CASE-NET: Deep Spatio-Temporal Representation Learning via Causal Attention and Channel Recalibration for Multivariate Time Series Classification

Fan Zhang[1] , **Yating Cui**[1] , **Hua Wang**[2,*]
[1]Shandong Technology and Business University
[2]Ludong University
zhangfan51@sina.com, 2025420269@sdtbu.edu.cn, hwa229@163.com

Abstract

Multivariate time series (MTS) classification is foundational to pervasive computing and financial analysis, yet existing multi-scale paradigms are often constrained by suboptimal representation fidelity. We identify two critical bottlenecks: temporal non-causality in standard encoders that induces temporal confounding in non-stationary dynamics, and the absence of explicit channel saliency mechanisms that allows noise to contaminate the latent space. To address these challenges, we propose the Causal Attention and Spatio-temporal Encoder Network (CASE-Net), an architecture designed for structural manifold pre-conditioning. CASE-Net synergizes a Causal Temporal Encoder, which enforces physical arrow-of-time constraints via masked self-attention and causal convolutions, with an Adaptive Channel Recalibration module functioning as an information bottleneck to suppress detrimental noise. Comprehensive evaluations across six heterogeneous domains demonstrate that CASE-Net establishes new state-of-the-art benchmarks on four tasks, achieving a peak accuracy of 98.6% on the AWR dataset and superior robustness in non-stationary regimes.

1 Introduction

The surge in IoT and wearable data makes MTS classification a core AI challenge, ranging from clinical diagnostics to activity recognition [Morid *et al.*, 2023; Liu *et al.*, 2024a]. Unlike static data modalities, MTS exhibits profound Spatio-Temporal Complexity: temporal dimensions encapsulate non-stationary dynamics across multiple granularities, while spatial dimensions involve intricate, high-dimensional couplings between heterogeneous variables [Wang *et al.*, 2024b]. Extracting robust representations from these complex sequences is a core challenge.

Multi-scale analysis has proven effective in capturing cross-granularity patterns. Building upon this, recent breakthroughs in Multi-scale Disentanglement have successfully introduced scale-shared and scale-specific orthogonal constraints to mitigate inter-scale redundancy [Liu *et al.*, 2025], fostering a solid theoretical foundation for representation learning. However, these frameworks typically operate under the implicit assumption that the spatio-temporal features fed into the disentanglement module are sufficiently robust and noise-resilient. Critically, we argue that this assumption is often violated in real-world scenarios due to two representation bottlenecks in current feature extraction pipelines.

First, existing encoders often suffer from temporal blindness: recurrent models struggle with long-range dependencies, while bidirectional processing may introduce future-induced stochastic fluctuations into current representations. In physical systems, the arrow of time dictates that current states should not depend on future observations. For non-stationary MTS, such future observations often correspond to evolutionary noise or sensor artifacts rather than useful semantic context, leading to temporal confounding and distorted latent motifs.

Second, high-dimensional sensing systems also exhibit spatial channel heterogeneity. Variable importance is inherently dynamic and context-dependent, yet many architectures treat all channels with homogeneous priority and leave noise filtering to the downstream disentanglement module. This increases the burden on the disentanglement head and fundamentally limits the purity of the distilled features. CASE-NET mitigates these two bottlenecks by redefining the causal mask as a structural noise filter and introducing channel recalibration as a proactive spatial information bottleneck.

Beyond time-series analysis, compositional representation learning has also been actively explored in composed image and video retrieval, where models disentangle target semantics from modification cues to support robust cross-modal matching [Li *et al.*, 2025; Fu *et al.*, 2025; Huang *et al.*, 2025; Chen *et al.*, 2025; Zhang *et al.*, 2026c; Li *et al.*, 2026a; Li *et al.*, 2026b]. Inspired by this general principle of composition-aware representation learning, we investigate compositionality along temporal-scale and channel dimensions in MTS, where future-induced stochasticity and sensor heterogeneity become the dominant obstacles. Motivated by the insight that better representation facilitates better disentanglement, we propose CASE-NET. Our objective is to implement structural manifold pre-conditioning on the upstream representations to support high-fidelity feature decoupling.

*Corresponding author: hwa229@163.com

Our contributions are summarized as follows:

- Causal-Aware Temporal Reconstruction: We introduce a causal attention framework that explicitly enforces a physical arrow-of-time prior, filtering future-induced evolutionary noise to help preserve the integrity of latent action motifs. By coordinating local causal convolutions with global masked self-attention, the CASE encoder functions as a structural manifold pre-conditioner that is architecture-agnostic, empowering various downstream tasks as a high-fidelity front-end.
- Context-Adaptive Spatial Recalibration: We implement a channel-wise information bottleneck via adaptive recalibration (SE module) to execute proactive manifold pre-conditioning. By suppressing non-discriminative spatial artifacts before the disentanglement stage, this mechanism encourages high-saliency variables to propagate to the latent space, thereby enhancing the purity of the distilled features.
- Holistic Evaluation across Six Heterogeneous Domains: Unlike conventional approaches validated only in a single domain, CASE-NET was fully evaluated on six benchmark datasets with diverse physical characteristics and data distributions. Results show that CASE-NET established new state-of-the-art benchmarks on four tasks.

2 Related Work

2.1 Deep Temporal Modeling

Extracting temporal dependencies is central to MTS classification [Qin *et al.*, 2017; Zhu *et al.*, 2023]. While traditional recursive models like LSTMs and GRUs utilize inductive bias for sequential encoding, they often suffer from vanishing gradients and a critical paradigm limitation: Temporal Blindness. Most recurrent architectures lack temporal causality, leading to spurious future-information leakage that compromises reliability in sensitive domains like clinical diagnostics. Although recent works attempt to encode causality through transfer entropy [Sun *et al.*, 2024] or spatiotemporal information transformation [Cai *et al.*, 2025], these often incur heavy computational costs. CASE-NET mitigates this by re-engineering the encoder with causal convolutions and masked self-attention, enforcing temporal arrows to capture global dynamic motifs with high fidelity.

Recent forecasting models further exploit multi-offset temporal interactions, semantic fusion, and non-ideal data utilization to strengthen multivariate dependency learning [Zhang *et al.*, 2026a; Zhang *et al.*, 2026b; Qiu *et al.*, 2025b; Wang *et al.*, 2026b]. Nevertheless, these methods primarily emphasize predictive dependency modeling, whereas CASE-NET focuses on classification-oriented representation purification under temporal causality. In addition, ensuring high-fidelity representations remains a challenge, as explored by Zhang et al. [Zhang *et al.*, 2022] through time-frequency consistency. In contrast, CASE-NET enforces a causal inductive bias to improve manifolds, especially in non-stationary states. This distinguishes it from latent SDE-Attention [Fang *et al.*, 2025], which focuses on irregular sampling, as we optimize the upstream representation manifold for deterministic MTS.

Recently, Xu et al. [Xu *et al.*, 2025] proposed a spatio-temporal fusion architecture for dynamic ECN analysis, using Granger-causality-based discovery to model the evolution of brain networks. While their approach focuses on learning the explicit causal structure from observational data, CASE-NET leverages a causal inductive bias through architectural masking. We argue that for non-stationary classification tasks, enforcing causality as a structural prior is more effective for filtering future-induced stochastic noise than explicit causal discovery, which can be sensitive to latent confounding factors.

2.2 Channel Saliency and Information Bottleneck

Attention mechanisms have shifted MTS modeling toward dynamic recalibration. Recent architectures have introduced cross-channel attention [Hao and Cao, 2020; Zhang *et al.*, 2025] or association adapters [Lin *et al.*, 2025] to handle multivariate dependencies. Recent developments such as iTransformer [Liu *et al.*, 2024b] have rethought the role of channel dependencies by inverting the transformer structure. However, many frameworks still rely on a suboptimal assumption of Spatial Channel Homogeneity, treating sensors with uniform priority regardless of their varying Signal-to-Noise Ratios [Cheng *et al.*, 2023; Huang *et al.*, 2023]. Consistent with feature-centric explanation models like CAFO [Kim *et al.*, 2024], CASE-NET leverages Information Bottleneck theory and integrates an adaptive recalibration mechanism to execute proactive feature-cleaning, helping high-saliency variables propagate to the latent space for robust disentanglement.

2.3 Multi-Scale Analysis and Disentanglement Fidelity

Recent milestones like DisMS-TS have introduced multi-scale disentanglement to mitigate inter-scale redundancy [Liu *et al.*, 2025]. At the benchmark and optimization levels, recent studies have emphasized fair time-series evaluation, non-ideal data utilization, and scalable optimization for time-series modeling [Qiu *et al.*, 2024; Wang *et al.*, 2026b; Wang *et al.*, 2026a; Qiu *et al.*, 2025a]. Unlike multi-scale token mixing [Zhong *et al.*, 2025], hierarchical temporal-frequency fusion [Song *et al.*, 2025], or multi-view meta-learning [Zhang *et al.*, 2024] that prioritize feature fusion, the efficacy of decoupling is fundamentally contingent upon the fidelity of underlying representations. Existing frameworks often overlook feature extractor quality, creating a performance bottleneck when processing noise-contaminated or temporally inconsistent inputs. Departing from efforts focused solely on downstream decoupling logic, CASE-NET focuses on optimizing the quality of the upstream representation manifold through structural pre-conditioning. By providing causal-aware spatio-temporal priors, our model bridges the gap between raw sensing data and decoupled representations, establishing new performance ceilings across heterogeneous domains.

3 Methodology

3.1 Overall Architecture

CASE-NET enhances classification through three stages (Fig. 1): Multi-scale Projection, the CASE Encoder, and decomposition. The CASE encoder is weight-shared across all S branches. This allows the model to learn a scale-invariant causal-spatial transformation while maintaining efficiency.

Formal Rationale: We define the CASE encoder not merely as a collection of components, but as a synergistic information bottleneck and a manifold pre-conditioner designed to address the fundamental problem of distorted input manifolds in existing frameworks. In this paper, structural manifold preconditioning refers to representation-level conditioning before the shared-specific disentanglement head, implemented through causal temporal masking and bottleneck-style channel recalibration. Existing multi-scale encoders often suffer from temporal blindness and channel homogeneity, which introduce future-information leakage and spatial noise. Such contaminated inputs force the downstream disentanglement loss ($\mathcal{L}_{diff}$) to search within an excessively complex hypothesis space. By improving the input manifold through structural pre-conditioning before it enters the sensitive disentanglement backend, we significantly narrow the search space of the downstream loss function and promote high-resolution representations.

3.2 Generic Multi-Scale Projection

To capture the polymorphic feature mapping of MTS at different granularities, we construct a multi-scale input space consisting of S branches. Specifically, given the input $\mathbf{X} \in \mathbb{R}^{N \times L}$, we utilize a set of convolutional operators with heterogeneous receptive fields and strides to project it as a multi-resolution view set $\mathcal{V} = \{\mathbf{V}_s\}_{s=1}^{S}$, where:

$$\mathbf{V}_s = \sigma(\text{Conv1D}_s(\mathbf{X})) \in \mathbb{R}^{C_s \times L_s} \tag{1}$$

Non-uniform sampling here establishes multi-granularity representations, providing context for the CASE encoder. Importantly, this stage only initializes scale-specific views, while the subsequent weight-shared CASE encoder imposes a unified causal-spatial transformation across all resolutions.

3.3 Causal Adaptive Spatio-Temporal Encoder

The CASE encoder serves as the core architecture for spatio-temporal feature extraction and is designed to solve the representation fidelity bottleneck in multi-scale decoupling. Specifically, the encoder implements structural preconditioning by integrating two complementary modules: Causal Temporal Attention and Channel Adaptive Recalibration. This dual-constrained design encourages the resulting latent space to be both physically consistent and semantically sparse.

Causal Temporal Attention

The CTA module is designed to overcome the bottlenecks of traditional recurrent models in long-range dependency modeling while imposing a robust temporal causality constraint. By coupling local causal operators (Dilated Convolutions) with global masked self-attention, it promotes autoregressive consistency ($\mathbf{h}_t = f(\{x_\tau\}_{\tau \le t})$). This configuration mitigates future information leakage while capturing multi-scale, long-range dependencies across the sequence.

In contrast to bidirectional transformers or explicit discovery models [Xu *et al.*, 2025], CASE-NET recognizes that in non-stationary regimes (e.g., NASDAQ), future observations often represent stochastic evolutionary noise relative to the current state. By enforcing causality, our CTA module functions as a Structural Noise Filter. This helps the encoder prioritize the capture of underlying action motifs rather than fitting transient signal fluctuations, effectively executing Temporal De-confounding through architectural constraints. By enforcing this Manifold Pre-conditioning, we narrow the downstream hypothesis space and help the representation manifold remain invariant to future-induced noise, preventing data leakage within the architecture.

To instantiate local causal modeling, we introduce dilated convolutional layers with asymmetric left padding. For an input sequence $\mathbf{X}$, the temporal dimension is padded with $L_{pad} = (k-1) \times d$ zeros on the left, where k denotes the kernel size and d is the dilation rate. The causal operator is defined as:

$$\mathbf{y}_t = \sigma\left(\sum_{i=0}^{k-1} \mathbf{W}_i \cdot \mathbf{x}_{t-d \cdot i} + \mathbf{b}\right) \tag{2}$$

This operator forces the convolution kernel to slide only over current and historical trajectories, extracting local dynamics with high causal fidelity. Such a structured inductive bias is crucial for non-stationary sequences, since it keeps the transition probability $P(x_t|x_{<t})$ undisturbed by future observations.

To complement local causal filtering, we further integrate multi-head self-attention with an upper-triangular causal mask $\mathbf{M} \in \mathbb{R}^{L \times L}$. The mask is defined as:

$$M_{ij} = \begin{cases} 0, & i \ge j \\ -\infty, & i < j \end{cases} \tag{3}$$

Given linear projection matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$, the masked attention reconstructs the temporal manifold by:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V} \tag{4}$$

Channel Adaptive Recalibration

To complement the temporal consistency established by CTA, this module uses the Information Bottleneck principle to address Channel Heterogeneity. Due to varying signal-to-noise ratios (SNRs) across sensors, the raw temporal representation $\mathbf{H}$ is frequently contaminated by non-discriminative channel noise.

The recalibration module complements CTA by constraining the channel dimension of the feature manifold. While Eq. (4) filters noise along the temporal axis through causality, channel recalibration suppresses redundant or low-saliency variables along the spatial axis, encouraging only discriminative and physically consistent spatio-temporal motifs to be propagated to the disentanglement head.

Given the temporal representation $\mathbf{H} \in \mathbb{R}^{B \times N \times L}$, we first summarize the global response of each channel using a

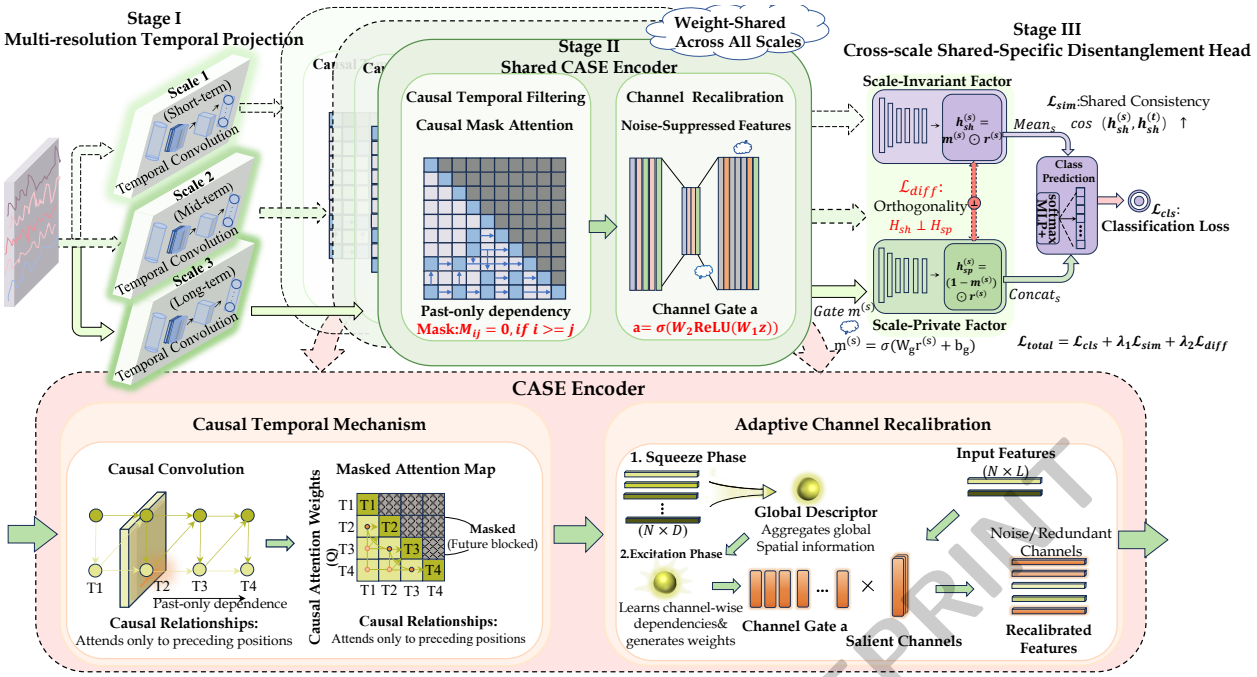

Figure 1: The hierarchical framework of CASE-NET, illustrating: (1) multi-scale representation initialization via parallel branches; (2) a *weight-shared* CASE encoder (2 layers) that processes all branches synchronously, constrained by physical laws and SNR; and (3) inter-scale redundancy elimination via orthogonal constraints.

Global Average Pooling operator $\mathcal{F}_{sq}(\cdot)$. For the n-th channel, the descriptor is computed as:

$$z_n = \mathcal{F}_{sq}(\mathbf{h}_n) = \frac{1}{L}\sum_{t=1}^{L} h_{n,t} \tag{5}$$

Although early states in a causal sequence have limited receptive fields, the later hidden states progressively accumulate historical context. Therefore, Eq. (5) can be viewed as a compact summary of the feature evolution trajectory rather than a naive temporal average, providing stable anchors for estimating channel saliency under non-stationary dynamics.

The channel descriptor $\mathbf{z}$ is then passed through a bottleneck excitation gate with dimensionality reduction and expansion $N \to N/r \to N$:

$$\mathbf{a} = \sigma\left(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1 \cdot \mathbf{z})\right) \tag{6}$$

where $\sigma(u) = 1/(1+\exp(-u))$ is the element-wise sigmoid function, $\mathbf{a} \in [0,1]^N$ denotes the channel saliency weights, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the reduction and expansion weights. The bottleneck ratio r encourages the model to learn essential inter-variable dependencies in a compact latent space.

Finally, the recalibrated representation is obtained by applying the learned channel saliency weights to the temporal representation:

$$\tilde{\mathbf{H}} = \mathbf{a} \odot \mathbf{H} \tag{7}$$

where $\odot$ denotes element-wise multiplication. This variable-level filtering mechanism improves the separability and noise-resilience of the representation before shared-specific disentanglement. In this way, the downstream decomposition head receives a cleaner manifold rather than being forced to suppress temporal and channel noise simultaneously.

3.4 Multi-Scale Feature Decomposition and Prediction

We adopt a shared-specific disentanglement head inspired by multi-scale disentanglement learning [Liu *et al.*, 2025] to evaluate the downstream utility and representational purity of the manifolds produced by the CASE encoder. This head explicitly separates each recalibrated scale representation into scale-shared and scale-private factors, enabling inter-scale redundancy reduction while preserving scale-dependent discriminative details.

While CASE-NET is designed to be compatible with various discriminative heads, we instantiate the decomposition module with an adaptive shared-specific gate. For each scale s, the disentanglement gate $\mathbf{m}^{(s)}$ is generated from the recalibrated representation $\tilde{\mathbf{H}}^{(s)}$:

$$\begin{aligned} \mathbf{m}^{(s)} &= \sigma\left(\mathbf{W}_g \tilde{\mathbf{H}}^{(s)} + \mathbf{b}_g\right), \\ \mathbf{H}_{sh}^{(s)} &= \mathbf{m}^{(s)} \odot \tilde{\mathbf{H}}^{(s)}, \\ \mathbf{H}_{sp}^{(s)} &= \left(1 - \mathbf{m}^{(s)}\right) \odot \tilde{\mathbf{H}}^{(s)}. \end{aligned} \tag{8}$$

Here, $\mathbf{m}^{(s)}$ adaptively allocates each feature dimension to the scale-shared or scale-private subspace, $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable gate parameters, and $\odot$ denotes element-wise multiplication. The resulting shared and scale-private factors play complementary roles. The shared representation $\mathbf{H}_{sh}$ is encouraged to capture stationary rhythms and scale-invariant semantic cores, such as basal cardiac or neural oscillatory patterns. In contrast, $\mathbf{H}_{sp}$ preserves non-stationary residual dynamics,

including high-frequency transients and localized paroxysmal events that may be suppressed during global aggregation. This complementary allocation prevents scale-invariant semantics and scale-private residuals from collapsing into a single entangled representation.

The efficacy of this decomposition is fundamentally predicated on the manifold pre-conditioning provided by the CASE encoder. By furnishing a noise-filtered and causally-consistent input, we demonstrate that the backend can achieve a higher degree of Feature Orthogonality with minimal information leakage. The final prediction $\hat{y}$ is derived by fusing these descriptors:

$$\hat{y} = \text{Softmax}\left(\text{MLP}\left(\mathcal{G}\left(\mathbf{H}_{sh} \oplus \mathbf{H}_{sp}\right)\right)\right) \tag{9}$$

where $\oplus$ denotes feature concatenation, $\mathcal{G}(\cdot)$ denotes global aggregation over temporal and scale dimensions, and the MLP outputs class logits before the Softmax normalization. This design combines scale-invariant semantics with scale-private discriminative details for final classification.

3.5 Synergistic Manifold Optimization

We formulate a multi-objective functional $\mathcal{L}_{total}$ to supervise the latent manifold evolution:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{sim} + \lambda_2 \mathcal{L}_{diff} \tag{10}$$

where $\mathcal{L}_{cls}$ is the task-specific Negative Log-Likelihood applied to the fused representation $[\mathbf{H}_{sh} \oplus \mathbf{H}_{sp}]$, and λ_1, λ_2 are non-negative balancing coefficients controlling the contributions of cross-scale consistency and shared-private orthogonality, respectively. To capture invariant rhythmic motifs, $\mathcal{L}_{sim}$ minimizes the cosine distance between shared representations across resolutions [Chen *et al.*, 2023; Chen *et al.*, 2024]:

$$\mathcal{L}_{sim} = \mathbb{E}_{s_1 \neq s_2}\left[1 - \frac{\langle \mathbf{H}_{sh}^{(s_1)}, \mathbf{H}_{sh}^{(s_2)} \rangle}{\|\mathbf{H}_{sh}^{(s_1)}\|_2 \|\mathbf{H}_{sh}^{(s_2)}\|_2}\right] \tag{11}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product after flattening or pooling the corresponding representations. This inductive bias identifies the universal physiological core and filters stochastic noise inherent to specific resolutions. Conversely, $\mathcal{L}_{diff}$ promotes Feature Orthogonality by penalizing the cross-covariance between $\mathbf{H}_{sh}$ and $\mathbf{H}_{sp}$ via the Frobenius Norm [Liu *et al.*, 2025]:

$$\mathcal{L}_{diff} = \left\|\mathbf{H}_{sh}^{\top}\mathbf{H}_{sp}\right\|_F^2 \tag{12}$$

By driving representational diversity, this constraint prunes the hypothesis space, suggesting that $\mathbf{H}_{sp}$ captures unique residual dynamics (e.g., pathological paroxysms) with less overlap with the global semantic anchor. Together, the three objectives couple task discrimination with representation geometry: $\mathcal{L}_{cls}$ preserves label separability, $\mathcal{L}_{sim}$ enforces cross-scale invariance, and $\mathcal{L}_{diff}$ prevents shared-private feature leakage.

3.6 Computational Complexity Analysis

The computational overhead of CASE-NET is primarily dominated by the CASE encoder. The CTA module incurs a complexity of $\mathcal{O}(L^2 \cdot D)$ due to masked self-attention and $\mathcal{O}(L \cdot k \cdot D)$ for causal convolutions. The adaptive recalibration adds a negligible $\mathcal{O}(N^2/r)$.

Although the CTA module has $\mathcal{O}(L^2)$ attention complexity, it is fully parallelizable and remains practical for long-range sequences. Empirically, CASE-NET uses about 2.1M parameters and maintains competitive throughput, supporting its feasibility for real-time MTS applications.

The detailed step-by-step execution of the end-to-end training procedure is summarized in Algorithm S1 in the Supplementary Material.

4 Experiments

4.1 Experimental Setup

Datasets

We evaluate CASE-NET on six benchmark classification datasets following the protocol of DisMS-TS [Liu *et al.*, 2025], including AWR (Articulary Word Recognition), FM (Finger Movements), and NASDAQ for market regime classification. The NASDAQ task is window-level classification rather than next-step price prediction, preventing look-ahead leakage (see Table S1 in Supplementary Material).

Evaluation Metrics

We employ Accuracy (Acc), Macro-averaged F1-score (F1), and Matthews Correlation Coefficient (MCC) to comprehensively assess performance.

Baseline Methods

We compare CASE-NET with eight representative baselines: LSTNet [Lai *et al.*, 2018], DTP [Lee *et al.*, 2021], TS-TCC [Eldele *et al.*, 2021], PatchTST [Nie *et al.*, 2022], MAGNN [Chen *et al.*, 2023], Pathformer [Chen *et al.*, 2024], TimeMixer [Wang *et al.*, 2024a], and DisMS-TS [Liu *et al.*, 2025]. These baselines cover recurrent and pooling-based classifiers, contrastive representation learning, Transformer-based temporal modeling, graph-based multivariate modeling, and recent multi-scale architectures.

Implementation Details

We implement CASE-NET in PyTorch and report the average results over five independent runs. All models are optimized with Adam [Kingma and Ba, 2014], and key hyperparameters, including learning rate, batch size, hidden dimension, scale depth, and loss weights, are selected on a 20% hold-out validation set. Dropout, layer normalization, and early stopping are applied for regularization. For each dataset, the validation split is used only for hyperparameter selection and early stopping, while the test split remains strictly held out for final reporting. The best checkpoint is selected according to validation accuracy, and all metrics are computed on the corresponding test set to reduce the influence of a single favorable initialization.

4.2 Comparative Analysis

Table 1 shows CASE-NET achieves SOTA performance on HAR, NASDAQ, AWR, and Epilepsy. On NASDAQ, CASE-NET leads by 7.05% via Causal Attention, which enforces temporal constraints to block future noise. For Epilepsy, SE

Methods	Metrics	HAR	ISRUC-S3	NASDAQ	AWR	FM	Epilepsy
LSTNet	ACC (%)	87.52 ± 1.25	80.19 ± 1.21	35.34 ± 1.60	91.73 ± 0.88	51.60 ± 3.01	62.57 ± 1.08
	F1 (%)	86.58 ± 1.28	77.93 ± 1.12	25.29 ± 2.18	91.33 ± 0.93	51.59 ± 3.00	60.36 ± 1.72
	MCC	0.846 ± 0.014	0.746 ± 0.016	0.058 ± 0.022	0.914 ± 0.009	0.032 ± 0.060	0.556 ± 0.011
DTP	ACC (%)	92.09 ± 0.88	74.23 ± 1.04	34.25 ± 0.85	94.87 ± 1.20	50.00 ± 1.98	42.57 ± 2.46
	F1 (%)	92.10 ± 0.85	72.38 ± 1.05	22.78 ± 1.14	94.82 ± 1.23	49.98 ± 1.99	41.14 ± 2.17
	MCC	0.905 ± 0.010	0.671 ± 0.012	0.046 ± 0.015	0.946 ± 0.012	0.020 ± 0.039	0.301 ± 0.036
TS-TCC	ACC (%)	90.62 ± 0.59	77.63 ± 1.18	35.52 ± 1.26	89.65 ± 0.58	51.60 ± 1.62	64.84 ± 1.14
	F1 (%)	90.66 ± 0.62	75.50 ± 1.17	24.96 ± 1.52	89.57 ± 0.60	51.33 ± 1.80	63.26 ± 1.42
	MCC	0.898 ± 0.006	0.735 ± 0.016	0.054 ± 0.017	0.889 ± 0.007	0.008 ± 0.032	0.552 ± 0.015
PatchTST	ACC (%)	93.34 ± 0.32	81.05 ± 0.81	36.51 ± 1.06	97.47 ± 0.57	53.40 ± 3.61	68.68 ± 1.56
	F1 (%)	93.28 ± 0.34	79.14 ± 0.80	25.50 ± 1.27	97.45 ± 0.64	52.83 ± 3.68	67.26 ± 1.72
	MCC	0.919 ± 0.003	0.767 ± 0.015	0.056 ± 0.016	0.965 ± 0.005	0.106 ± 0.071	0.609 ± 0.020
MAGNN	ACC (%)	92.45 ± 0.58	79.07 ± 0.86	38.12 ± 1.13	93.33 ± 0.61	52.20 ± 2.48	51.42 ± 2.20
	F1 (%)	92.52 ± 0.64	77.51 ± 0.85	29.52 ± 1.46	93.36 ± 0.75	51.32 ± 1.74	49.55 ± 2.69
	MCC	0.912 ± 0.008	0.749 ± 0.012	0.079 ± 0.014	0.930 ± 0.006	0.044 ± 0.055	0.430 ± 0.027
Pathformer	ACC (%)	94.88 ± 0.76	82.52 ± 0.78	37.24 ± 0.89	98.43 ± 1.13	55.60 ± 1.85	69.18 ± 0.88
	F1 (%)	94.92 ± 0.74	81.27 ± 0.83	27.96 ± 1.62	98.41 ± 1.15	54.92 ± 1.90	67.28 ± 0.85
	MCC	0.938 ± 0.008	0.771 ± 0.012	0.078 ± 0.016	0.981 ± 0.013	0.121 ± 0.041	0.614 ± 0.010
TimeMixer	ACC (%)	95.53 ± 0.65	83.50 ± 1.04	38.42 ± 0.75	97.53 ± 0.98	54.20 ± 3.73	67.28 ± 1.19
	F1 (%)	95.60 ± 0.64	82.18 ± 1.12	29.68 ± 0.97	97.54 ± 1.01	53.67 ± 3.85	65.65 ± 1.35
	MCC	0.947 ± 0.005	0.790 ± 0.014	0.088 ± 0.012	0.972 ± 0.010	0.082 ± 0.074	0.584 ± 0.015
DisMS-TS	ACC (%)	96.47 ± 0.37	**85.37±0.51**	39.71 ± 0.76	98.27 ± 0.86	**61.00±3.16**	71.30 ± 0.71
	F1 (%)	96.43 ± 0.40	**84.20±0.54**	32.69 ± 0.83	98.23 ± 0.90	**60.67±3.10**	70.70 ± 1.18
	MCC	0.957 ± 0.004	**0.829±0.008**	0.104 ± 0.012	0.976 ± 0.009	**0.228±0.065**	0.644 ± 0.006
CASE-NET	ACC (%)	**96.52 ± 0.35**	84.38 ± 0.31	**46.76 ± 0.38**	**98.60 ± 0.91**	58.67 ± 3.82	**73.48 ± 0.57**
	F1 (%)	**96.47 ± 0.37**	82.33 ± 0.35	**33.26 ± 0.36**	**98.49 ± 0.73**	58.36 ± 4.09	**73.35 ± 0.56**
	MCC	**0.960 ± 0.030**	0.794 ± 0.040	**0.110 ± 0.004**	**0.985 ± 0.008**	0.173 ± 0.082	**0.669 ± 0.007**

Table 1: Full performance comparison on six benchmark datasets. Results are reported as Mean ± Std across 5 independent runs. Bold and underline indicate the best and second-best mean results, respectively.

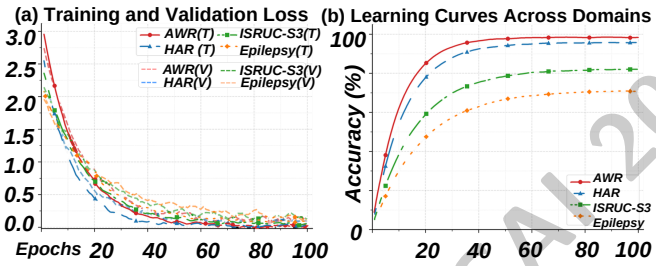

Figure 2: Training and validation loss and learning curves across four diverse domains.

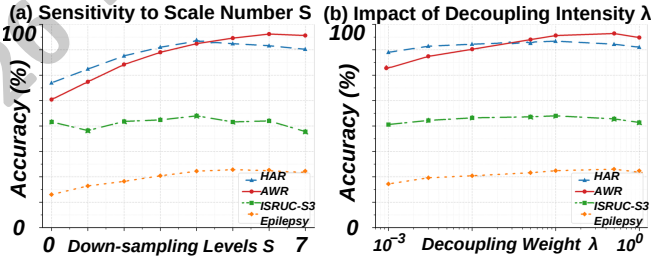

Figure 3: Sensitivity analysis of down-sampling levels S and decoupling intensity λ.

recalibration provides noise resistance against artifacts, consistent with [Sun *et al.*, 2024], yielding a 2.18% gain. On HAR and AWR, CASE-NET achieves higher mean performance than bidirectional baselines. Unlike BERT-style models where future context clarifies semantics, high-frequency MTS future frames often introduce stochastic noise (e.g., sensor artifacts) inducing temporal confounding. By causal masking, our model acts as a directional filter, encouraging the encoder to capture discriminative action motifs rather than transient signal fluctuations through structural regularization.

Though ranking second on ISRUC-S3 and FM, CASE-NET remains competitive. DisMS-TS's seasonal decomposition better suits the extreme periodicity of ISRUC-S3 ($L = 3000$), while its simpler logic offers stronger regularization in the low-data FM regime ($N = 284$).

4.3 Ablation Study

We define the *w/o Causal* variant as the bidirectional version of our architecture, removing the temporal mask $\mathbf{M}$ to evaluate component efficacy. In Table 2, this variant on NASDAQ drops from 46.76% to 38.20% (-8.56%), suggesting that bidirectional context in non-stationary tasks may introduce future-induced stochastic noise rather than useful information.

Similarly, HAR performance decreases to 91.25% (-5.27%), confirming that unconstrained bidirectional context can blur discriminative temporal boundaries. The causal prior also improves Epilepsy and AWR, demonstrating its stabilizing effect under noisy and non-stationary dynamics. To isolate the contribution of the CASE encoder, we replace the disentanglement backend with a linear MLP. As shown in Table 2, this variant remains competitive, while the full model achieves the best results, indicating that causal and spatial pri-

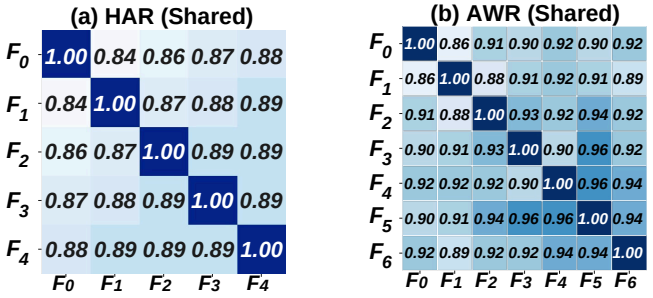

Figure 4: Cross-scale correlation heatmaps of shared features F_{sha}, equipped with quantitative colorbars.

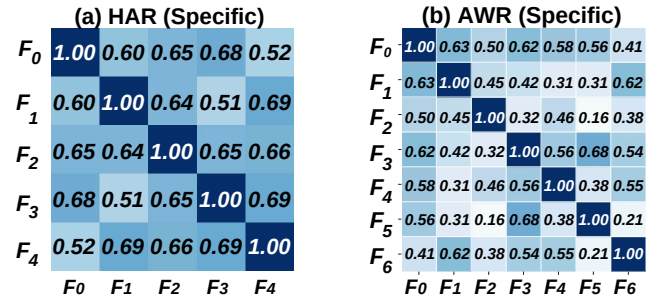

Figure 5: Correlation heatmaps of specific features F_{spe}, demonstrating lower and more dispersed cross-scale correlations than shared features.

ors jointly improve multi-scale representation learning.

Model Variant	HAR	NASDAQ	AWR	Epilepsy
Baseline (Vanilla GRU)	89.42 ± .22	36.50 ± .45	92.15 ± .18	64.20 ± .45
w/o SE (Only Causal)	93.10 ± .18	44.15 ± .32	96.48 ± .14	68.50 ± .31
w/o Causal (Only SE)	91.25 ± .15	38.20 ± .28	94.30 ± .21	67.40 ± .38
CASE-NET + MLP Head	95.10 ± .12	45.30 ± .25	97.80 ± .16	71.90 ± .42
CASE-NET (Full)	**96.52 ± .35**	**46.76 ± .38**	**98.60 ± .91**	**73.48 ± .57**

Table 2: Ablation study results (Mean ± Std %). Bold denotes best performance.

4.4 Hyperparameter Analysis

Impact of Scale Levels S

As depicted in Fig. 3(a), different datasets require different optimal scale depths: $S = 6$ for AWR and $S = 4$ for HAR. The performance drop at lower S values, particularly for high-cardinality tasks like AWR where $S \geq 5$ is required, does not indicate model instability. Instead, it reflects the necessity of resolution matching. A sufficient scale depth helps ensure an adequate temporal receptive field for domain-specific hierarchical motifs.

Sensitivity to Decoupling Intensity λ

As shown in Fig. 3(b), model performance remains stable across three orders of magnitude for the decoupling weight λ. This high consistency suggests that the performance gains originate from the inherent structural design of CASE-NET rather than from heuristic parameter fine-tuning.

4.5 Convergence and Stability Analysis

Fig. 2 shows stable optimization across domains: training and validation losses decrease consistently and converge within about 40 epochs, while accuracy curves remain smooth across four modalities. These results indicate that the causal-spatial encoder provides robust optimization behavior for heterogeneous MTS data. The close tracking between training and validation curves further suggests that the causal mask and channel bottleneck act as structural regularizers rather than merely increasing model capacity.

4.6 Representation Interpretability and Disentanglement

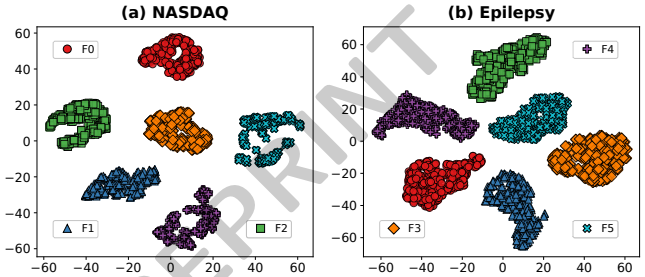

Figure 6: t-SNE visualization of scale-specific representations F_{spe} on the NASDAQ and Epilepsy datasets. Colors and markers denote different scale-specific factors F_0–F_5.

For visualization, we denote features derived from $\mathbf{H}_{sh}$ and $\mathbf{H}_{sp}$ as F_{sha} and F_{spe}, with F_i indexing scale branches; Fig. 4 and Fig. 5 show strong cross-scale correlations for F_{sha} and lower, more dispersed correlations for F_{spe}, indicating effective disentanglement. The t-SNE projections in Fig. 6 further visualize scale-specific representations on NASDAQ and Epilepsy, which respectively represent non-stationary financial dynamics and noisy biomedical signals. The different scale-specific factors form compact and distinguishable clusters, suggesting that CASE-NET learns structured scale-private manifolds rather than collapsing multi-scale information into an entangled latent space. In particular, the NASDAQ visualization supports the role of causal temporal filtering in non-stationary regime classification, while the Epilepsy visualization reflects the benefit of channel recalibration in suppressing noisy artifacts.

5 Conclusion

CASE-NET mitigates the representation bottleneck in multi-scale MTS classification by combining causal temporal modeling with channel recalibration before shared-specific disentanglement. Experiments across six heterogeneous domains suggest that the proposed causal-spatial pre-conditioning improves robustness in non-stationary and noisy settings. The results further indicate that causal temporal filtering and channel recalibration play complementary roles: the former stabilizes temporal dynamics, while the latter suppresses channel-wise noise before disentanglement. Future work will explore large-scale spatio-temporal pre-training for broader sensing applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. U24A20219, 62272281, and U24A20328; the Yantai Natural Science Foundation under Grant No. 2024JCYJ034; and the Youth Innovation Technology Project of Higher School in Shandong Province under Grant No. 2023KJ212.

References

[Cai *et al.*, 2025] Sihua Cai, Hao Peng, Rui Liu, and Pei Chen. Causal-oriented representation learning for time-series forecasting based on the spatiotemporal information transformation. *Communications Physics*, 8(1):242, 2025.

[Chen *et al.*, 2023] Ling Chen, Donghui Chen, Zongjiang Shang, Binqing Wu, Cen Zheng, Bo Wen, and Wei Zhang. Multi-scale adaptive graph neural network for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10748–10761, 2023.

[Chen *et al.*, 2024] Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. *arXiv preprint arXiv:2402.05956*, 2024.

[Chen *et al.*, 2025] Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Haokun Wen, and Weili Guan. Hud: Hierarchical uncertainty-aware disambiguation network for composed video retrieval. In *Proceedings of the ACM International Conference on Multimedia*, page 6143–6152, 2025.

[Cheng *et al.*, 2023] Yunyao Cheng, Peng Chen, Chenjuan Guo, Kai Zhao, Qingsong Wen, Bin Yang, and Christian S Jensen. Weakly guided adaptation for robust time series forecasting. *Proceedings of the VLDB Endowment*, 17(4):766–779, 2023.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

[Fang *et al.*, 2025] Yuting Fang, Qouc Le Gia, and Flora Salim. Sde-attention: Latent attention in sde-rnns for irregularly sampled time series with missing data. *arXiv preprint arXiv:2511.23238*, 2025.

[Fu *et al.*, 2025] Zhiheng Fu, Zixu Li, Zhiwei Chen, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Pair: Complementarity-guided disentanglement for composed image retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Hao and Cao, 2020] Yifan Hao and Huiping Cao. A new attention mechanism to classify multivariate time series. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020.

[Huang *et al.*, 2023] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. *Advances in Neural Information Processing Systems*, 36:46885–46902, 2023.

[Huang *et al.*, 2025] Qinlei Huang, Zhiwei Chen, Zixu Li, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Median: Adaptive intermediate-grained aggregation network for composed image retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Kim *et al.*, 2024] Jaeho Kim, Seok-Ju Hahn, Yoontae Hwang, Junghye Lee, and Seulki Lee. Cafo: Feature-centric explanation on time series classification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1372–1382, 2024.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lai *et al.*, 2018] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018.

[Lee *et al.*, 2021] Dongha Lee, Seonghyeon Lee, and Hwanjo Yu. Learnable dynamic temporal pooling for time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8288–8296, 2021.

[Li *et al.*, 2025] Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5101–5109, 2025.

[Li *et al.*, 2026a] Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23373–23381, 2026.

[Li *et al.*, 2026b] Zixu Li, Yupeng Hu, Zhiwei Chen, Mingyu Zhang, Zhiheng Fu, and Liqiang Nie. Conesep: Cone-based robust noise-unlearning compositional network for composed image retrieval, 2026.

[Lin *et al.*, 2025] Xiao Lin, Zhichen Zeng, Tianxin Wei, Zhining Liu, and Hanghang Tong. Cats: Mitigating correlation shift for multivariate time series classification. *arXiv preprint arXiv:2504.04283*, 2025.

[Liu *et al.*, 2024a] Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, pages 31–40. IEEE, 2024.

[Liu *et al.*, 2024b] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for

time series forecasting. *arXiv preprint arXiv:2310.06625*, 2024.

[Liu *et al.*, 2025] Zhipeng Liu, Peibo Duan, Binwu Wang, Xuan Tang, Qi Chu, Changsheng Zhang, Yongsheng Huang, and Bin Zhang. Disms-ts: Eliminating redundant multi-scale features for time series classification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10817–10826, 2025.

[Morid *et al.*, 2023] Mohammad Amin Morid, Olivia R Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems*, 14(1):1–29, 2023.

[Nie *et al.*, 2022] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. arxiv 2022. *arXiv preprint arXiv:2211.14730*, 2022.

[Qin *et al.*, 2017] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

[Qiu *et al.*, 2024] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. In *Proc. VLDB Endow.*, pages 2363–2377, 2024.

[Qiu *et al.*, 2025a] Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. Dbloss: Decomposition-based loss function for time series forecasting. In *NeurIPS*, 2025.

[Qiu *et al.*, 2025b] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1185–1196, 2025.

[Song *et al.*, 2025] Yaqi Song, Rujie Wan, Li Li, Wanyu Wang, and Haonan Xing. A multi-scale temporal-frequency fusion network based on mlp for long-term time series forecasting. *International Journal of Machine Learning and Cybernetics*, 16(5):3943–3954, 2025.

[Sun *et al.*, 2024] Jie Sun, Jie Xiang, Yanqing Dong, Bin Wang, Mengni Zhou, Jiuhong Ma, and Yan Niu. Deep learning for epileptic seizure detection using a causal-spatio-temporal model based on transfer entropy. *Entropy*, 26(10):853, 2024.

[Wang *et al.*, 2024a] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024.

[Wang *et al.*, 2024b] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024.

[Wang *et al.*, 2026a] Hua Wang, Jinghao Lu, and Fan Zhang. Eeo-tfv: Escape-explore optimizer for web-scale time-series forecasting and vision analysis. In *Proceedings of the ACM Web Conference 2026*, pages 7271–7282, 2026.

[Wang *et al.*, 2026b] Hua Wang, Jinghao Lu, and Fan Zhang. Idealtsf: Can non-ideal data contribute to enhancing the performance of time series forecasting models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26224–26232, 2026.

[Xu *et al.*, 2025] Faming Xu, Yiding Wang, Gang Qu, Vince D Calhoun, Julia M Stephen, Tony W Wilson, Yu-Ping Wang, and Chen Qiao. A deep spatio-temporal architecture for dynamic ecn analysis with granger causality based causal discovery. *Pattern Recognition*, page 112346, 2025.

[Zhang *et al.*, 2022] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

[Zhang *et al.*, 2024] Liang Zhang, Jianping Zhu, Bo Jin, and Xiaopeng Wei. Multiview spatial-temporal meta-learning for multivariate time series forecasting. *Sensors (Basel, Switzerland)*, 24(14):4473, 2024.

[Zhang *et al.*, 2025] Zhenglin Zhang, Tengfei Wang, Zian Hu, Li-Zhuang Yang, and Hai Li. Multivariate time series approach integrating cross-temporal and cross-channel attention for dysarthria detection from speech. *Neurocomputing*, page 130708, 2025.

[Zhang *et al.*, 2026a] Fan Zhang, Shiming Fan, and Hua Wang. Time-tk: A multi-offset temporal interaction framework combining transformer and kolmogorov-arnold networks for time series forecasting. In *Proceedings of the ACM Web Conference 2026*, pages 7495–7506, 2026.

[Zhang *et al.*, 2026b] Fan Zhang, Shiming Fan, and Hua Wang. Timesaf: Towards llm-guided semantic asynchronous fusion for time series forecasting. *arXiv preprint arXiv:2604.12648*, 2026.

[Zhang *et al.*, 2026c] Mingyu Zhang, Zixu Li, Zhiwei Chen, Zhiheng Fu, Xiaowei Zhu, Jiajia Nie, Yinwei Wei, and Yupeng Hu. Hint: Composed image retrieval with dual-path compositional contextualized network. *arXiv preprint arXiv:2603.26341*, 2026.

[Zhong *et al.*, 2025] Shuhan Zhong, Weipeng Zhuo, Sizhe Song, Guanyao Li, Zhongyi Yu, and S-H Gary Chan. Mtm: A multi-scale token mixing transformer for irregular multivariate time series classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4074–4085, 2025.

[Zhu *et al.*, 2023] Chenglong Zhu, Xueling Ma, Weiping Ding, and Jianming Zhan. Long-term time series forecasting with multilinear trend fuzzy information granules for lstm in a periodic framework. *IEEE Transactions on Fuzzy Systems*, 32(1):322–336, 2023.