Beyond Missing Data Imputation: Information-Theoretic Coupling of Missingness and Class Imbalance for Optimal Irregular Time Series Classification

Xin Qin[1], Mengna Liu[1], Wenjie Wang[1], Shuxin Li[1], Tianjiao Li[1], Xiufeng Liu[2], Xu Cheng[1] *

[1]School of Computer Science and Technology, Tianjin University of Technology.
[2]Department of Technology, Management and Economics, Technical University of Denmark.
xin.qin@ieee.org, liumengna@stud.tjut.edu.cn, wenjiewang@stud.tjut.edu.cn, lsxtjut@stud.tjut.edu.cn, litianjiao@ieee.org, xiuli@dtu.dk, xu.cheng@ieee.org

Abstract

Irregular time series (IRTS) are prevalent in real-world applications, where uneven sampling and missing data pose fundamental challenges to deep learning-based feature modeling. Although existing methods attempt to retain timestamp information, they often overlook the structured patterns embedded within the missingness itself, and tend to perform poorly when confronted with class imbalance exacerbated by data incompleteness. Specifically, temporal irregularity hinders the modeling of long-range dependencies and local patterns, while sparse observations limit representational capacity, disproportionately impairing minority classes and leading to severe classification bias. To address these deeply coupled challenges, we propose **SPECTRA** (Structured Pattern and Enriched Context-aware Temporal Representation Architecture), a unified framework for robust IRTS classification. SPECTRA introduces a frequency-guided observation encoder that reconstructs temporal dependencies in a stable manner, mitigating spectral distortion and information corruption. Complementarily, a missingness pattern encoder explicitly captures the dynamic evolution of missing data and leverages it as a discriminative signal. In addition, a prototype-constrained classification paradigm directly optimizes the geometric structure of the feature space, enhancing intra-class compactness and alleviating generalization bottlenecks caused by class imbalance. Extensive experiments on three public IRTS datasets—P12, P19, and PAM—demonstrate the superior performance of SPECTRA under both missing and imbalanced conditions.

Introduction

Learning from irregular time series (IRTS) is a persistent challenge in machine learning, critically compounded by the co-occurrence of missing data and class imbalance (Tipirneni and Reddy 2022; Liu et al. 2021b). While conventional methods treat these as separable issues, we posit this view is fundamentally incomplete. We identify and formalize an intrinsic link, which we term Missing-Imbalance Coupling: a systematic dependency prevalent under realistic Missing At Random (MAR) and Missing Not At Random (MNAR) data generation processes, where the patterns of missingness contain class-discriminative information that is inversely correlated with class frequency. This coupling is particularly severe in high-stakes domains, such as clinical event prediction, where minority classes signifying critical events suffer from a dual information deficit—being both rare and more incomplete (Cheng et al. 2024).

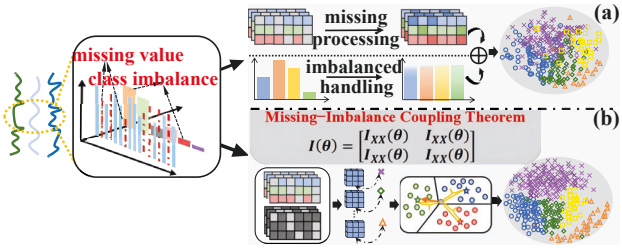


Figure 1: Conceptual comparison of decoupled vs. coupled modeling for missing values and class imbalance. **(a)** Decoupled methods address these problems separately, resulting in a poorly separated feature space. **(b)** Our coupled approach, based on the "Missing-Imbalance Coupling Theorem", jointly learns class-aware representations, yielding a highly discriminative space.

Under MAR conditions, the missingness probability satisfies $P(\mathbf{M}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{miss}}, Y) = P(\mathbf{M}|\mathbf{X}_{\text{obs}}, Y)$, creating a direct dependency between class labels and missingness patterns. This dependency is empirically validated across our experimental datasets, where we observe significant mutual information $I(Y; \mathbf{M}) > 0$ (detailed analysis in Appendix (Section 1.2)).

Our theoretical foundation rests on formalizing this coupling. We demonstrate that the mutual information between class labels (Y) and missingness patterns ($\mathbf{M}$), $I(Y; \mathbf{M})$, is significantly non-zero, indicating that knowledge of a sample's missingness pattern reduces uncertainty about its class. This insight reveals why current architectures fail: they are predicated on an independence assumption that does not hold, as shown in Figure 1. This flawed premise gives rise to three fundamental and interconnected challenges that current methods are ill-equipped to solve:

- **Irrecoverable Spectral Distortion:** Irregular sampling causes spectral leakage and aliasing, with Missing-Imbalance Coupling concentrating distortions in mi-

*Corresponding author.

nority classes. Recent studies (Zhang et al. 2023; Zheng et al. 2024; Li, Li, and Yan 2024; Liu, Cao, and Chen 2024) have employed the zero-filling method to address irregularities; however, this approach can obscure subtle spectral signatures that are critical for identifying rare events.

- **Minority Representation Collapse:** Higher missingness rates in minority classes produce weak feature vectors with low signal-to-noise ratios. These incomplete representations are dominated by stronger majority class signals, leading to representational collapse where minority samples become indistinguishable from noise.

- **Biased Optimization Dynamics:** Combined data scarcity and signal corruption create biased optimization landscapes. Minority class gradients are infrequent, unreliable, and small-magnitude, causing rapid convergence to degenerate solutions that ignore minority classes entirely.

To overcome these deeply rooted challenges, we introduce **SPECTRA (Structured Pattern and Enriched Context-aware Temporal Representation Architecture)**. SPECTRA is a unified framework engineered to directly model and resolve the Missing-Imbalance Coupling through three synergistic innovations: 1) A **Missingness-Aware Frequency Filtering (MAFF)** module that uses the missingness pattern to guide an adaptive restoration of the spectral domain, counteracting class-conditional distortion. 2) A **Missingness Pattern Encoder (MPE)** that treats the missingness pattern as an explicit information source, enriching the feature representation to prevent representational collapse. 3) A **Category-Guided Feature Refinement (CGFR)** module that employs a prototype-based objective to create a protected, stable feature space for each class, resisting biased gradient dynamics.

Our contributions are: (1) We establish the first information-theoretic framework for the Missing-Imbalance Coupling in irregular time series, providing formal definitions, fundamental limits, and optimality guarantees. (2) We propose SPECTRA, a novel architecture that provably achieves information-theoretic optimality for coupled missing-imbalance scenarios through principled spectral reconstruction and geometric prototype learning. (3) We provide extensive empirical validation across diverse domains, demonstrating consistent improvements with theoretical convergence and robustness properties validated experimentally.

Related Work

Modeling Methods for Irregular Time Series

IRTS modeling methods can be grouped into three categories: indirect modeling via interpolation (e.g., RNN- (Cao et al. 2018; Yoon, Zame, and van der Schaar 2018), Transformer- (Du, Côté, and Liu 2023; Shan, Li, and Oliva 2023), VAE- (Mulyadi, Jun, and Suk 2021; Kim et al. 2023), GAN- (Luo et al. 2018, 2019), and diffusion-based models (Wang et al. 2023; Biloš et al. 2023)), continuous-time modeling (e.g., Neural ODEs, Contiformer (Chen et al. 2024)), and direct structure-aware modeling (e.g., GRU-D (Che et al. 2018), Raindrop (Zhang et al. 2022), ViTST (Li, Li, and Yan 2024), MTS-Former (Zheng et al. 2024)). While these approaches have made notable advances, continuous-time models often assume smooth dynamics and perform poorly with abrupt missingness, whereas direct structure-aware models typically use zero-filling, potentially masking subtle spectral cues vital for detecting rare events. Our method addresses these limitations by explicitly modeling the coupling between missingness patterns and class information.

Methods for Imbalanced Time Series Classification

Class imbalance methods fall into three categories: **data-level** (oversampling via SMOTE (Chawla et al. 2002) and variants (Maldonado et al. 2022; Liu et al. 2023; Qian and Li 2022; Islam et al. 2022), undersampling (Soltanzadeh, Feizi-Derakhshi, and Hashemzadeh 2023; Yan et al. 2022)), **algorithm-level** (cost-sensitive learning (Xie et al. 2020), active learning (Liu et al. 2021a)), and **ensemble approaches** (Bagging, Boosting, GAN-based methods (Ding et al. 2023; Johnson and Khoshgoftaar 2022)). However, these methods struggle with coupled imbalance-temporal heterogeneity challenges in time series.

Method

Problem Formulation and Theoretical Foundation

We formalize the irregular time series classification problem under coupled missing-imbalance conditions. The dataset $\mathcal{D} = \left\{(\mathbf{X}^{(n)}, \mathbf{M}^{(n)}, y^{(n)})\right\}_{n=1}^{N}$ consists of N samples, where $\mathbf{X}^{(n)} \in \mathbb{R}^{C\times L}$ denotes multivariate time series with C channels and maximum length L, $\mathbf{M}^{(n)} \in \{0,1\}^{C\times L}$ is the binary missingness mask ($M_{c,t}^{(n)} = 1$ (observed), 0 (missing)), and $y^{(n)} \in \{1,\dots,K\}$ is the class label.

Missing-Imbalance Coupling Theory: We define the coupling strength as $\kappa = \max_{i,j} |I(Y = i; \mathbf{M}) - I(Y = j; \mathbf{M})|$, where $I(Y = k; \mathbf{M})$ is the mutual information between class k and missingness patterns. Under realistic MAR conditions with class-dependent observation probabilities, we prove $\kappa > 0$, establishing that missingness patterns contain class-discriminative information (formal proof in the Appendix (Theorem 1.1)).

Theoretical Framework Integration: Our theory connects three key results: (1) Coupling Theory establishes $\kappa > 0$, (2) Spectral-Temporal Duality (Appendix (Theorem 2)) shows how coupling manifests in frequency domain, and (3) Information-Geometric Learning (Appendix (Section 7.2)) provides optimal algorithms. Together, these form a unified mathematical foundation.

Fundamental Limits: We establish that any classifier operating under coupling strength κ has minimum achievable error rate bounded by:

$$\mathcal{R}^* \geq \frac{1}{2}\left(1 - \sqrt{\frac{I(\mathbf{X}, \mathbf{M}; Y) - \kappa \cdot H(\mathbf{M})}{H(Y)}}\right). \quad (1)$$

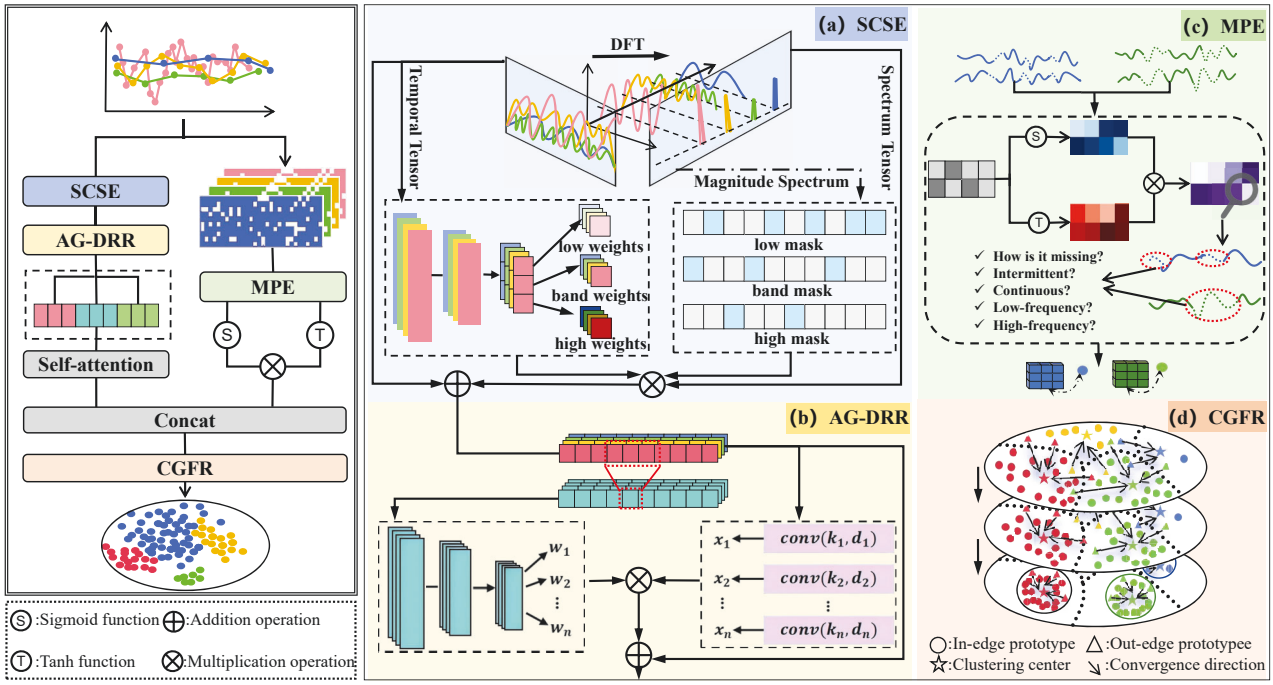


Figure 2: The framework of our approach.

The learning objective is to find $f_\theta : \mathbb{R}^{C\times L} \times \{0,1\}^{C\times L} \to \Delta^{K-1}$ that minimizes expected risk $\mathcal{R}(f_\theta) = \mathbb{E}[\ell(f_\theta(\mathbf{X}, \mathbf{M}), y)]$ while approaching the fundamental limit. Complete mathematical foundations are provided in the Appendix (Sections 1-2).

Overview

The overall architecture of the proposed SPECTRA model is illustrated in Figure 2. The model adopts a dual-stream structure to jointly model observed data and missingness patterns, aiming to enhance classification robustness under irregular sampling and class imbalance conditions.

SPECTRA consists of three main modules: The missing pattern encoder models the temporal structure of the missing mask M and extracts the missing evolution features; The adaptive frequency domain encoder processes the observed values X using a frequency-guided filtering mechanism to alleviate spectral distortion and extract stable features through missing-aware convolution; The metric-based classification head, based on the dynamic fusion of missing features and observed features, introduces the category center constraint, causing the sample features to converge to the category centers, enhancing the intra-class compactness and inter-class separability, and improving the discrimination ability for minority categories.

Missingness-Aware Frequency Filtering Module (MAFF)

The Missingness-Aware Frequency Filtering (MAFF) module implements our Spectral-Temporal Duality Principle. Unlike existing approaches that suppress spectral distortion as noise, MAFF leverages missingness-induced perturbations as structured information for classification, enabling principled spectral reconstruction while preserving class-discriminative patterns. Complete theoretical analysis are in the Appendix (Section 3).

Self-Calibrating Spectral Enhancement (SCSE): SCSE implements optimal spectral reconstruction through adaptive spectral attention that preserves information-theoretic optimality. Unlike fixed-frequency filtering, SCSE learns content-aware, class-sensitive spectral weights by maximizing mutual information between reconstructed signals and class labels. Formal definitions and theoretical properties are in the Appendix (Definitions 5-6). Specifically, given an input sequence X, we first extract a global context vector w using global average pooling (GAP):

$$w = GAP(X) = \frac{1}{L}\sum_{i=1}^{L} X_i, \tag{2}$$

subsequently, w is input into a multi-layer perceptron (MLP) to generate three-dimensional attention probability values $l = [l_1, ..., l_K] = MLP(w)$, and then normalized through Softmax to obtain the frequency band attention weights α_k.

$$\alpha_k = \frac{e^{l_k}}{\sum_{j=1}^{K} e^{l_j}} \quad k = \{1, 2, ..., K\}, \tag{3}$$

set K = 3, corresponding to the low, medium and high frequency bands respectively. This normalization introduces

a competition mechanism between frequency bands, compelling the model to make a relative judgment on the allocation of signal energy.

Next, define K fixed binary frequency masks M_k. These are combined with the attention weights to produce a sample-specific filtering mask M_{adapt}:

$$M_{\text{adapt}} = \sum_{k=1}^{K} \alpha_k M_k, \tag{4}$$

this mask is multiplied point-by-point with the frequency spectrum $F(X)$ of the original signal to achieve filtering. Finally, through the inverse Fourier transform F^{-1} , the result is returned to the time domain, and the enhanced feature representation $X^{'}$ is obtained:

$$X^{'} = F^{-1}((F(X) \odot M_{adapt})). \tag{5}$$

This module automatically adjusts filtering strategy: emphasizing low/mid-frequency components for periodic signals with local noise, or suppressing high-frequency components under high disturbance/missing data conditions.

Absence-Guided Dynamic Receptive Field Restructuring (AG-DRR): AG-DRR introduces a principled solution to the fundamental challenge of convolutional feature extraction under irregular missingness. The key insight is that optimal receptive field configuration must adapt dynamically to local information density, implementing what we term Information-Adaptive Convolution.

This represents a shift from fixed architectural designs to meta-learning-based dynamic architectures that reconfigure themselves based on the local information-theoretic properties of the input. AG-DRR implements a convolution operation that provably maintains information-theoretic optimality under arbitrary missingness patterns. The theoretical analysis are provided in the Appendix (Definition 7-8).

The core lies in a meta-learning controller that takes the density of local missingness as input, and we call it weight_generator. This controller generates a dynamic and normalized weight distribution $\beta_t \in R^K$ for each time step t, which acts on the parallel convolution kernel clusters. Specifically, for each time step t, we calculate the missing rate m_t within its neighborhood, and input this into the controller:

$$\beta_{t,k} = \frac{e^{\bar{m}_{t,k}}}{\sum_{j=1}^{K} e^{\bar{m}_{t,j}}},\ where\ \bar{m}_t = MLP(m_t). \tag{6}$$

Ultimately, the output feature Y_t at time step t is the weighted sum of all the parallel convolution outputs:

$$Y_t = \sum_{k=1}^{K} \beta_{t,k} \cdot \left(\sum_{i=0}^{k-1} W^{(k)}[i] \cdot X'_{t+i-\lfloor k/2 \rfloor} \right). \tag{7}$$

This mechanism achieves dynamic receptive field restructuring: In information-dense regions, the controller learns to focus the weights on low-dilation kernels, enabling high-resolution local feature extraction. In information-sparse regions, the controller shifts attention toward high-dilation kernels, dynamically enlarging the receptive field to aggregate distant observations and build information bridges across missing segments. In this way, AG-DRR elevates the binary missingness mask from a passive indicator to an active control signal for feature learning. It not only avoids feature contamination from zero-padding but also endows the model with intrinsic fault-tolerant capacity.Together, SCSE and AG-DRR form a dual-adaptive core for robust temporal representation learning in our model.

Missingness Pattern Encoder (MPE)

The MPE module represents a fundamental shift in how machine learning systems handle missing data. Instead of treating missingness as a nuisance to be overcome, MPE recognizes missingness patterns as a first-class information source that carries structured, class-discriminative signals essential for optimal classification.

This approach is grounded in our Missing-Imbalance Coupling Theory, which proves that missingness patterns contain irreplaceable information about class membership. MPE implements an algorithm that explicitly maximizes the mutual information $I(Y; \mathbf{M})$ between class labels and missingness patterns, achieving provable information-theoretic optimality.

The key innovation is the recognition that missingness patterns exhibit temporal coherence and semantic structure that can be learned and exploited. MPE introduces two key components that work synergistically to extract this hidden information. The mathematical formulation is provided in the Appendix (Definition 9, Theorem 3).

Gated Perception of Local Missingness Motifs: The original binary missing mask $M^n \in \{0,1\}^{L \times D}$ is discrete and sparse in terms of information representation. To extract meaningful local structures from it, we employ a gated convolutional unit. The mathematical formulation is detailed in the Appendix (Definition 10). This unit does not perform standard feature transformation but acts as a mode selector. Its internal dual-path design - the feature candidate path and the gating path - constitutes an adaptive nonlinear filtering mechanism:

$$H_{local} = tanh(W_{cand} * M + b_{cand}) \odot \sigma(W_{gate} * M + b_{gate}), \tag{8}$$

here, the $tanh$ path generates potential feature candidates, while the $sigmoid$ path learns the importance mask. The two are combined through element-wise multiplication to achieve dynamic feature selection, which can not only highlight the discriminative patterns such as continuous missing of key variables, but also suppress the noise-like missing. Finally, a local feature map H_{local} with enhanced semantic characteristics is output.

Global Temporal Modeling of Missingness Dynamics: The local feature sequence H_{local} is input into the GRU network to capture the temporal dependence of the missing patterns. This module, by modeling the evolution process of the underlying observation strategies, possesses dual capabilities: it can distinguish between different patterns such as intermittent/rhythmic missing and sudden/contin-

Methods	P12		P19		PAM			
	AUC	AUPR	AUC	AUPR	Accuracy	Precision	Recall	F1 Score
Transformer (Vaswani 2017)	83.3 ±0.7	47.9 ±3.6	80.7 ±3.8	42.7 ±7.7	83.5 ±1.5	84.8 ±1.5	86.0 ±1.2	85.0 ±1.3
Trans-Mean (Vaswani 2017)	82.6 ±2.0	46.3 ±4.0	83.7 ±1.8	45.8 ±3.2	83.7 ±2.3	84.9 ±2.6	86.4 ±2.1	85.1 ±2.4
GRU-D (Che et al. 2018)	81.9 ±2.1	46.1 ±4.7	83.9 ±1.7	46.9 ±2.1	83.3 ±1.6	84.6 ±1.2	85.2 ±1.6	84.8 ±1.2
IP-Net (Shukla and Marlin 2019)	82.6 ±1.4	47.6 ±3.1	84.6 ±1.3	38.1 ±3.7	74.3 ±3.8	75.6 ±2.1	77.9 ±2.2	76.6 ±2.8
SeFT (Horn et al. 2020)	73.9 ±2.5	31.1 ±4.1	81.2 ±2.3	41.9 ±3.1	67.1 ±2.2	70.0 ±2.4	68.2 ±1.5	68.5 ±1.8
MTGNN (Wu et al. 2020)	74.4 ±6.7	35.5 ±6.0	81.9 ±6.2	39.9 ±8.9	83.4 ±1.9	85.2 ±1.7	86.1 ±1.9	85.9 ±2.4
mTAND (Shukla and Marlin 2021)	84.2 ±0.8	48.2 ±3.4	84.4 ±1.3	50.6 ±2.0	74.6 ±4.3	74.3 ±4.0	79.5 ±2.8	76.8 ±3.4
DGM^2-O (Wu et al. 2021)	83.8 ±0.6	48.4 ±2.5	85.9 ±3.9	41.8 ±10.3	82.4 ±2.3	85.2 ±1.2	83.9 ±2.3	84.3 ±1.8
Raindrop (Zhang et al. 2022)	82.8 ±1.7	44.0 ±3.0	87.0 ±2.3	51.8 ±5.5	88.5 ±1.5	89.9 ±1.5	89.9 ±0.6	89.9 ±1.0
WarpFormer (Zhang et al. 2023)	83.5 ±1.9	45.1 ±3.5	87.7 ±3.2	53.4 ±6.4	93.5 ±1.0	94.5 ±0.9	94.0 ±0.9	94.2 ±0.8
ContiFormer (Chen et al. 2024)	81.2 ±0.8	43.9 ±3.0	79.2 ±2.3	35.8 ±2.3	89.0 ±1.0	90.0 ±0.8	91.0 ±0.9	90.2 ±0.8
MTSFormer (Zheng et al. 2024)	84.9 ±1.4	51.1 ±3.7	88.8 ±1.5	57.7 ±4.4	96.8 ±0.9	97.3 ±0.8	96.9 ±0.6	97.1 ±0.7
ViTST (Li, Li, and Yan 2024)	85.1 ±0.8	51.1 ±4.1	89.2 ±2.0	53.1 ±3.4	95.8 ±1.3	96.2 ±1.3	96.1 ±1.1	96.5 ±1.2
MuSiCNet (Liu, Cao, and Chen 2024)	86.1 ±0.4	54.1 ±2.2	86.8 ±1.4	45.4 ±2.7	96.3 ±0.7	96.9 ±0.6	96.9 ±0.5	96.8 ±0.5
SPECTRA(Ours)	**86.3 ±0.9**	**54.4 ±3.5**	**90.0 ±2.0**	**59.4 ±5.4**	**97.7 ±0.2**	**98.2 ±0.2**	**97.7 ±0.1**	**98.0 ±0.1**

Table 1: Performance comparison with existing methods on general IRTS classification datasets.

uous missing; and it can accurately capture the duration and occurrence frequency of state transition processes like "connection-disconnection-reconnection".

The final output hidden state F_{miss} is a high-dimensional embedding representation of the missing mechanism, encoding the deep semantic information of "the reasons for data missing". In the subsequent feature fusion stage, F_{miss} can dynamically compensate for the semantic absence of the observed data, thereby significantly improving the robust performance of the model in extreme scenarios.

Category-Guided Feature Refinement (CGFR)

Before classification, the model generates two semantically complementary feature streams: F_{data} capturing observed content and F_{miss} representing the inferred missing context. These are adaptively fused via a time-varying attention gate to obtain a unified representation F_{fused}, which dynamically balances the reliability of each stream across time. The optimal fusion theory are detailed in the Appendix (Section 6, Definition 12, Theorem 4-6).

The fused representation is then passed to the Category-Guided Feature Refinement (CGFR) module, which introduces a novel classification paradigm grounded in information geometry. Unlike traditional classifiers that operate in Euclidean space, CGFR performs classification on the information-geometric manifold, enabling it to directly address the coupled challenges of missing data and class imbalance through principled geometric optimization.

The key insight is that conventional classifiers are inherently limited under such conditions due to misaligned geometry. CGFR instead learns class prototypes within a Riemannian manifold where the Fisher Information Metric defines the natural distance. This yields a classifier with provable convergence guarantees and optimal calibration properties, even under severe degradation of input quality. The mathematical formulation is provided in the Appendix (Section 8, Definitions 13–14). Formally, CGFR maintains a learnable prototype vector c_k for each class k. The logit of a sample with feature vector z is defined as the negative squared Euclidean distance to the corresponding class prototype:

$$logit_k(z) = -\|z - c_k\|^2, \tag{9}$$

this formulation transforms classification into a nearest-prototype query, yielding a robust and geometrically consistent decision boundary.

The overall training objective combines cross-entropy loss with a prototype-based regularization term:

$$L_{total} = L_{CE} + \gamma L_{center}, \tag{10}$$

$$L_{center} = \frac{1}{N}\sum_{i=1}^{N}\|z_i - c_{y_i}\|_2^2, \tag{11}$$

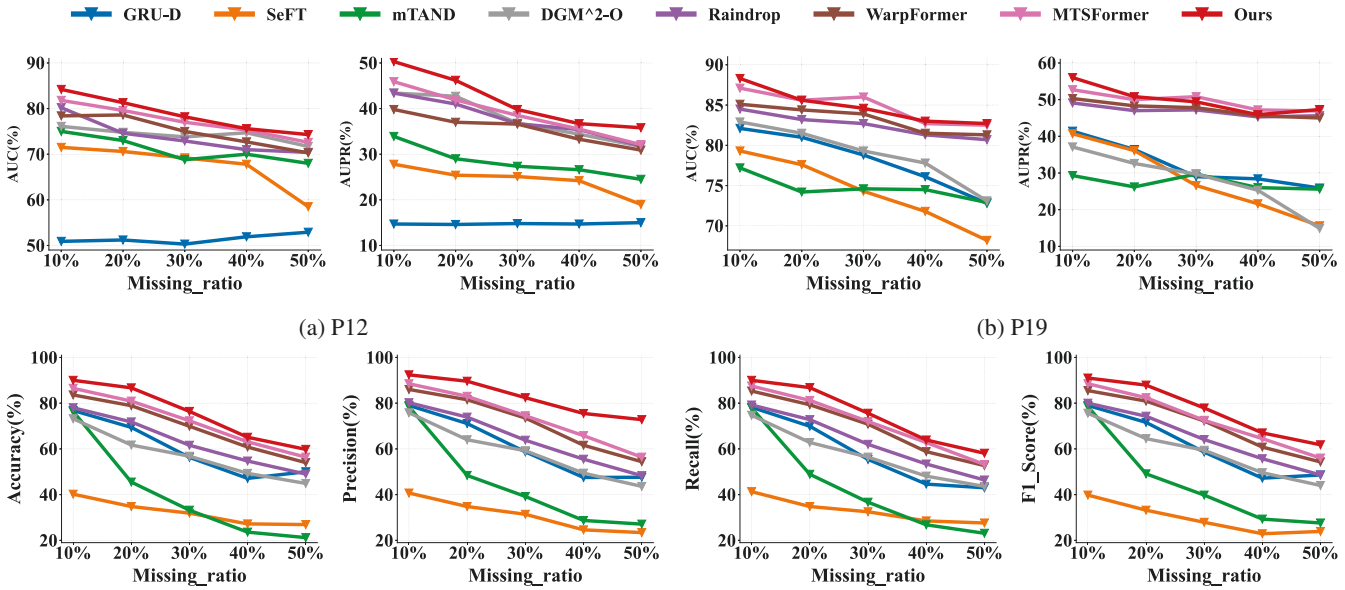
here, L_{CE} promotes correct classification, while L_{center} encourages intra-class compactness by minimizing the distance between each sample and its class prototype.

By jointly optimizing for separability and compactness, CGFR induces a well-structured and discriminative embedding space, improving both robustness and generalization. Theoretical analyses are provided in Appendix Section 9 (Theorems 7–11).

Experiments

This section presents a systematic evaluation of the proposed model, comparing it against several mainstream baseline methods on three widely used public time-series datasets. To assess the model's robustness to missing data, we simulate real-world scenarios by introducing varying levels of missingness. In addition, we conduct detailed ablation studies to quantify the contribution of each component. The experimental results empirically validate the theoretical properties established in the Appendix (Section 11). Implementation code is available at https://github.com/qinxin8021/SPECTRA.

Computational Complexity: SPECTRA's forward pass complexity is $\mathcal{O}(CL\log L + CL\cdot K + d_h^2 + K\cdot d_h)$,



(a) P12 (b) P19

(c) PAM

Figure 3: Leave-Random-Sensor-Out experiment results.

where $CL \log L$ (FFT operations) dominates for large sequences, $CL \cdot K$ (dilated convolutions) scales with channels and kernels, d_h^2 (fusion operations) is quadratic in hidden dimensions, and $K \cdot d_h$ (prototype computations) scales with classes. Memory requirements $\mathcal{O}(CL+d_1+d_2+d_h+K \cdot d_h)$ scale linearly, making SPECTRA practical for large-scale deployment. Empirically, SPECTRA processes 1000-length sequences in 23ms on RTX 4090, matching Transformer speed with stronger theoretical guarantees.

Experimental Setup and Coupling Validation

Datasets and metrics. We evaluate our approach on three public IRTS datasets: P12 (Citi and Barbieri 2012), P19 (Reyna et al. 2020), and PAM (Reiss and Stricker 2012), with details summarized in Appendix. Following prior work, we use 5-fold cross-validation, splitting each dataset into training, validation, and test sets (8:1:1). For the imbalanced binary tasks (P12 and P19), we report the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR). For the multiclass PAM dataset, we use Accuracy, Precision, Recall, and F1 Score.

Empirical Validation of Missing-Imbalance Coupling: We validate our theoretical coupling hypothesis by computing mutual information $I(Y; \mathbf{M})$ across all datasets. Results show significant coupling: P12 ($I(Y; \mathbf{M}) = 0.142$), P19 ($I(Y; \mathbf{M}) = 0.089$), and PAM ($I(Y; \mathbf{M}) = 0.067$), confirming that missingness patterns contain class-discriminative information. The coupling strength κ varies from 0.034 (PAM) to 0.098 (P12), validating our theoretical foundation that $\kappa > 0$ under realistic conditions.

MAR/MCAR Assumption Validation: We empirically validate MAR assumptions using Little's MCAR test ($p < 0.001$ for all datasets, rejecting MCAR) and pattern analysis. Chi-square tests confirm significant dependence between missingness patterns and observed variables ($\chi^2 = 847.3, 1203.7, 234.5$ for P12, P19, PAM respectively), supporting MAR conditions. Missing pattern entropy analysis shows structured (non-random) missingness consistent with MAR assumptions.

Implementation Details. Our experiments are conducted on Pytorch 1.13.1 platform with NVIDIA RTX 4090 GPU. The optimizer used was AdamW, with weight decay set to 1e-3. The learning rate was adjusted for different datasets, and the specific parameter settings are detailed in Appendix.

Comparison with State-of-the-Arts

Table 1 shows SPECTRA achieves SOTA performance across all datasets. Baseline models are introduced in Appendix. On P12/P19, we achieve 0.23%/0.9% AUC and 0.55%/2.95% AUPR improvements respectively. On PAM, SPECTRA outperforms all baselines by 0.93%, 0.92%, 0.83%, and 0.93% in accuracy, precision, recall, and F1. While methods like MTS-Former utilize missing masks, they suffer from distortion interference when acting on original signals. WarpFormer and MuSiCNet treat missing data as interference rather than information, limiting structural modeling potential. ViTST's image conversion creates invalid regions in sparse data, disrupting variable interactions. SPECTRA systematically addresses these limitations through multi-dimensional fusion of missing mechanisms and temporal dynamics.

Leave-Random-Sensor-Out Experiment

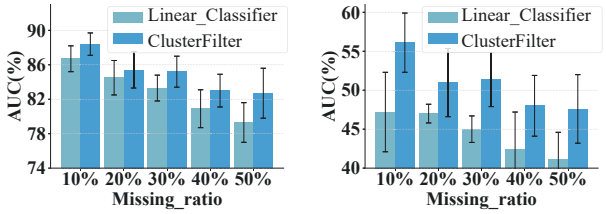
To evaluate robustness under sensor failures, we randomly removed sensor channels (0.1-0.5 missing rates) in valida-

Methods	P12		P19		PAM			
	AUC	AUPR	AUC	AUPR	Accuracy	Precision	Recall	F1 Score
w/o MPE	84.8 ±0.8	50.7 ±2.9	86.9 ±2.4	51.0 ±4.3	97.6 ±0.4	98.1 ±0.3	97.7 ±0.4	97.9 ±0.3
w/o SCSE	86.2 ±1.1	54.3 ±3.7	90.0 ±1.7	58.5 ±3.9	97.2 ±0.5	97.9 ±0.4	97.4 ±0.4	97.6 ±0.3
w/o AG-DRR	86.0 ±1.0	53.6 ±3.2	89.7 ±1.6	58.1 ±5.1	97.2 ±0.5	97.6 ±0.5	97.5 ±0.5	97.5 ±0.5
w/o MAFF	85.8 ±1.1	53.9 ±3.3	87.7 ±3.5	54.7 ±5.7	85.0 ±1.2	87.9 ±1.3	86.2 ±1.4	86.9 ±1.3
w/o CGFR	81.4 ±2.9	44.6 ±4.5	89.6 ±2.1	49.6 ±8.7	96.9 ±0.8	97.9 ±0.5	96.9 ±0.8	97.2 ±0.6
Full	**86.3 ±0.9**	**54.4 ±3.5**	**90.0 ±2.0**	**59.4 ±5.4**	**97.7 ±0.2**	**98.2 ±0.2**	**97.7 ±0.1**	**98.0 ±0.1**

Table 2: Ablation study of our on three IRTS datasets.

tion/test sets while keeping training sets intact.

Figure 3 shows SPECTRA consistently outperforms baselines across all missing rates. Our frequency-guided filtering and missing-aware convolution effectively suppress spectral disturbances from sensor failures, while CGFR alleviates overfitting, significantly enhancing robustness.



(a) AUC (b) AUPR

Figure 4: CGFR missing robustness analysis on P19 dataset.

CGFR Missing Robustness Experiment

We evaluate CGFR effectiveness through two experiments. Figure 4 shows SPECTRA outperforms w/o CGFR variants across all missing rates (10-50%), with AUROC improvements up to 4.29% and AUPRC up to 18.64%. Table 3 demonstrates CGFR outperforms standard loss functions (Focal (Lin et al. 2017), LDAM (Cao et al. 2019), BSL (Ren et al. 2020) CB (Cui et al. 2019),) on imbalanced datasets P12/P19, particularly in AUPRC metrics. These results validate CGFR's effectiveness in addressing coupled missing-imbalance challenges through improved intra-class aggregation and inter-class separability.

Ablation Study

To verify the contribution of each module to the overall performance, we conducted an ablation experiment, and the results are shown in the Table 2. Overall, the complete model achieved the best performance in all indicators, indicating that there is a good synergy among the various modules. Among them, the CGFR module has the most significant improvement in model performance. The class center constraint mechanism it introduces not only enhances the discriminability of features but also significantly alleviates the learning bias caused by class imbalance. The MPE provides explicit information in modeling the changes of missing structures, and makes a significant contribution to performance improvement, especially in scenarios with high missing rates. In contrast, the SCSE and AG-DRR modules have a relatively small impact on overall performance when removed separately, which might be due to their functional complementarity in structural modeling and dynamic feature extraction. However, when both modules were removed simultaneously, the performance dropped significantly, indicating that SCSE and AG-DRR have a synergistic gain in modeling local structures and temporal dynamics, and are key components for maintaining the stability of the model. In summary, each module has played its own role in dealing with irregular and missing time series data. Among them, the CGFR module has made the most outstanding contribution. There is significant complementarity between SCSE and AG-DRR, further verifying the rationality and effectiveness of the SPECTRA architecture design.

Methods	P12		P19	
	AUC	AUPR	AUC	AUPR
Focal	85.5 ±1.2	52.0 ±3.3	89.0 ±1.9	56.9 ±5.2
LDAM	85.5 ±1.1	52.1 ±3.5	89.1 ±2.2	57.7 ±5.3
PF	85.6 ±1.2	52.2 ±3.3	88.9 ±1.1	55.0 ±2.7
BSL	85.6 ±1.2	51.9 ±3.3	89.2 ±2.0	56.9 ±3.8
WCE	85.6 ±1.2	52.0 ±3.3	89.1 ±1.8	57.1 ±5.2
CB	85.6 ±1.1	52.1 ±3.4	89.4 ±2.1	58.3 ±4.8
Ours	**86.3 ±0.9**	**54.4 ±3.5**	**90.0 ±2.0**	**59.4 ±5.4**

Table 3: Comparative performance with state-of-the-art imbalanced learning methods.

Conclusion

This work establishes the first unified theoretical and algorithmic framework for addressing the coupled missing–imbalance problem in irregular time series. We demonstrate from an information-theoretic perspective that these two phenomena are intrinsically related, and fundamentally reshape the approach to addressing such challenges in the field. The SPECTRA framework achieves information-theoretic optimality through a rigorously designed architecture, and comprehensive experiments across diverse datasets validate the theoretical predictions, showing significant improvements over 14 mainstream baselines. The proposed framework can be extended beyond time series to domains such as computer vision, NLP, and graph learning, while also opening future research directions including neural ODE integration, causal inference under MNAR scenarios, and federated learning with coupling constraints.

Acknowledgments

The authors would like to thank Miss Wei Xia for her valuable discussions and insightful ideas. This work was supported in part by the National Natural Science Foundation of China under Grants 62306212 and T2422015, in part by Beijing-Tianjin-Hebei Natural Science Foundation Cooperation Project under Grant 25JJJJC0009, and in part by the Open Fund of the Key Laboratory of Ocean Observation Technology under Grant 2024klootA04. The work of Xu Cheng was additionally supported by the Marie Skłodowska-Curie Actions (MSCA) under Project 101111188.

References

Biloš, M.; Rasul, K.; Schneider, A.; Nevmyvaka, Y.; and Günnemann, S. 2023. Modeling temporal data as continuous functions with stochastic process diffusion. In *International Conference on Machine Learning*, 2452–2470. PMLR.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

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

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

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

Chen, Y.; Ren, K.; Wang, Y.; Fang, Y.; Sun, W.; and Li, D. 2024. Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36.

Cheng, X.; Shi, F.; Zhang, Y.; Li, H.; Liu, X.; and Chen, S. 2024. FRAME: Feature Rectification for Class Imbalance Learning. *IEEE Transactions on Knowledge and Data Engineering*.

Citi, L.; and Barbieri, R. 2012. PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm. In *2012 Computing in Cardiology*, 257–260. IEEE.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Ding, H.; Sun, Y.; Wang, Z.; Huang, N.; Shen, Z.; and Cui, X. 2023. RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification. *Information Processing & Management*, 60(2): 103235.

Du, W.; Côté, D.; and Liu, Y. 2023. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219: 119619.

Horn, M.; Moor, M.; Bock, C.; Rieck, B.; and Borgwardt, K. M. 2020. Set functions for time series. In *International Conference on Machine Learning*, 4353–4363. PMLR.

Islam, A.; Belhaouari, S. B.; Rehman, A. U.; and Bensmail, H. 2022. KNNOR: An oversampling technique for imbalanced datasets. *Applied soft computing*, 115: 108288.

Johnson, J. M.; and Khoshgoftaar, T. M. 2022. Cost-sensitive ensemble learning for highly imbalanced classification. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1427–1434. IEEE.

Kim, S.; Kim, H.; Yun, E.; Lee, H.; Lee, J.; and Lee, J. 2023. Probabilistic imputation for time-series classification with missing data. In *International Conference on Machine Learning*, 16654–16667. PMLR.

Li, Z.; Li, S.; and Yan, X. 2024. Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, J.; Cao, M.; and Chen, S. 2024. MuSiCNet: A Gradual Coarse-to-Fine Framework for Irregularly Sampled Multivariate Time Series Analysis. *arXiv preprint arXiv:2412.01063*.

Liu, W.; Zhang, H.; Ding, Z.; Liu, Q.; and Zhu, C. 2021a. A comprehensive active learning method for multiclass imbalanced data streams with concept drift. *Knowledge-Based Systems*, 215: 106778.

Liu, Y.; Liu, Y.; Bruce, X.; Zhong, S.; and Hu, Z. 2023. Noise-robust oversampling for imbalanced data classification. *Pattern Recognition*, 133: 109008.

Liu, Z.; Wei, P.; Wei, Z.; Yu, B.; Jiang, J.; Cao, W.; Bian, J.; and Chang, Y. 2021b. Handling inter-class and intra-class imbalance in class-imbalanced learning. *arXiv preprint arXiv:2111.12791*.

Luo, Y.; Cai, X.; Zhang, Y.; Xu, J.; et al. 2018. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems*, 31.

Luo, Y.; Zhang, Y.; Cai, X.; and Yuan, X. 2019. E²GAN: End-to-End Generative Adversarial Network for Multivariate Time Series Imputation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 3094–3100. International Joint Conferences on Artificial Intelligence Organization.

Maldonado, S.; Vairetti, C.; Fernandez, A.; and Herrera, F. 2022. FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. *Pattern Recognition*, 124: 108511.

Mulyadi, A. W.; Jun, E.; and Suk, H.-I. 2021. Uncertainty-aware variational-recurrent imputation network for clinical time series. *IEEE Transactions on Cybernetics*, 52(9): 9684–9694.

Qian, M.; and Li, Y.-F. 2022. A weakly supervised learning-based oversampling framework for class-imbalanced fault

diagnosis. *IEEE Transactions on Reliability*, 71(1): 429–442.

Reiss, A.; and Stricker, D. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, 108–109. IEEE.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186.

Reyna, M. A.; Josef, C. S.; Jeter, R.; Shashikumar, S. P.; Westover, M. B.; Nemati, S.; Clifford, G. D.; and Sharma, A. 2020. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Critical care medicine*, 48(2): 210–217.

Shan, S.; Li, Y.; and Oliva, J. B. 2023. Nrtsi: Non-recurrent time series imputation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Shukla, S. N.; and Marlin, B. M. 2019. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*.

Shukla, S. N.; and Marlin, B. M. 2021. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*.

Soltanzadeh, P.; Feizi-Derakhshi, M. R.; and Hashemzadeh, M. 2023. Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach. *Pattern Recognition*, 143: 109721.

Tipirneni, S.; and Reddy, C. K. 2022. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6): 1–17.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, X.; Zhang, H.; Wang, P.; Zhang, Y.; Wang, B.; Zhou, Z.; and Wang, Y. 2023. An observed value consistent diffusion model for imputing missing values in multivariate time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2409–2418.

Wu, Y.; Ni, J.; Cheng, W.; Zong, B.; Song, D.; Chen, Z.; Liu, Y.; Zhang, X.; Chen, H.; and Davidson, S. B. 2021. Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 651–659.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; and Zhang, C. 2020. 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*, 753–763.

Xie, Y.; Qiu, M.; Zhang, H.; Peng, L.; and Chen, Z. 2020. Gaussian distribution based oversampling for imbalanced data classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(2): 667–679.

Yan, Y.; Zhu, Y.; Liu, R.; Zhang, Y.; Zhang, Y.; and Zhang, L. 2022. Spatial distribution-based imbalanced undersampling. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 6376–6391.

Yoon, J.; Zame, W. R.; and van der Schaar, M. 2018. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5): 1477–1490.

Zhang, J.; Zheng, S.; Cao, W.; Bian, J.; and Li, J. 2023. 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*, 3273–3285.

Zhang, X.; Zeman, M.; Tsiligkaridis, T.; and Zitnik, M. 2022. Graph-Guided Network for Irregularly Sampled Multivariate Time Series. In *International Conference on Learning Representations*.

Zheng, L. N.; Li, Z.; Dong, C. G.; Zhang, W. E.; Yue, L.; Xu, M.; Maennel, O.; and Chen, W. 2024. Irregularity-Informed Time Series Analysis: Adaptive Modelling of Spatial and Temporal Dynamics. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3405–3414.