# RESOPATH: DISENTANGLING MULTI-SCALE CONTINUOUS PATHS WITH MISSINGNESS-AWARE NCDES FOR IRREGULAR TIME SERIES FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Irregularly sampled multivariate time series play a pivotal role across science and engineering, yet non-uniform observation intervals and severe missingness make continuous dynamic modeling challenging. Existing neural controlled differential equation based methods collapse sparse observations into single interpolated paths, conflating multi-scale patterns with synthetic smoothing artifacts and ignoring observation reliability. In this paper, we introduce **ResoPath**, an end-to-end framework that constructs multi**reso**lution continuous control **path**s by disentangling coarse temporal trends from fine-scale fluctuations. ResoPath couples wavelet decomposition with frequency-conditioned denoising to construct continuous control paths that drive multiscale latent dynamics through vector fields conditioned on missingness-aware information. Extensive comparisons with diverse baselines across five real-world datasets demonstrate that ResoPath achieves state-of-the-art accuracy and strong robustness in the irregular multivariate time series forecasting task.

## 1 INTRODUCTION

Irregularly sampled multivariate time series, or IMTS, are common in practice. In healthcare, ICU vital signs and electronic records are observed at times set by clinical need rather than a fixed schedule. In finance, trade records themselves arrive at uneven timestamps (Qiu et al., 2026). Accurate forecasting from such sparse data is essential for real-world decision making (Liu et al., 2026). This sparse data also carries multi-scale information, since time series exhibit different patterns at different sampling granularities, with fine-grained views revealing local variations and coarse-grained views revealing trend and periodicity (Li et al., 2025; 2026). Consequently, achieving accurate and robust forecasting from such sparse, multi-scale data remains an open challenge.

Among recent approaches to IMTS forecasting, two lines of research are particularly dominant. The first constructs discrete local representations through patching or temporal aggregation, extending techniques successful on regularly sampled time series (Nie et al., 2023; Zhang et al., 2024; Liu et al., 2025; Luo et al., 2025; Zhang et al., 2023). While improving computational efficiency and capturing temporal dependencies, such patch-level aggregation often blurs fine-grained temporal dynamics, particularly when observations are highly sparse or unevenly distributed.

A complementary line models temporal evolution directly in continuous time using neural differential equations (Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020; Biloš et al., 2021). Specifically, neural controlled differential equations (NCDE) construct a continuous control path from irregular observations via interpolation to drive latent dynamics. However, this process typically results in a single, deterministic trajectory that integrates all observations into a unified representation, potentially conflating coarse trends with fine fluctuations and failing to account for missing data effectively.

Under highly sparse and irregular observations, as shown in Figure 1, NCDE-based models face three related challenges: (1) **Interpolation Distortions.** NCDEs require constructing a continuous driving path from discrete observations. Across large observation gaps, standard interpolation often creates overly smooth or monotonic trajectories, biasing the learned vector field toward these artificial assumptions rather than the underlying system dynamics (Liu et al., 2026). (2) **Uncertainty-**



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

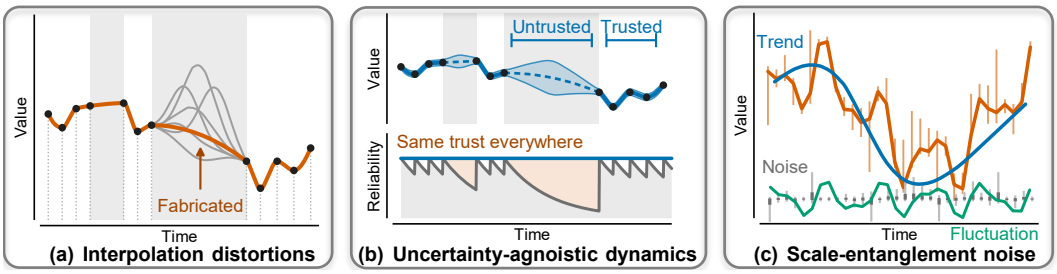



Figure 1: Challenges of NCDE-based models under sparse and irregular observations. **(a)**: Interpolation fabricates a single smooth, monotone path as if it were certain. **(b)**: Once interpolated, the model sees one continuous curve and cannot tell observed segments from guessed ones. **(c)**: A single path entangles trend, local fluctuation, and interpolation-induced noise.

**Agnostic Dynamics.** Existing CDEs treat genuine measurements and interpolated values indiscriminately, regardless of the gap length. Consequently, the model evolves its latent state with uniform confidence, failing to account for the decaying reliability of the driving path across extended intervals. (3) **Scale Entanglement Noise.** A single driving path conflates macro-level trends with local fluctuations, hindering scale-specific representation. Concurrently, the high-frequency oscillations induced by sparse interpolation destabilize the numerical solver and exacerbate overfitting (Oh et al., 2025). In short, conventional single-path continuous models are ill-equipped to manage sparsity-induced uncertainty while simultaneously disentangling multi-scale dynamics.

In this work, we propose ResoPath, a continuous-time missingness-aware multi-scale NCDE-based method for forecasting irregular multivariate time series. ResoPath first employs a frequency-conditioned wavelet frame denoising module to construct denoised and scale-specific control paths. These paths then drive parallel multi-scale CDEs, with each scale governed by a dedicated latent vector field. To account for the reliability of irregular observations, we further introduce a missingness-aware dynamics function that continuously modulates each vector field using observation-gap and recency information. The resulting multi-scale latent states are aggregated and propagated through an autonomous latent ODE, enabling forecasts at arbitrary future timestamps without requiring future interpolation.

The main contributions are as follows:

- We propose **ResoPath**, the first multi-scale NCDE framework for irregular time series that decouples local fluctuations from macro trends, eliminating reliance on a single interpolated path.
- We design a robust wavelet frame denoising module coupled with a missingness-aware dynamics function, effectively suppressing interpolation artifacts across subbands while adapting vector fields to observation uncertainty.
- Extensive experiments on five real-world benchmarks demonstrate that ResoPath consistently outperforms state-of-the-art baselines under sparse and irregular sampling conditions.

## 2 Problem Formulation

We denote an irregularly-sampled multivariate time series as $\mathcal{D} = \{(t_i, \mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$, where $t_i \in \mathbb{R}^+$ are non-uniform timestamps satisfying $t_0 \leq t_1 < \cdots < t_N \leq T$, $\mathbf{x}_i \in \mathbb{R}^D$ denotes observations across $D$ channels, and $\mathbf{m}_i \in \{0,1\}^D$ is the corresponding observation mask, with $m_{i,d} = 1$ indicating that the $d$-th variable is measured at $t_i$ and $m_{i,d} = 0$ otherwise. The sampling interval $\Delta t_i = t_i - t_{i-1}$ varies arbitrarily across the observation window $[t_0, T]$. Given the partially observed trajectory $\mathcal{D}$ within $[t_0, T]$, sparse multivariate time series forecasting aims to learn a dynamics model that predicts future system states $\hat{\mathbf{Y}} = \{\hat{\mathbf{x}}(T + \tau_k)\}_{k=1}^{K}$ at arbitrary query timestamps $\{T + \tau_k\}_{k=1}^{K} \subset [T, T + H]$. Unlike discrete models that regress over fixed grids, continuous-time NCDEs (Kidger et al., 2020) drive latent dynamics through a continuous control

path $X : [t_0, T] \to \mathbb{R}^{D+1}$ fitted onto discrete observations via interpolation. Standard interpolation schemes, however, are prone to amplifying observation noise into spurious path derivatives; ResoPath instead denoises this construction through frequency-aware denoising to jointly tackle severe irregularity, missingness, and multi-scale dynamics.

## 3 METHODOLOGY

In this section, we present the comprehensive formulation of the proposed ResoPath model for IMTS forecasting. Figure 2 illustrates the overall pipeline.

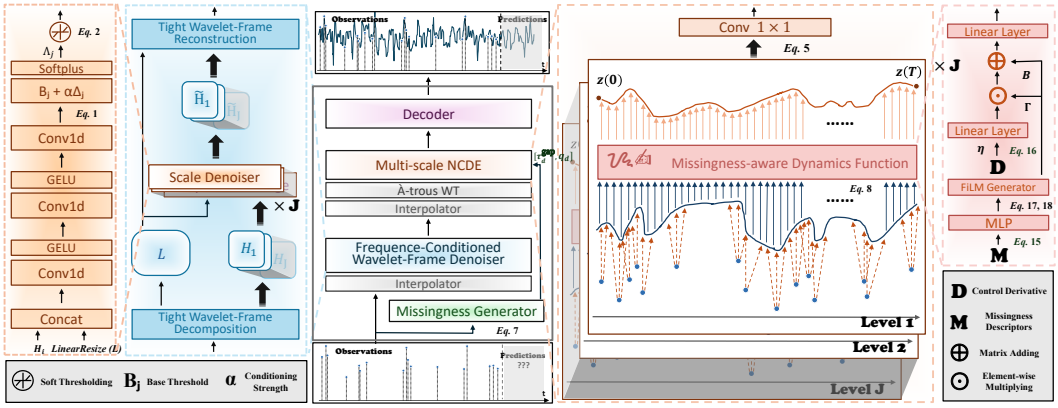



Figure 2: Architecture of ResoPath.

Given an irregularly sampled input, we first apply wavelet frame decomposition to obtain scale-specific coefficients $\{\mathbf{c}^{(j)}\}_{j=0}^{J}$, separating fast local fluctuations from coarse trends without relying on a single interpolated path. A **Frequency-Conditioned Wavelet Frame Denoising (FCWFD)** module (Section 3.1) then adaptively suppresses interpolation-induced noise within each subband before smooth control paths $\{\widetilde{X}^{(j)}(t)\}_{j=0}^{J}$ are constructed. These denoised paths drive a **Multi-Scale NCDE (MS-NCDE)** (Section 3.2), in which each scale evolves through an independent, jointly trained vector field, mitigating the monotonic interpolation artifacts of single-path NCDEs. Each scale-specific vector field is parameterized as a **Missingness-Aware Dynamics Function (MaDF)** (Section 3.3), which conditions on the observation mask and the time gap via feature-wise linear modulation (FiLM) to attenuate the hidden state of synthetic derivatives during unobserved intervals. The resulting multi-scale latent states are fused and decoded to produce forecasts $\hat{\mathbf{Y}} = \{\hat{\mathbf{x}}(T + \tau_k)\}_{k=1}^{K}$ at arbitrary future timestamps (Section 3.4).

### 3.1 FREQUENCY-CONDITIONED WAVELET FRAME DENOISING

Standard interpolation schemes construct continuous paths across unobserved intervals, inadvertently amplifying observation noise and propagating artificial fluctuations into the path derivatives $\dot{X}(t)$. In NCDEs, where latent states are driven by infinitesimal path increments $\mathrm{d}X(t)$, such perturbations act as spurious forcing signals that destabilize numerical integration and lead to overfitting (Oh et al., 2025). To resolve this, we introduce the **Frequency-Conditioned Wavelet Frame Denoising (FCWFD)** module, which eliminates interpolation-induced artifacts in the frequency domain before continuous path construction.

**Wavelet Frame Decomposition.** Rather than relying on standard orthogonal wavelets, which are sensitive to shifts and boundary perturbations, we employ a *tight wavelet frame system* in $\mathbf{L}^2(\mathbb{R}^D)$ generated by a set of framelets $\Psi = \{\psi_1, \ldots, \psi_L\}$ (Leonardi & Van De Ville, 2013). Under a discrete quasi-affine system $\mathcal{G}_J(\Psi)$ (Dong et al., 2017), the frame satisfies the tight frame identity $\sum_{g \in \mathcal{G}_J} |\langle f, g \rangle|^2 = \|f\|^2$, guaranteeing energy preservation and shift-invariance (detailed frame constructions are provided in Appendix A). Given the initial discrete observations, a $J$-level quasi-affine wavelet frame transform decomposes the trajectory into $J+1$ multi-resolution coefficient sub-

bands $\{c^{(j)}\}_{j=0}^{J}$, where $c^{(0)} \in \mathbb{R}^{D\times T_0}$ denotes the coarse approximation trend and $c^{(j)} \in \mathbb{R}^{D\times T_j}$ ($j = 1, \ldots, J$) represent fine-grained detail components across dyadic scales. Unlike non-redundant orthogonal wavelets where a localized error can corrupt the entire scale, the overcomplete (redundant) frame distributes signal energy across overlapping atoms. This structural redundancy acts as an error buffer, preventing sparse observation noise from dominating local path derivatives.

**Frequency-Conditioned Thresholding.** Unlike conventional filters that uniformly attenuate high frequencies, FCWFD adaptively shrinks noise while preserving sharp temporal changes that carry important information. For each detail subband $j \in \{1, \ldots, J\}$, we compute a scale and channel-dependent threshold matrix $\Lambda^{(j)} \in \mathbb{R}^{D\times T_j}$:

$$\begin{aligned} \tilde{e}_j &= g_{\theta^j}(|H_j|, \tilde{L}_j), \quad \tilde{L}_j := \mathcal{R}_j(|L_J|), \\ \Lambda^{(j)} &= \mathrm{softplus}(B_j + \alpha \tilde{e}_j). \end{aligned} \tag{1}$$

Here, $\tilde{L}_j$ is derived from the lowest-frequency global trend, specifically $|L_J| = |c^{(0)}|$, which is resampled to length $T_j$ by the resizing function $\mathcal{R}_j(\cdot)$. The term $\tilde{e}_j$ is then computed by $g_{\theta^j}$, a conditional network that takes the local subband energy $|H_j|$ and the resampled global trend $\tilde{L}_j$ as inputs. Finally, $\Lambda^{(j)}$ is determined by adding this data-dependent term $\alpha\tilde{e}_j$ to a learnable base threshold $B_j \in \mathbb{R}^{D\times T_j}$, where $\alpha > 0$ controls the influence of this conditioning. The denoised coefficients $\tilde{c}^{(j)}$ are then obtained via soft shrinkage:

$$\tilde{c}^{(j)} = \mathrm{sign}\big(c^{(j)}\big) \odot \max\Big(|c^{(j)}| - \Lambda^{(j)}, 0\Big), \quad j = 1, \ldots, J, \tag{2}$$

while setting $\Lambda^{(0)} = \mathbf{0}$ ($\tilde{c}^{(0)} = c^{(0)}$) to preserve global evolutionary trends.

**Continuous Control Path Construction.** After frequency-aware denoising, we fit continuous cubic Hermite splines with bounded variation onto each denoised subband $\tilde{c}^{(j)}$ ($j = 0, \ldots, J$), producing a collection of denoised, continuously differentiable control paths $\{\tilde{X}^{(j)}(t)\}_{j=0}^{J}$.

> **Remark 1 (Preventive vs. Post-hoc Denoising).** *FCWFD fundamentally diverges from standard input preprocessing. In NCDEs, trajectories evolve as* $\mathrm{d}z(t) = f_\theta(z(t))\,\mathrm{d}X(t)$. *Errors in* $\tilde{X}(t)$ *propagate directly to the path velocity* $\dot{X}(t)$*, distorting the driving vector field throughout integration. By regularizing coefficients in the frame domain prior to continuous spline interpolation, FCWFD ensures bounded-variation driving paths with well-behaved derivatives, suppressing spurious oscillations while isolating multi-scale dynamical regimes.*

### 3.2 Multi-Scale Neural Controlled Differential Equations

Rather than collapsing multi-resolution signals into a single interpolated path, we propagate the frequency decomposition from FCWFD directly into the continuous latent space. Concretely, each denoised component $\tilde{X}^{(j)}(t)$ is augmented with time to yield the driving control path $X^{(j)}(t) = [\tilde{X}^{(j)}(t), t] \in \mathbb{R}^{D+1}$ for $j \in \{0, \ldots, J\}$.

**Scale-Specific Continuous Evolution.** For each scale $j$, we define a dedicated continuous latent state $z^{(j)}(t) \in \mathbb{R}^{d_h}$ governed by its own vector field $f_{\theta_j}^{(j)} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h\times(D+1)}$:

$$z^{(j)}(t) = z^{(j)}(t_0) + \int_{t_0}^{t} f_{\theta_j}^{(j)}\big(z^{(j)}(s)\big)\,\mathrm{d}X^{(j)}(s), \quad z^{(j)}(t_0) = \zeta_j\big(X^{(j)}(t_0)\big), \tag{3}$$

where $\zeta_j$ denotes a scale-specific initialization network. Equivalently, the joint latent state $Z(t) = [z^{(0)}(t), \ldots, z^{(J)}(t)] \in \mathbb{R}^{(J+1)d_h}$ evolves through a compact block-diagonal vector field:

$$\mathrm{d}Z(t) = \mathcal{F}_\Theta\big(Z(t)\big)\,\mathrm{d}X(t), \quad \mathcal{F}_\Theta(Z) = \mathrm{BlockDiag}\Big(f_{\theta_0}^{(0)}\big(z^{(0)}\big), \ldots, f_{\theta_J}^{(J)}\big(z^{(J)}\big)\Big), \tag{4}$$

where $X(t) = [X^{(0)}(t), \ldots, X^{(J)}(t)] \in \mathbb{R}^{(J+1)(D+1)}$. Crucially, this block-diagonal formulation prevents premature cross-scale interference during continuous integration, ensuring that the approximation branch exclusively governs coarse macro trends while detail branches independently capture fine-grained fluctuations. The *regularity and well-posedness* of each scale-specific NCDE are provided in Appendix C.

**Cross-Scale Latent Aggregation.** Upon integrating each scale-specific branch up to the observation horizon $T$, we introduce cross-scale interaction to fuse the disentangled temporal regimes. The terminal states $Z_T = [z^{(0)}(T), \dots, z^{(J)}(T)] \in \mathbb{R}^{(J+1)d_h}$ are aggregated into a unified historical representation $\bar{z}(T) \in \mathbb{R}^{d_h}$ via:

$$\bar{z}(T) = W_{\text{agg}} Z_T + b_{\text{agg}}, \tag{5}$$

where $W_{\text{agg}} \in \mathbb{R}^{d_h \times (J+1)d_h}$ and $b_{\text{agg}} \in \mathbb{R}^{d_h}$ act as a $1 \times 1$ convolution across the scale dimension.

### 3.3 Missingness-Aware Dynamics Function

While multi-scale NCDE assigns scale-dedicated vector fields to the denoised paths, standard NCDEs assume uniform driving reliability. Over prolonged unobserved intervals, reconstructed path derivatives $\dot{X}^{(j)}(t)$ reflect interpolation priors rather than authentic system dynamics. Thus, we propose the *Missingness-Aware Dynamics Function* (MaDF), which continuously modulates the latent vector field's sensitivity to driving increments based on observation uncertainty:

$$z^{(j)}(t) = z^{(j)}(t_0) + \int_{t_0}^{t} \mathcal{F}_{\theta_j}^{(j)}\big(z^{(j)}(s),\, \eta(s)\big)\, \mathrm{d}X^{(j)}(s), \quad j = 0, \dots, J, \tag{6}$$

where $\eta(t) \in \mathbb{R}^{d_m}$ is a continuous observation-history embedding. Here, $\mathrm{d}X^{(j)}(t)$ strictly supplies the driving increments, while $\eta(t)$ regulates how the dynamical system responds to them.

**Observation-History Descriptors.** For channel $d \in \{1, \dots, D\}$, let $\delta_d(t)$ denote the elapsed time since its most recent observation, and $s_d(t) \in \{0, 1\}$ indicate whether channel $d$ has been observed by time $t$. Normalized by a characteristic timescale $h_d > 0$ (estimated from training statistics), we construct two complementary descriptors: a log-compressed interval $\tau_d^{\text{gap}}(t)$ and an exponentially decaying recency score $q_d(t)$:

$$\tau_d^{\text{gap}}(t) = \log\left(1 + \frac{\delta_d(t)}{h_d + \epsilon}\right), \quad q_d(t) = s_d(t) \exp\left(-\frac{\delta_d(t)}{h_d + \epsilon}\right), \tag{7}$$

where $\tau_d^{\text{gap}}(t)$ captures the magnitude of temporal gaps, while $q_d(t) \in [0, 1]$ bounds observation reliability. Concatenating these over all channels yields $M(t) = [\tau_1^{\text{gap}}(t), \dots, \tau_D^{\text{gap}}(t),\, q_1(t), \dots, q_D(t)]^\top \in \mathbb{R}^{2D}$. We interpolate $M(t)$ into a continuous descriptor path and map it via a shared encoder to obtain the embedding $\eta(t) = E_m(M(t)) \in \mathbb{R}^{d_m}$.

**FiLM-Modulated Dynamics.** We inject $\eta(t)$ into each scale-specific vector field via Feature-wise Linear Modulation (FiLM). For scale $j$, a dedicated modulation network $G_j$ generates scaling and shifting parameters $(\boldsymbol{\Gamma}_j(t), \boldsymbol{B}_j(t)) = G_j(\eta(t)) \in \mathbb{R}^{r_j}$. Given intermediate hidden features $h_j(t) = H_j(z^{(j)}(t))$, the modulated vector field is parameterized as:

$$\mathcal{F}_\theta^{(j)}\big(z^{(j)}(t),\, \eta(t)\big) = \mathrm{mat}\left(P_j\big(\boldsymbol{\Gamma}_j(t) \odot h_j(t) + \boldsymbol{B}_j(t)\big)\right) \in \mathbb{R}^{d_h \times (D+1)}, \tag{8}$$

where $P_j$ is a linear projection and $\mathrm{mat}(\cdot)$ reshapes features into matrix form. Descriptor construction and network details are provided in Appendix B.

By decoupling driving increments $\mathrm{d}X^{(j)}(t)$ from FiLM-modulated responsiveness $\mathcal{F}_{\theta_j}^{(j)}$, MaDF dynamically attenuates unreliable synthetic derivatives during prolonged gaps without violating the standard CDE formulation. Moreover, because $G_j$ is scale-specific, the model adaptively differentiates the impact of missingness across slowly varying macro trends and rapid local fluctuations.

### 3.4 Unified Latent Rollout and Forecasting Objective

Once the denoised and scale-specific control paths are obtained, the resulting scale-specific terminal states are then aggregated according to Eq.(5) to obtain a unified historical representation, i.e., $\bar{z}(T) \in \mathbb{R}^{d_h}$. This representation serves as the boundary condition for future prediction.

**Autonomous latent rollout** Beyond the observation boundary $T$, the future control paths are unavailable by definition. Consequently, extending the historical CDE beyond $T$ would require extrapolating the observed control paths and could introduce additional interpolation or extrapolation artifacts. Instead, we transition from data-driven controlled dynamics to an autonomous latent dynamical system. Denoted by $u(t) \in \mathbb{R}^{d_h},\ t \in [T, T+H]$, the future latent trajectory initialized at $u(T) = \bar{z}(T)$. We model its continuous-time evolution through an autonomous neural ODE,

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = \mathcal{F}_\theta\big(u(t)\big), \qquad u(T) = \bar{z}(T),$$

where $\mathcal{F}_\theta(\cdot)$ is obtained from Eq. (8). The resulting trajectory satisfies $u(t) = \bar{z}(T) + \int_T^t \mathcal{F}_\theta\left(u(s)\right) \mathrm{d}s,\ t \in [T, T+H]$. This formulation cleanly separates historical data assimilation from future latent evolution. Therefore, all information extracted from irregular and partially observed historical data is compressed into the boundary representation $\bar{z}(T)$, after which the model performs forecasting without requiring future interpolation or artificial observation paths.

**Arbitrary-horizon prediction** Let $\{T+\tau_k\}_{k=1}^K,\ 0 < \tau_1 < \cdots < \tau_K \le H$, be an arbitrary set of future query timestamps. Since $u(t)$ is defined continuously, then predictions can be evaluated directly at given timestamps via

$$\hat{\mathbf{x}}(T+\tau_k) = \pi_\eta(u(T+\tau_k))\,, \qquad k = 1, \ldots, K,$$

where $\pi_\eta : \mathbb{R}^{d_h} \to \mathbb{R}^D$ is a learnable decoder. Hence, the forecasting resolution is determined by the query set, rather than by the architecture or by a predefined discretization interval. The same learned latent trajectory can therefore be queried at regular or non-uniform future timestamps.

**Masked forecasting objective** Let $\mathbf{x}(T+\tau_k) \in \mathbb{R}^D$ be the ground-truth future observation and let $\mathbf{m}(T+\tau_k) \in \{0,1\}^D$ be its corresponding observation mask. To avoid assigning loss to unavailable targets, we define the masked forecasting loss as:

$$\mathcal{L}_{\text{forecast}} = \frac{\displaystyle\sum_{k=1}^{K} \|\mathbf{m}(T+\tau_k) \odot (\hat{\mathbf{x}}(T+\tau_k) - \mathbf{x}(T+\tau_k))\|_2^2}{\displaystyle\sum_{k=1}^{K} \|\mathbf{m}(T+\tau_k)\|_1 + \epsilon}, \tag{9}$$

where $\epsilon > 0$ prevents numerical instability when the number of observed future entries is small. This objective evaluates the model only on valid future measurements while normalizing by the number of available target entries, ensuring that the contribution of each forecasting sample is not dominated by differences in the amount of future observations.

**End-to-end optimization** All components of the proposed framework are trained jointly through the forecasting objective in Eq.(9). Let $\Delta_{\text{all}}$ be the collection of trainable parameters, including the FCWFD gating networks, scale-specific MaDF vector fields, cross-scale aggregation parameters, autonomous rollout dynamics, and decoder parameters. The model is optimized as:

$$\Delta_{\text{all}}^* = \arg\min_{\Delta_{\text{all}}} \mathcal{L}_{\text{forecast}}. \tag{10}$$

Gradients are propagated end-to-end through both the historical controlled integration and the future latent ODE, using either adjoint sensitivity methods or backpropagation. This architectural separation enables the model to produce forecasts at arbitrary continuous query times without requiring future interpolation, cleanly decoupling past data assimilation from future latent dynamics.

# 4 Experiments

## 4.1 Experimental Setup

**Baselines.** To conduct a comprehensive and rigorous evaluation, we benchmark against 18 representative baselines spanning two major paradigms based on their continuous latent-state formulation:

- **ODE-based models:** Latent-ODE (Rubanova et al., 2019), ODE-RNN (Rubanova et al., 2019), ODE-GRU (Rubanova et al., 2019), ODE-LSTM (Lechner & Hasani, 2020), GRU-ODE-Bayes (De Brouwer et al., 2019), Augmented-ODE (Dupont et al., 2019), ACE-ODE (Jhin et al., 2021a), and DeNOTS (Kuleshov et al., 2026).
- **CDE-based models:** NCDE (Kidger et al., 2020), Log-NCDE (Walker et al., 2024), AN-CDE (Jhin et al., 2021b), TE-NCDE (Seedat et al., 2022), BNCDE (Hess et al., 2024), DualDynamics (Oh et al., 2025), NRDE (Morrill et al., 2021), LEAP (Jhin et al., 2023), CTRL (Liu et al., 2024), and Diff-MN (Zhang et al., 2026).

Baseline hyperparameters follow publicly available implementations in the original papers. Detailed configurations and hyperparameter analyses of our model are provided in Appendix F.

**Datasets.** We employ five real-world irregular multivariate time series datasets, covering climate, biomechanics, and healthcare. These datasets exhibit diverse characteristics, including different numbers of variables, sampling frequencies, and missingness rates.

- **USHCN** (Menne et al., 2016) is a long-term climate dataset with 5 variables, containing 1, 114 US weather stations over the four-year period from 1996 to 2000.
- **Human Activity** (Vidulin et al., 2010) with 12 time-varying variables includes biomechanics data from five individuals performing physical activity, which are recorded in millisecond level.
- **The PhysioNet 2012** (**PhysioNet' 12**) (Silva et al., 2012) contains clinical records with 41 varaibles from 12,000 adult ICU stays extracted from MIMIC-II, which are collected during first 48 hours after ICU admission.
- **MIMIC-III v1.4** (Johnson et al., 2016) with 96 variables contains data from more than 40, 000 ICU patients between 2001 and 2012.
- **MIMIC-IV v3.1** (Johnson et al., 2024) with 107 variables, built on MIMIC-III, consists of data for over 65, 000 ICU patients and over 200, 000 emergency department patients, and generally provides finer-grained temporal records than MIMIC-III with minute-level timestamps.

All datasets are chronologically split into training, validation, and test sets with an 8:1:1 ratio. Full preprocessing and implementation details are deferred to Appendix D and E.

### 4.2 Main Results

Table 1 reports the forecasting performance across all five benchmarks. ResoPath achieves the lowest overall average rank of 2.4, substantially outperforming prior leading baselines including CTRL (4.6) and LEAP (5.1), and establishes the best results on 7 of the 10 dataset–metric configurations.

Specifically, on the high-frequency Human Activity benchmark, ResoPath yields significant error reductions, cutting MSE by 7.78% relative to LEAP and MAE by 8.26% relative to CTRL. Across the clinical benchmarks (MIMIC-III, MIMIC-IV, and PhysioNet'12), ResoPath consistently sets new state-of-the-art results, steadily outperforming competitive continuous-time models like TE-CDE and CTRL while surpassing vanilla NCDE by a pronounced margin. These consistent gains empirically substantiate the benefit of our multi-scale trajectory disentanglement and missingness-aware vector field modulation in handling sparse, irregularly observed trajectories.

In contrast, on the climate dataset USHCN, no individual continuous-time model establishes a decisive margin, as all baselines exhibit substantially elevated variance in performance. Notably, generative alternatives such as Diff-MN perform well on low-dimensional climate series but degrade severely when scaling to high-dimensional clinical trajectories (e.g., MIMIC-III and MIMIC-IV). ResoPath, by contrast, retains exceptional predictive stability across complex multivariate settings, striking a superior balance between expressiveness and generalization.

### 4.3 Ablation Study

ResoPath is composed of three modular components for IMTS representation learning: FCWFD, MS-NCDE, and MaDFs. Since these components are non-sequential and each targets a distinct

Table 1: Mean ± STD for MSE and MAE, reported in units of $10^{-1}$. The best result is in **bold**, and the runner-up is underlined. Ranks use unrounded means (lower is better), with lower STD breaking exact mean ties. Average rank is computed over the ten dataset–metric pairs.

| Dataset | USHCN | | Human Activity | | PhysioNet'12 | | MIMIC-III | | MIMIC-IV | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | MSE×$10^{-1}$ | MAE×$10^{-1}$ | MSE×$10^{-1}$ | MAE×$10^{-1}$ | MSE×$10^{-1}$ | MAE×$10^{-1}$ | MSE×$10^{-1}$ | MAE×$10^{-1}$ | MSE×$10^{-1}$ | MAE×$10^{-1}$ | Rank |
| latent-ODE | 3.35 ± 1.09 | 3.21 ± 0.16 | 2.08 ± 0.99 | 3.03 ± 0.81 | 3.78 ± 0.04 | 4.23 ± 0.02 | 6.72 ± 0.11 | 5.37 ± 0.04 | 5.03 ± 0.27 | 4.89 ± 0.14 | 9.3 |
| ODE-RNN | 3.39 ± 0.96 | 3.71 ± 0.10 | 2.50 ± 0.67 | 4.11 ± 0.47 | 3.60 ± 0.05 | 5.03 ± 0.01 | 6.24 ± 0.95 | 5.75 ± 0.02 | 5.39 ± 0.43 | 5.22 ± 0.08 | 11.7 |
| ODE-GRU | 3.57 ± 4.19 | 3.86 ± 0.36 | 2.37 ± 0.24 | 3.93 ± 0.15 | 5.08 ± 0.26 | 5.58 ± 0.08 | 6.94 ± 0.49 | 5.52 ± 0.71 | 5.20 ± 0.76 | 4.71 ± 0.28 | 13.0 |
| ODE-LSTM | 3.59 ± 8.65 | 4.48 ± 0.55 | 2.68 ± 1.37 | 2.87 ± 0.68 | 3.81 ± 0.65 | 4.76 ± 0.73 | 7.13 ± 4.21 | 5.84 ± 0.10 | 6.71 ± 0.94 | 5.01 ± 0.46 | 13.6 |
| GRU-ODE-Bayes | 4.27 ± 3.85 | 3.15 ± 0.04 | 2.14 ± 0.52 | 2.60 ± 0.35 | 4.11 ± 0.04 | 4.37 ± 0.07 | 6.57 ± 0.03 | 4.97 ± 0.03 | 4.70 ± 0.24 | 4.54 ± 0.08 | 9.5 |
| Augmented-ODE | **1.83 ± 0.04** | **2.92 ± 0.08** | 0.98 ± 0.03 | 2.31 ± 0.07 | 3.63 ± 0.05 | 4.23 ± 0.04 | 7.53 ± 0.28 | 5.62 ± 0.08 | 7.73 ± 0.54 | 6.20 ± 0.46 | 8.7 |
| ACE-ODE | 2.32 ± 0.18 | 2.97 ± 0.06 | 1.12 ± 0.11 | 2.48 ± 0.14 | 3.76 ± 0.03 | 4.22 ± 0.02 | 6.99 ± 0.18 | 5.49 ± 0.08 | 5.12 ± 0.37 | 4.96 ± 0.10 | 7.9 |
| DeNOTS | 4.73 ± 0.71 | 6.37 ± 0.61 | 1.15 ± 0.07 | 2.55 ± 0.09 | 4.67 ± 0.03 | 5.38 ± 0.02 | 5.71 ± 0.11 | 5.17 ± 0.09 | 3.93 ± 0.92 | 4.83 ± 0.02 | 10.9 |
| NCDE | 3.07 ± 0.78 | 3.16 ± 0.77 | 1.63 ± 0.74 | 2.83 ± 0.21 | 4.11 ± 0.06 | 4.24 ± 0.03 | 5.77 ± 0.85 | 5.05 ± 0.06 | 4.18 ± 0.78 | 4.48 ± 0.35 | 7.6 |
| Log-NCDE | 3.06 ± 0.33 | 3.70 ± 0.29 | 1.30 ± 0.12 | 2.66 ± 0.13 | 4.20 ± 0.02 | 4.55 ± 0.01 | 5.91 ± 0.03 | 4.95 ± 0.04 | 4.13 ± 0.02 | 4.37 ± 0.02 | 8.0 |
| ANCDE | 3.48 ± 0.72 | 4.30 ± 0.74 | 4.70 ± 1.56 | 4.89 ± 0.89 | 6.54 ± 1.34 | 6.01 ± 0.74 | 9.74 ± 0.03 | 6.57 ± 0.02 | 9.85 ± 0.31 | 7.38 ± 0.17 | 17.4 |
| NRDE | 4.02 ± 0.11 | 3.41 ± 0.19 | 2.50 ± 0.57 | 3.71 ± 0.39 | 4.29 ± 0.09 | 4.63 ± 0.08 | 5.89 ± 0.04 | 4.91 ± 0.01 | 4.20 ± 0.05 | 4.40 ± 0.04 | 10.2 |
| LEAP | 3.89 ± 2.30 | 3.71 ± 0.76 | 0.90 ± 0.05 | 2.20 ± 0.09 | 3.62 ± 0.03 | 4.19 ± 0.01 | 5.58 ± 0.28 | 4.79 ± 0.15 | 3.92 ± 0.46 | 4.27 ± 0.33 | 5.1 |
| CTRL | 5.34 ± 1.78 | 3.35 ± 0.10 | 0.96 ± 0.06 | 2.18 ± 0.08 | 3.60 ± 0.02 | **4.01 ± 0.02** | 5.49 ± 0.04 | 4.42 ± 0.00 | 3.96 ± 0.01 | 4.09 ± 0.01 | 4.6 |
| TE-CDE | 4.68 ± 2.34 | 5.10 ± 0.99 | 1.11 ± 0.06 | 2.47 ± 0.09 | 3.82 ± 0.08 | 4.31 ± 0.05 | 5.42 ± 0.03 | 4.61 ± 0.02 | 3.79 ± 0.03 | 4.06 ± 0.02 | 7.3 |
| BNCDE | 3.17 ± 0.51 | 3.96 ± 0.33 | 3.41 ± 0.26 | 4.31 ± 0.14 | 5.83 ± 0.11 | 5.64 ± 0.08 | 9.47 ± 0.17 | 6.33 ± 0.12 | 7.82 ± 0.17 | 6.21 ± 0.11 | 16.2 |
| DualDynamics | 15.86 ± 2.09 | 8.75 ± 0.87 | 2.88 ± 0.35 | 4.14 ± 0.29 | 5.51 ± 0.34 | 5.48 ± 0.21 | 6.72 ± 0.22 | 5.36 ± 0.08 | 5.70 ± 0.26 | 5.32 ± 0.11 | 15.5 |
| Diff-MN | 4.20 ± 8.20 | 4.93 ± 0.51 | 0.97 ± 0.02 | 2.19 ± 0.04 | 3.69 ± 0.01 | 4.15 ± 0.02 | 7.60 ± 1.14 | 5.63 ± 0.48 | 7.80 ± 0.68 | 6.16 ± 0.12 | 11.1 |
| **ResoPath** | 3.48 ± 0.74 | 3.24 ± 0.23 | **0.83 ± 0.05** | **2.00 ± 0.07** | **3.60 ± 0.00** | 4.14 ± 0.01 | **5.36 ± 0.05** | **4.33 ± 0.03** | **3.73 ± 0.20** | **4.01 ± 0.15** | **2.4** |

Table 2: Ablation results of ResoPath on five datasets evaluated by MSE ± STD $\times 10^{-1}$.

| Method | USHCN | Human Activity | PhysioNet'12 | MIMIC-III | MIMIC-IV |
|---|---|---|---|---|---|
| **ResoPath** | $\mathbf{3.48 \pm 0.74}$ | $\mathbf{0.83 \pm 0.05}$ | $\mathbf{3.59 \pm 0.00}$ | $\mathbf{5.36 \pm 0.05}$ | $\mathbf{3.73 \pm 0.20}$ |
| w/o FCWFD | $4.17 \pm 0.55$ | $1.06 \pm 0.02$ | $3.84 \pm 0.03$ | $5.77 \pm 0.02$ | $3.81 \pm 0.01$ |
| rp FCWFD | $4.05 \pm 0.61$ | $0.92 \pm 0.02$ | $3.80 \pm 0.02$ | $5.61 \pm 0.00$ | $3.75 \pm 0.02$ |
| rp$^{(1)}$ MS-NCDE | $4.69 \pm 0.53$ | $1.73 \pm 0.03$ | $4.89 \pm 0.04$ | $6.34 \pm 0.04$ | $4.72 \pm 0.05$ |
| rp$^{(2)}$ MS-NCDE | $3.94 \pm 0.82$ | $1.26 \pm 0.03$ | $4.17 \pm 0.05$ | $5.58 \pm 0.03$ | $3.84 \pm 0.03$ |
| rp MaDFs | $4.69 \pm 0.39$ | $1.32 \pm 0.04$ | $4.27 \pm 0.05$ | $5.81 \pm 0.08$ | $4.40 \pm 0.04$ |

challenge, namely interpolation noise, scale entanglement, and missingness, respectively, we design five variants to isolate their individual contributions. Table 2 reports the results on five public IMTS datasets. **w/o FCWFD** removes the FCWFD module entirely; **rp FCWFD** replaces its learnable thresholding with a non-learnable hard threshold; **rp$^{(1)}$ MS-NCDE** collapses the joint multi-scale system into a single-layer NCDE, while **rp$^{(2)}$ MS-NCDE** instead replaces the scale-specific vector fields with a simple mixture-of-experts (MoE) parameterization; and **rp MaDFs** substitutes the FiLM-based missingness conditioning with a plain MLP.

Removing or weakening **FCWFD** consistently degrades performance across all five datasets, e.g., raising MSE from 3.48 to 4.17 on USHCN and from 0.83 to 1.06 on Human Activity, confirming both the necessity of frequency-conditioned denoising and the benefit of its learnable thresholding over a fixed rule. The two MS-NCDE variants incur even larger losses, most notably **rp$^{(1)}$ MS-NCDE** (MSE from 0.83 to 1.73 on Human Activity and from 5.36 to 6.34 on MIMIC-III), showing that modeling multiple scales jointly, rather than collapsing them into a single path or routing them through an undifferentiated MoE, is essential for expressiveness and for mitigating interpolation-induced monotonic bias. Replacing FiLM with a plain MLP in **rp MaDFs** produces comparable degradation (e.g., MSE from 3.73 to 4.40 on MIMIC-IV), confirming that multiplicative modulation is markedly more effective than naive concatenation at correcting missingness-related artifacts.

Taken together, these results confirm that all three components contribute complementary and largely non-substitutable improvements to ResoPath's overall performance.

### 4.4 Efficiency and Solver Sensitivity Analysis

**Evaluation Protocol.** To assess computational efficiency, we measure the number of function evaluations (NFE) alongside forward inference latency across continuous-time models. NFE serves as a direct, solver-aware proxy for the dynamical complexity of the parameterized vector fields. For a rigorous comparison, all continuous baselines are evaluated on identical hardware, batch sizes, and integration schemes, testing both adaptive and fixed-step solvers across Human Activity and PhysioNet'12 (Figure 3).

**Adaptive Solver Dynamics (Dopri5).** As shown in Figures 3(a) and 3(c), tightening the error tolerance (`rtol`/`atol`) generally escalates NFE and computational runtime across all models. ResoPath consistently establishes an optimal Pareto frontier: it maintains the lowest prediction error across the entire tolerance spectrum while incurring significantly lower latency than competitive continuous models such as CTRL and NCDE. Crucially, ResoPath exhibits minimal error degradation even at coarse tolerance thresholds ($10^{-2}/10^{-4}$), demonstrating that its superior efficiency stems from inherently well-conditioned latent trajectories rather than solver-specific tuning.

**Robustness to Fixed-Step Discretization.** Under fixed-step solvers (Figures 3(b) and 3(d)), ResoPath retains remarkable numerical stability. Switching from the higher-order RK4 solver to the lower-cost Midpoint scheme introduces negligible predictive degradation for ResoPath, whereas baseline models often suffer noticeable performance drops. This confirms that our multi-scale vector-field formulation successfully mitigates stiff dynamics, enabling resource-efficient inference with coarser numerical integration.

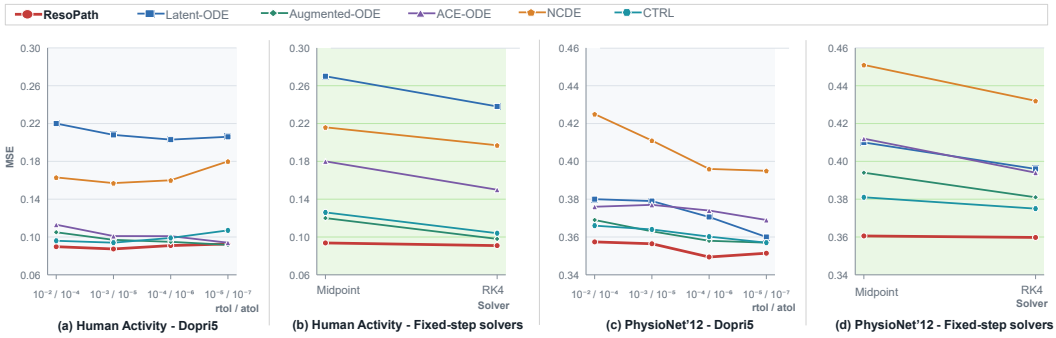



Figure 3: NFE analysis of ResoPath and other continuous-time models. Forecasting error (MSE) is plotted against solvers and their configurations for each model. Marker size indicates inference time.

## 5 Conclusion

This paper introduces ResoPath, a continuous-time framework designed for IMTS forecasting, which resolves the limitations of interpolation-based control paths in existing NCDEs. ResoPath achieves this by disentangling coarse temporal trends and denoising fine-scale fluctuations through multi-scale continuous paths and missingness-aware dynamics. This innovative approach effectively mitigates interpolation-induced biases and enhances expressiveness, demonstrating strong forecasting performance and lower NFE cost across five diverse IMTS benchmarks. However, while ResoPath exhibits robustness to varying NFE, its reliance on numerical integration still incurs substantial runtime overhead compared to compact forecasting architectures. These computational considerations highlight the need for future research into more efficient numerical solvers to complement advancements in forecasting accuracy.

### AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES


Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural ODEs. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b21f9f98829dea9a48fd8aaddc1f159d-Abstract.html.

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

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*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/455cb2657aaa59e32fad80cb0b65b9dc-Abstract.html.

Bin Dong, Zuowei Shen, and Peichu Xie. Image restoration: A general wavelet frame based model and its asymptotic analysis. *SIAM Journal on Mathematical Analysis*, 49(1):421–445, 2017.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/21be9a4bd4f81549a9d1d241981cec3c-Abstract.html.

David Freedman and Persi Diaconis. On the histogram as a density estimator: $L_2$ theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57:453–476, 1981. doi: 10.1007/BF01025868. URL https://doi.org/10.1007/BF01025868.

Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uwO71a8wET.

Sheo Yon Jhin, Minju Jo, Taeyong Kong, Jinsung Jeon, and Noseong Park. ACE-NODE: Attentive co-evolving neural ordinary differential equations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 736–745, 2021a. doi: 10.1145/3447548.3467419. URL https://doi.org/10.1145/3447548.3467419.

Sheo Yon Jhin, Heejoo Shin, Seoyoung Hong, Minju Jo, Solhee Park, Noseong Park, Seungbeom Lee, Hwiyoung Maeng, and Seungmin Jeon. Attentive neural controlled differential equations for time-series classification and forecasting. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 250–259, 2021b. doi: 10.1109/ICDM51629.2021.00035. URL https://doi.org/10.1109/ICDM51629.2021.00035.

Sheo Yon Jhin, Minju Jo, Seungji Kook, and Noseong Park. Learnable path in neural controlled differential equations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8014–8022, 2023. doi: 10.1609/aaai.v37i7.25969. URL https://ojs.aaai.org/index.php/AAAI/article/view/25969.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III clinical database. *PhysioNet*, 2016. doi: 10.13026/C2XW26. URL https://physionet.org/content/mimiciii/1.4/. Version 1.4.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, 2024. doi: 10.13026/kpb9-mt58. URL https://physionet.org/content/mimiciv/3.1/. Version 3.1.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Ilya Kuleshov, Evgenia Romanenkova, Vladislav Zhuzhel, Galina Boeva, Evgeni Vorsin, and Alexey Zaytsev. DeNOTS: Stable deep neural ODEs for time series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SFoDJZ1sSk.

Mathias Lechner and Ramin Hasani. Learning long-term dependencies in irregularly-sampled time series. *arXiv preprint arXiv:2006.04418*, 2020. doi: 10.48550/arXiv.2006.04418. URL https://arxiv.org/abs/2006.04418.

Nora Leonardi and Dimitri Van De Ville. Tight wavelet frames on multislice graphs. *IEEE Transactions on Signal Processing*, 61(13):3357–3367, 2013.

Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. CrossAD: Time series anomaly detection with cross-scale associations and cross-window modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4605. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c9658e8c20879632cb1cfca91d80ceb7-Abstract-Conference.html.

Boyuan Li, Zhen Liu, Yicheng Luo, and Qianli Ma. Learning recursive multi-scale representations for irregular multivariate time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JEIDxiTWzB.

Jiexi Liu, Meng Cao, and Songcan Chen. TimeCHEAT: A channel harmony strategy for irregularly sampled multivariate time series analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18861–18869, 2025. doi: 10.1609/aaai.v39i18.34076. URL https://ojs.aaai.org/index.php/AAAI/article/view/34076.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23873–23881, 2026. doi: 10.1609/aaai.v40i28.39563. URL https://ojs.aaai.org/index.php/AAAI/article/view/39563.

Zihan Liu, Bowen Du, Junchen Ye, Xianqing Wen, and Leilei Sun. An NCDE-based framework for universal representation learning of time series. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4623–4633, 2024. doi: 10.24963/ijcai.2024/511. URL https://www.ijcai.org/proceedings/2024/0511.

Yicheng Luo, Bowen Zhang, Zhen Liu, and Qianli Ma. Hi-Patch: Hierarchical patch GNN for irregular multivariate time series. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41494–41519, 2025. URL https://proceedings.mlr.press/v267/luo25r.html.

M. Menne, C. Williams, and R. Vose. Long-term daily and monthly climate records from stations across the contiguous United States (U.S. Historical Climatology Network) (NDP-019). Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, 2016. URL https://doi.org/10.3334/CDIAC/CLI.NDP019.

James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7829–7838, 2021. URL https://proceedings.mlr.press/v139/morrill21b.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

YongKyung Oh, Dong-Young Lim, and Sungil Kim. DualDynamics: Synergizing implicit and explicit methods for robust irregular time series analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19730–19739, 2025. doi: 10.1609/aaai.v39i18.34173. URL https://ojs.aaai.org/index.php/AAAI/article/view/34173.

Xiangfei Qiu, Kangjia Yan, Xvyuan Liu, Xingjian Wu, and Jilin Hu. Bridging time and frequency: A joint modeling framework for irregular multivariate time series forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.00582.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5321–5331, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/42a6845a557bef704ad8ac9cb4461d43-Abstract.html.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19497–19521, 2022. URL https://proceedings.mlr.press/v162/seedat22b.html.

Ikaro Silva, George Moody, Daniel J. Scott, Leo A. Celi, and Roger G. Mark. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology challenge 2012. In *Computing in Cardiology*, volume 39, pp. 245–248, 2012. URL https://physionet.org/files/challenge-2012/1.0.0/papers/0245.pdf.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012. URL https://www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf.

Vedrana Vidulin, Mitja Luštrek, Boštjan Kaluža, Rok Piltaver, and Jana Krivec. Localization data for person activity. UCI Machine Learning Repository, 2010. URL https://doi.org/10.24432/C57G8X.

Benjamin Walker, Andrew Donald Mcleod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The lie brackets make a difference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49822–49844, 2024. URL https://proceedings.mlr.press/v235/walker24a.html.

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

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

Xu Zhang, Junwei Deng, Chang Xu, Hao Li, and Jiang Bian. Diff-MN: Diffusion parameterized MoE-NCDE for continuous time series generation with irregular observations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.13534.

## A Preliminaries on Wavelet Frame Systems

In this section, we briefly introduce the concept of wavelet frames. A set $\mathcal{G} = \{g_l : l \in \mathbb{Z}\} \subset \mathbf{L}_2(\mathbb{R}^D)$ is called a frame of $\mathbf{L}_2(\mathbb{R}^D)$ if

$$A\|f\|^2_{\mathbf{L}_2(\mathbb{R}^D)} \leq \sum_{l\in\mathbb{Z}} |\langle f, g_l\rangle|^2 \leq B\|f\|^2_{\mathbf{L}_2(\mathbb{R}^D)}, \qquad \forall f \in \mathbf{L}_2(\mathbb{R}^D),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathbf{L}_2(\mathbb{R}^D)$. We refer to $\mathcal{G}$ as a tight frame if it is a frame with $A = B = 1$.

For any given frame $\mathcal{G}$ of $\mathbf{L}_2(\mathbb{R}^D)$, there exists a dual frame $\widetilde{\mathcal{G}} = \{\tilde{g}_l : l \in \mathbb{Z}\}$ of $\mathcal{G}$ such that

$$f = \sum_{l \in \mathbb{Z}} \langle f, g_l \rangle \tilde{g}_l, \qquad \forall f \in \mathbf{L}_2(\mathbb{R}^D),$$

in which the function $\tilde{g}_l$ is a finitely linear combination of $g_L(\cdot - k), k \in \mathbb{Z}^d$.

For given $\Psi = \{\psi_1, \cdots, \psi_L\} \subset \mathbf{L}_2(\mathbb{R}^D)$, the corresponding quasi-affine system $\mathcal{G}^J(\Psi), J \in \mathbb{Z}$ generated by $\Psi$ is defined by the collection of the dilations and the shifts of $\Psi$ as

$$\mathcal{G}^J(\Psi) = \{\psi_{ljk} : 1 \le l \le L; j \in \mathbb{Z}, k \in \mathbb{Z}^D\}, \tag{11}$$

where $\psi_{ljk}$ is defined by

$$\psi_{ljk} = \begin{cases} 2^{\frac{jD}{2}} \psi_l\left(2^j \cdot -k\right), & j \ge J; \\ 2^{(j-\frac{J}{2})D} \psi_l\left(2^j \cdot -2^{j-J}k\right), & j < J. \end{cases} \tag{12}$$

When $\mathcal{G}^J(\Psi)$ forms a (tight) frame of $\mathbf{L}_2(\mathbb{R}^D)$, each function $\psi_l$ is called a (tight) framelet and the whole system $\mathcal{G}^J(\Psi)$ is called a (tight) wavelet frame system. The quasi-affine system is obtained by over sampling the traditional affine (wavelet) system

$$\bar{\mathcal{G}}(\Psi) := \left\{2^{\frac{jD}{2}} \psi_l\left(2^j \cdot -k\right) : 1 \le l \le L, j \in \mathbb{Z}, k \in \mathbb{Z}^D\right\}$$

starting from level $j = J - 1$ and downward. Hence, the whole quasi-affine system is a $2^{-J}$-shift-invariant system. It was shown in that a traditional affine system $\bar{\mathcal{G}}(\Psi)$ is a tight frame of $\mathbf{L}_2(\mathbb{R}^D)$ if and only if the corresponding quasi-affine counterpart $\mathcal{G}^J(\Psi)$ is a tight frame of $\mathbf{L}_2(\mathbb{R}^D)$.

Note that in the literature, the affine system $\bar{\mathcal{G}}(\Psi)$ is commonly used, which corresponds to the decimated wavelet (frame) transforms. The quasi-affine system, which corresponds to the so-called undecimated wavelet (frame) transforms. When a quasi-affine system is $\mathcal{G}^J(\Psi)$ selected, the discretely observed signal data is assumed to be sampled at level $J$ and the associated undecimated wavelet (frame) decomposition starts at level $J$ and goes downward. Here, we only discuss the quasi-affine system Eq.(11), since it works better in signal recovery and its connection to variational models and PDEs is more natural than the affine system.

The constructions of framelets $\Psi$, which are desirably (anti)symmetric and compactly supported functions, are usually based on a multiresolution analysis (MRA) that is generated by some refinable function $\phi$. Specifically, starting from $\phi = B_r := \int_0^1 B_{r-1}(\cdot - t)\mathrm{d}t, r > 1$, where $B_1$ is the indicator function of interval $[0, 1]$, an univariate tight framelet $\Psi$ can be explicitly constructed. Furthermore, the $d$-dimensional tight framelet is constructed by the tensor product of the univariate tight framelets, and then the quasi-affine system $\mathcal{G}^J(\Psi)$ is obtained via Eq.(12).

## B Missingness-aware Dynamics Function: Construction and Parameterization

This appendix details the missingness descriptors and FiLM parameterization used in Section 3.3. We use the hidden-state and control notation $Z^{(j)}(t) \in \mathbb{R}^{d_h}$ and $X^{(j)}(t) \in \mathbb{R}^{d_u}$ from the multi-scale NCDE formulation.

**Observation history and initialization.** For channel $c$, the observation indicator is

$$m_c(t_i) = \begin{cases} 1, & \text{if channel } c \text{ is observed at } t_i, \\ 0, & \text{otherwise.} \end{cases}$$

Let $s_c(t) = 1$ if at least one observation of channel $c$ is available at or before $t$, and $s_c(t) = 0$ otherwise. When $s_c(t) = 1$, define

$$t_c^{\text{last}}(t) = \max\{t_i \le t : m_c(t_i) = 1\}, \qquad \delta_c(t) = t - t_c^{\text{last}}(t). \tag{13}$$

Before the first observation, we set $\delta_c(t) = t - t_0$ and $q_c(t) = 0$. In Equations (7), the logarithm compresses long gaps, while $q_c$ measures recency relative to $h_c$. At a new observation, $\delta_c = 0$ and $q_c = 1$; the zero-recency initialization prevents a channel with no observations from being treated as recently observed.

**Channel-specific timescale.** Let $t_{c,r}$ denote the $r$-th observed timestamp of channel $c$. We collect the positive consecutive gaps $g_{c,r} = t_{c,r} - t_{c,r-1} > 0$ within each training sequence into $\mathcal{G}_c$. Following the Freedman–Diaconis rule (Freedman & Diaconis, 1981), we estimate

$$h_c = \frac{2\ \mathrm{IQR}(\mathcal{G}_c)}{|\mathcal{G}_c|^{1/3}}. \tag{14}$$

This quantity serves as a data-dependent normalization and decay timescale. We use the median positive gap when the interquartile range is zero, and clip the estimate to a predefined positive range. The timescales are estimated from the training set.

**Descriptor interpolation.** The elapsed-time descriptors reset at observations and are therefore not themselves continuous. We evaluate them at the observation knots and form

$$\overline{\mathbf{M}}_i = [\boldsymbol{\tau}^{\mathrm{gap}}(t_i); \mathbf{q}(t_i)], \qquad \mathbf{M}(t) = \mathcal{I}_m\big(\{(t_i, \overline{\mathbf{M}}_i)\}_i\big)(t) \in \mathbb{R}^{2C}, \tag{15}$$

where $\mathcal{I}_m$ uses the interpolation scheme for the input controls. The interpolated path $\mathbf{M}(t)$ provides conditioning values between knots; the exact elapsed-time and recency definitions apply to the descriptors before interpolation.

**Shared missingness encoder.** A two-layer encoder maps the missingness path to a common conditioning representation:

$$\boldsymbol{\eta}(t) = E_m(\mathbf{M}(t)) = W_{m,2}\phi\big(W_{m,1}\mathbf{M}(t) + b_{m,1}\big) + b_{m,2} \in \mathbb{R}^{D_m}, \tag{16}$$

where $\phi$ is a nonlinear activation. The encoder parameters are shared across all scales.

**Scale-specific FiLM generator.** The intermediate vector-field features are

$$\mathbf{h}_j(t) = H_j(Z^{(j)}(t)) = \phi\big(W_{j,\mathrm{in}}Z^{(j)}(t) + b_{j,\mathrm{in}}\big) \in \mathbb{R}^{D_h}. \tag{17}$$

Each branch transforms the shared missingness embedding into its own modulation parameters:

$$\begin{aligned} \mathbf{a}_j(t) &= \mathrm{SiLU}\big(W^F_{j,1}\,\mathrm{LayerNorm}(\boldsymbol{\eta}(t)) + b^F_{j,1}\big), \\ \boldsymbol{\gamma}_j(t) &= W^F_{j,\gamma}\mathbf{a}_j(t) + b^F_{j,\gamma}, \\ \boldsymbol{\beta}_j(t) &= W^F_{j,\beta}\mathbf{a}_j(t) + b^F_{j,\beta}. \end{aligned} \tag{18}$$

We bound the multiplicative and additive modulation using

$$\boldsymbol{\Gamma}_j(t) = \mathbf{1} + \alpha_\gamma \tanh(\boldsymbol{\gamma}_j(t)), \qquad \mathbf{B}_j(t) = \alpha_\beta \tanh(\boldsymbol{\beta}_j(t)), \tag{19}$$

where nonnegative $\alpha_\gamma$ and $\alpha_\beta$ control the maximum deviation from identity scaling and zero shift, respectively. Both modulation vectors have dimension $D_h$.

## C Lipschitz Regularity and Well-posedness of MS-NCDE

**Regularity and well-posedness.** Following standard NCDE regularity analysis, we consider a finite interval $I = [t_0, T]$. For the joint state $\mathbf{Z} = (\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)})$, define $\|\mathbf{Z}\|_{\max} = \max_l \|\mathbf{z}^{(l)}\|_2$.

**Assumption 1** (Scale-wise Lipschitz regularity)**.** Each control $\tilde{X}^{(l)}$ is continuous and piecewise $C^1$ on $I$. Each vector field $f^{(l)}_{\theta_l}(\mathbf{z}, t)$ is jointly continuous and globally Lipschitz in $\mathbf{z}$, uniformly in $t$:

$$\big\|f^{(l)}_{\theta_l}(\mathbf{z}, t) - f^{(l)}_{\theta_l}(\mathbf{w}, t)\big\|_{\mathrm{op}} \le K_l\|\mathbf{z} - \mathbf{w}\|_2, \qquad l = 1, \ldots, L. \tag{20}$$

**Theorem 1** (Existence and uniqueness of MS-NCDE)**.** *Under Assumption 1, the joint block-diagonal vector field is globally $K$-Lipschitz in the joint state, where $K = \max_l K_l$. For every initial state, the MS-NCDE in Equation (3) admits a unique solution on $I$. For fixed vector fields, this solution depends locally Lipschitz continuously on the initial state and the controls in 1-variation. In particular, two solutions driven by the same controls satisfy*

$$\|\mathbf{Z} - \overline{\mathbf{Z}}\|_{\infty,\max} \le e^{KV} \|\mathbf{Z}(t_0) - \overline{\mathbf{Z}}(t_0)\|_{\max}, \qquad V = \max_l \|\tilde{X}^{(l)}\|_{1\text{-}var;I}, \tag{21}$$

*where* $\|\mathbf{Z}\|_{\infty,\max} = \sup_{t\in I} \|\mathbf{Z}(t)\|_{\max}$.

We prove Theorem 1 for the parallel system in Section 3.2. Throughout, the learned parameters are fixed. For brevity, write $X_l = \tilde{X}^{(l)}$ and $f_l = f_{\theta_l}^{(l)}$. The vector fields may depend on time through a fixed continuous missingness-conditioning path.

**Notation.** For a continuous path $X : I \to \mathbb{R}^m$, its total variation is

$$\|X\|_{1\text{-var};I} = \sup_{\Pi} \sum_{i=1}^{n} \|X(t_i) - X(t_{i-1})\|_2,$$

where the supremum is over finite partitions $\Pi = \{t_0 < \cdots < t_n = T\}$. For the piecewise $C^1$ controls used here, this equals $\int_I \|\dot{X}(t)\|_2 \, dt$. We equip stacked state and control vectors with the maximum of their blockwise Euclidean norms; $\|\cdot\|_{\mathrm{op},\max}$ denotes the corresponding induced matrix norm. Set

$$V_l(t) = \int_{t_0}^{t} \|\dot{X}_l(s)\|_2 \, ds, \qquad V_l = V_l(T), \qquad A_l = \sup_{t\in I} \|f_l(\mathbf{0}, t)\|_{\mathrm{op}}.$$

The constants $A_l$ are finite by continuity on the compact interval.

*Proof. Lipschitz regularity of the joint vector field.* For $\mathbf{Z} = (\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)})$, define

$$\mathbf{F}_\theta(\mathbf{Z}, t) = \mathrm{BlockDiag}\big(f_1(\mathbf{z}^{(1)}, t), \ldots, f_L(\mathbf{z}^{(L)}, t)\big).$$

The induced norm of a block-diagonal matrix is the maximum of its block operator norms. Thus,

$$\begin{aligned} \|\mathbf{F}_\theta(\mathbf{Z}, t) - \mathbf{F}_\theta(\mathbf{W}, t)\|_{\mathrm{op},\max} &= \max_l \|f_l(\mathbf{z}^{(l)}, t) - f_l(\mathbf{w}^{(l)}, t)\|_{\mathrm{op}} \\ &\le \max_l K_l \|\mathbf{z}^{(l)} - \mathbf{w}^{(l)}\|_2 \\ &\le K \|\mathbf{Z} - \mathbf{W}\|_{\max}. \end{aligned} \tag{22}$$

This bounds the matrix-valued vector field. After multiplication by the control derivatives, the driven ODE has time-dependent Lipschitz modulus at most $\max_l K_l \|\dot{X}_l(t)\|_2$.

*Global existence and uniqueness on $I$.* For each branch, define

$$G_l(\mathbf{z}, t) = f_l(\mathbf{z}, t)\dot{X}_l(t).$$

Its state dependence and growth satisfy, almost everywhere,

$$\begin{aligned} \|G_l(\mathbf{z}, t) - G_l(\mathbf{w}, t)\|_2 &\le K_l \|\dot{X}_l(t)\|_2 \|\mathbf{z} - \mathbf{w}\|_2, \\ \|G_l(\mathbf{z}, t)\|_2 &\le \big(A_l + K_l \|\mathbf{z}\|_2\big) \|\dot{X}_l(t)\|_2. \end{aligned}$$

Both coefficients are integrable over $I$. The standard initial-value theorem therefore gives a unique local absolutely continuous solution; equivalently, one may solve on each smooth spline segment and concatenate at the knots (Teschl, 2012). The integral equation and Gronwall's inequality give

$$\|\mathbf{z}^{(l)}(t)\|_2 \le \big(\|\mathbf{z}^{(l)}(t_0)\|_2 + A_l V_l(t)\big) e^{K_l V_l(t)}. \tag{23}$$

Since $V_l < \infty$, the solution cannot diverge in finite time and extends uniquely over all of $I$. Stacking the $L$ branch solutions gives the unique MS-NCDE solution.

*Dependence on the initial state.* For two solutions with the same vector fields and controls, let $e_l(t) = \|\mathbf{z}^{(l)}(t) - \overline{\mathbf{z}}^{(l)}(t)\|_2$. Subtracting their integral equations yields

$$e_l(t) \le e_l(t_0) + K_l \int_{t_0}^{t} e_l(s)\|\dot{X}_l(s)\|_2 \, ds.$$

Applying Gronwall's inequality to each branch gives $e_l(t) \le e_l(t_0)e^{K_l V_l(t)}$. Taking the maximum over branches and time proves Equation (21). The branchwise argument is essential: it uses the maximum branch variation without replacing it by the variation of the stacked path.

*Dependence on the control paths.* Now let $\overline{\mathbf{z}}^{(l)}$ be driven by another control $\overline{X}_l$, with the same $f_l$. Decompose the difference as

$$\begin{aligned}\mathbf{z}^{(l)}(t) - \overline{\mathbf{z}}^{(l)}(t) = \mathbf{z}^{(l)}(t_0) - \overline{\mathbf{z}}^{(l)}(t_0) \\ + \int_{t_0}^{t} \left[f_l(\mathbf{z}^{(l)}(s), s) - f_l(\overline{\mathbf{z}}^{(l)}(s), s)\right] dX_l(s) \\ + \int_{t_0}^{t} f_l(\overline{\mathbf{z}}^{(l)}(s), s)\, d(X_l - \overline{X}_l)(s).\end{aligned}$$

Let $M_l = \sup_{t\in I} \|f_l(\overline{\mathbf{z}}^{(l)}(t), t)\|_{\mathrm{op}}$, which is finite by continuity and Equation (23). The variation bound for the last integral and Gronwall's inequality then imply

$$\|e_l\|_\infty \le e^{K_l V_l} \left[e_l(t_0) + M_l \|X_l - \overline{X}_l\|_{\text{1-var};I}\right].$$

To obtain a uniform Lipschitz constant, restrict both initial states to $\|\mathbf{Z}(t_0)\|_{\max}, \|\overline{\mathbf{Z}}(t_0)\|_{\max} \le r$ and both sets of controls to maximum branch variation at most $R$. With $A = \max_l A_l$, Equation (23) gives

$$M_l \le M_{r,R} := A + K(r + AR)e^{KR}.$$

Consequently,

$$\begin{aligned}\|\mathbf{Z} - \overline{\mathbf{Z}}\|_{\infty,\max} \le e^{KR}\big[\|\mathbf{Z}(t_0) - \overline{\mathbf{Z}}(t_0)\|_{\max} \\ + M_{r,R} \max_l \|X_l - \overline{X}_l\|_{\text{1-var};I}\big].\end{aligned} \tag{24}$$

Thus the solution map is Lipschitz on every set with bounded initial states and bounded control variation, proving the stated local Lipschitz dependence. The initial-state term also accounts for changes in any initialization computed from the input. □

**Verification for the FiLM-conditioned vector fields.** For the MaDF parameterization in Appendix B, assume that the hidden activation $\phi$ is globally $L_\phi$-Lipschitz and all learned weights are finite. For a fixed continuous conditioning path $\boldsymbol{\eta}(t)$, the FiLM parameters depend on $t$ but not on the hidden state. The bound $\|\mathbf{\Gamma}_l(t)\|_\infty \le 1 + \alpha_\gamma$ yields

$$\|\mathbf{h}_l^{\mathrm{cond}}(\mathbf{z}, t) - \mathbf{h}_l^{\mathrm{cond}}(\mathbf{w}, t)\|_2 \le (1 + \alpha_\gamma)L_\phi \|W_{l,\mathrm{in}}\|_2 \|\mathbf{z} - \mathbf{w}\|_2.$$

The additive FiLM shift cancels in this difference. After the output projection and reshaping, using the Frobenius norm to bound the matrix operator norm gives the valid choice

$$K_l = \|W_{l,\mathrm{out}}\|_2 (1 + \alpha_\gamma) L_\phi \|W_{l,\mathrm{in}}\|_2. \tag{25}$$

Continuity of the conditioning path and the network operations gives joint continuity in the state and time. Hence the described MaDF satisfies Assumption 1 under these activation and parameter conditions.

The stability bounds compare trajectories with fixed parameters and the same conditioning path. If the observation mask or its derived conditioning path changes, the vector fields change as well, and an additional term controlling that change is required. The bounds apply on any finite interval over which the controls and conditioning are defined and satisfy the assumption.

## D Dataset Preprocessing and Forecasting Protocols

**Shared Setup.** Records are partitioned into approximately $80\%/10\%/10\%$ training/validation/test splits, using a fixed shuffled split for USHCN and ordered splits for the other four datasets. Missing and padded entries are zero-filled with explicit observation masks, and evaluation involves only observed targets.

**USHCN.** Its five variables are retained as well as the observation masks without additional feature scaling. Timestamps are divided by the maximum benchmark timestamp. The initial 150 benchmark time units provide the history for predicting the next 3 recorded timestamps.

**Human Activity.** Its 12 positional channels are used from four three-dimensional sensors. Recordings are split before extracting windows with a 3,000 ms history, a 300 ms forecast interval, and a 4,000 ms stride. Timestamps are normalized but positional values remain unscaled.

**P12.** The sets A–C are combined using the hourly preprocessing and a 36-channel configuration. Features are standardized, timestamps are min–max normalized, and values with $|z| \geq 5$ are marked as missing. Entirely missing channels are removed. A nominal 36-hour history predicts the next 3 available timestamps.

**MIMIC-III.** We follow the supplied v1.4 preprocessing pipeline. The first 48 hours relative to each admission's earliest recorded event are retained, and measurements are aggregated at 30-minute resolution. Also, duplicate laboratory values are averaged; input, output, and prescription quantities are summed. Admissions with fewer than 50 variable–timestamp observations are excluded. The resulting benchmark contains 96 channels and uses precomputed feature-wise standardized values, with a 36-hour history predicting the next 3 recordeds.

**MIMIC-IV.** The v3.1 pipeline maintains the first 48 hours relative to each admission's earliest recorded event, converts timestamps to integer minutes, and merges coincident records by channel-wise maximum. Features and timestamps are standardized and min–max normalized respectively. The values with $|z| \geq 5$ are treated as missing. The 107-channel configuration uses a nominal 36-hour history to predict the next 3 available timestamps.

## E Implementation Details

All experiments run on a single NVIDIA RTX 4090 GPU with 24 GB of memory. The environment uses Python 3.12 and PyTorch 2.7.0. The AdamW optimizer is used with a learning rate of $10^{-3}$ and a weight decay of $10^{-4}$. The batch size is set to 16 for USHCN and Human Activity, and 32 for P12, MIMIC-III, and MIMIC-IV. All models are trained for 300 epochs with a maximum of 10 early stoppings. Each experiment is repeated 5 times with different random seeds from $\{2027, 2028, 2029, 2030, 2031\}$ to ensure statistical significance, and calculate the mean and standard deviation of the results. The performance metrics are MSE and MAE, which are computed only on the observed target values.The loss function is the MSE.

## F Hyperparameter Analysis

We focus on three key hyperparameters in our model: (1) the number of denoising scales, (2) the number of NCDE scales, and (3) the choice of ODE solver. Fig 4 illustrates how the MSE varies as each of these hyperparameters is changed on the P12 and Human Activity datasets. Notably, adding more denoising scales does not always improve accuracy, as excessive denoising can smooth out informative fine-grained variations along with the noise. Similarly, increasing the number of NCDE scales does not yield consistent gains and can even degrade performance beyond a moderate setting, indicating that the optimal scale granularity is dataset-dependent. For the ODE solver, the adaptive-step Dopri5 solver with a moderately tight tolerance generally matches or outperforms the fixed-step RK4 and Midpoint solvers, while further tightening the tolerance brings little additional benefit, suggesting that the underlying dynamics are not stiff and a coarser tolerance already offers a good accuracy–efficiency trade-off.

Table 3: Hyperparameter configuration of ResoPath.

| **Component** | **Hyperparameter** | **Setting** |
|---|---|---|
| Multi-scale NCDE | Wavelet detail levels | $J = 3$, producing $J + 1 = 4$ independent NCDE trajectories |
| | Hidden dimension | 128 for every scale-specific NCDE |
| | Vector-field width | 256 |
| | Initial-state projection | Independent linear projection for each scale |
| | Multi-scale decomposition | Undecimated à trous transform with stationary Haar low-pass filter $[0.5, 0.5]$ |
| Missingness module | Missingness features | Time-gap and confidence trajectories, giving $2C$ missingness channels for $C$ variables |
| | Missingness embedding dimension | 64 |
| | Missingness encoder width | 128 |
| | FiLM generator width | 128 |
| | FiLM bounds | $\alpha_\gamma = 1.0$ and $\alpha_\beta = 1.0$ |
| | Numerical epsilon | $1 \times 10^{-6}$ |
| | Fallback gap width | 1.0 |
| Wavelet denoising | Denoising levels | 2 |
| | Threshold-predictor channels | 32 |
| | Wavelet and boundary mode | Haar wavelet with symmetric padding |
| | Initial threshold | 0.1 |
| | Conditioning strength | 0.1 |
| Interpolation and solver | Control interpolation | Natural cubic spline |
| | Missingness interpolation | Piecewise linear interpolation |
| | ODE solver | Midpoint method |
| | Absolute/relative tolerances | $\mathrm{atol} = \mathrm{rtol} = 1 \times 10^{-3}$ |
| | Adjoint differentiation | Disabled |
| Prediction head | Readout width | 128 with SiLU activation |
| | Scale fusion | Learnable $1 \times 1$ fusion across the four scale-specific hidden trajectories |
| | Output projection | Shared linear temporal projection from the input length to the forecasting horizon |
| | Dropout | 0.02 before temporal projection |
| Optimization | Loss | Masked mean-squared error |
| | Optimizer | AdamW |
| | Initial learning rate | $1 \times 10^{-3}$ |
| | Learning-rate schedule | Delayed step decay: multiplier 1 for $e < 2$ and $0.8^{e-2}$ thereafter |
| | Maximum epochs | 300 |
| | Early-stopping patience | 10 epochs |
| | Validation interval | Every epoch |
| | Independent runs | 5 runs using seeds 2027–2031 |

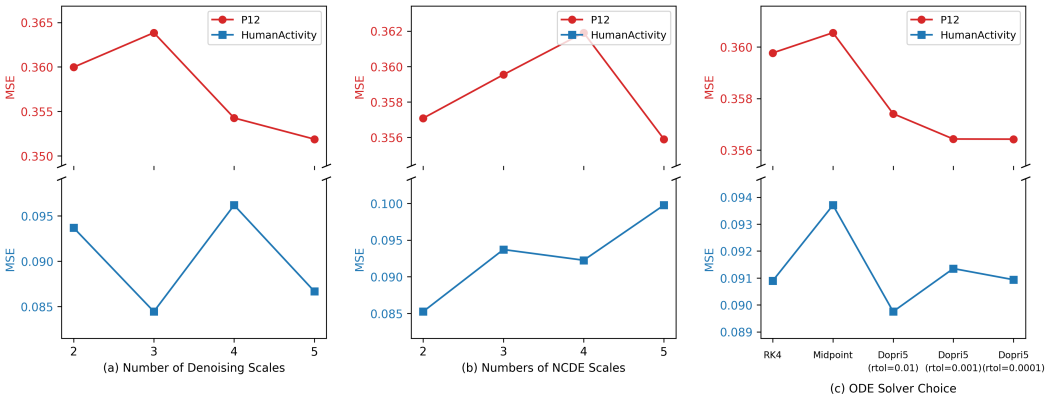



Figure 4: Parameter sensitivity analysis.