Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MMCont-Net: Towards Full-Scale Context Modeling for Irregular Series Forecasting

Qinghua Ren
School of Computer Science and Communication Engineering, Jiangsu University
Yanlin Wu
School of Computer Science and Communication Engineering, Jiangsu University
Dalong Gao
School of Computer Science and Communication Engineering, Jiangsu University
Lijian Gao
ljgao@stmail.ujs.edu.cn

School of Computer Science and Communication Engineering, Jiangsu University

Research Article

Keywords: Irregular Multivariate Time Series, Spectral Analysis, Full-scale Context modeling, Dynamic Graph Topology Diffusion, Spatiotemporal Modeling

Posted Date: June 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9008630/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MMCont-Net: Towards Full-Scale Context Modeling for Irregular Series Forecasting

Qinghua Ren[1], Yanlin Wu[1], Dalong Gao[1], Lijian Gao[1*]

[1]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, Jiangsu, China.

*Corresponding author(s). E-mail(s): ljgao@stmail.ujs.edu.cn;
Contributing authors: qinghua.ren@ujs.edu.cn;
2212308040@stmail.ujs.edu.cn; 2222408096@stmail.ujs.edu.cn;

Abstract

Forecasting irregularly sampled multivariate time series (IMTS) remains a fundamental challenge in contemporary time series analysis, owing to the inherent non-uniformity of sampling and pronounced inter-variable asynchrony. While recent advances in context learning achieves remarkable performance, microscopic signal distortion and macroscopic spatial isolation issues still limit the effectiveness of irregular series Forecasting. In this paper, we propose MMCont-Net (Micro–Macro Context Modeling Network), MMCont-Net (Micro–Macro Context Modeling Network), a novel framework that pioneers full-scale context modeling. Specifically, MMCont-Net first captures the micro-level intra-patch context within localized signal segments, then models the micro-level cross-patch context across time, and finally reasons over the macro-level inter-variable context across the observations. By this way, MMCont-Net achieves coherent modeling through progressive context integration. Extensive experiments on multiple benchmark datasets, including PhysioNet, Human Activity, and USHCN, demonstrate that MMCont-Net consistently achieves lower forecasting errors while maintaining superior computational efficiency and robustness, demonstrating the importance of the proposed full-scale context modeling paradigm.

Keywords: Irregular Multivariate Time Series; Spectral Analysis; Full-scale Context modeling; Dynamic Graph Topology Diffusion; Spatiotemporal Modeling

1 Introduction

Multivariate time series (MTS) forecasting plays a foundational role in analyzing and decision-making for complex systems [1], such as intensive care units (ICUs) patient monitoring and large-scale meteorological networks. However, in real-world scenarios, data acquisition is often governed by stochastic or event-driven sampling mechanisms [2, 3], leading to Irregular Multivariate Time Series (IMTS) issue. IMTS is characterized by non-uniformly distributed observation timestamps and pronounced asynchrony across variables [4], where the collected observations rarely conform to the idealized assumption of regular sampling.

For instance, in clinical settings, the sampling frequency of physiological indicators (e.g., heart rate or blood pressure) adapts dynamically to patient conditions. Specifically, critical periods may trigger dense and high-frequency measurements, whereas other intervals remain extremely sparse. While such irregularity implicitly encodes valuable semantic information (e.g., urgency or stability), it simultaneously introduces substantial noise and uncertainty, posing significant challenges for reliable context modeling, i.e., potential relationships intra- or inter-observations. Such an intrinsic irregularity of IMTS demands a modeling paradigm capable of coherently integrating context across all scales, that is, from fine-grained signal perception to global observations interaction.

Currently, prior context learning work for IMTS problems are majorly focused on the two separated scales: microscopic signal perception level and macroscopic observations interaction level. Firstly, at the microscopic signal level, most prior approaches primarily rely on time-domain evolution, learning temporal dependencies by tracking changes in observed values over time[5, 6]. However, under irregular sampling, this time-centric perspective often suffers from pronounced spectral blindness: they struggle to distinguish sampling-induced high-frequency artifacts from intrinsic periodic or transient dynamics of the underlying signal. While recent frequency-enhanced approaches (e.g., Fourier-based decomposition and autocorrelation modeling [7, 8]) are proposed, they are primarily designed for regular data and often process time and frequency domains in a decoupled, sequential manner. As a result, they fail to achieve a unified receptive field that jointly balances local nonlinear perturbations and global spectral invariance, limiting their ability to model fine-grained signal semantics context under sampling-induced distortions.

Secondly, at the macroscopic observation level, IMTS introduces a critical spatial isolation challenge. Conventional spatiotemporal graph neural networks (ST-GNNs) [9] commonly assume synchronized snapshots, where all variables in the system are observed simultaneously at each time step, enabling spatial information propagation through graph convolutions. This assumption collapses under IMTS due to cross-variable asynchrony. Specifically, physically adjacent nodes may be unobserved within a given temporal window, leaving them in a state of semantic absence. Even approaches that incorporate dynamic graph structures [10, 11] are typically constrained to shallow, local interactions and fail to reconstruct high-order, global spatiotemporal dependencies. As a result, the macroscopic interaction information among asynchronous variables are obscured by fragmented observation gaps.

The coexistence of microscopic spectral blindness and macroscopic spatial isolation reveals a deeper gap in current IMTS research: the lack of full-scale contextual modeling. Existing methods often perform within a single scale, lacking a unified framework that coherently bridges fine-grained spectral perception and global topological diffusion. In practice, microscopic signal perception forms the foundation for reliable macroscopic relation modeling, while large spatial interactions are crucial for compensating sparse and irregular observations. Without such cross-scale contextual alignment, models are limited in capturing the continuous evolution of system states along non-uniform temporal flows. Therefore, the core challenge remains how to organically integrate localized spectral enhancement with global dynamic diffusion, and thereby close the representational gap from microscopic signal units to macroscopic system structures.

To this end, we propose MMCont-Net (Micro-Macro Context Modeling Network), a unified framework that pioneers full-scale context modeling for IMTS forecasting. MMCont-Net progressively integrates context from micro-level signal units to macro-level variable interactions, enabling a faithful reconstruction of system dynamics from sparse and asynchronous observations. Our key contribution can be concluded as follows.

- To effectively model the context information, we propose a unified full-scale context modeling framework, referred to as MMCont-Net, which gradually learns the micro-level intra-patch context, micro-level cross-patch context, and macro-level inter-variables context. This design captures a continuous context from local details to global interactions.
- To mitigate microscopic spectral blindness, our MMCont-Net integrates a frequency-domain branch to capture periodic patterns and a gated time-domain branch to filter local noise. Meanwhile, MMCont-Net performs a dynamic graph topology diffusion module to explore the dynamic and multi-hop diffusion pathways over inferred graph topologies, and thereby breaks the asynchronous isolation by propagating information beyond immediately missing neighbors.
- Comprehensive experiments on the well-adopted datasets (i.e., PhysioNet Challenge 2012, Human Activity, USHCN and MIMIC) show that our MMCont-Net greatly outperforms the current methods, which validates the effectiveness and necessity of the proposed full-scale context modeling paradigm.

2 Related Work

2.1 From Discrete Imputation to Continuous-Time Context Modeling

Early efforts on irregular multivariate time series (IMTS) mainly focused on discrete-time context modeling to irregular observations. Representative approaches followed imputation-then-modeling paradigms, where missing values were first filled before applying standard recurrent architectures, such as BRITS [12] and GAIN [13]. Although effective in practice, such two-stage pipelines may introduce imputation bias and distort intrinsic temporal dynamics by enforcing artificial regularity.

To enable end-to-end modeling under irregular sampling, subsequent works incorporated time information directly into sequence models. GRU-D [14] introduces time-dependent decay or gating mechanisms to modulate hidden states according to elapsed time. More principled continuous-time formulations were later proposed via Neural Ordinary Differential Equations (Neural ODEs) [15], with extensions such as ODE-RNN [6], Latent ODE [6], and Neural CDEs [5] modeling latent states as continuous trajectories. While these methods naturally handle asynchronous observations, they often incur high computational cost due to numerical solvers and primarily operate in the time domain, limiting their ability to capture long-range semantic context.

2.2 Attention-Based Continuous-Time Context Learning

The success of Transformers in modeling long-range dependencies has motivated attention-based approaches for IMTS. A central challenge lies in extending discrete positional encodings to continuous time. Time-aware embedding and interpolation methods such as Time2Vec [16] and mTAN [2] inject continuous timestamp information into attention mechanisms, enabling flexible handling of irregular observations. ContiFormer [17] further integrates continuous-time dynamics into attention computation, allowing non-local temporal interactions through integral operators.

In parallel, a line of research has explored frequency-enhanced and decomposition-based Transformer architectures. Autoformer [7] and FEDformer [8] exploit autocorrelation or Fourier-based attention to disentangle global trends from high-frequency components, while TimesNet [18] reshapes sequences into two-dimensional representations to capture multi-period dependencies. Crossformer [19] introduces dimension-wise segmentation and hierarchical attention for cross-variable modeling, and iTransformer [20] inverts temporal and variate dimensions to strengthen inter-variable interactions. However, these models are primarily designed for regularly sampled data and rely on pre-alignment when applied to IMTS.

Recent studies have attempted to improve scalability through patch-based representations. PatchTST [21] segments long sequences into fixed-length patches, and extensions such as T-PatchGNN [22] adapt this idea to irregular settings. Nevertheless, fixed-size patching can introduce semantic misalignment under irregular sampling, as patches containing the same number of observations may correspond to vastly different physical durations. To alleviate this issue, Pathformer [23] proposes adaptive multi-scale pathways, while PrimeNet [24] introduces irregularity-aware pre-training that explicitly encodes physical time gaps.

2.3 Spatiotemporal Graph Learning under Asynchronous Observations

Modeling macro-level inter-variable context is crucial for multivariate time series forecasting, and spatiotemporal graph neural networks (ST-GNNs) have been widely adopted for this purpose. Early methods such as Graph WaveNet [10], MTGNN [25], and StemGNN [26] learn adaptive or spectral graph structures to capture latent spatial

correlations. Dynamic graph learning further enables time-varying topology inference in non-stationary environments.

However, most ST-GNNs implicitly assume synchronous observations across variables, which is violated in irregular and asynchronous settings. To address this limitation, RainDrop [3] constructs sample-specific sensor graphs conditioned on observation availability, while GraFITi [11] unifies spatial and temporal dependencies through bipartite attention, maintaining connectivity via temporal history when spatial neighbors are missing. Hybrid approaches such as MTGODE [27] and STG-NCDE [28] further combine continuous-time dynamics with graph message passing.

Beyond deterministic modeling, recent works have explored generative approaches to handle uncertainty in highly sparse IMTS. CSDI [29] employs score-based diffusion models conditioned on observed values for probabilistic imputation, and PriSTI [30] incorporates spatiotemporal graph structures into the diffusion process. While effective for imputing missing values, these methods incur substantial computational overhead and are not explicitly optimized for forecasting continuous system evolution.

3 Methodology

The core design of MMCont-Net adheres to a full-scale context modeling paradigm that progresses from micro-sensing to macro-diffusion.

Specifically, the complex dynamics inherent in irregular multivariate time series (IMTS) are explicitly disentangled into three progressively structured dimensions: micro-level spectral signal perception, sequence-level continuous-time evolution, and macro-level dynamic graph diffusion. The overall architecture of MMCont-Net is illustrated in Fig. 1. Following this principled dataflow—from localized fine-grained signal characteristics to global topological interactions—this section systematically elucidates how the proposed model incrementally reconstructs robust system-level spatiotemporal representations from sparse and non-uniform observations.

3.1 Data Preprocessing

Irregular Multivariate Time Series Representation. An irregular multivariate time series (IMTS) with N variables is represented as a set of observations $\mathcal{O} = \{o_n\}_{n=1}^{N}$. For the n-th variable, the observation sequence consists of L_n irregularly sampled points, denoted as $o_n = \{(t_{n,i}, v_{n,i})\}_{i=1}^{L_n}$, where $t_{n,i} \in \mathbb{R}^+$ is the continuous timestamp and $v_{n,i} \in \mathbb{R}$ is the corresponding value. Due to the inherent irregularity of real-world sampling, both sequence lengths L_n and timestamps $\{t_{n,i}\}$ vary across variables, resulting in pronounced inter-variable asynchrony and temporal heterogeneity.

Transformable Patching Strategy. Conventional fixed-sample segmentation disrupts the temporal coherence of underlying micro-level physical processes in irregularly sampled data. To extract robust local representations while preserving physical temporal continuity, we propose a transformable patching strategy based on fixed physical time windows [22]. The strategy is applied independently to each univariate sequence; we illustrate it using the n-th variable o_n.

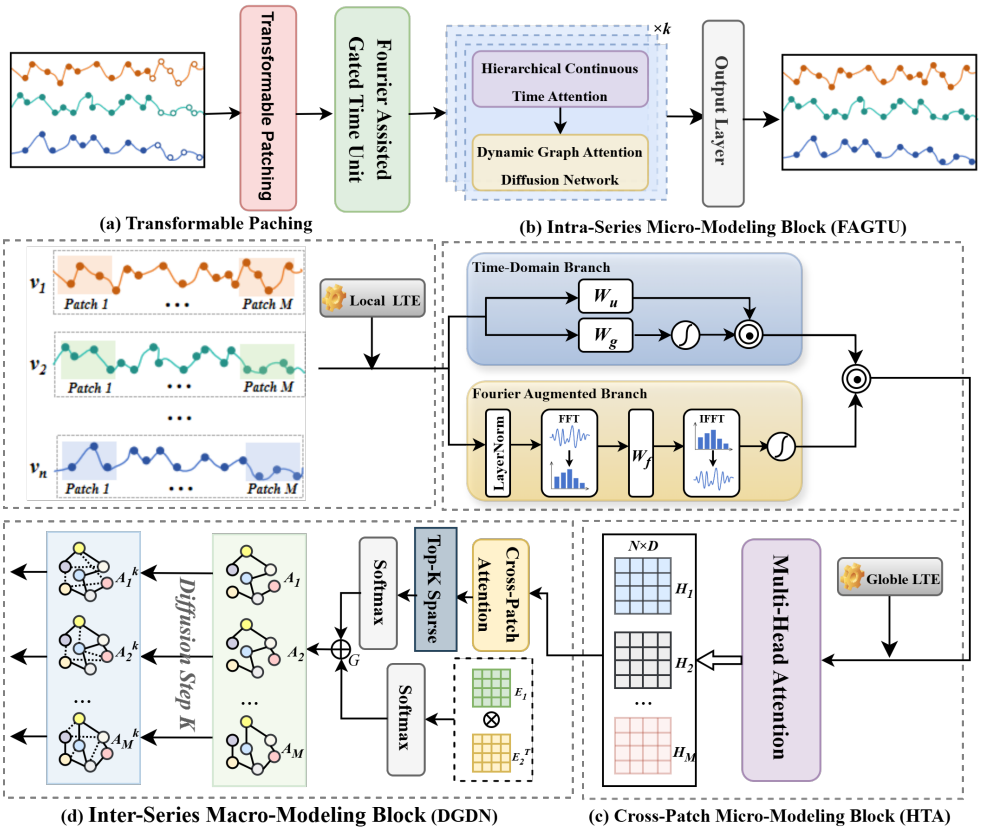


Fig. 1: Overview of MMCont-Net. The framework initially segments irregular multivariate time series into aligned semantic patches via (a) Transformable Patching. Subsequently, the representation is refined by (b) Fourier-Assisted Gated Temporal Unit (FAGTU) through a dual-branch spectral-temporal fusion mechanism. Finally, the deep features are processed by (c) Hierarchical Continuous-Time Attention (HTA) and (d) Dynamic Graph Attention Diffusion Network (DGDN) to jointly model continuous temporal dynamics and evolving spatial dependencies under a unified framework.

We define a sequence of sliding physical time windows on the continuous time axis. The m-th window $\mathcal{T}_m$ is given by

$$\mathcal{T}_m = [t_0 + (m-1)S, t_0 + (m-1)S + \tau), \tag{1}$$

where t_0 is the global starting time serving as an absolute temporal reference, S is the sliding stride controlling overlap between adjacent windows, and τ is the fixed physical window duration (e.g., two hours) that determines the micro-level modeling granularity.

The univariate sequence o_n is then partitioned into M transformable patches $\{\mathcal{P}_m^{(n)}\}_{m=1}^{M}$, where the m-th patch is

$$\mathcal{P}_m^{(n)} = \{(t_{n,i}, v_{n,i}) \in o_n \mid t_{n,i} \in \mathcal{T}_m\}. \tag{2}$$

Each patch $\mathcal{P}_m^{(n)}$ spans a constant physical duration τ, but contains a variable number of observations $N_m = |\mathcal{P}_m^{(n)}|$ depending on local sampling density. This design enforces strict physical-time alignment across all variables, effectively mitigating inter-variable asynchrony without resorting to extensive missing-value imputation that would otherwise cause sequence-length explosion. The resulting patches provide a standardized, temporally coherent input for subsequent micro-level intra-patch context modeling.

Continuous Learnable Time Encoding. To preserve fine-grained temporal information potentially obscured by windowing and to equip the model with explicit time awareness, we introduce a learnable continuous time encoding [16]. For any timestamp t, the encoding maps it to a D_t-dimensional vector:

$$\phi(t)_k = \begin{cases} \omega_0 t + \beta_0 & \text{if } k = 0, \\ \sin(\omega_k t + \beta_k) & \text{if } 1 \le k < D_t, \end{cases} \tag{3}$$

where ω_k and β_k are learnable parameters. The linear term captures non-periodic trends, while sinusoidal terms model latent periodic patterns.

Within each patch $\mathcal{P}_m^{(n)}$, we augment every observation by concatenating its value v_i with its time encoding $\phi(t_i)$:

$$\mathbf{Z}_m^{(n)} = [v_i \parallel \phi(t_i)]_{i=1}^{N_m} \in \mathbb{R}^{N_m \times D_{\text{in}}}, \tag{4}$$

where $\parallel$ denotes concatenation and $D_{\text{in}} = D_t + 1$. The resulting temporally aware feature matrices $\mathbf{Z}_m^{(n)}$ serve as standardized inputs for microscopic spectral modeling (Section 3.2). The shared encoding function $\phi(\cdot)$ is later reused in the macro-level module (Section 3.3) to facilitate coherent cross-scale context integration.

3.2 Fourier-Assisted Gated Temporal Unit

While the transformable patching strategy (Section 3.1) achieves physical-time alignment at the window level, the observation matrix $\mathbf{Z}_m^{(n)} \in \mathbb{R}^{N_m \times D_{\text{in}}}$ within each micro-level patch still suffers from sparsity and high-frequency artifacts induced by irregular sampling. These distortions give rise to the microscopic spectral blindness challenge, severely hindering reliable extraction of intrinsic semantic context. To overcome this and enable robust micro-level intra-patch spectral sensing, we propose the Fourier-Assisted Gated Temporal Unit (FAGTU). FAGTU is designed based on spectral-temporal duality [18], jointly reasoning across frequency and time domains to recover coherent dynamics from noise-corrupted sparse observations. The unit comprises two complementary parallel branches: a gated time-domain branch for local noise filtering and a frequency-domain branch for periodic pattern capture.

Gated Time-Domain Branch. This branch captures local sequential dependencies while dynamically suppressing transient noise arising from sensor imperfections or environmental fluctuations. The patch matrix $\mathbf{Z}_m^{(n)}$ is first projected into a latent space, yielding the time-domain hidden representation $\mathbf{H}_t^{(m)}$ as

$$\mathbf{H}_t^{(m)} = (\mathbf{Z}_m^{(n)}\mathbf{W}_u) \odot \sigma(\mathbf{Z}_m^{(n)}\mathbf{W}_g), \tag{5}$$

where $\mathbf{W}_u, \mathbf{W}_g \in \mathbb{R}^{D_{\text{in}} \times D}$ are learnable content and gate projection matrices (biases omitted for clarity), $\sigma(\cdot)$ is the sigmoid activation producing adaptive gates in $(0, 1)$, and $\odot$ denotes element-wise multiplication. This gating mechanism selectively attenuates anomalous fluctuations while preserving salient local features, guided by the micro-level intra-patch context [31].

Frequency-Domain Branch. Purely time-domain modeling is susceptible to local optima due to temporal discontinuities and sparsity in irregular sampling, which obscure global signal structure. To address this aspect of microscopic spectral blindness, the frequency-domain branch explicitly captures global periodic patterns across the entire physical window. The input patch matrix $\mathbf{Z}_m^{(n)}$ is first stabilized via layer normalization and projected to the hidden dimension. It is then transformed into the frequency domain using the Fast Fourier Transform (FFT), modulated by learnable spectral weights, and reconstructed via inverse FFT. The global frequency-augmented representation $\mathbf{H}_f^{(m)}$ is obtained by extracting the real component:

$$\mathbf{H}_f^{(m)} = \text{Re}\left(\mathcal{F}^{-1}\left(\mathcal{F}(\text{LN}(\mathbf{Z}_m^{(n)}\mathbf{W}_f)) \odot \mathbf{W}_{\text{spec}}\right)\right), \tag{6}$$

where $\mathbf{W}_f \in \mathbb{R}^{D_{\text{in}} \times D}$ is the projection matrix, $\mathbf{W}_{\text{spec}} \in \mathbb{C}^{F \times D}$ (with F the padded frequency length) is a learnable complex-valued spectral modulator, and $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote 1D FFT and IFFT along the temporal dimension. This design provides a unified global receptive field over the physical duration τ, bridging irregular temporal gaps, suppressing sampling-induced high-frequency artifacts, and enhancing latent periodic invariances [8].

Spectral-Temporal Adaptive Fusion. To coherently integrate global periodic priors with local time-domain features, we employ an adaptive fusion mechanism that uses the frequency branch as a data-driven spectral gate [32]. The fused micro-level representation $\hat{\mathbf{H}}^{(m)}$ is computed as

$$\hat{\mathbf{H}}^{(m)} = \text{LN}\left((\mathbf{H}_t^{(m)}\mathbf{W}_o) \odot \sigma(\mathbf{H}_f^{(m)})\right), \tag{7}$$

where $\mathbf{W}_o \in \mathbb{R}^{D \times D}$ is a learnable output projection. This cascaded modulation allows frequency-informed gates $\sigma(\mathbf{H}_f^{(m)})$ to selectively reweight time-domain features, yielding intra-patch representations that jointly preserve local nonlinear dynamics and enforce global spectral consistency [33]. Through this complementary dual-branch design, FAGTU effectively mitigates microscopic spectral blindness and produces robust micro-level intra-patch context.

Dynamic Aggregation and Patch Embedding. The fused variable-length representation $\hat{\mathbf{H}}^{(m)} \in \mathbb{R}^{N_m \times D}$ must be aggregated into a fixed-dimensional patch embedding to interface with subsequent cross-patch modeling. Naive pooling risks bias from non-uniform observation densities [34]. Instead, we adopt attention-based dynamic aggregation. Importance weights for each observation are computed as

$$\alpha_{m,i} = \frac{\exp(\hat{\mathbf{h}}_{m,i}\mathbf{w}_a^\top)}{\sum_{j=1}^{N_m} \exp(\hat{\mathbf{h}}_{m,j}\mathbf{w}_a^\top)}, \tag{8}$$

where $\hat{\mathbf{h}}_{m,i}$ is the i-th row of $\hat{\mathbf{H}}^{(m)}$ and $\mathbf{w}_a \in \mathbb{R}^D$ is a learnable attention vector. The aggregated feature is then obtained via weighted summation. Additionally, a binary validity indicator $\delta_m \in \{0, 1\}$ encodes patch availability ($\delta_m = 1$ if $N_m > 0$, else 0). The final patch embedding is

$$\mathbf{h}_m^{(n)} = \left(\sum_{i=1}^{N_m} \alpha_{m,i} \hat{\mathbf{h}}_{m,i} \right) \parallel \delta_m. \tag{9}$$

Across all M patches of the n-th variable, these embeddings form a temporally ordered sequence $\mathbf{H}_{\text{seq}}^{(n)} = [\mathbf{h}_1^{(n)}, \mathbf{h}_2^{(n)}, \ldots, \mathbf{h}_M^{(n)}]^\top \in \mathbb{R}^{M \times (D+1)}$. By producing physically aligned, fixed-length embeddings from irregular observations, this stage establishes a coherent foundation for subsequent micro-level cross-patch and macro-level inter-variable context modeling [35].

3.3 Hierarchical Continuous-Time Attention

Following micro-level intra-patch context extraction by FAGTU, the resulting patch embeddings $\{\mathbf{h}_m^{(n)}\}_{m=1}^M$ for each variable provide robust, physically aligned local representations. To further capture micro-level cross-patch temporal dependencies and bridge toward macro-scale interactions, we propose the Hierarchical Continuous-Time Attention (HTA) module. Unlike conventional Transformers that rely on discrete positional encodings, HTA explicitly incorporates physical temporal continuity into the attention mechanism through global absolute anchoring and local relative biasing.

Global Absolute Time Anchoring. The patch embedding $\mathbf{h}_m^{(n)} \in \mathbb{R}^{D+1}$ (Eq. 9) encapsulates refined micro-scale semantics but lacks explicit global temporal positioning. To inject macro-aware temporal context while reusing the shared continuous encoding framework (Eq. 3), we compute the absolute center timestamp t_m of the m-th physical window and map it to a global temporal embedding $\mathbf{e}_t^{(m)} = \text{Linear}(\phi(t_m)) \in \mathbb{R}^D$. This embedding is additively fused with the projected patch feature:

$$\tilde{\mathbf{h}}^{(m,n)} = (\mathbf{h}_m^{(n)} \mathbf{W}_{\text{proj}}) + \mathbf{e}_t^{(m)}, \tag{10}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{(D+1)\times D}$ is a learnable projection matrix. The resulting $\tilde{\mathbf{h}}^{(m,n)}$ explicitly anchors each patch within the global physical timeline, mitigating semantic distortions from irregular inter-patch intervals [36, 37].

Continuous Relative Temporal Biasing. Building on global anchoring, HTA models fine-grained cross-patch evolutionary dynamics via relative physical intervals. For each variable n, we assemble the anchored representations into a sequence tensor $\mathbf{H}^{(n)} = [\tilde{\mathbf{h}}^{(1,n)}, \ldots, \tilde{\mathbf{h}}^{(M,n)}]^\top \in \mathbb{R}^{M \times D}$. A multi-head self-attention mechanism is then applied. For the h-th head, queries, keys, and values are obtained as

$$\begin{aligned} \mathbf{Q}^{(h,n)} &= \mathbf{H}^{(n)} \mathbf{W}_Q^{(h)}, \\ \mathbf{K}^{(h,n)} &= \mathbf{H}^{(n)} \mathbf{W}_K^{(h)}, \\ \mathbf{V}^{(h,n)} &= \mathbf{H}^{(n)} \mathbf{W}_V^{(h)}, \end{aligned} \tag{11}$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{D \times D_h}$.

To enforce continuous temporal awareness, we introduce a learnable physical interval bias. For any patch pair (i, j) with center timestamps t_i and t_j, the relative distance $\Delta t_{i,j} = t_i - t_j$ is mapped to a scalar bias $b_{i,j} = \psi(\Delta t_{i,j})$ via an MLP $\psi(\cdot)$. This bias is added to the scaled dot-product attention:

$$\mathbf{A}_{i,j}^{(h,n)} = \mathrm{Softmax}_j \left(\frac{\mathbf{Q}_i^{(h,n)} (\mathbf{K}_j^{(h,n)})^\top + b_{i,j}}{\sqrt{D_h}} \right). \tag{12}$$

This relative biasing mechanism weights historical dependencies according to actual physical time elapsed, effectively compensating for irregular sampling gaps and enabling coherent micro-level cross-patch context propagation.

The refined sequence for variable n is obtained by concatenating heads, projecting, and applying a residual connection with layer normalization:

$$\mathbf{H}_{\text{cross}}^{(n)} = \mathrm{LN} \left(\mathbf{H}^{(n)} + \left(\Big\Vert_{h=1}^{N_h} \mathbf{A}^{(h,n)} \mathbf{V}^{(h,n)} \right) \mathbf{W}_O \right), \tag{13}$$

where $\mathbf{W}_O \in \mathbb{R}^{N_h D_h \times D}$. The resulting $\mathbf{H}_{\text{cross}}^{(n)} \in \mathbb{R}^{M \times D}$ integrates spectral fidelity from intra-patch processing with continuous cross-patch temporal logic, providing a standardized, temporally coherent representation for subsequent macro-level inter-variable context modeling via dynamic graph diffusion.

3.4 Dynamic Graph Attention Diffusion Network

The macro-scale dynamics of complex systems are not solely governed by the independent evolution of individual variables, but are fundamentally shaped by intricate mechanisms of spatial co-evolution across variables. In the context of irregular multivariate time series (IMTS), however, modeling such spatial dependencies is confronted with a severe asynchronous isolation challenge: owing to the extreme inconsistency in sampling times across variables, the synchronized snapshot assumption—namely, that all nodes are observed at the same time—underpinning conventional spatiotemporal graph neural networks is no longer valid. Consequently, inter-variable information flow

becomes fragmented under sparse and misaligned observations. To break this sparsity-induced spatial isolation at the macro level and to reconstruct disrupted cross-variable interactions, this paper proposes the Dynamic Graph Attention Diffusion Network (DGDN). By learning a data-driven dynamic topology and coupling it with a multi-hop graph diffusion mechanism, DGDN enables robust inference of global system states under asynchronous observation regimes [38].

Dynamic Graph Topology Learning. To comprehensively capture latent topological relationships among variables, the graph structure is decomposed into two complementary components: (i) an adaptive static adjacency that reflects invariant properties such as spatial proximity or physical constraints, and (ii) a dynamic data-dependent adjacency that captures transient correlations induced by instantaneous system states.

Formally, let $\mathbf{H}_{\text{macro}}^{(n)} \in \mathbb{R}^{M \times D}$ denote the macro-level representation sequence for the n-th variable ($n \in \{1, \ldots, N\}$), obtained from the preceding HTA module.To model the spatial co-evolution at the m-th time step, we collect the representations of all N variables at this specific patch index to construct the spatial node feature matrix $\mathbf{X}_m$:

$$\mathbf{X}_m = \left[\mathbf{h}_1^{(m)}, \mathbf{h}_2^{(m)}, \ldots, \mathbf{h}_N^{(m)}\right]^\top \in \mathbb{R}^{N \times D}, \tag{14}$$

where $\mathbf{h}_n^{(m)} = \mathbf{H}_{\text{macro}}^{(n)}[m, :] \in \mathbb{R}^D$ corresponds to the m-th row vector of the n-th variable's HTA output. This rearrangement effectively transforms the data view from temporal evolution (sequence-wise) to spatial topology (node-wise), serving as the input for the subsequent dynamic graph learning.

First, to model inherent constraints (e.g., spatial proximity) that remain stable over time, we employ two learnable node embedding dictionaries $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N \times d}$. The static adjacency matrix is derived as:

$$\mathbf{A}_{\text{adp}} = \text{Softmax}\left(\text{ReLU}(\mathbf{E}_1 \mathbf{E}_2^\top)\right), \tag{15}$$

which represents the global invariant structure. Complementing this static prior, we capture transient correlations induced by instantaneous system states by applying a multi-head self-attention mechanism on the current feature matrix $\mathbf{X}_m$. To mitigate noise introduced by sparse observations, we introduce a Sparsity-enforcing Top-K Operator, retaining only the k most significant connections. The dynamic adjacency matrix is formalized as [10]:

$$\mathbf{A}_{\text{dat}}^{(m)} = \text{RowNorm}\left(\text{TopK}\left(\frac{(\mathbf{X}_m \mathbf{W}_Q)(\mathbf{X}_m \mathbf{W}_K)^\top}{\sqrt{d}}, k\right)\right), \tag{16}$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D \times d}$ are projection matrices. Finally, to dynamically balance long-term patterns and sudden state changes, we propose an Edge-wise Gated Fusion mechanism. We treat the static and dynamic adjacencies as two channels of topological information and learn a gate matrix $\mathbf{G}^{(m)} \in \mathbb{R}^{N \times N}$ via a lightweight projection:

$$\mathbf{G}^{(m)} = \sigma\left(\text{Linear}\left(\text{Concat}(\mathbf{A}_{\text{adp}}, \mathbf{A}_{\text{dat}}^{(m)})\right)\right), \tag{17}$$

here, the concatenation is performed along the feature dimension, mapping $\mathbb{R}^{N\times N\times 2} \to \mathbb{R}^{N\times N}$. The final unified adjacency matrix $\mathbf{A}_{\text{final}}^{(m)}$ is obtained via adaptive interpolation:

$$\mathbf{A}_{\text{dyn}}^{(m)} = \mathbf{G}^{(m)} \odot \mathbf{A}_{\text{adp}} + (\mathbf{1} - \mathbf{G}^{(m)}) \odot \mathbf{A}_{\text{dat}}^{(m)}. \tag{18}$$

where $\odot$ denotes the Hadamard product (element-wise multiplication) and $\mathbf{1}$ is an all-ones matrix. This adaptive fusion ensures that the learned topology robustly adapts to both the inherent static structure and the time-varying semantic dependencies [39].

Multi-Hop Graph Diffusion Convolution. In IMTS settings, the sparsity of graph structures poses a critical challenge, as the direct neighbors of a given variable may be entirely unobserved within a particular patch (i.e., their features are zero-valued or masked). Under such circumstances, standard single-hop graph convolutions (1-hop GCNs) result in broken information propagation paths. To address this issue, we introduce a graph diffusion process on the dynamic graph [40, 41].

Formally, let $\mathcal{T}^{(m)} = \mathbf{A}_{\text{dyn}}^{(m)}$ denote the state transition matrix derived from the unified graph at step m. The diffusion process is modeled as a discrete random walk, where the information propagation at the k-th step is recursively defined as $\mathbf{X}^{(k)} = \mathcal{T}^{(m)}\mathbf{X}^{(k-1)}$. This implies that after k steps, the effective reachability is governed by the k-th power of the transition matrix, i.e., $\mathbf{X}^{(k)} = (\mathcal{T}^{(m)})^k\mathbf{X}_m$.

Based on this, the graph diffusion convolution is defined as the learnable aggregation of these multi-scale diffusion states. It convolves the input features with a finite power series of the transition matrix to capture spatial dependencies across varying neighborhood ranges:

$$\mathbf{H}_{\text{spatial}}^{(m)} = \text{LN}\left(\mathbf{X}_m + \sum_{k=1}^{K}\left((\mathcal{T}^{(m)})^k\mathbf{X}_m\right)\mathbf{\Theta}_k\right), \tag{19}$$

where:$\mathbf{X}_m$ serves as the residual identity connection, preserving the original signal fidelity against over-smoothing. $(\mathcal{T}^{(m)})^k$ explicitly models the k-hop transition probability, allowing the model to bridge semantic voids by propagating signals through k intermediate nodes. $\mathbf{\Theta}_k \in \mathbb{R}^{D\times D}$ is the learnable filter for the k-th diffusion scale, and K is the diffusion depth [42].

Through this mechanism, MMCont-Net effectively propagates weak local signals across the entire graph, systematically reconstructing the macro-scale system states even under unstructured and asynchronous observation conditions.

3.5 Forecasting and Optimization.

By stacking L layers of HTA and DGDN, the model iteratively refines the spatiotemporal features. Specifically, we collect the output $\mathbf{H}_{\text{spatial}}$ from the final layer across all M time steps to reconstruct the comprehensive variable-level representation sequence for the n-th variable, denoted as $\mathbf{H}_{\text{hist}}^{(n)} \in \mathbb{R}^{M\times D}$. This tensor encapsulates the rich context of historical observations, ready for downstream decoding. Given that forecasting for IMTS typically necessitates inference at non-uniformly distributed future moments, we propose a Time-Query Based Decoding mechanism. Specifically, considering a set of future prediction queries $\mathbf{Q} = \{q_j\}_{j=1}^{T_{\text{pred}}}$, where q_j denotes the j-th

continuous future timestamp, we reuse the pre-trained global time encoder to generate corresponding future time embeddings $\Phi(\mathbf{Q}) \in \mathbb{R}^{T_{\text{pred}} \times D}$. This step provides precise temporal positioning information for the decoder. To construct the decoder input, the historical sequence $\mathbf{H}_{\text{hist}}^{(n)}$ is first aggregated via mean pooling to obtain a fixed-length context vector $\mathbf{h}_{\text{hist}}^{(n)} \in \mathbb{R}^D$. This context vector is then broadcast along the prediction horizon and concatenated with the corresponding time embeddings. The final predictions are subsequently derived via a Multi-Layer Perceptron (MLP) mapping:

$$\hat{y}_{n,j} = \text{MLP}([\mathbf{h}_{\text{hist}}^{(n)} \parallel \phi(q_j)]) \in \mathbb{R} \tag{20}$$

where $\mathbf{h}_{\text{hist}}^{(n)} = \text{Mean}(\mathbf{H}_{\text{hist}}^{(n)})$ represents the aggregated historical hidden state of the n-th variable, and $\phi(q_j)$ signifies the time embedding at the j-th prediction moment. This architectural design ensures that the model can dynamically tailor its output according to the specific prediction time q_j, thereby adapting to irregular forecasting requirements.

Since missing observations are ubiquitous in real-world IMTS scenarios, the direct application of standard Mean Squared Error (MSE) would compel the model to fit invalid padding values, thereby undermining the validity of gradients. To address this, we adopt Masked MSE as the objective function for end-to-end optimization. During the training phase, reconstruction errors are computed exclusively for spatiotemporal nodes possessing authentic observations. Formally, let Ω denote the set of valid observations, defined as $\Omega = \{(n, j) \mid m_{n,j} = 1\}$, where $m_{n,j}$ is a binary mask indicating the presence of a ground truth label $y_{n,j}$ for the n-th variable at time q_j. The loss function is defined as:

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{(n,j)\in\Omega} (y_{n,j} - \hat{y}_{n,j})^2 \tag{21}$$

This objective function effectively circumvents the noise introduced by missing value padding, enabling MMCont-Net to learn robust spatiotemporal evolutionary patterns under sparse supervision.

4 Experiments

In this section, we conduct extensive experiments on three real-world datasets spanning the medical, biomechanical, and meteorological domains to comprehensively evaluate the performance of MMCont-Net in Irregular Multivariate Time Series (IMTS) forecasting tasks. Furthermore, we verify the effectiveness and robustness of the core modules via ablation studies and parameter sensitivity analysis.

4.1 Experimental Setup

4.1.1 Datasets

To evaluate the effectiveness and generalization of MMCont-Net across diverse domains with varying degrees of irregularity and sparsity, we conduct extensive experiments on four representative benchmark datasets for irregular multivariate time series (IMTS) forecasting:

- **PhysioNet Challenge 2012 [43]:** This dataset consists of electronic health records (EHR) from 12,000 ICU patient stays, featuring 41 irregularly sampled physiological variables (e.g., heart rate, blood pressure) recorded during the first 48 hours of admission. Following standard protocols, we use observations from the initial 24 hours as historical input to forecast vital signs over the subsequent 24 hours.
- **Human Activity [44]:** This dataset captures multimodal sensor readings from wearable devices placed at four body locations (ankles, chest, belt, etc.), yielding 12 variables that represent 3D spatial coordinates and accelerations. Sampling irregularity arises from sensor-specific characteristics and subject motion. We segment sequences into 4000 ms windows and use the first 3000 ms to predict trajectories for the remaining 1000 ms.
- **USHCN (U.S. Historical Climatology Network) [45]:** This large-scale meteorological dataset comprises daily records of 5 core climate variables (e.g., precipitation, maximum/minimum temperature) from stations across the United States, exhibiting substantial missing values due to operational factors. We use historical observations from the past 24 months to forecast the upcoming month's climate indices.
- **MIMIC-III [46]:** This large-scale critical care database contains de-identified records from over 40,000 ICU patients. Following the preprocessing pipeline of Biloš et al. [47], we obtain 23,457 samples with 96 irregularly sampled physiological variables collected over the first 48 hours of admission. Consistent with the PhysioNet protocol, we use the initial 24-hour observations to forecast values in the subsequent 24-hour period.

4.1.2 Evaluation Metrics

Given the inherent sparsity of IMTS data, we adopt masked evaluation metrics, computing prediction errors exclusively on spatiotemporal nodes with valid ground truth observations to avoid interference from padding values. Specifically, we employ Mean Squared Error (MSE) and Mean Absolute Error (MAE):

$$\text{MSE} = \frac{1}{|\Omega|} \sum_{(n,t)\in\Omega} (y_{n,t} - \hat{y}_{n,t})^2$$

$$\text{MAE} = \frac{1}{|\Omega|} \sum_{(n,t)\in\Omega} |y_{n,t} - \hat{y}_{n,t}|$$

4.1.3 Implementation Details

All experiments were conducted on a server equipped with a GeForce RTX 3090 GPU and dual Intel Xeon Silver 4316 CPUs. The model was optimized using the Adam optimizer with an initial learning rate set to 1×10^{-3}. Regarding the patch window size τ, we adopted dataset-specific configurations: 8 hours for PhysioNet, 300ms for Human Activity, and 2 months for USHCN. To mitigate overfitting, early stopping was employed if the validation loss did not decrease for 20 consecutive epochs.

Table 1: Forecasting performance comparison against state-of-the-art baselines across four benchmark datasets. The results are reported in terms of MSE and MAE. Bold indicates the best performance, and underlined denotes the second best.

Algorithm	PhysioNet		Human Activity		USHCN		MIMIC	
	MSE$\times 10^{-3}$	MAE$\times 10^{-2}$	MSE$\times 10^{-3}$	MAE$\times 10^{-2}$	MSE$\times 10^{-1}$	MAE$\times 10^{-1}$	MSE$\times 10^{-2}$	MAE$\times 10^{-2}$
DLinear	41.86±0.05	15.52±0.03	4.03±0.01	4.21±0.01	6.21±0.00	3.88±0.02	4.90±0.00	16.29±0.05
TimesNet	16.48±0.11	6.14±0.03	3.12±0.01	3.56±0.02	5.58±0.05	3.60±0.04	5.88±0.08	13.62±0.07
PatchTST	12.00±0.23	6.02±0.14	4.29±0.14	4.80±0.09	5.75±0.01	3.57±0.02	3.78±0.03	12.43±0.10
Crossformer	6.66±0.11	4.81±0.11	4.29±0.20	4.89±0.17	5.25±0.04	3.27±0.09	2.65±0.10	9.56±0.29
Graph WaveNet	6.04±0.28	4.41±0.11	2.89±0.03	3.40±0.05	5.29±0.04	3.16±0.09	2.93±0.09	10.50±0.15
MTGNN	6.26±0.18	4.46±0.07	3.03±0.03	3.53±0.03	5.39±0.05	3.34±0.02	2.71±0.23	9.55±0.65
StemGNN	6.86±0.28	4.76±0.19	8.81±0.37	6.90±0.02	5.75±0.09	3.40±0.09	1.73±0.02	7.71±0.11
CrossGNN	7.22±0.36	4.96±0.12	3.03±0.10	3.48±0.08	5.66±0.04	3.53±0.05	2.95±0.16	10.82±0.21
FourierGNN	6.84±0.35	4.65±0.12	2.99±0.02	3.42±0.02	5.82±0.06	3.62±0.07	2.55±0.03	10.22±0.08
GRU-D	5.59±0.09	4.08±0.05	2.94±0.05	3.53±0.06	5.54±0.38	3.40±0.28	1.76±0.03	7.53±0.09
SeFT	9.22±0.18	5.40±0.08	12.20±0.17	8.43±0.07	5.80±0.19	3.70±0.11	1.87±0.01	7.84±0.08
RainDrop	9.82±0.08	5.57±0.06	14.92±0.14	9.45±0.05	5.78±0.22	3.67±0.17	1.99±0.03	8.27±0.07
Warpformer	5.94±0.35	4.21±0.12	2.79±0.04	3.39±0.03	5.25±0.05	3.23±0.05	1.73±0.04	7.58±0.13
mTAND	6.23±0.24	4.51±0.17	3.22±0.07	3.81±0.07	5.33±0.05	3.26±0.10	1.85±0.06	7.73±0.13
Latent-ODE	6.05±0.57	4.23±0.26	3.34±0.11	3.94±0.12	5.62±0.03	3.60±0.12	1.89±0.19	8.11±0.52
CRU	8.56±0.26	5.16±0.09	6.97±0.78	6.30±0.47	6.09±0.17	3.54±0.18	1.97±0.02	7.93±0.19
Neural Flow	7.20±0.07	4.67±0.04	4.05±0.13	4.46±0.09	5.35±0.05	3.25±0.05	1.87±0.05	8.03±0.19
T-PATCHGNN	4.98±0.08	3.72±0.03	2.66±0.03	3.15±0.02	5.00±0.04	3.08±0.04	1.69±0.03	7.22±0.09
GraTITi	6.02±0.06	3.73±0.03	2.73±0.03	3.14±0.02	5.09±0.03	3.09±0.04	1.71±0.02	6.94±0.03
TimeCHEAT	5.05±0.08	3.89±0.04	4.06±0.08	4.65±0.04	**4.42±0.04**	3.10±0.04	1.70±0.01	7.40±0.09
HyperIMTS	**4.59±0.03**	4.50±0.04	4.36±0.06	4.12±0.08	5.21±0.03	3.21±0.03	1.72±0.03	7.23±0.04
Ours	4.77±0.05	**3.51±0.02**	**2.60±0.01**	**3.09±0.02**	4.93±0.03	**2.81±0.05**	**1.60±0.02**	**6.72±0.03**

All experiments were repeated five times, and the mean values along with standard deviations are reported.

4.2 Main Results and Benchmark Comparison

4.2.1 Baseline

To establish a thorough benchmark for the under-explored IMTS forecasting task, we incorporate twenty-one relevant baselines for a fair comparison, covering the SOTA models from: (i) MTS forecasting: DLinear [48], TimesNet [18], PatchTST [21], Crossformer [19], GraphwaveNet [10], MTGNN [25], StemGNN [26], CrossGNN [49], and FourierGNN [50]; (ii) IMTS classification: GRU-D [14], SeFT [51], RainDrop [3], and Warpformer [52]; (iii) IMTS interpolation: mTAND [2]; and (iv) IMTS forecasting: Latent-ODE [6], CRU [53], Neural Flows [54], T-PATCHGNN [22], GraTITi [11], TimeCHEAT [55], and HyperIMTS [56].

4.2.2 Performance Insights

As evidenced in Table 1, MMCont-Net consistently outperforms all baselines, achieving the top rank in 6 out of 8 evaluation metrics across the four benchmark datasets.

- **Superiority over General SOTA Methods.** Compared with general-purpose state-of-the-art approaches such as PatchTST, MMCont-Net delivers substantial gains, particularly on the large-scale MIMIC-III dataset, where it reduces mean squared error (MSE) by approximately 57% relative to PatchTST. This marked improvement underscores the drawbacks of methods that forcibly align irregular observations to regular grids—a common prerequisite for channel-independent models—which often incurs sequence-length explosion and severe interpolation artifacts. In contrast, the proposed transformable patching strategy (Section 3.1) natively preserves physical temporal continuity without extensive imputation, enabling more faithful micro-level intra-patch context extraction.
- **Advantage over Specialized IMTS Baselines.** Even against the strongest IMTS-specific baseline (e.g., T-PATCHGNN), MMCont-Net maintains clear superiority on the highly challenging MIMIC-III dataset. Although both approaches employ patch-based representations, the proposed full-scale context modeling—integrating frequency-assisted gated temporal processing (FAGTU), continuous-time cross-patch attention (HTA), and dynamic graph topology diffusion—more effectively mitigates microscopic spectral blindness and macroscopic spatial isolation, reconstructing coherent spatiotemporal dependencies from severely asynchronous observations.
- **Robustness and Generalization.** While certain baselines (e.g., HyperIMTS on PhysioNet and TimeCHEAT on USHCN) occasionally achieve marginally lower MSE in isolated cases, MMCont-Net consistently records the lowest mean absolute error (MAE) across *all* four datasets. This superior MAE performance reflects greater robustness to outliers and extreme values, affirming the stability and generalization of the proposed progressive full-scale context integration paradigm across diverse domains, irregularity levels, and sparsity regimes.

4.3 Ablation Studies

To quantitatively evaluate the contribution of each constituent module of MMCont-Net, we devised a comprehensive set of variants. As shown in Table 2, these variants are categorized into three groups based on the structural hierarchy of the model:

4.3.1 Group (a): Analysis of Micro-level FAGTU

- **r/p FAGTU:** We substitute the entire Fourier-Assisted Gated Temporal Unit (FAGTU) with Masked Global Average Pooling (MGAP). This variant validates the efficacy of the spectral-temporal dual-domain feature extraction mechanism in generating high-dimensional semantic embeddings from variable-length sequences.
- **w/o Fourier:** We eliminate only the Fourier augmented branch within FAGTU, reducing it to a standard Gated Linear Unit (GLU). This setup assesses the specific contribution of frequency-domain information to global periodicity modeling and noise suppression.

Table 2: Ablation study of key modules within MMCont-Net. The table reports the performance impact of removing or replacing major components such as FAGTU, HTA, and DGDN across three datasets.

Ablation		PhysioNet		Human Activity		USHCN		MIMIC	
		MSE×10^{-3}	MAE×10^{-2}	MSE×10^{-3}	MAE×10^{-2}	MSE×10^{-1}	MAE×10^{-1}	MSE×10^{-2}	MAE×10^{-2}
MMCont-Net (Full)		**4.77±0.05**	**3.51±0.02**	**2.60±0.01**	**3.09±0.02**	**4.93±0.05**	**2.81±0.05**	**1.60±0.02**	**6.72±0.03**
(a)	r/p FAGTU	5.03±0.07	3.66±0.07	2.65±0.01	3.15±0.02	4.95±0.02	2.97±0.15	1.62±0.02	6.76±0.01
	w/o Fourier	4.98±0.10	3.68±0.08	2.62±0.02	3.13±0.02	4.96±0.02	2.87±0.03	1.63±0.03	6.75±0.03
(b)	r/p HTA	5.12±0.08	3.66±0.05	2.63±0.02	3.14±0.01	4.97±0.02	2.83±0.02	1.72±0.04	6.81±0.03
	w/o LTE	5.31±0.14	3.83±0.11	2.66±0.02	3.14±0.01	4.98±0.02	2.92±0.06	1.76±0.06	6.86±0.05
(c)	w/o DGDN	10.53±0.04	5.16±0.01	2.79±0.01	3.14±0.03	5.09±0.02	2.88±0.08	1.86±0.07	6.97±0.02
	w/o $\mathbf{A}_{\text{dat}}$	5.76±0.04	3.98±0.10	2.77±0.02	3.14±0.02	4.98±0.03	2.96±0.09	1.70±0.03	6.83±0.02
	w/o $\mathbf{A}_{\text{adp}}$	10.25±0.07	5.08±0.05	2.73±0.01	3.11±0.01	5.03±0.03	3.06±0.11	1.80±0.06	6.92±0.05
	w/o K-step	5.89±0.05	4.12±0.03	2.76±0.03	3.13±0.02	5.00±0.04	2.89±0.08	1.75±0.05	6.85±0.03

4.3.2 Group (b): Analysis of Continuous Temporal Modeling

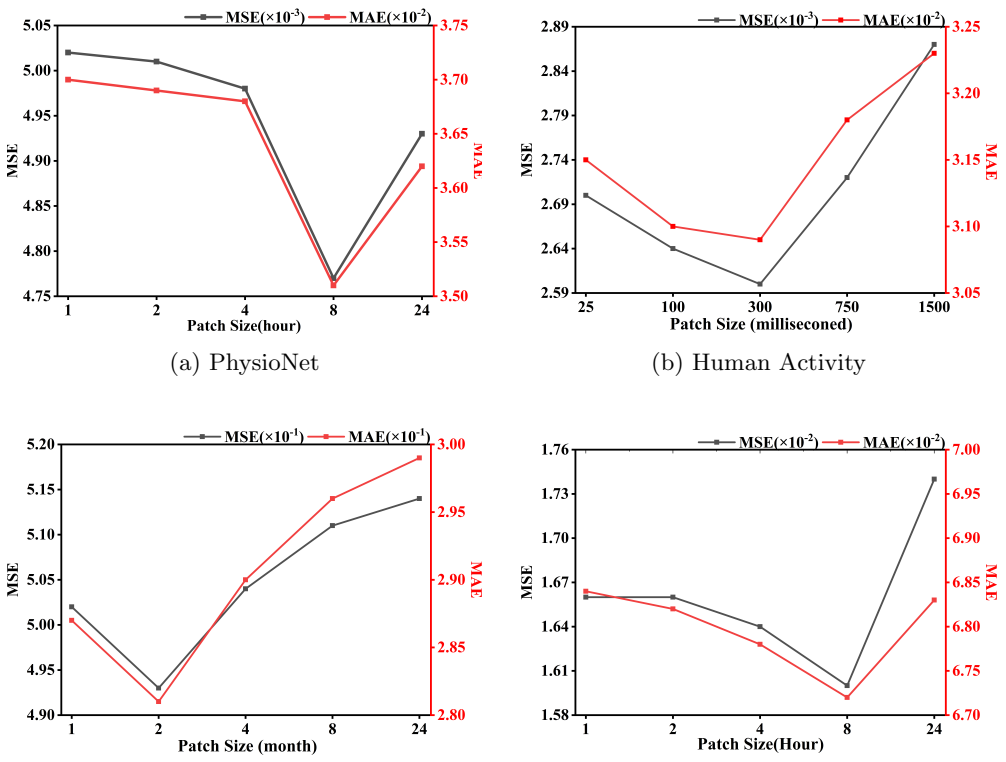
- **r/p HTA:** The Hierarchical Continuous-time Attention (HTA) is replaced by standard Multi-Head Attention. This variant analyzes the impact of explicitly encoding physical time intervals Δt on long-range dependency modeling compared to discrete position-based attention.
- **w/o LTE:** We exclude both Global and Local Learnable Time Embeddings (LTE) from the HTA module. This evaluates the critical role of absolute temporal awareness in modeling periodic and non-uniformly sampled data.

4.3.3 Group (c): Analysis of Macro-level DGDN

- **w/o DGDN:** We remove the entire Dynamic Graph Diffusion Network (DGDN), thereby rendering the variables informationally isolated. This allows us to quantify the overall contribution of cross-variable spatial interactions.
- **w/o $\mathbf{A}_{dat}$:** We substitute the data-driven dynamic graph in DGDN with a static graph structure by removing the attention-based adjacency matrix $\mathbf{A}_{dat}$, retaining only the adaptive prior matrix $\mathbf{A}_{adp}$. This verifies the necessity of capturing time-varying correlations induced by instantaneous system states.
- **w/o $\mathbf{A}_{adp}$:** We remove the static prior branch (the adaptive adjacency $\mathbf{A}_{adp}$) and the gating fusion mechanism, preserving only the data-driven dynamic feature graph $\mathbf{A}_{dat}$. This analyzes the complementary role of static constraints (e.g., spatial proximity) in stabilizing the dynamic graph learning.
- **w/o K-step:** We set the diffusion step $K = 0$ to bypass the multi-hop information propagation process while retaining the graph structure construction. This isolates the impact of deep neighborhood aggregation and evaluates whether immediate neighbor information suffices for spatial reasoning.

4.4 Parameter Sensitivity Analysis

To investigate the impact of hyperparameter configurations on the performance of MMCont-Net, we conducted comprehensive sensitivity experiments on three datasets, focusing on two pivotal hyperparameters: the patch window size (Patch Size, L) and



(a) PhysioNet (b) Human Activity

(c) USHCN (d) MIMIC

Fig. 2: Effect of different patch sizes across four datasets.

the hidden dimension (D). Figures 2 and 3 illustrate the trends of MSE and MAE under varying parameter settings, respectively.

4.4.1 Impact of Patch Window Size

The window size L dictates the physical temporal resolution of the transformation patching strategy, thereby influencing the model's capability to capture local temporal semantics. The experimental results (Fig. 2) demonstrate the following:

- **PhysioNet (Optimal $\tau \approx 8$ hours):** Performance peaks around an 8-hour window. Physiological signals typically exhibit clinically meaningful cycles and trends over several hours (e.g., circadian influences on heart rate and blood pressure). Shorter windows (e.g., 2–4 hours) fragment these patterns, leading to loss of local semantic coherence, while excessively long windows dilute sensitivity to acute events.
- **Human Activity (Optimal $\tau \approx 300$ ms):** The lowest error occurs at very fine-grained windows of approximately 300 ms. This aligns with the high-frequency nature of motion sensor data, where rapid postural transitions and instantaneous

accelerations dominate; micro-level patches of this scale best capture transient dynamics critical for activity recognition.

- **USHCN (Optimal $\tau \approx 2$ months):** The model achieves optimal performance with 2-month windows. Long-term climate processes are governed by seasonal cycles, and this macro-scale granularity effectively encodes persistent trends while filtering short-term meteorological noise.
- **MIMIC-III (Optimal $\tau \approx 8$ hours):** Similar to PhysioNet, performance is strongest in the 8-hour range. The large number of variables (96) and severe irregularity in critical-care monitoring benefit from windows that balance detailed intra-patch spectral sensing with sufficient observational density to mitigate sparsity-induced distortion.

Overall, patch size selection reflects a trade-off between temporal resolution and semantic integrity. The optimal τ is inherently task-dependent, tied to the intrinsic timescales and sampling characteristics of each domain. The transformable patching strategy enables flexible adaptation to these diverse regimes without imputation artifacts.

4.4.2 Impact of Hidden Dimension

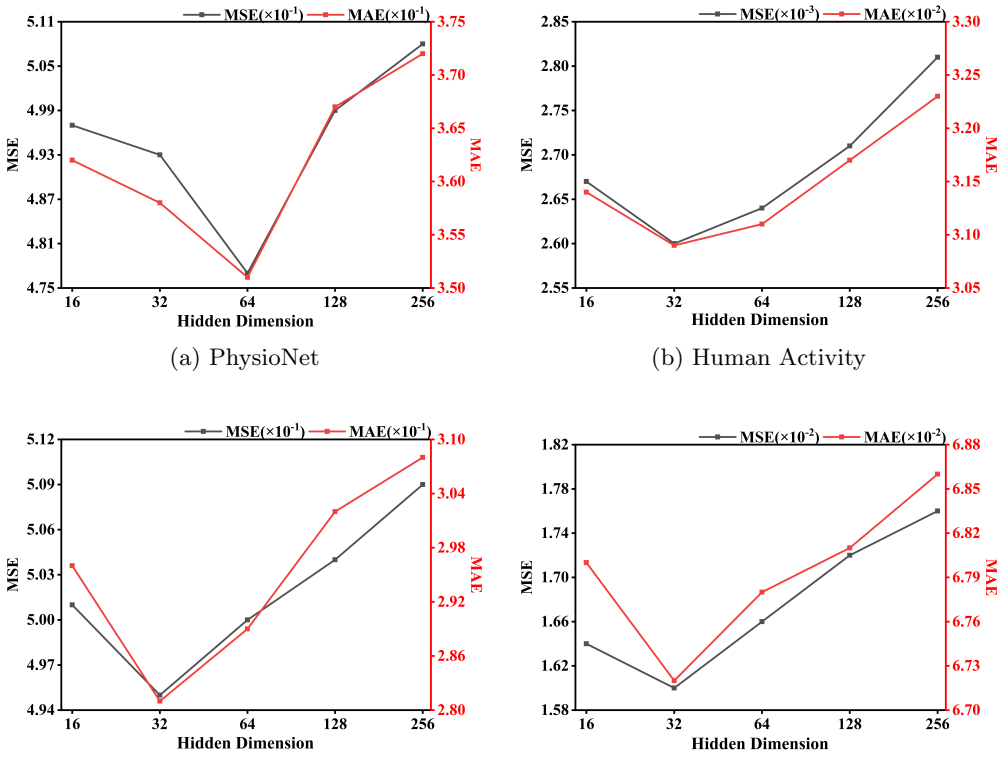
The hidden dimension D governs the feature representation capacity of the model. The experimental results (Fig. 3) reveal:

- **PhysioNet (Optimal $D = 64$):** This dataset, with 41 variables and complex physiological interdependencies across 12,000 samples, benefits distinctly from larger representational capacity. Performance improves consistently as D increases, reaching optimal results around 64.
- **MIMIC-III, Human Activity, and USHCN (Optimal $D = 32$):** In contrast, these three datasets achieve peak performance at the relatively modest dimension $D = 32$. Further increasing D (e.g., to 64, 128, or 256) yields no improvement and often leads to higher error, indicative of overfitting due to over-parameterization on these systems—despite MIMIC-III's high variable count (96), its severe sparsity and irregularity appear to favor more constrained capacity to avoid fitting noise.

These results highlight that MMCont-Net's hidden dimension should be tuned according to dataset characteristics: larger dimensions are advantageous for datasets exhibiting dense, complex inter-variable couplings (as in PhysioNet), while moderate dimensions are preferable for highly sparse or lower-effective-dimensionality regimes to maintain generalization and prevent overfitting.

4.5 Efficiency Analysis

Fig. 4 shows that MMCont-Net achieves substantially higher computational efficiency than all competing methods. ODE-based models (e.g., Latent-ODE and CRU) incur prohibitive costs, with per-epoch training time on the order of 10^3 seconds due to repeated numerical integration. In contrast, MMCont-Net reduces training time by one to two orders of magnitude, enabled by transformable patching that avoids point-wise



(a) PhysioNet (b) Human Activity

(c) USHCN (d) MIMIC

Fig. 3: Effect of different hidden dimensions.

recurrence. Compared to pre-alignment RNN methods (e.g., Raindrop and Warpformer), MMCont-Net effectively mitigates sequence length explosion as the number of variables increases. Moreover, even relative to the strongest graph-based baseline T-PatchGNN, MMCont-Net achieves faster training and lower inference latency. These results demonstrate that MMCont-Net offers an excellent trade-off between modeling capacity and computational efficiency for large-scale IMTS forecasting.

5 Conclusion

In conclusion, irregular multivariate time series (IMTS) forecasting is severely constrained by microscopic spectral blindness and macroscopic spatial isolation arising from non-uniform sampling and cross-variable asynchrony. To address these fundamental challenges, we propose MMCont-Net, a unified framework that pioneers progressive full-scale context modeling: it first captures micro-level intra-patch context through spectral-temporal dual-branch processing, then models cross-patch dependencies via continuous physical-time aware attention, and finally reconstructs macro-level inter-variable interactions using dynamic graph topology diffusion with multi-hop

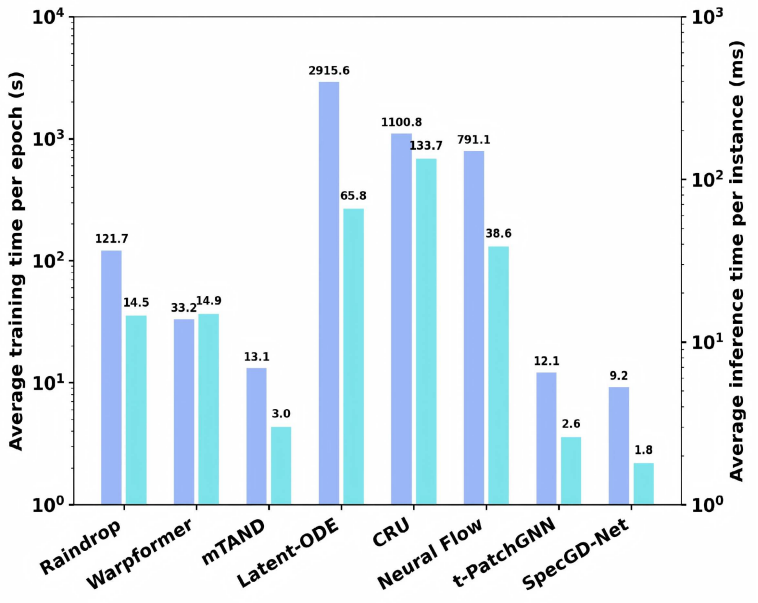


Fig. 4: Efficiency comparison of training and inference.

propagation. By coherently integrating context across all scales without relying on extensive imputation or synchronized assumptions, MMCont-Net faithfully recovers coherent system dynamics from sparse, asynchronous observations. Extensive experiments on benchmark datasets including PhysioNet, Human Activity, USHCN, and MIMIC-III consistently demonstrate superior forecasting accuracy, computational efficiency, and robustness over state-of-the-art baselines. This work establishes an effective new paradigm for IMTS modeling in complex real-world systems, with promising future extensions to multi-modal irregular signals and continuous-time generative tasks.

References

[1] Lim, B., Zohren, S.: Time-series forecasting with deep learning: a survey. Philosophical transactions of the royal society a: mathematical, physical and engineering sciences **379**(2194) (2021)

[2] Shukla, S.N., Marlin, B.M.: Multi-time attention networks for irregularly sampled time series. arXiv preprint arXiv:2101.10318 (2021)

[3] Zhang, X., Zeman, M., Tsiligkaridis, T., Zitnik, M.: Graph-guided network for irregularly sampled multivariate time series. arXiv preprint arXiv:2110.05357 (2021)

[4] Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J.,

Liu, X., Marcus, J., Sun, M., *et al.*: Scalable and accurate deep learning with electronic health records. NPJ digital medicine **1**(1), 18 (2018)

[5] Kidger, P., Morrill, J., Foster, J., Lyons, T.: Neural controlled differential equations for irregular time series. Advances in neural information processing systems **33**, 6696–6707 (2020)

[6] Rubanova, Y., Chen, R.T., Duvenaud, D.K.: Latent ordinary differential equations for irregularly-sampled time series. Advances in neural information processing systems **32** (2019)

[7] Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in neural information processing systems **34**, 22419–22430 (2021)

[8] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., Jin, R.: Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In: International Conference on Machine Learning, pp. 27268–27286 (2022). PMLR

[9] Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875 (2017)

[10] Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121 (2019)

[11] Yalavarthi, V.K., Madhusudhanan, K., Scholz, R., Ahmed, N., Burchert, J., Jawed, S., Born, S., Schmidt-Thieme, L.: Grafiti: Graphs for forecasting irregularly sampled time series. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 16255–16263 (2024)

[12] Cao, W., Wang, D., Li, J., Zhou, H., Li, L., Li, Y.: Brits: Bidirectional recurrent imputation for time series. Advances in neural information processing systems **31** (2018)

[13] Yoon, J., Jordon, J., Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: International Conference on Machine Learning, pp. 5689–5698 (2018). PMLR

[14] Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Scientific reports **8**(1), 6085 (2018)

[15] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)

[16] Kazemi, S.M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S.,

Smyth, C., Poupart, P., Brubaker, M.: Time2vec: Learning a vector representation of time. arXiv preprint arXiv:1907.05321 (2019)

[17] Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., Li, D.: Contiformer: Continuous-time transformer for irregular time series modeling. Advances in Neural Information Processing Systems **36**, 47143–47175 (2023)

[18] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M.: Timesnet: Temporal 2d-variation modeling for general time series analysis. arXiv preprint arXiv:2210.02186 (2022)

[19] Zhang, Y., Yan, J.: Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: The Eleventh International Conference on Learning Representations (2023)

[20] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625 (2023)

[21] Nie, Y.: A time series is worth 64words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730 (2022)

[22] Zhang, W., Yin, C., Liu, H., Zhou, X., Xiong, H.: Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In: Forty-first International Conference on Machine Learning (2024)

[23] Chen, P., Zhang, Y., Cheng, Y., Shu, Y., Wang, Y., Wen, Q., Yang, B., Guo, C.: Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. arXiv preprint arXiv:2402.05956 (2024)

[24] Chowdhury, R.R., Li, J., Zhang, X., Hong, D., Gupta, R.K., Shang, J.: Primenet: Pre-training for irregular multivariate time series. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 7184–7192 (2023)

[25] Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: Connecting the dots: Multivariate time series forecasting with graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 753–763 (2020)

[26] Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., *et al.*: Spectral temporal graph neural network for multivariate time-series forecasting. Advances in neural information processing systems **33**, 17766–17778 (2020)

[27] Jin, M., Zheng, Y., Li, Y.-F., Chen, S., Yang, B., Pan, S.: Multivariate time series forecasting with dynamic graph neural odes. IEEE Transactions on Knowledge and Data Engineering **35**(9), 9168–9180 (2022)

[28] Choi, J., Choi, H., Hwang, J., Park, N.: Graph neural controlled differential equations for traffic forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 6367–6374 (2022)

[29] Tashiro, Y., Song, J., Song, Y., Ermon, S.: Csdi: Conditional score-based diffusion models for probabilistic time series imputation. Advances in neural information processing systems **34**, 24804–24816 (2021)

[30] Liu, M., Huang, H., Feng, H., Sun, L., Du, B., Fu, Y.: Pristi: A conditional diffusion framework for spatiotemporal imputation. In: 2023 IEEE 39th International Conference on Data Engineering (ICDE), pp. 1927–1939 (2023). IEEE

[31] Lim, B., Arık, S.Ö., Loeff, N., Pfister, T.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. International journal of forecasting **37**(4), 1748–1764 (2021)

[32] Zhang, D., Li, B., Zhao, Z., Nie, F., Gao, J., Li, X.: Fusad: Time-frequency fusion with adaptive denoising for general time series analysis. arXiv preprint arXiv:2512.14078 (2025)

[33] Li, Z., Tan, Y., Cheng, X., Qin, Y.: Ftmixer: Frequency and time domain representations fusion for time series forecasting. IEEE Signal Processing Letters (2025)

[34] Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International Conference on Machine Learning, pp. 3744–3753 (2019). PMLR

[35] Chen, X., Zheng, X., Mollaysa, A., Schürch, M., Allam, A., Krauthammer, M.: Two-stage aggregation with dynamic local attention for irregular time series. arXiv preprint arXiv:2311.07744 (2023)

[36] Kim, B., Lee, J.-G.: Continuous-time linear positional embedding for irregular time series forecasting. arXiv preprint arXiv:2409.20092 (2024)

[37] Yu, J., Feng, Z., Lu, J., Cai, T., Zhou, D.: Time-aware attention for enhanced electronic health records modeling. arXiv preprint arXiv:2507.14847 (2025)

[38] Hajisafi, A., Siampou, M.D., Azarijoo, B., Shahabi, C.: Wavegnn: Modeling irregular multivariate time series for accurate predictions. arXiv preprint arXiv:2412.10621 (2024)

[39] He, Z., Xu, L., Yu, J., Wu, X.: Dynamic multi-fusion spatio-temporal graph neural network for multivariate time series forecasting. Expert Systems with Applications **241**, 122729 (2024)

[40] Gasteiger, J., Weißenberger, S., Günnemann, S.: Diffusion improves graph

learning. Advances in neural information processing systems **32** (2019)

[41] Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926 (2017)

[42] Luo, X., Zhu, C., Zhang, D., Li, Q.: Dynamic graph convolutional network with attention fusion for traffic flow prediction. arXiv preprint arXiv:2302.12598 (2023)

[43] Silva, I., Moody, G., Scott, D.J., Celi, L.A., Mark, R.G.: Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In: 2012 Computing in Cardiology, pp. 245–248 (2012). IEEE

[44] Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Tr"oster, G., Mill'an, J.d.R., Roggen, D.: The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. Pattern Recognition Letters **34**(15), 2033–2042 (2013)

[45] Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G.: An overview of the global historical climatology network-daily database. Journal of Atmospheric and Oceanic Technology **29**(7), 897–910 (2012)

[46] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)

[47] Biloš, M., Sommer, J., Rangapuram, S., Januschowski, T., Günnemann, S.: Neural flows: Efficient alternative to neural odes. arXiv: Learning,arXiv: Learning (2021)

[48] Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 11121–11128 (2023)

[49] Huang, Q., Shen, L., Zhang, R., Ding, S., Wang, B., Zhou, Z., Wang, Y.: Cross-gnn: Confronting noisy multivariate time series via cross interaction refinement. Advances in Neural Information Processing Systems **36**, 46885–46902 (2023)

[50] Yi, K., Zhang, Q., Fan, W., He, H., Hu, L., Wang, P., An, N., Cao, L., Niu, Z.: Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. Advances in neural information processing systems **36**, 69638–69660 (2023)

[51] Horn, M., Moor, M., Bock, C., Rieck, B., Borgwardt, K.: Set functions for time series. In: International Conference on Machine Learning, pp. 4353–4363 (2020). PMLR

[52] Zhang, J., Zheng, S., Cao, W., Bian, J., Li, J.: 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)

[53] Schirmer, M., Eltayeb, M., Lessmann, S., Rudolph, M.: Modeling irregular time series with continuous recurrent units. In: International Conference on Machine Learning, pp. 19388–19405 (2022). PMLR

[54] Biloš, M., Sommer, J., Rangapuram, S.S., Januschowski, T., Günnemann, S.: Neural flows: Efficient alternative to neural odes. Advances in neural information processing systems **34**, 21325–21337 (2021)

[55] Liu, J., Cao, M., Chen, S.: Timecheat: A channel harmony strategy for irregularly sampled multivariate time series analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 18861–18869 (2025)

[56] Li, B., Luo, Y., Liu, Z., Zheng, J., Lv, J., Ma, Q.: Hyperimts: Hypergraph neural network for irregular multivariate time series forecasting. arXiv preprint arXiv:2505.17431 (2025)