonly adopt light-tailed Gaussian (or related) error models for tractability, yet such assumptions can be fragile under model misspecification and lead to unreliable uncertainty statements [19]. Even nominal Gaussian prediction intervals may substantially under-cover true outcomes when the assumed model fails to match the data-generating mechanism, resulting in systematic miscalibration [20]. This issue is particularly pronounced for stepwise time series, where abrupt regime shifts violate smooth-evolution assumptions and can lead to systematically miscalibrated uncertainty estimates under light-tailed likelihoods.

These observations motivate the use of heavier-tailed likelihoods as a robust alternative. Student-t models have long been employed to avoid underestimating extreme deviations commonly missed by Gaussian assumptions [21], and Student-t-based process models improve robustness under heavy-tailed noise in nonlinear dynamical systems [22]. More broadly, heavy-tailed behavior can arise from underlying data-generating and aggregation mechanisms and is well documented in classical heavy-tail analyses [23], [24]. Together, these results suggest that heavy-tailed likelihood choices, including Laplace as a moderate alternative and Student-t for stronger tail behavior, provide a more suitable foundation for uncertainty quantification under sparse stepwise dynamics.

A third challenge concerns evaluation objectives. Despite the growing interest in probabilistic forecasting, model assessment in many learning-based pipelines remains dominated by point prediction metrics such as RMSE. While informative for mean accuracy, such metrics are insensitive to distributional properties and fail to capture calibration and uncertainty quality, which are central to probabilistic forecasts.

This mismatch is particularly problematic for stepwise time series. In long stable regimes, accurate mean prediction can be achieved even when uncertainty is poorly characterized, masking deficiencies in probabilistic modeling. As a result, models optimized or selected based solely on point-wise error may exhibit misleadingly strong performance while producing unreliable uncertainty estimates around regime transitions.

These limitations motivate the use of proper scoring rules that directly evaluate predictive distributions. In this work, we adopt the continuous ranked probability score (CRPS) as the primary evaluation criterion, as it jointly assesses calibration and sharpness and provides a principled measure of probabilistic forecast quality [2], [9], complemented by point and interval-based metrics. Fig. 2 provides an overview of the proposed dual-path Neural CDE framework. The architecture consists of two complementary components: (1) a trend pathway that captures smooth, long-term temporal dynamics, and (2) an event pathway that encodes abrupt, infrequent changes. These representations are jointly modeled in continuous time and coupled with likelihood-based probabilistic forecasting to produce calibrated distributional predictions. These three challenges motivate our design choices in representation, likelihood modeling, and evaluation, which we summarize as the following contributions:

- We propose a dual-path Neural CDE framework that explicitly disentangles smooth trend dynamics from abrupt event-driven changes, addressing the structural mismatch between stepwise time series and conventional single-stream temporal representations.

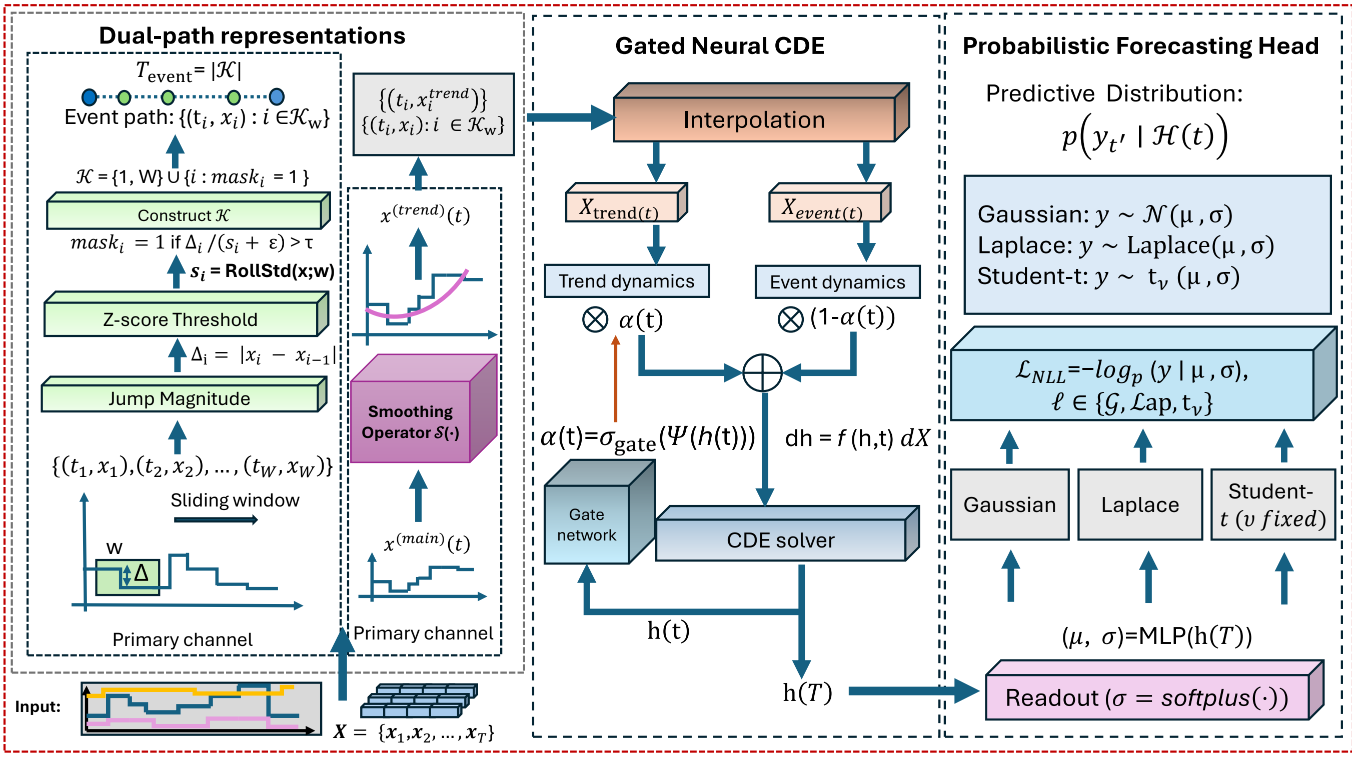

Fig. 2. Overview of the proposed architecture, organized into three modules. **(A) Dual-path representations:** a trend path summarizes persistent stepwise regimes via smoothing, while an event path sparsifies the primary channel by retaining endpoints and detected change points. **(B) Gated Neural CDE:** both paths are interpolated into continuous-time controls and integrated by a gated Neural CDE, producing a terminal latent summary $h(T)$. **(C) Probabilistic forecasting head:** a readout maps $h(T)$ to distribution parameters and supports interchangeable likelihoods (Gaussian, Laplace, Student-t); training uses the corresponding negative log-likelihood. Here, W denotes the local window length used in the within-window construction.

- We systematically investigate likelihood-based probabilistic modeling under sparse stepwise dynamics, showing that heavy-tailed likelihoods improve distributional accuracy under regime-shifted dynamics, while calibrated interval reliability remains subject to a sharpness trade-off.
- We adopt a CRPS-first evaluation protocol to assess probabilistic forecast quality, highlighting that improvements in uncertainty modeling are not captured by point prediction metrics alone and establishing a principled basis for model comparison under regime-shifted dynamics.

II. Method

A. Problem Setup

We consider the problem of probabilistic forecasting for time series exhibiting sparse stepwise dynamics. Let $\{(t_i, \mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote an observed sequence, where $t_i \in \mathbb{R}^+$ represents time, $\mathbf{x}_i \in \mathbb{R}^d$ denotes a vector of temporal covariates, and $y_i \in \mathbb{R}$ is the corresponding target variable. The covariates are asynchronously updated and remain piecewise-constant between updates, resulting in long near-constant regimes interspersed with rare abrupt regime changes.

Given historical observations up to time t, our goal is to model the conditional predictive distribution of future values at a target time $t' > t$. Rather than producing a single point estimate, we aim to learn a predictive distribution

$$p(y(t') \mid \mathcal{H}(t)),$$

where $\mathcal{H}(t)$ denotes the available observation history up to time t. This formulation enables explicit quantification of predictive uncertainty, which is essential for decision-making in environments subject to sudden, infrequent changes.

A key challenge in this setting arises from the mismatch between the temporal structure of stepwise series and conventional modeling assumptions. Most observations convey redundant information during stable regimes, while informative signals are concentrated around rare transition events. As a result, effective forecasting requires both capturing persistent trends over extended periods and accurately representing uncertainty around regime shifts.

To address these challenges, we adopt a continuous-time modeling framework that naturally accommodates irregular sampling and enables flexible representation of temporal dynamics. In the following sections, we introduce a representation and modeling framework that explicitly accounts for both long-term persistence and abrupt transitions, and integrates these components with likelihood-based probabilistic forecasting to model uncertainty under sparse stepwise dynamics.

B. Dual-Path Representation for Stepwise Dynamics

To address the structural mismatch between stepwise time series and conventional single-stream temporal representations, we introduce a dual-path representation that explicitly separates smooth trend dynamics from event-driven changes. The core idea is to encode persistent regimes and abrupt transitions through complementary temporal channels, enabling effective modeling of both long stable periods and rare regime shifts.

Given an observed sequence $\{(t_i, \mathbf{x}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$, we construct two continuous-time input paths via deterministic transformation operators:

$$\mathbf{x}^{(\text{trend})}(t) = \mathcal{S}(\mathbf{x})(t), \qquad \mathbf{x}^{(\text{event})}(t) = \mathcal{E}(\mathbf{x})(t), \tag{1}$$

where $\mathcal{S}(\cdot)$ denotes a smoothing operator that extracts slowly varying dynamics, and $\mathcal{E}(\cdot)$ extracts transition-related information.

The trend path captures long-term persistence by suppressing high-frequency fluctuations and emphasizing stable regime structure. In contrast, the event path focuses on abrupt temporal changes. Specifically, for the primary scalar channel $x(t_i)$, we define the normalized temporal increment

$$\Delta_i = |x(t_i) - x(t_{i-1})|, \qquad s_i = \operatorname{Std}\big(x(t_{\max(1,\, i-w+1)}), \ldots, x(t_i)\big), \quad i = 2, \ldots, N, \tag{2}$$

and identify an event at time t_i if

$$\frac{\Delta_i}{s_i + \varepsilon} > \tau, \tag{3}$$

where w denotes a rolling window length and τ is a fixed threshold. This criterion detects statistically significant deviations relative to the local temporal scale and serves as an implicit indicator of regime transitions.

To emphasize informative transitions and prevent frequent event-driven excitation during stable regimes, we perform *event-path sparsification* by retaining only a subset of informative time indices to construct the event-driven control path. Specifically, we define the retained index set as

$$\mathcal{K} = \{1, N\} \cup \left\{ i \in \{2, \ldots, N\} : \frac{\Delta_i}{s_i + \varepsilon} > \tau \right\}. \tag{4}$$

where the inclusion of the sequence endpoints ensures numerical well-posedness and preserves global temporal structure even in the absence of detected events. The resulting event-driven control path is given by

$$\mathbf{x}^{(\text{event})}(t) = \{(t_i, x(t_i))\}_{i \in \mathcal{K}}. \tag{5}$$

The two paths are jointly provided to the downstream model, allowing long-term trends and abrupt transitions to be represented simultaneously. As demonstrated in the ablation study, neither path alone is sufficient to capture sparse stepwise dynamics. The trend path dominates predictive performance during long stable regimes, whereas the event path provides complementary signals that become informative around regime transitions. Their integration therefore forms a necessary representational foundation for modeling stepwise time series.

However, the relative importance of the trend and event components is inherently time-varying. During extended constant segments, excessive reliance on event-driven signals may introduce noise, whereas around abrupt changes, ignoring event information can lead to delayed adaptation and mischaracterized uncertainty. An effective forecasting model must therefore be able to dynamically balance persistent regime-level dynamics and transient event-driven cues as the system evolves.

To accommodate this temporal variability, we introduce an adaptive gating mechanism that regulates the relative contributions of the two paths over time. The concrete realization of this gating mechanism is described in the subsequent continuous-time modeling framework.

C. Continuous-Time Modeling via Neural CDE with Gated Fusion

To realize the dual-path representation under irregular sampling, we adopt a continuous-time formulation based on Neural Controlled Differential Equations (Neural CDEs). Neural CDEs model latent dynamics driven by continuous control paths, making them well suited for irregularly sampled time series in which both trend evolution and event effects unfold asynchronously over time [25].

Given the trend and event control paths constructed in the previous section, we define a continuous-time latent dynamics governed by a gated Neural CDE. Each discrete control path is first transformed into a continuous-time control function via a deterministic interpolation operator, following standard practice in Neural CDEs. After interpolating each path into a continuous control function, the latent state $\mathbf{h}(t) \in \mathbb{R}^H$ evolves according to

$$\begin{aligned} \frac{d\mathbf{h}(t)}{dt} = {} & \alpha(t)\, f_\theta\big(\mathbf{h}(t), t\big)\, \dot{\mathbf{X}}^{(\text{trend})}(t) \\ & + \big(1 - \alpha(t)\big)\, g_\theta\big(\mathbf{h}(t), t\big)\, \dot{\mathbf{X}}^{(\text{event})}(t), \end{aligned} \tag{6}$$

with the initial condition

$$\mathbf{h}(0) = \phi_\theta\Big(\mathbf{X}^{(\text{trend})}(0)\Big), \tag{7}$$

where $\phi_\theta(\cdot)$ is a learnable initialization function that maps the initial trend control to the latent state space. Here, $\dot{\mathbf{X}}^{(\cdot)}(t)$ denotes the time-derivative of the interpolated control path. The functions f_θ and g_θ are learnable vector fields that map the latent state and control inputs to $\mathbb{R}^{H \times d_{\text{trend}}}$ and $\mathbb{R}^{H \times d_{\text{event}}}$, respectively, governing how trend and event information drive the latent dynamics.

The scalar gating function $\alpha(t) \in [0, 1]$ is learned jointly with the vector fields and enables time-varying fusion of the two paths. We parameterize the gate as

$$\alpha(t) = \sigma\big(\psi_\theta(\mathbf{h}(t))\big), \tag{8}$$

where $\psi_\theta(\cdot)$ denotes a learnable function and $\sigma(\cdot)$ is the sigmoid activation.

The terminal latent state $\mathbf{h}(T)$, where T denotes the end time of the conditioning window, summarizes the observation history and serves as the conditioning representation for likelihood-based probabilistic forecasting. We initialize the latent state from the trend path only, as it provides a stable representation of the prevailing regime, while the event path is sparse and intended to modulate the dynamics rather than define the initial state.

D. Likelihood-Based Probabilistic Forecasting

The continuous-time latent state $\mathbf{h}(T)$ produced by the gated Neural CDE serves as a compact summary of the historical information up to time T. To perform probabilistic forecasting, we map this latent representation to the parameters of a predictive distribution over future observations. Specifically, a lightweight output network maps the latent state to the parameters of the predictive distribution. For the Gaussian and Laplace likelihoods, the network outputs the location and scale parameters (μ, σ), while for the Student-*t* likelihood, we use a fixed degrees-of-freedom parameter ν.

We consider three likelihood families that differ in their distributional assumptions and tail behavior. First, the Gaussian likelihood assumes light-tailed residuals and is commonly used in probabilistic time-series models:

$$p_{\mathcal{N}}(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{9}$$

Second, the Laplace likelihood provides a symmetric heavy-tailed alternative that is more robust to moderate deviations:

$$p_{\mathcal{L}}(y \mid \mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|y-\mu|}{\sigma}\right). \tag{10}$$

Finally, to explicitly account for rare but substantial deviations induced by abrupt regime changes, we adopt the Student-*t* likelihood:

$$p_{\mathcal{T}}(y \mid \mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}, \tag{11}$$

where ν denotes the degrees of freedom controlling tail heaviness. In our experiments, we fix $\nu = 5$.

These likelihood choices allow the model to adapt to the effective heavy-tailed residual behavior induced by abrupt regime changes in stepwise time series.

E. Training Objective

Let $\mathcal{D} = \{(y_i, \mathbf{h}_i)\}_{i=1}^N$ denote the training set, where $\mathbf{h}_i = \mathbf{h}_i(T)$ is the terminal latent state produced by the gated Neural CDE for the i-th input sequence. Model parameters Θ are learned by minimizing the negative log-likelihood of the observed targets under the chosen predictive distribution:

$$\mathcal{L}(\Theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \mathbf{h}_i; \Theta), \tag{12}$$

where $p(\cdot)$ denotes the Gaussian, Laplace, or Student-*t* likelihood parameterized by the output network.

All components of the model, including the dual-path representation, the gated Neural CDE dynamics, and the likelihood parameters, are optimized jointly in an end-to-end manner.

III. Experiments

A. Experimental Setup

We study one-step-ahead probabilistic forecasting on a proprietary multivariate real-estate rental time series. The target is rental price (stepwise), and inputs include rental price with time-aligned covariates (commuting time and property type, one-hot encoded); covariates are forward-filled between updates. Let $W{=}20$ denote the input context length. For each start index i, we use the context segment $\{(t_j, \mathbf{x}_j)\}_{j=i}^{i+W-1}$ to predict the next-step target y_{i+W}.

We split the sequence chronologically, using the first 80% of timesteps for training and the last 20% for validation. Within each split, we generate examples by sliding a length-W context window along the segment and pairing it with the next-step target y_{i+W}; we keep only those indices i for which both the context $\{(t_j, \mathbf{x}_j)\}_{j=i}^{i+W-1}$ and the target y_{i+W} fall within the same split.

To capture long-term trends and rare regime shifts, we use dual-path inputs: a multichannel trend representation derived from a fixed smoothing-based preprocessing of the main series together with time-aligned covariates, and a raw event path from the primary target-associated feature, where event sparsification is applied only to this raw channel. All inputs and targets are MinMax-normalized using parameters fitted on the training split. Predictions are mapped back to the original target scale for reporting point metrics, and we additionally report probabilistic metrics including CRPS, NLL, cov@90, and wid@90.

To further emphasize abrupt regime shifts in the raw target-associated channel, we optionally compress the event-path input by retaining only informative change points within each conditioning window. Let $x(t)$ denote the normalized raw series in the event path. For consecutive timestamps $\{t_j\}$ in a window, we compute $\Delta_j = |x(t_j) - x(t_{j-1})|$ for $j \geq 1$ and normalize it by a local scale s_j estimated via a rolling standard deviation with window size w_{roll}. We retain t_j if $\Delta_j/(s_j + \epsilon) > \tau$, where ϵ is a small constant for numerical stability. The first and last timestamps of each window are always included, yielding a variable-length event control with $T_c \leq W$.

We evaluate both point and probabilistic forecasting quality. We report RMSE, CRPS, and NLL. For uncertainty quantification, we report 90% prediction interval coverage and width (Cov@90 and Width@90). For likelihood-based models, raw 90% intervals are obtained from predictive quantiles implied by the fitted distribution, and are further calibrated using cross-conformal calibration with a calibration subset carved out from the training split. We additionally report MSE, MAE, MAPE, and R^2 for a more complete point-forecast comparison.

B. Implementation Details

The proposed model is a dual-branch Neural CDE with a trend encoder and an event encoder, followed by gated fusion and a likelihood-specific output head. Each branch constructs a continuous control path via linear interpolation and evolves a latent state using `torchcde.cdeint`. We solve the resulting CDE using an adaptive Runge–Kutta method (e.g., `dopri5`) with tolerances $\texttt{rtol} = 10^{-4}$ and $\texttt{atol} = 10^{-5}$, which automatically adjusts internal step sizes. All timestamps are mapped to a normalized continuous domain $t \in [0, 1]$ using an equally spaced base grid shared across both trend and event paths.

The latent dimensionality is fixed to H=32. In each branch, the CDE vector field is parameterized by a two-layer MLP (hidden width 64) with ReLU and a final $\tanh$, producing a matrix-valued field in $\mathbb{R}^{H\times D}$ that is scaled by 0.05. Each encoder initializes z_0 from its first observation via a learnable linear map and outputs the terminal latent state. The two branch latents are linearly projected and fused by a dimension-wise sigmoid gate initialized to yield an equal (0.5/0.5) mixture, and the fused latent is scaled by 0.2 before the likelihood-specific output heads. We follow the event-path sparsification rule in Sec. III-A and use $w_{\text{roll}} = 16$ and $\tau = 2.0$ unless otherwise stated.

For likelihood-specific output heads, the predictive mean is produced by an MLP with LayerNorm and a final sigmoid activation to match the MinMax-normalized target range. The predictive scale is produced by a separate MLP that outputs an unconstrained raw value s. We parameterize the scale as $\sigma = \sigma_{\min} + \text{softplus}(s)$ to ensure positivity, and clip it to an upper bound $\sigma_{\max}$ for numerical stability when optimizing the heteroscedastic Gaussian likelihood (in normalized space, $\sigma_{\min} = 0.07$ and $\sigma_{\max} = 1.00$ in our experiments). For the Student-*t* likelihood, the degrees-of-freedom parameter ν is fixed to a constant during training.

All models are trained end-to-end using AdamW with learning rate 3×10^{-4} and weight decay 10^{-2}. Training is run for up to 50 epochs with gradient norm clipping at 1.0. Early stopping is applied based on validation RMSE with patience 8 and a relative minimum improvement threshold of 0.002. Training batches are shuffled, while validation is evaluated without shuffling under the chronological split.

When event sparsification is enabled, the event-path control has variable length across windows; we therefore use a custom collation function and set the batch size to 1 for both training and validation loaders.

C. Results

We report probabilistic forecasting quality using CRPS (lower is better), complemented by empirical coverage and average interval width at the 90% nominal level to quantify calibration and sharpness; RMSE is included to ensure that improvements in distributional quality are not achieved by degrading the mean prediction. As shown in Table I, the likelihood-based dual-path Neural CDE variants achieve the strongest distributional accuracy. In particular, Student-*t* CDE attains the lowest CRPS (7.86), outperforming Gaussian CDE (8.10), Laplace CDE (9.65), the strongest non-CDE probabilistic baseline (naïve Gaussian, 9.75), and all standard neural baselines. To control for the effect of the predictive distribution, we additionally equip GRU with a Student-*t* head. This improves GRU from CRPS 10.70 to 9.84 and NLL 4.91 to 4.23, showing that heavy-tailed likelihoods are beneficial for stepwise forecasting. However, Student-*t* CDE still yields substantially better CRPS and NLL (7.86/2.32 vs. 9.84/4.23), indicating that the gain cannot be explained by the likelihood choice alone.

Likelihood choices also exhibit a clear calibration–sharpness trade-off on 90% intervals. Gaussian CDE is conservative on raw intervals (Cov=0.99, Width=103.93), whereas Student-*t* CDE is much sharper but under-covers (Cov=0.70, Width=19.42). After cross-conformal calibration, all CDE variants reach coverage near the nominal level, but at the cost of wider intervals, especially for Student-*t* CDE (19.42 $\rightarrow$ 53.18). Thus, Student-*t* provides the best proper scoring performance, while calibrated interval sharpness remains a trade-off rather than a uniform improvement.

Table II confirms that these probabilistic gains do not come at the expense of mean prediction accuracy. Gaussian CDE achieves the best RMSE (16.74) and the highest R^2 (0.9593), compared with the best neural Gaussian-head baseline (RNN, RMSE=22.79, R^2=0.9043) and the best classical baseline (naïve Gaussian, RMSE=28.80, R^2=0.9287). In terms of likelihood fit, CDE models attain much lower NLL (Gaussian: 1.98; Laplace: 1.42; Student-*t*: 2.32) than neural Gaussian-head baselines (4.62–5.21) and the naïve Gaussian baseline (4.79), indicating a substantially better match between predicted distributions and observed outcomes; CQR does not define a likelihood and is excluded from NLL comparisons.

Taken together, the results show that the proposed dual-path Neural CDE family improves distributional accuracy (best CRPS=7.86) while maintaining strong mean prediction performance (best RMSE=16.74, R^2=0.9593), and achieves statistically calibrated intervals after cross-conformal adjustment (Cov@90 $\approx$ 0.90–0.91) with markedly tighter widths than classical alternatives.

D. Ablation Study

To examine the necessity of the proposed dual-path input design, we conduct an ablation study by comparing event-only, trend-only, and dual-path representations. As shown in Table III, the trend-only model already captures the dominant predictive signal, substantially outperforming the event-only variant, which performs poorly when used in isolation. This indicates that smooth trend dynamics constitute the primary component for point forecasting in sparse stepwise time series. However, incorporating event-driven information further yields consistent improvements across all evaluation metrics. The dual-path model achieves the best performance by effectively integrating stable trend representations with complementary event cues, providing a more informative representation foundation upon which probabilistic forecasting becomes effective.

E. Discussion

The experimental results reveal several insights into probabilistic forecasting for sparse stepwise time series. First, the discrepancy between point prediction metrics and CRPS

TABLE I
PROBABILISTIC FORECASTING PERFORMANCE ON THE STEPWISE TIME SERIES. LOWER IS BETTER FOR RMSE/CRPS/NLL AND INTERVAL WIDTH. **RAW INTERVALS** ARE OBTAINED FROM PREDICTIVE QUANTILES OF LIKELIHOOD-BASED MODELS. **CROSS-CONFORMAL INTERVALS** ARE PRODUCED BY A POST-HOC CROSS-CONFORMAL CALIBRATION APPLIED TO THE RAW INTERVALS. THE QUANTILE BASELINE USES CQR BY DESIGN (CONFORMAL-STYLE) AND THUS DOES NOT DEFINE AN NLL.

Model	**Point / Distribution Metrics**			**Raw Interval @90**		**Cross-conformal Interval @90**	
	RMSE↓	**CRPS↓**	**NLL↓**	**Cov↑**	**Width↓**	**Cov↑**	**Width↓**
Likelihood-based Dual-path Neural CDE							
Gaussian CDE	16.74	8.10	1.98	0.99	103.93	0.91	43.58
Laplace CDE	18.87	9.65	1.42	0.81	42.40	0.90	51.21
Student-t CDE (ν fixed)	19.02	7.86	2.32	0.70	19.42	0.91	53.18
Standard neural sequence baselines							
RNN	22.79	15.65	4.85	0.94	67.01	0.88	33.39
GRU (Gaussian)	24.07	10.70	4.91	0.94	73.73	0.92	39.21
GRU (Student-t)	24.07	9.84	4.23	0.95	76.93	0.92	41.78
LSTM	24.30	12.59	4.62	0.96	92.56	0.91	49.65
Transformer	27.11	13.66	5.21	0.95	70.45	0.88	50.91
Other probabilistic baselines							
Quantile baseline (CQR)	32.95	13.58	–	0.89	62.44	0.95	126.40
Naïve Gaussian	28.80	9.75	4.79	0.91	45.60	0.93	60.00
Constant Gaussian	162.01	111.77	7.38	0.34	212.98	0.61	180.55

TABLE II
POINT FORECASTING ACCURACY AND LIKELIHOOD-BASED FIT. LOWER IS BETTER FOR ALL ERROR METRICS AND NLL; HIGHER IS BETTER FOR R^2. MAPE IS REPORTED IN PERCENTAGE (%). QUANTILE BASELINE DOES NOT ADMIT A LIKELIHOOD AND THEREFORE HAS NO NLL.

Model	MSE↓	RMSE↓	MAE↓	MAPE (%)↓	R^2↑	NLL↓
Gaussian CDE	283.50	16.74	9.86	1.26	0.9593	1.98
Laplace CDE	355.91	18.87	11.15	2.14	0.9483	1.42
Student-t CDE	361.61	19.02	10.41	1.97	0.9475	2.32
RNN (Gauss)	519.36	22.79	14.52	2.91	0.9043	4.85
GRU (Gauss)	579.51	24.07	11.58	2.33	0.9166	4.91
LSTM (Gauss)	590.61	24.30	18.09	3.48	0.9100	4.62
Transformer (Gauss)	735.18	27.11	20.06	3.78	0.9137	5.21
Quantile (CQR)	1085.41	32.95	18.87	3.03	0.9067	–
Naïve Gaussian	829.35	28.80	11.79	1.31	0.9287	4.79
Constant Gaussian	25667.8	162.01	140.6	21.19	-1.21	7.38

TABLE III
ABLATION STUDY ON THE DUAL-PATH INPUT DESIGN FOR POINT FORECASTING.

Model	**MSE↓**	**RMSE↓**	**MAE↓**	**MAPE (%)↓**	R^2 ↑
Event-only	1259.81	35.49	23.15	4.28	0.8171
Trend-only	371.93	19.29	8.93	1.64	0.9500
Dual (trend + event)	**234.34**	**15.31**	**7.53**	**1.40**	**0.9660**

highlights that accurate mean estimation does not necessarily translate into high-quality uncertainty modeling. Gaussian likelihoods achieve the best point accuracy (RMSE=16.74), yet Student-t yields the best probabilistic accuracy under CRPS (7.86 vs. 8.10 for Gaussian), indicating that distributional fit rather than mean accuracy becomes decisive under regime-shifted dynamics.

Second, the calibration–sharpness analysis suggests that likelihood choice induces distinct uncertainty behaviors. On raw 90% intervals, Gaussian is strongly over-conservative (Cov=0.99) with very wide intervals (Width=103.93), whereas Student-t is sharp (Width=19.42) but under-covers (Cov=0.70), and Laplace lies between them (Cov=0.81, Width=42.40). After cross-conformal calibration, all likelihood-based CDE variants become well calibrated near the nominal level (Cov $\approx$ 0.90–0.91), but exhibit different sharpness (Gaussian Width=43.58 vs. 51.21 for Laplace and 53.18 for Student-t). These patterns support the use of heavy-tailed likelihoods in settings with abrupt, infrequent changes: Student-t attains the best CRPS despite mild under-coverage, suggesting that its advantage is not driven by interval inflation but by better alignment to the effective heavy-tailed residual behavior induced by regime shifts. The NLL–CRPS mismatch is also informative: although Laplace achieves the lowest NLL (1.42), it does not achieve the best CRPS (9.65 vs. 7.86 for Student-t), which is consistent with the two metrics emphasizing different aspects of probabilistic quality. NLL is dominated by local log-density at realized outcomes and can favor sharper distributions under moderately heavy-tailed residuals, whereas CRPS evaluates the entire predictive distribution as a proper scoring rule and is less dominated by a small number of extreme errors; under rare but substantial regime-shift deviations, Student-t better accommodates tail events, improving CRPS even when its log-likelihood is not minimal. Notably, removing extreme observations does not resolve miscalibration and can even exacerbate over-coverage, indicating that uncertainty inflation is driven by structural mismatch rather than data-level outliers; this further motivates likelihood choices that remain robust under regime shifts.

Third, the ablation study clarifies the role of representation structure in enabling effective probabilistic modeling. The trend-only variant captures the dominant predictive signal, reflecting the persistence of stable regimes over extended periods. The event-only model is ineffective in isolation, but it encodes complementary information associated with abrupt transitions. Integrating both components yields consistent improvements across point prediction metrics and provides a structural foundation upon which likelihood-based uncertainty modeling becomes effective; without this decomposition, the benefits of advanced probabilistic modeling are substantially diminished.

Overall, these findings suggest that probabilistic forecasting for sparse stepwise time series requires both an appropriate distributional assumption and a representation that disentangles smooth dynamics from abrupt events. The proposed dual-path Neural CDE framework addresses these requirements jointly, improving probabilistic forecasting under regime-shifted dynamics, while calibrated interval sharpness remains subject to a trade-off.

IV. Conclusion

We presented a dual-path Neural CDE framework for probabilistic forecasting of sparse stepwise time series. By explicitly separating smooth trend dynamics from event-driven variations, the model provides a structured representation for likelihood-based uncertainty modeling. Empirically, heavy-tailed likelihoods improve distributional accuracy under regime-shifted dynamics: the Student-t variant achieves the best CRPS, while the Gaussian variant attains the best RMSE, illustrating that optimal probabilistic fit and mean accuracy may favor different residual assumptions. Ablation results further show that the trend pathway captures the dominant predictive signal, whereas the event pathway contributes complementary transition information, and combining both yields the most consistent overall performance. Overall, our findings suggest that effective probabilistic forecasting for real-world stepwise series requires jointly designing the representation and the predictive distribution, while calibration quality and interval sharpness remain subject to a trade-off.

Acknowledgement

The authors used ChatGPT (OpenAI) for limited language polishing during manuscript preparation. All technical content, experimental design, analysis, and final wording were verified by the authors, who take full responsibility for the paper.

References

[1] S. L. Joslyn and J. E. LeClerc, "Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error." *Journal of experimental psychology: applied*, vol. 18, no. 1, p. 126, 2012.

[2] L. Held, S. Meyer, and J. Bracher, "Probabilistic forecasting in infectious disease epidemiology: the 13th armitage lecture," *Statistics in medicine*, vol. 36, no. 22, pp. 3443–3460, 2017.

[3] Z.-Y. Chen, "The value of probabilistic forecasting in emergency medical resource planning under uncertainty," *Kybernetes*, vol. 52, no. 6, pp. 1962–1975, 2023.

[4] Y. Yang, D. Yin, H. Xue, and F. Salim, "Stoat: Spatial-temporal probabilistic causal inference network," in *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, 2025, pp. 499–510.

[5] Y. Jiao, A. Sharma, A. Ben Abdallah, T. M. Maddox, and T. Kannampallil, "Probabilistic forecasting of surgical case duration using machine learning: model development and validation," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1885–1893, 2020.

[6] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[7] G. Muradoğlu and D. Önkal, "An exploratory analysis of portfolio managers' probabilistic forecasts of stock prices," *Journal of Forecasting*, vol. 13, no. 7, pp. 565–578, 1994.

[8] J.-H. Böse, V. Flunkert, J. Gasthaus, T. Januschowski, D. Lange, D. Salinas, S. Schelter, M. Seeger, and Y. Wang, "Probabilistic demand forecasting at scale," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1694–1705, 2017.

[9] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 69, no. 2, pp. 243–268, 2007.

[10] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arxiv," *arXiv preprint arXiv:1803.01271*, vol. 10, 2018.

[11] Y. Liu, H. Wu, J. Wang, and M. Long, "Non-stationary transformers: Exploring the stationarity in time series forecasting," *Advances in neural information processing systems*, vol. 35, pp. 9881–9893, 2022.

[12] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[13] K. Lehnertz, L. Zabawa, and M. R. R. Tabar, "Characterizing abrupt transitions in stochastic dynamics," *New Journal of Physics*, vol. 20, no. 11, p. 113043, 2018.

[14] A. O. Aydınhan, P. N. Kolm, J. M. Mulvey, and Y. Shu, "Identifying patterns in financial markets: extending the statistical jump model for regime identification," *Annals of Operations Research*, pp. 1–37, 2024.

[15] M. Shepperd and M. Cartwright, "Predicting with sparse data," *IEEE Transactions on Software Engineering*, vol. 27, no. 11, pp. 987–998, 2002.

[16] U. Evci, Y. Ioannou, C. Keskin, and Y. Dauphin, "Gradient flow in sparse neural networks and how lottery tickets win," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 6, 2022, pp. 6577–6586.

[17] C. Vox, D. Broneske, I. M. Shaikat, and G. Saake, "Data streams: Investigating data structures for multivariate asynchronous time series prediction problems." in *ICPRAM*, 2023, pp. 686–696.

[18] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. 1–29, 2023.

[19] D. T. Frazier, D. J. Nott, and C. Drovandi, "Synthetic likelihood in misspecified models," *Journal of the American Statistical Association*, vol. 120, no. 550, pp. 884–895, 2025.

[20] H. Papadopoulos, "Guaranteed coverage prediction intervals with gaussian process regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9072–9083, 2024.

[21] C. S. Wong, W.-S. Chan, and P. Kam, "A student t-mixture autoregressive model with applications to heavy-tailed financial data," *Biometrika*, vol. 96, no. 3, pp. 751–760, 2009.

[22] J. Prüher, F. Tronarp, T. Karvonen, S. Särkkä, and O. Straka, "Student-t process quadratures for filtering of non-linear systems with heavy-tailed noise," in *2017 20th International Conference on Information Fusion (Fusion)*. IEEE, 2017, pp. 1–8.

[23] C. Schluter and M. Trede, "Weak convergence to the student and laplace distributions," *Journal of Applied Probability*, vol. 53, no. 1, pp. 121–129, 2016.

[24] T. C. Aysal and K. E. Barner, "Second-order heavy-tailed distributions and tail analysis," *IEEE transactions on signal processing*, vol. 54, no. 7, pp. 2827–2832, 2006.

[25] P. Kidger, J. Morrill, J. Foster, and T. Lyons, "Neural controlled differential equations for irregular time series," *Advances in neural information processing systems*, vol. 33, pp. 6696–6707, 2020.