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.

TFMG: Time-Frequency Multi-View Graph Learning for Multivariate Time-Series Anomaly Detection

Hongjun Tian
Harbin Engineering University
Songtao Yang
Harbin Engineering University
Hongtao Song
songhongtao@hrbeu.edu.cn

Harbin Engineering University

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11337770/v1

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

Additional Declarations: No competing interests reported.

TFMG: Time-Frequency Multi-View Graph Learning for Multivariate Time-Series Anomaly Detection

Hongjun Tian[1], Songtao Yang[1], Hongtao Song[1*]

[1*]College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, Heilongjiang, China.

*Corresponding author(s). E-mail(s): songhongtao@hrbeu.edu.cn;
Contributing authors: hongjuntian@hrbeu.edu.cn;
songtaoyang@hrbeu.edu.cn;

Abstract

Multivariate time-series anomaly detection is essential for reliable industrial control and intelligent operation and maintenance. Anomalies may manifest as variable values deviating from their normal ranges or as disruptions of inter-variable dependency patterns. Graph-based methods can explicitly model cross-variable dependencies, but most approaches learn a single relation graph within each observation window, making it difficult to distinguish heterogeneous dependencies associated with trends, periodic patterns, and transient disturbances. Since inter-variable relations may vary across frequency scales, frequency-domain analysis helps characterize scale-dependent dependencies. However, frequency-domain methods mainly focus on frequency decomposition and time-frequency feature fusion, while rarely distinguishing inter-variable dependencies across frequency scales at the relation-structure level. To address these issues, we propose Time-Frequency Multi-View Graph Learning (TFMG). TFMG treats variables as nodes and jointly learns one directed time-domain relation graph and multiple directed frequency-band relation graphs to characterize integrated window-level dependencies and variable couplings at different frequency scales. After band-wise graph propagation, adaptive band-attention fusion uses band self-attention to exchange relation information across scales and cross-domain attention to assign state-dependent weights to frequency-band relation evidence using temporal summaries as queries. Hierarchical gated relation injection then performs time-frequency relation fusion and position-wise graph-context injection. Finally, a prediction-step-aware forecasting decoder generates multi-step forecasts, and

anomalies are identified from prediction deviations. Experiments on four multivariate time-series anomaly detection benchmarks show that TFMG attains the highest mean Affiliation F1 (Aff-F1) on each dataset among the compared methods, supporting its effectiveness across different data scales and application settings.

Keywords: multivariate time series; anomaly detection; time-frequency modeling; graph neural networks; multi-view learning

1 Introduction

With the continued development of sensing, control, and communication technologies, modern systems continuously generate growing volumes of high-dimensional time-series data[1–3]. Timely identification of observations that deviate from normal operating patterns is essential to ensuring system reliability and reducing failure risk. However, real-world multivariate time series are typically nonlinear, nonstationary, high-dimensional, and strongly coupled; anomalies may manifest either as shifts in the temporal dynamics of individual variables or as disruptions of normal inter-variable coordination [3], [10], [16]. Moreover, anomalous events are scarce and heterogeneous, and confirming them often requires domain expertise, making sufficiently comprehensive large-scale annotations difficult to obtain [1], [2], [4]. Consequently, many unsupervised and self-supervised approaches learn normal operating patterns and identify anomalies from the deviation of test observations from those patterns [4], [9–12], [31]. Their effectiveness depends not only on accurately characterizing the temporal evolution of each variable but also on preserving cross-variable dependencies under normal operation.

Existing deep anomaly detection methods can effectively characterize complex temporal dynamics through reconstruction and forecasting [5–10]. However, multivariate anomalies do not always produce conspicuous numerical deviations in individual variables. When an anomaly primarily manifests as a mismatch in variable coordination, joint encoding or latent-space interaction alone may not adequately characterize the structural change. Graph learning has therefore become an important means of explicitly modeling cross-variable dependencies [10–12], [16]: it treats variables as nodes and organizes their information exchange through learnable relation structures and graph propagation. Recent studies have further advanced relation modeling through data-driven graph construction, causal relations, structural priors, and topology enhancement [41–44]. Nevertheless, these methods mainly investigate which relations exist between variables and how they change with the observed state, while seldom distinguishing relation evidence associated with different dynamic scales within the same observation window.

An observation window generally contains signal components with different rates of change. Frequency decomposition provides another structured representation of the same sequence [19], [20], distributing slow variations and rapid fluctuations across

different frequency ranges. Amplitude co-variation, energy differences, and phase relationships between variables may vary across frequency ranges [21], [23]. Accordingly, constructing a single relation graph from a full-spectrum or unified temporal representation may compress dependency evidence from different dynamic scales into one adjacency structure. For example, two variables may exhibit strong coordination at low frequencies but weaker or qualitatively different associations under high-frequency disturbances. This observation does not imply that different bands must form mutually independent topologies; rather, it motivates examining whether they preserve shared relations while providing scale-discriminative structural information.

Existing time-frequency methods mainly exploit frequency information through spectral encoding, frequency decomposition, spectral reconstruction, and time-frequency feature fusion [20], [22], [31]. Recent studies have also explored common spectral latent spaces to encode time-series patterns across heterogeneous domains [25]. Although these approaches broaden the representation space of time series, they typically emphasize spectral feature extraction, correlation selection, or feature fusion; different frequency scales are rarely modeled as explicit variable-relation views that can be propagated and compared independently. Moreover, even when multiple band representations are available, the relevance of their relation evidence may vary across operating states, as the dynamic scales most informative for future behavior may change with the current state. Fixed averaging, direct concatenation, or state-independent fusion may therefore fail to fully exploit relation evidence across frequency scales. This gives rise to two related questions: how to construct band-specific relation views over a shared variable set, and how to aggregate those views adaptively according to the current temporal state without erasing scale differences.

To address these questions, we propose TFMG. TFMG jointly learns one time-domain relation graph and multiple frequency-band relation graphs over a shared set of variable nodes to capture integrated window-level dependencies and frequency-scale-specific relation evidence, respectively. After graph propagation, adaptive band-attention fusion aggregates the frequency-band relations according to the current temporal state, and hierarchical gated relation injection then fuses them with the time-domain relation and injects the fused relation into the temporal representation; anomalies are ultimately identified from multi-step forecasting errors. Unlike existing approaches that primarily fuse frequency information as auxiliary features, TFMG introduces frequency scales at the relation-structure level, allowing cross-variable dependencies at different dynamic scales to be explicitly distinguished and adaptively exploited.

The main contributions of this paper are summarized as follows:

1. We propose TFMG, a time-frequency multi-view graph learning framework. Over a shared set of variable nodes, TFMG jointly constructs a time-domain relation graph and multiple frequency-band relation graphs, extending frequency scales from feature representation to explicit relation modeling so that cross-variable dependencies at different dynamic scales can be represented and propagated separately.
2. We propose an adaptive band-attention fusion mechanism. Band self-attention models interactions among graph-propagated band representations, while cross-domain attention uses variable-level temporal summaries as queries to dynamically

regulate the contribution of each band-relation view. Relation evidence at different frequency scales can therefore be adaptively incorporated into temporal modeling according to the operating state.

3. We design a hierarchical gated relation injection mechanism. A two-stage gated residual structure first fuses time-domain and frequency-domain relation representations and then adaptively injects graph context into temporal features at each historical position. The enhanced temporal representation is subsequently used for multi-step forecasting, with prediction residuals serving as anomaly evidence.
4. We conduct extensive experiments on multiple public multivariate time-series benchmarks. The results show that TFMG outperforms a range of advanced anomaly detection methods, while ablation studies and multi-view relation-structure analysis further validate the effectiveness of frequency-band relation modeling and adaptive band-attention fusion.

The remainder of this paper is organized as follows. Section 2 reviews multivariate time series anomaly detection, graph-based relation learning, and time-frequency joint modeling. Section 3 introduces the overall architecture of TFMG and its key modules. Section 4 reports the experimental setup, performance comparison, ablation study, parameter sensitivity analysis, and relation interpretability analysis. Section 5 concludes the paper and discusses future research directions.

2 Related Work

2.1 Multivariate Time Series Anomaly Detection

Multivariate time-series anomaly detection aims to identify behavior that deviates from normal operating patterns across multiple correlated observations. With the development of deep learning, existing studies can be broadly categorized into three main paradigms: reconstruction, forecasting, and representation learning. Reconstruction-based methods learn the latent structure of normal sequences and identify anomalies from discrepancies between inputs and reconstructions [5–9], [13–15]. Forecasting-based methods estimate future states from historical observations and use the residuals between forecasts and observations as anomaly evidence [10–12]. One-class learning, domain adaptation, and contrastive learning have also been employed to improve the discriminability of normal representations and generalization across scenarios [26–32]. Together, these lines of research have substantially strengthened the modeling of complex temporal dynamics and diverse anomaly patterns.

More recently, convolutional networks, recurrent networks, and Transformers have further improved the modeling of local variation, long-range dependencies, and periodic patterns [33–37], while general-purpose forecasting architectures have provided new temporal backbones for residual-based anomaly detection [38–40]. Nevertheless, these approaches primarily emphasize temporal patterns, latent representations, and sequential dependencies, whereas multivariate anomalies may also arise from changes in normal coordination between variables. Because joint encoding or latent-space interaction alone does not explicitly constrain inter-variable information propagation,

cross-variable relation modeling has become an important direction in multivariate time-series anomaly detection [10–12], [16–18].

2.2 Graph Learning for Multivariate Relations

Graph learning represents variables or sensors in multivariate time series as nodes and describes their dependencies explicitly through edges, enabling structured information propagation and relation aggregation with graph convolution or graph attention [10–12], [16], [46]. Compared with feature fusion performed solely in a latent space, an explicit graph more transparently specifies the paths along which variables exchange information. Research has evolved from static graphs based on predefined topology, domain knowledge, or statistical associations to adaptive graphs estimated from data-driven similarities, learnable node representations, and dynamic mechanisms, thereby accommodating changes in variable relations across operating states [16–18], [44], [45].

Building on this foundation, recent studies have enhanced the representation of variable relations through directional dependencies, structural priors, multi-view spatio-temporal modeling, and topology enhancement [41–44]. Relation selection, neighborhood refinement, and weak-connection suppression have also been introduced to reduce the interference of noisy edges during graph propagation [47], [48]. These advances have substantially improved the capacity of graph models to capture complex cross-variable relations and to adapt adjacency structures to current observations. Overall, however, existing graph-learning methods mainly focus on adjacency construction, dynamic updating, and propagation quality. Relation differences associated with distinct dynamic scales within the same temporal window remain underexplored, and dependencies at different temporal scales are usually compressed into a unified relation structure.

2.3 Time-Frequency Joint Modeling

Frequency analysis provides a structured representation of time series from the perspective of variation scale. Existing studies commonly use the Fourier transform, wavelet transform, or other time-frequency decompositions [19], [20] to map raw sequences into the frequency domain and enrich temporal representations with spectral magnitudes, periodic patterns, or local frequency components. Such methods support anomaly detection through frequency-domain encoding, spectral reconstruction, and time-frequency feature fusion [24]; finer frequency partitioning has also been used to improve the representation of different anomaly patterns [21], [22]. Meanwhile, some recent methods consider channel correlations in different frequency regions of multivariate data and adaptively select cross-variable information according to frequency scale [21], [23]. These studies demonstrate that frequency information can complement purely temporal representations in modeling periodicity and scale-dependent variation.

Despite broadening the representation space of time-series anomaly detection, these methods still focus primarily on frequency-feature extraction, periodic-pattern modeling, and cross-channel information selection. Different frequency scales are usually treated as feature components and subsequently concatenated, weighted, or fused by attention, rather than organized as explicit variable-relation structures that can

be propagated, compared, and updated independently. In a multivariate system, dependency strengths, synchronization patterns, and interaction modes may differ substantially across frequency ranges. Explicitly modeling cross-variable dependencies at different frequency scales at the relation level and exploiting their complementary evidence therefore remain insufficiently explored.

3 Method

3.1 Problem Formulation

Given an N-dimensional multivariate time series as input, let t denote the beginning of a prediction window, and let ω and p denote the historical-window length and prediction length, respectively. For variable i, the observations over the interval $[t-\omega, t-1]$ are denoted by $\mathbf{x}_i$, and all variable windows are stacked row-wise to form $\mathbf{X}$. TFMG takes $\mathbf{X}$ as input and outputs the N-variable, p-step prediction matrix $\widehat{\mathbf{Y}}$.

$$\mathbf{x}_i = [x_i^{t-\omega}, \ldots, x_i^{t-1}]^\top \in \mathbb{R}^{\omega} \tag{1}$$

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times \omega} \tag{2}$$

$$\widehat{\mathbf{Y}} = \mathcal{F}_\theta(\mathbf{X}) = [\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_N]^\top \in \mathbb{R}^{N \times p} \tag{3}$$

Here, $\widehat{\mathbf{y}}_i = [\widehat{y}_{i,1}, \ldots, \widehat{y}_{i,p}]^\top$ is the prediction sequence of variable i, and $\nu \in \{1, \ldots, p\}$ denotes a particular prediction step. Only normal windows are used during training. At test time, anomaly scores are computed from the deviations between predictions and the corresponding observations. In the following, the symbols i, j, ℓ, ν, k index the receiving variable, sending variable, historical position, prediction step, and frequency-band index, respectively. Encoder parameters are shared across channels, and information between variables is exchanged through explicit graph propagation.

3.2 Overall Framework

The overall framework of TFMG is shown in Fig. 1. Given a historical window, TFMG first constructs temporal and frequency-domain representations through two parallel branches. The time-domain branch extracts position-wise temporal features and learns a time-domain relation graph, while the frequency branch decomposes the input into multiple frequency bands and learns frequency-band relation graphs. Adaptive band-attention fusion then aggregates the frequency-band relation representations according to the current temporal state. Finally, hierarchical gated relation injection fuses the aggregated frequency-domain relation with the time-domain relation and injects the fused relation into the main temporal pathway, after which the prediction-step-aware forecasting decoder produces multi-step forecasts; forecast residuals are used to compute anomaly scores.

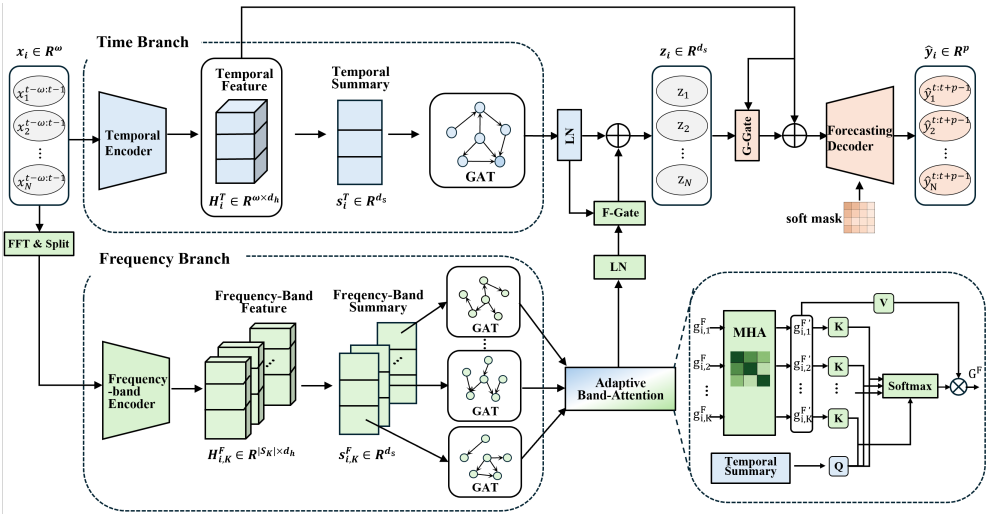

Fig. 1 Overall architecture of TFMG. Temporal and frequency-band relation representations are learned in parallel and subsequently integrated for multi-step forecasting and anomaly scoring.

3.3 Time-Domain Relation Modeling

3.3.1 Time-Domain Representation

The time-domain branch takes the variable window $\mathbf{x}_i$ as input, with encoder parameters shared across variables. The temporal encoder $\mathcal{T}$ consists of stacked dilated residual convolutions, positional encoding, and temporal self-attention. The convolutions extract local variations, whereas attention captures long-range dependencies within the window, producing features over ω historical positions, denoted by $\mathbf{H}_i^{\mathrm{T}}$.

$$\mathbf{H}_i^{\mathrm{T}} = \mathcal{T}(\mathbf{x}_i) \in \mathbb{R}^{\omega \times d_h} \tag{4}$$

To compress position-level states into graph node descriptors, temporal pooling first uses a shared scoring network to estimate weights over historical positions, then computes a weighted sum of $\mathbf{H}_i^{\mathrm{T}}$ and projects it into a d_s-dimensional relation space. We denote this operation by $\mathrm{AttPool}_{\mathrm{T}}$. The resulting temporal summary $\mathbf{s}_i^{\mathrm{T}}$ is used for time-domain relation graph construction and subsequent cross-domain queries. Here, d_h and d_s denote the temporal-feature and relation-summary dimensions, respectively.

$$\mathbf{s}_i^{\mathrm{T}} = \mathrm{AttPool}_{\mathrm{T}}\left(\mathbf{H}_i^{\mathrm{T}}\right) \in \mathbb{R}^{d_s} \tag{5}$$

The position-wise features $\mathbf{H}_i^{\mathrm{T}}$ are preserved as the main temporal pathway for later graph-context injection and prediction, while $\mathbf{s}_i^{\mathrm{T}}$ is used for relation modeling and cross-domain attention.

3.3.2 Time-Domain Graph Learning

The normalized temporal summary is denoted by $\mathbf{u}_i^{\mathrm{T}}$. Time-domain edge scores are computed by a scaled dot product between the summaries, where d_g is the

edge-projection dimension. Distinct query and key projections make the relation asymmetric, and e_{ij}^{T} measures the dependence of node i on node j. This yields a window-conditioned directed time-domain relation graph whose edge scores characterize asymmetric inter-variable dependencies under the current operating state.

$$\mathbf{u}_i^{\mathrm{T}} = \mathrm{LN}\left(\mathbf{s}_i^{\mathrm{T}}\right) \tag{6}$$

$$e_{ij}^{\mathrm{T}} = d_g^{-1/2}\left(\mathbf{W}_{\mathrm{Q}}^{\mathrm{T}}\mathbf{u}_i^{\mathrm{T}}\right)^{\top}\left(\mathbf{W}_{\mathrm{K}}^{\mathrm{T}}\mathbf{u}_j^{\mathrm{T}}\right) \tag{7}$$

To suppress low-confidence time-domain connections, the candidate edge scores of each node together with a learnable null-edge score are passed through Sparsemax [49], yielding a sparse directed time-domain adjacency matrix. Weak connections are assigned exactly zero weight, while the null edge represents retaining the node's own state without receiving external messages. Consequently, the number of effective neighbors is determined adaptively according to the relation evidence in the current window. The resulting nonzero adjacency weights further serve as logarithmic priors in subsequent graph attention.

The model then performs graph attention network (GAT) propagation on the time-domain relation graph to aggregate information from related variables. The incoming-neighbor set of each node contains the variables with positive time-domain adjacency weights. GAT first computes attention weights over these neighbors and then performs weighted aggregation:

$$\alpha_{ij}^{\mathrm{T}} = \mathrm{softmax}_{j\in\mathcal{N}_i^{\mathrm{T}}}\left(\mathrm{LeakyReLU}\left(\mathbf{a}_{\mathrm{T}}^{\top}\left[\mathbf{W}_{\mathrm{T}}\mathbf{u}_i^{\mathrm{T}}\|\mathbf{W}_{\mathrm{T}}\mathbf{u}_j^{\mathrm{T}}\right]\right) + \lambda_{\mathrm{A}}\log\left(\max\left(A_{ij}^{\mathrm{T}},\epsilon\right)\right)\right) \tag{8}$$

$$\mathbf{g}_i^{\mathrm{T}} = \mathrm{LN}\left(\mathbf{R}_{\mathrm{T}}\mathbf{u}_i^{\mathrm{T}} + \sum_{j\in\mathcal{N}_i^{\mathrm{T}}}\alpha_{ij}^{\mathrm{T}}\mathbf{W}_{\mathrm{T}}\mathbf{u}_j^{\mathrm{T}}\right) \tag{9}$$

The projection, attention, and residual parameters of the time-domain GAT are learned independently from those of the frequency-domain GAT. The adjacency-prior coefficient controls the contribution of the learned graph weight, and a positive numerical-stability constant prevents taking the logarithm of zero.

3.4 Frequency-Band Relation Modeling

3.4.1 Frequency-Band Representation

The frequency branch operates in parallel with the temporal branch and constructs spectral evidence for band relation learning. The preprocessing operator $\mathcal{P}_{\mathrm{F}}$ applies mean removal and a Hann window normalized by mean-square energy to $\mathbf{x}_i$. A real fast Fourier transform then produces the complex spectrum $\mathbf{R}_i^{\mathrm{F}}$. For a real sequence of length ω, the number of nonredundant frequency bins is $F = \lfloor \omega/2 \rfloor + 1$.

$$\mathbf{R}_i^{\mathrm{F}} = \mathrm{rFFT}\left(\mathcal{P}_{\mathrm{F}}(\mathbf{x}_i)\right) \in \mathbb{C}^F \tag{10}$$

From the complex spectrum $\mathbf{R}_i^{\mathrm{F}}$, the model constructs a frequency-bin descriptor sequence $\mathbf{D}_i$. Its channels comprise log magnitude, log power, and sign-preserving compressed real and imaginary parts, while retaining spectral intensity and phase direction. The F bins are partitioned in frequency order into K contiguous, non-overlapping bands. The band boundaries are denoted by q_k. We use an approximately octave-like partition with finer resolution at low frequencies and ensure after rounding that each band contains at least one bin.

$$\mathcal{S}_k = [q_{k-1}, q_k) \cap \mathbb{Z}, 0 = q_0 < \cdots < q_K = F \tag{11}$$

For the kth band, the band encoder $\mathcal{E}_k^{\mathrm{F}}$ first uses a frequency-bin projection shared across bands to align the feature space, and then uses a band-specific convolutional adapter to extract local spectral shapes. Encoding is performed independently within each variable without mixing information across variables, yielding the spectral representation sequence of variable i, denoted by $\mathbf{H}_{i,k}^{\mathrm{F}}$, in that band. Each row is a hidden vector for one frequency position after local spectral enhancement.

$$\mathbf{H}_{i,k}^{\mathrm{F}} = \mathcal{E}_k^{\mathrm{F}}\left(\mathbf{D}_i[\mathcal{S}_k]\right) \in \mathbb{R}^{|\mathcal{S}_k| \times d_h} \tag{12}$$

To obtain node summaries for graph construction, the model concatenates the within-band mean, standard deviation, and attention-pooled representation of $\mathbf{H}_{i,k}^{\mathrm{F}}$ and applies a linear projection. Attention pooling assigns weights from the hidden frequency-bin representations, emphasizing informative bins while retaining their true frequency positions. The resulting $\mathbf{s}_{i,k}^{\mathrm{F}}$ serves as the node input to the kth frequency-band relation graph. Here, $\mathbf{p}_{\mathcal{S}_k}$ denotes the normalized absolute positions of the bins in band $\mathcal{S}_k$ within the complete nonredundant spectrum.

$$\mathbf{s}_{i,k}^{\mathrm{F}} = \mathbf{W}_{\mathrm{s}}^{\mathrm{F}}\left[\mathrm{Mean}\left(\mathbf{H}_{i,k}^{\mathrm{F}}\right); \mathrm{Std}\left(\mathbf{H}_{i,k}^{\mathrm{F}}\right); \mathrm{AttPool}_{\mathrm{F}}\left(\mathbf{H}_{i,k}^{\mathrm{F}}, \mathbf{p}_{\mathcal{S}_k}\right)\right] \tag{13}$$

3.4.2 Frequency-Band Graph Learning

The normalized frequency-band summary is denoted by $\mathbf{u}_{i,k}^{\mathrm{F}}$:

$$\mathbf{u}_{i,k}^{\mathrm{F}} = \mathrm{LN}\left(\mathbf{s}_{i,k}^{\mathrm{F}}\right) \tag{14}$$

The edge scores of the kth frequency-band relation graph are computed exclusively from complex-spectrum relations within that band. For variables i and j, the model uses amplitude similarity, energy difference, phase coherence, and signed phase lag to describe spectral-shape co-variation, relative energy, phase synchronization, and phase direction. A shared edge-scoring network ψ_{F} maps these statistics to directed edge scores.

Specifically, let $\mathbf{a}_{i,k}$ denote the log-magnitude vector of variable i in band k, let $E_{i,k}$ denote its mean spectral energy within the band, and define the inter-variable phase difference as $\Delta\phi_{ij}(f) = \phi_i(f) - \phi_j(f)$. The weighted cosine and sine means of the phase difference are denoted by $C_{ij,k}$ and $S_{ij,k}$, respectively. Here, $w_{ij,k}(f)$ is a nonnegative weight obtained by normalizing the geometric mean of the two variables' magnitudes at each frequency bin.

$$C_{ij,k} = \sum_{f \in \mathcal{S}_k} w_{ij,k}(f) \cos\left(\Delta\phi_{ij}(f)\right) \tag{15}$$

$$S_{ij,k} = \sum_{f \in \mathcal{S}_k} w_{ij,k}(f) \sin\left(\Delta\phi_{ij}(f)\right), \qquad \sum_{f \in \mathcal{S}_k} w_{ij,k}(f) = 1 \tag{16}$$

Here, $\epsilon > 0$ is a numerical-stability constant. Based on these statistics, the four measures are defined as follows:

$$\mathrm{sim}_{ij,k}^{\mathrm{amp}} = \frac{\mathbf{a}_{i,k}^{\top} \mathbf{a}_{j,k}}{\|\mathbf{a}_{i,k}\|_2 \|\mathbf{a}_{j,k}\|_2 + \epsilon} \tag{17}$$

$$\Delta E_{ij,k} = \tanh\left(\log\left(E_{i,k} + \epsilon\right) - \log\left(E_{j,k} + \epsilon\right)\right) \tag{18}$$

$$\mathrm{coh}_{ij,k}^{\mathrm{pha}} = \sqrt{C_{ij,k}^2 + S_{ij,k}^2} \tag{19}$$

$$\mathrm{lag}_{ij,k}^{\mathrm{pha}} = S_{ij,k} \tag{20}$$

$$\mathbf{r}_{ij,k}^{\mathrm{F}} = [\mathrm{sim}_{ij,k}^{\mathrm{amp}}; \Delta E_{ij,k}; \mathrm{coh}_{ij,k}^{\mathrm{pha}}; \mathrm{lag}_{ij,k}^{\mathrm{pha}}] \tag{21}$$

$$e_{ij,k}^{\mathrm{F}} = \psi_{\mathrm{F}}\left(\mathbf{r}_{ij,k}^{\mathrm{F}}\right) \tag{22}$$

Swapping i and j changes the signed energy-difference and phase-lag terms. The relation vector is therefore not constrained to be symmetric, allowing the edge-scoring network to learn asymmetric scores; in general, $e_{ij,k}^{\mathrm{F}} \neq e_{ji,k}^{\mathrm{F}}$. Each band consequently yields a window-conditioned directed score matrix $\mathbf{E}_k^{\mathrm{F}}$ that explicitly represents variable dependencies at that frequency scale.

Each frequency-band relation graph applies the same adaptive sparsification and propagation mechanism defined for the time-domain relation graph. The candidate frequency-band scores together with the learnable null-edge score are passed through Sparsemax, yielding a sparse directed frequency-band adjacency matrix. Weak connections are assigned zero weight, and the resulting nonzero adjacency weights serve as logarithmic priors in frequency-band graph attention.

For each frequency band, the band GAT follows Eqs. (8) and (9), replacing the temporal summary, adjacency matrix, neighbor set, attention coefficients, and GAT parameters with their frequency-domain counterparts. The resulting attention and aggregation equations are

$$\begin{aligned}\alpha_{ij,k}^{\mathrm{F}} = \mathrm{softmax}_{j \in \mathcal{N}_{i,k}^{\mathrm{F}}}\big(&\,\mathrm{LeakyReLU}\big(\mathbf{a}_{\mathrm{F}}^{\top}\left[\mathbf{W}_{\mathrm{F}}\mathbf{u}_{i,k}^{\mathrm{F}} \| \mathbf{W}_{\mathrm{F}}\mathbf{u}_{j,k}^{\mathrm{F}}\right]\big) \\ &+ \lambda_{\mathrm{A}} \log\big(\max\big(A_{ij,k}^{\mathrm{F}}, \epsilon\big)\big)\big)\end{aligned} \tag{23}$$

$$\mathbf{g}_{i,k}^{\mathrm{F}} = \mathrm{LN}\left(\mathbf{R}_{\mathrm{F}}\mathbf{u}_{i,k}^{\mathrm{F}} + \sum_{j \in \mathcal{N}_{i,k}^{\mathrm{F}}} \alpha_{ij,k}^{\mathrm{F}} \mathbf{W}_{\mathrm{F}}\mathbf{u}_{j,k}^{\mathrm{F}}\right) \tag{24}$$

where $\mathcal{N}_{i,k}^{\mathrm{F}} = \{j \mid A_{ij,k}^{\mathrm{F}} > 0\}$. The frequency-domain GAT parameters are shared across bands, producing one frequency-band relation representation for each variable and frequency band.

3.5 Adaptive Band-Attention Fusion

3.5.1 Cross-Band Interaction

Adaptive band-attention fusion is applied after graph propagation. The K frequency-band relation representations of the same variable are ordered by frequency and combined with the band positional embedding $\mathbf{E}^{\mathrm{B}}$ to form the band sequence $\mathbf{G}_i^{\mathrm{B}}$. Band multi-head self-attention, denoted by $\mathrm{MHA}_{\mathrm{B}}$ with H_{B} attention heads, models interactions across bands. A residual connection then produces the interaction-enhanced sequence $\mathbf{G}_i^{\mathrm{F}\prime}$, whose kth position is denoted by $\mathbf{g}_{i,k}^{\mathrm{F}\prime}$.

$$\mathbf{G}_i^{\mathrm{B}} = \left[\mathbf{g}_{i,1}^{\mathrm{F}}, \ldots, \mathbf{g}_{i,K}^{\mathrm{F}}\right] + \mathbf{E}^{\mathrm{B}} \tag{25}$$

$$\mathbf{G}_i^{\mathrm{F}\prime} = \mathbf{G}_i^{\mathrm{B}} + \mathrm{MHA}_{\mathrm{B}}\left(\mathrm{LN}\left(\mathbf{G}_i^{\mathrm{B}}\right)\right) \tag{26}$$

3.5.2 Cross-Domain Band Weighting

Cross-domain attention uses the temporal summary $\mathbf{s}_i^{\mathrm{T}}$ as the query and the interaction-enhanced frequency-band relation representation $\mathbf{g}_{i,k}^{\mathrm{F}\prime}$ as the key to estimate the contribution of each frequency-band relation view under the current state. Here, $\mathbf{W}_{\mathrm{Q},h}^{\mathrm{x}}$ and $\mathbf{W}_{\mathrm{K},h}^{\mathrm{x}}$ are the query and key projections of attention head h, H_{x} is the number of heads, d_{x} is the dimension of each head, and b_k is the band bias:

$$\beta_{i,k} = \frac{1}{H_{\mathrm{x}}} \sum_{h=1}^{H_{\mathrm{x}}} \mathrm{softmax}_k\left(\frac{(\mathbf{W}_{\mathrm{Q},h}^{\mathrm{x}} \mathbf{s}_i^{\mathrm{T}})^{\top} (\mathbf{W}_{\mathrm{K},h}^{\mathrm{x}} \mathbf{g}_{i,k}^{\mathrm{F}\prime})}{\sqrt{d_{\mathrm{x}}}} + b_k\right) \tag{27}$$

The attention scores are normalized across the band dimension to generate state-dependent weights over the band relation representations for each variable. A larger weight indicates that the corresponding band better matches the current temporal summary. The model then uses these weights to aggregate the value representations of the bands, yielding the frequency-domain relation representation.

The weights $\beta_{i,k}$ are normalized over the K bands. The interaction-enhanced band representations are mapped by the shared value projection $\mathcal{V}_{\mathrm{x}}$ and then combined by a weighted sum to obtain the aggregated frequency-domain relation $\mathbf{g}_i^{\mathrm{F}}$ for variable i. Band self-attention models relation interaction, whereas cross-domain attention performs state-conditioned view weighting; the two operations thus serve relation integration and adaptive view weighting, respectively.

$$\mathbf{g}_i^{\mathrm{F}} = \sum_{k=1}^{K} \beta_{i,k} \mathcal{V}_{\mathrm{x}}(\mathbf{g}_{i,k}^{\mathrm{F}\prime}) \tag{28}$$

3.6 Hierarchical Gated Relation Injection

The reliability of relation evidence extracted from different frequency bands may vary across operating states, and frequency-band relation graphs may contain relations that are irrelevant to the current state or redundant with those captured by the time-domain relation graph. Direct fusion can therefore introduce redundant information and interfere with prediction. Moreover, window-level graph representations and position-wise temporal features operate at different granularities. To address these issues, TFMG employs hierarchical gated relation injection: the first stage controls the contribution of frequency-domain relations during time-frequency fusion, while the second adaptively regulates the injection of graph context at each historical position. This design preserves complementary time-frequency dependencies while suppressing irrelevant or redundant relational information.

3.6.1 Time-Frequency Relation Fusion

The first stage, F-Gate, fuses the temporal and frequency-domain relation representations. Let $\mathbf{t}_i = \mathrm{LN}(\mathbf{g}_i^{\mathrm{T}})$ and $\mathbf{f}_i = \mathrm{LN}(\mathbf{g}_i^{\mathrm{F}})$ denote their normalized forms. The gating network outputs a scalar frequency-residual gate for each variable, while $\mathbf{W}_{\mathrm{FT}}$ maps the frequency-domain relation into the temporal relation space. The fused representation $\mathbf{z}_i$ is then passed to G-Gate. Here, σ denotes the sigmoid function, and the scalar gate lies in $(0, 1)$.

$$\gamma_i^{\mathrm{F}} = \sigma\left(\mathrm{MLP}_{\mathrm{F}}\left([\mathbf{t}_i; \mathbf{f}_i; \mathbf{t}_i \odot \mathbf{f}_i]\right)\right) \tag{29}$$

$$\mathbf{z}_i = \mathbf{t}_i + \gamma_i^{\mathrm{F}} \mathbf{W}_{\mathrm{FT}} \mathbf{f}_i \tag{30}$$

3.6.2 Graph-Context Injection

The second stage, G-Gate, injects the fused relation representation into the main temporal pathway. The fused representation $\mathbf{z}_i$ is transformed by a shared projection and layer normalization to obtain the graph context $\mathbf{c}_i$. Each historical position computes an independent scalar gate from its temporal state and $\mathbf{c}_i$; the gate is broadcast over the feature dimension to regulate the injection strength of relation information at that temporal position.

$$\mathbf{c}_i = \mathrm{LN}\left(\mathbf{W}_{\mathrm{c}} \mathbf{z}_i\right) \tag{31}$$

$$\gamma_{i,\ell}^{\mathrm{G}} = \sigma\left(\mathrm{MLP}_{\mathrm{G}}\left([\mathbf{h}_{i,\ell}^{\mathrm{T}}; \mathbf{c}_i]\right)\right) \tag{32}$$

$$\mathbf{h}_{i,\ell}^{\mathrm{E}} = \mathbf{h}_{i,\ell}^{\mathrm{T}} + \gamma_{i,\ell}^{\mathrm{G}} \mathbf{c}_i \tag{33}$$

3.7 Prediction-Step-Aware Forecasting Decoder

For each variable, the decoder independently reads the enhanced temporal representation $\mathbf{H}_i^{\mathrm{E}} = [\mathbf{h}_{i,1}^{\mathrm{E}}, \ldots, \mathbf{h}_{i,\omega}^{\mathrm{E}}]$ and produces p forecasts beginning at time t in a single

pass. For prediction step ν, the terminal historical state $\mathbf{h}_{i,\omega}^{\mathrm{E}}$ is projected by $\mathbf{W}_{\mathrm{q}}$ and combined with a learnable step embedding $\mathbf{e}_{\nu}^{\mathrm{step}}$ to form the query:

$$\mathbf{q}_{i,\nu} = \mathbf{W}_{\mathrm{q}}\mathbf{h}_{i,\omega}^{\mathrm{E}} + \mathbf{e}_{\nu}^{\mathrm{step}} \tag{34}$$

This construction allows different forecast horizons to attend to distinct historical information. To introduce a temporal-distance prior beyond content matching, the model defines the scale-normalized relative distance between prediction step ν and historical position ℓ as follows:

$$d_{\nu,\ell} = \frac{\omega + \nu - \ell}{\omega} \tag{35}$$

A larger distance indicates that the historical position is farther from the target time. The mask network concatenates the historical state, prediction-step query, scale-normalized relative distance, and its exponentially decayed value to generate a position-dependent continuous soft mask:

$$m_{i,\nu,\ell} = \sigma\left(\mathrm{MLP}_{\mathrm{m}}\left(\left[\mathbf{h}_{i,\ell}^{\mathrm{E}}; \mathbf{q}_{i,\nu}; d_{\nu,\ell}; \mathrm{e}^{-d_{\nu,\ell}}\right]\right)\right) \tag{36}$$

Here, $\mathrm{MLP}_{\mathrm{m}}$ is the mask network and σ is the sigmoid function. During decoding, the soft mask enters the content score of attention head h as a logarithmic prior:

$$\xi_{i,\nu,\ell}^{h} = \frac{\left(\mathbf{W}_{\mathrm{Q},h}^{\mathrm{dec}}\mathbf{q}_{i,\nu}\right)^{\top}\left(\mathbf{W}_{\mathrm{K},h}^{\mathrm{dec}}\mathbf{h}_{i,\ell}^{\mathrm{E}}\right)}{\sqrt{d_{\mathrm{dec}}}} + \lambda_{\mathrm{m}}\log\left(m_{i,\nu,\ell} + \epsilon\right) \tag{37}$$

$\mathbf{W}_{\mathrm{Q},h}^{\mathrm{dec}}$ and $\mathbf{W}_{\mathrm{K},h}^{\mathrm{dec}}$ are the query and key projections, d_{dec} is the dimension of each head, λ_{m} controls the mask-prior strength, and $\epsilon > 0$ is a numerical-stability constant. A smaller mask induces a larger negative bias and thus reduces the attention weight of the corresponding historical position. Attention is then normalized over historical positions and used to aggregate the value representations of all heads. The multi-head output is fused residually with the query, followed by a standard feed-forward residual block and the prediction head to obtain $\widehat{y}_{i,\nu}$.

3.8 Training Objective and Anomaly Scoring

3.8.1 Training Objective

The model is trained only on normal windows, with the mean squared error between the prediction matrix $\widehat{\mathbf{Y}}$ and the corresponding observations as the primary objective.

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{pN}\sum_{\nu=1}^{p}\sum_{i=1}^{N}\left(\widehat{y}_{i,\nu} - x_{i}^{t+\nu-1}\right)^{2} \tag{38}$$

When $K > 1$, the model applies a diversity regularization term to the off-diagonal elements of the frequency-band adjacency matrices to discourage different frequency-band relation graphs from converging to nearly identical topologies. Let $\mathbf{v}_k$ be the vectorized form of $\mathbf{A}_k^{\mathrm{F}}$ after removing the diagonal. A penalty is imposed when the stabilized cosine similarity between two frequency-band relation graphs exceeds the margin δ_{div}, and λ_{div} controls the weight of this regularizer in the total objective.

$$\mathcal{L}_{\text{div}} = \frac{2}{K(K-1)} \sum_{1 \le k < k' \le K} \left[\max\left(0, \cos_\epsilon(\mathbf{v}_k, \mathbf{v}_{k'}) - \delta_{\text{div}}\right)\right]^2 \tag{39}$$

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}} \tag{40}$$

3.8.2 Anomaly Scoring

At test time, TFMG measures the deviation of the current state from normal patterns using the discrepancy between predictions and observations. For each variable and prediction step, the model computes the squared residual between the prediction and the observation at the corresponding target time in the same normalized space as the input:

$$e_{i,\nu} = \left(\widehat{y}_{i,\nu} - x_i^{t+\nu-1}\right)^2 \tag{41}$$

The variable-level squared residual measures the prediction deviation of the corresponding variable at the target time. A larger value indicates that the current observation is more difficult to explain using the temporal dynamics and variable dependencies learned from normal states.

To mitigate differences in variable scale and forecast horizon, robust calibration statistics for the residual of each variable and prediction step are estimated from the normal validation set and kept fixed during testing. A target time may receive multiple overlapping pieces of evidence from sliding windows and multi-step forecasts; prediction-step reliability is estimated from validation errors, and the evidence for each variable is aggregated using a weighted log-sum-exp. The maximum over variables then serves as the time-point anomaly score, highlighting a pronounced deviation in any variable.

4 Experiments

The experiments address four questions:
RQ1, whether TFMG can achieve competitive detection performance across different data scales and application scenarios;
RQ2, how the main relation-modeling and fusion components contribute to detection performance;
RQ3, whether the model is sensitive to key structural and training hyperparameters;
RQ4, whether the learned time-domain relation graph and frequency-band relation graphs provide distinguishable and complementary relation evidence.

4.1 Datasets and Experimental Setup

4.1.1 Datasets

We use four public multivariate time-series anomaly detection benchmarks: SMD for Internet server monitoring, MSL for spacecraft telemetry, WADI for water-distribution control, and GECCO for drinking-water quality monitoring. Their markedly different

Table 1 Statistics of the four multivariate time series datasets

Dataset	Train points	Test points	Variables	Anomaly rate
SMD	708,405	708,420	38	4.1563%
MSL	58,317	73,729	55	10.5332%
WADI	153,315	34,561	123	5.7811%
GECCO	69,260	69,261	9	1.0540%

data distributions and characteristics enable evaluation across multi-entity telemetry, high-dimensional industrial processes, and low-dimensional sensor systems. The statistics of the four datasets are summarized in Table 1.

SMD (Server Machine Dataset) [50]: Introduced with OmniAnomaly, SMD records approximately five weeks of 38 operating metrics from each of 28 servers at a large Internet company. Each server forms an independent entity; the first and second halves of its sequence form the standard training and test splits. Experts labeled anomalous intervals and affected variables from failure records, capturing changes in server loads and internal metric dependencies.

MSL (Mars Science Laboratory) [51]: MSL contains real telemetry from NASA/JPL's Curiosity rover, with anomalies labeled by mission experts. Its 27 telemetry channels form independent multivariate sequences. Each time point in the standard version used here comprises one target telemetry value and 54 encoded command features (55 dimensions). The anomalies include point and contextual cases, enabling evaluation of nonstationary telemetry dynamics and relation changes.

WADI (Water Distribution) [52]: Collected on the iTrust water-distribution testbed at the Singapore University of Technology and Design, WADI contains 14 days of normal operation and 2 days of attack operation. Its 123 sensor and actuator variables were recorded continuously. Fifteen attacks on sensors, actuators, or control processes produced coordinated deviations across multiple process variables. We use the officially updated version to assess anomaly detection in high-dimensional, strongly coupled industrial processes.

GECCO [53]: From the 2018 GECCO Industrial Challenge, GECCO records water quality and operating states in a real drinking-water system at one-minute intervals. The nine retained input variables comprise water temperature, pH, oxidation-reduction potential, conductivity, turbidity, two chlorine-dioxide concentrations, and two water-line flow rates. It provides a low-dimensional industrial setting for assessing relation modeling and detection stability.

4.1.2 Experimental Settings

We use the official training and test sets for all datasets and process each entity independently. The normal training sequence is chronologically split into 80% for training and 20% for validation, while the test set remains unchanged. The validation set is used for model selection, anomaly-score calibration, and threshold determination. All sequences are normalized using the mean and standard deviation estimated from the training portion. Unless otherwise stated, all methods use a history length of $\omega = 64$

and a sliding stride of 1, with identical data partitions, target-time indices, threshold candidates, and evaluation procedures. For forecasting-based methods, the prediction length is set to $p = 6$. For methods without an explicit forecasting horizon, their model-specific output and scoring configurations follow the original implementations. TFMG is implemented in PyTorch and trained on an NVIDIA Quadro RTX 8000 GPU using Adam with an initial learning rate of 1×10^{-3}, a batch size of 64, and up to 30 epochs. The comparison experiments use independent runs with random seeds 42, 43, and 44, and report the mean and sample standard deviation across the three runs; the ablation and parameter-sensitivity experiments use random seed 42. Both temporal-feature and node-summary dimensions are set to 64, the default number of frequency bands is 3, and the band-graph diversity weight is 0.05. These settings are shared across all four datasets.

4.1.3 Baselines

To evaluate TFMG comprehensively, we compare it with 13 representative baselines for multivariate time-series anomaly detection. The conventional and probabilistic baselines are iForest [54] and OmniAnomaly [50], which use isolation-based scoring and a stochastic latent-variable model, respectively. Graph-based baselines include MTAD-GAT [10], GDN [11], GCAD [41], and LaGraph [42], which capture inter-variable dependencies through graph attention, learned dependency graphs, Granger-causality-based modeling, and Laplacian-guided graph learning, respectively. STAMP [43] and TFT-GCN [24] provide complementary spatiotemporal and time-frequency approaches: STAMP combines multi-perspective spatiotemporal attention with forecasting and reconstruction, whereas TFT-GCN incorporates time-frequency representations into graph convolution. Anomaly Transformer [36] identifies anomalies from discrepancies between prior and series associations. Among general-purpose forecasting models, DLinear [38], PatchTST [39], and iTransformer [40] represent decomposition-based linear forecasting, patch-wise channel-independent modeling, and variate-token modeling, respectively; their anomaly scores are computed from forecast residuals. CATCH [21] uses frequency patching and channel-aware masking to model fine-grained spectral patterns and cross-channel dependencies.

The core architectures, optimization objectives, and model-specific anomaly-scoring mechanisms of the baselines follow their original papers and public implementations; only the data interfaces, task outputs, and target-time indices are adapted. Except for settings required by individual models, the data partitions, input and scoring time ranges, threshold candidates, and evaluation procedures are identical to those used for TFMG.

4.1.4 Evaluation Metrics

Because time-series anomalies usually occur as continuous events and point adjustment can yield overly optimistic performance [55], [56], we adopt the more stringent Affiliation metric [57]. It merges consecutive anomalous points into events and computes Affiliation Precision (Aff-P) and Affiliation Recall (Aff-R) from the bidirectional temporal proximity between predicted and ground-truth events within their associated temporal regions; their harmonic mean is denoted by Affiliation F1 (Aff-F1).

Unlike point adjustment, Affiliation does not expand a local hit within an anomalous interval into a correct prediction for the entire interval and thus evaluates event localization and coverage more conservatively. For the multi-entity datasets SMD and MSL, thresholds and metrics are determined independently for each entity and then averaged using the official test-sequence lengths as weights. Dataset-level Aff-F1 is therefore a weighted aggregation of entity-level Aff-F1 rather than a value recomputed from aggregated Aff-P and Aff-R. WADI and GECCO each contain a single entity, so their reported results correspond directly to that entity. All methods follow the same threshold protocol, and Aff-P, Aff-R, and Aff-F1 are reported for each dataset.

4.2 Overall Performance Comparison

Table 2 summarizes Aff-P, Aff-R, and Aff-F1 for TFMG and the 13 baselines on the four datasets. Each entry reports the mean $\pm$ sample standard deviation from independent runs with random seeds 42, 43, and 44. For ease of comparison, methods are arranged by row and datasets are grouped by column.

As shown in Table 2, TFMG attains the highest mean Aff-F1 on all four datasets. Relative to the best baseline on each dataset, its absolute gains in mean Aff-F1 are 0.62, 2.21, 1.62, and 1.03 percentage points on SMD, MSL, WADI, and GECCO, respectively, with the largest gain on MSL. The four datasets differ substantially in dimensionality, system structure, data distribution, and application setting, covering server monitoring, spacecraft telemetry, high-dimensional water-distribution control, and low-dimensional water-quality monitoring. TFMG achieves the highest mean Aff-F1 in each of these heterogeneous settings, suggesting that its observed advantage extends across the evaluated data scales and application settings. These findings are consistent with the benefit of jointly modeling temporal and multi-band relations for multivariate time series with differing characteristics.

Inspection of Aff-P and Aff-R provides a complementary view. TFMG attains the highest mean Aff-P on MSL and the highest mean Aff-R on SMD and WADI, indicating a favorable balance between event-localization accuracy and event coverage. On GECCO, neither component metric has the highest mean, yet TFMG achieves the highest mean Aff-F1, indicating a more balanced combination of precision and recall. This pattern is particularly clear on WADI, where several baselines exhibit high Aff-P but substantially lower Aff-R. TFMG achieves the highest mean Aff-F1 with an Aff-P of 0.7098 ± 0.0284 and an Aff-R of 0.9320 ± 0.0263, suggesting fewer missed anomalous events in this high-dimensional, strongly coupled industrial setting. Overall, TFMG shows a consistent advantage in mean Aff-F1 across datasets with differing characteristics, supporting its applicability and detection stability across the evaluated settings.

4.3 Ablation Study

To analyze the contributions of the core components, we use the complete TFMG as the reference, remove or replace the target components in each variant, and keep the remaining architecture and evaluation procedure unchanged. The four ablated variants are defined as follows.

Table 2 Detection performance comparison between TFMG and the baselines on four datasets

Method	SMD			MSL			WADI			GECCO		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
IForest	0.5967 ±0.0065	0.7851 ±0.0041	0.6718 ±0.0024	0.6431 ±0.0068	0.9506 ±0.0018	0.7559 ±0.0050	0.6565 ±0.0102	0.7037 ±0.0146	0.6791 ±0.0043	0.4866 ±0.0072	0.4579 ±0.2086	0.4563 ±0.0992
OmniAnomaly	0.7487 ±0.0114	0.8980 ±0.0083	0.8091 ±0.0029	0.6982 ±0.0241	0.9450 ±0.0135	0.7910 ±0.0112	0.8168 ±0.1173	0.6251 ±0.1249	0.6946 ±0.0484	0.8923 ±0.0091	0.7892 ±0.1527	0.8330 ±0.0860
MTAD-GAT	0.7973 ±0.0064	0.8982 ±0.0077	0.8329 ±0.0080	0.7327 ±0.0249	0.9378 ±0.0092	0.8106 ±0.0146	0.7048 ±0.0354	0.8072 ±0.0384	0.7521 ±0.0303	0.8298 ±0.0232	0.8265 ±0.1354	0.8254 ±0.0747
GDN	0.8031 ±0.0036	0.9020 ±0.0085	0.8422 ±0.0036	0.7182 ±0.0131	0.9416 ±0.0091	0.8063 ±0.0104	0.6053 ±0.0122	0.8086 ±0.0752	0.6917 ±0.0358	0.8779 ±0.0286	0.8928 ±0.1536	0.8799 ±0.0767
GCAD	0.6801 ±0.0128	0.6867 ±0.0302	0.6614 ±0.0149	0.7362 ±0.0050	0.9058 ±0.0128	0.7912 ±0.0049	0.7716 ±0.0190	0.7738 ±0.0152	0.7726 ±0.0132	0.8366 ±0.0367	0.3725 ±0.1414	0.5064 ±0.1508
LaGraph	0.8460 ±0.0157	0.8618 ±0.0265	0.8333 ±0.0079	0.7658 ±0.0158	0.9522 ±0.0087	0.8357 ±0.0154	0.8942 ±0.0937	0.4455 ±0.0889	0.5876 ±0.0605	0.8501 ±0.0179	0.7325 ±0.0100	0.7868 ±0.0057
STAMP	**0.8474 ±0.0034**	0.8732 ±0.0065	0.8495 ±0.0052	0.6842 ±0.0101	0.9520 ±0.0064	0.7856 ±0.0053	0.5985 ±0.0528	0.8360 ±0.0940	0.6975 ±0.0686	0.8580 ±0.0145	0.9815 ±0.0022	0.9155 ±0.0074
TFT-GCN	0.7858 ±0.0106	0.8266 ±0.0012	0.7910 ±0.0051	0.7044 ±0.0142	0.9293 ±0.0240	0.7883 ±0.0014	0.6366 ±0.0297	0.9057 ±0.0580	0.7462 ±0.0045	0.8407 ±0.0029	**0.9848 ±0.0003**	0.9071 ±0.0018
Anomaly Transformer	0.8474 ±0.0024	0.8807 ±0.0104	0.8502 ±0.0036	0.7171 ±0.0120	**0.9550 ±0.0089**	0.8089 ±0.0101	**0.9787 ±0.0139**	0.4159 ±0.0046	0.5838 ±0.0070	**0.9008 ±0.0084**	0.1299 ±0.0012	0.2271 ±0.0016
DLinear	0.8089 ±0.0069	0.9016 ±0.0046	0.8458 ±0.0024	0.7247 ±0.0070	0.9309 ±0.0020	0.8027 ±0.0038	0.7594 ±0.0261	0.8220 ±0.0174	0.7891 ±0.0088	0.8694 ±0.0111	0.9790 ±0.0055	0.9209 ±0.0038
PatchTST	0.8164 ±0.0131	0.8927 ±0.0118	0.8462 ±0.0020	0.7293 ±0.0172	0.9372 ±0.0045	0.8084 ±0.0101	0.7066 ±0.0216	0.7685 ±0.0153	0.7360 ±0.0076	0.8418 ±0.0069	0.9830 ±0.0001	0.9069 ±0.0040
iTransformer	0.8261 ±0.0069	0.8793 ±0.0151	0.8428 ±0.0077	0.7326 ±0.0064	0.9421 ±0.0062	0.8107 ±0.0016	0.6882 ±0.0162	0.8388 ±0.0523	0.7558 ±0.0310	0.8361 ±0.0394	0.9833 ±0.0001	0.9034 ±0.0233
CATCH	0.8231 ±0.0064	0.8811 ±0.0072	0.8425 ±0.0046	0.7422 ±0.0036	0.9307 ±0.0131	0.8126 ±0.0077	0.7054 ±0.0703	0.7375 ±0.0287	0.7203 ±0.0478	0.8376 ±0.0019	0.9830 ±0.0000	0.9045 ±0.0011
TFMG	0.8095 **±0.0094**	**0.9216 ±0.0042**	**0.8564 ±0.0052**	**0.7959 ±0.0171**	0.9526 ±0.0049	**0.8578 ±0.0131**	0.7098 ±0.0284	**0.9320 ±0.0263**	**0.8053 ±0.0080**	0.8896 ±0.0079	0.9769 ±0.0060	**0.9312 ±0.0033**

Note: P, R, and F1 denote Aff-P, Aff-R, and Aff-F1, respectively. Values are the mean ± sample standard deviation from three independent runs with random seeds 42, 43, and 44. Bold and underlined entries indicate the highest and second-highest unrounded means, respectively. For SMD and MSL, entity-level metrics are first aggregated within each run using official test-sequence lengths as weights and then summarized across seeds; Aff-F1 is not recomputed from the displayed mean Aff-P and Aff-R.

w/o-ABAF: removes the adaptive band-attention fusion module and replaces it with a linear aggregation layer over the GAT outputs from all frequency bands.

w/o-FBG: removes the frequency-band graph component, including the frequency-band relation graphs and their GAT propagation, while retaining frequency-band encoding and time-domain relation graph construction.

w/o-Graph: removes the time-domain relation graph, all frequency-band relation graphs, and all GAT modules, thereby eliminating explicit graph-based relation modeling.

w/o-HG: removes hierarchical gating from the hierarchical gated relation injection mechanism by replacing the two learnable gates with fixed equal-weight fusion and uniform position-wise injection.

As shown in Table 3, the complete TFMG achieves a higher Aff-F1 than every ablated variant on all four datasets, indicating that each component contributes to the overall performance. w/o-ABAF replaces adaptive band-attention fusion with non-adaptive aggregation while retaining the frequency-band relation graphs and GAT; its Aff-F1 decreases by 0.33–1.35 percentage points, with the largest reduction observed on WADI. This indicates that relation evidence from different frequency bands is better fused adaptively than through simple linear aggregation. w/o-FBG removes the frequency-band relation graphs and their GAT propagation while retaining frequency-band encoding and the time-domain relation graph; its Aff-F1 decreases by 0.28–1.41 percentage points, with larger reductions on SMD and GECCO. These results indicate that frequency features alone do not provide the same performance gains as within-band cross-variable propagation; the frequency-band relation graphs convert frequency information into scale-specific relation evidence. w/o-Graph additionally removes the time-domain relation graph and frequency-band relation graphs, reducing Aff-F1 by 1.81–6.37 percentage points and yielding the largest degradation on SMD, MSL, and GECCO. Its performance is also consistently below that of w/o-FBG, indicating that the time-domain relation graph provides additional benefits by modeling integrated window-level dependencies, even in the absence of frequency-band graph propagation. On WADI, this variant also remains markedly inferior to the complete model, confirming the importance of explicit graph propagation for high-dimensional coupled sequences. w/o-HG retains both domain-specific relation representations but uses fixed equal-weight fusion and uniform position-wise injection; its Aff-F1 decreases by 0.92–6.51 percentage points. Its WADI Aff-F1 of 0.7344 is slightly lower than the 0.7358 of w/o-Graph, suggesting that fixed injection may weaken the benefits of graph representations. Overall, frequency-band relation graphs provide scale-specific relation evidence, the time-domain relation graph preserves integrated window-level dependencies, adaptive band-attention fusion coordinates the contributions of relation evidence from different frequency bands, and hierarchical gated relation injection controls how the fused relation information is adaptively injected into the main temporal pathway. Together, the complementary roles of these modules support the performance of the complete TFMG.

Table 3 Ablation results for the core TFMG components

Method	SMD			MSL			WADI			GECCO		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
w/o-ABAF	0.8128	0.9201	0.8578	0.7952	0.9542	0.8587	0.6716	0.9474	0.7860	0.8760	0.9785	0.9244
w/o-FBG	0.8031	0.9084	0.8522	0.8011	0.9472	0.8603	0.6825	0.9410	0.7912	0.8534	**0.9828**	0.9136
w/o-Graph	**0.8270**	0.8858	0.8429	0.7564	0.9424	0.8314	**0.7246**	0.7473	0.7358	0.8541	0.9729	0.9096
w/o-HG	0.8000	**0.9300**	0.8532	0.7735	**0.9557**	0.8444	0.6206	0.8992	0.7344	0.8572	**0.9828**	0.9157
TFMG	0.8200	0.9182	**0.8624**	**0.8048**	0.9496	**0.8631**	0.6901	**0.9500**	**0.7995**	**0.8808**	0.9798	**0.9277**

Note: P, R, and F1 denote Aff-P, Aff-R, and Aff-F1, respectively. Bold values indicate the best results. The three metrics for SMD and MSL are entity-level averages weighted by official test-sequence length.

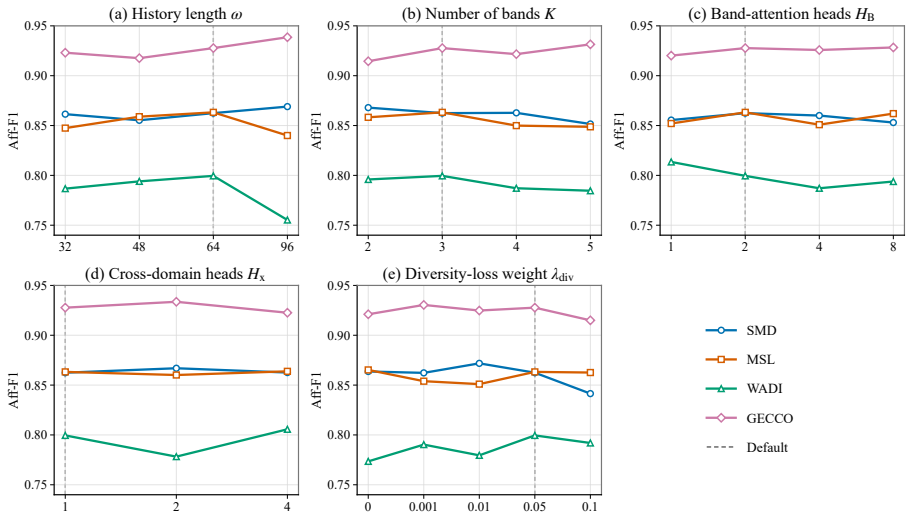

Fig. 2 Parameter sensitivity of TFMG on the four datasets: (a) history length ω, (b) number of frequency bands K, (c) band self-attention heads H_{B}, (d) cross-domain attention heads H_{x}, and (e) band-graph diversity-loss weight λ_{div}

4.4 Parameter Sensitivity Analysis

To examine the sensitivity of TFMG to key structural and training hyperparameters, we conduct one-factor-at-a-time experiments on SMD, MSL, WADI, and GECCO. We vary the history length ω, number of bands K, number of band self-attention heads H_{B}, number of cross-domain attention heads H_{x}, and band-graph diversity-loss weight λ_{div}, while keeping all other settings fixed. The candidate values are $\omega \in \{32, 48, 64, 96\}$, $K \in \{2, 3, 4, 5\}$, $H_{\mathrm{B}} \in \{1, 2, 4, 8\}$, $H_{\mathrm{x}} \in \{1, 2, 4\}$, and $\lambda_{\mathrm{div}} \in \{0, 0.001, 0.01, 0.05, 0.1\}$, respectively, and the default settings are $\omega = 64$, $K = 3$, $H_{\mathrm{B}} = 2$, $H_{\mathrm{x}} = 1$, and $\lambda_{\mathrm{div}} = 0.05$. The results are shown in Fig. 2.

Figure 2(a) shows that the effect of window length depends on the dataset: $\omega = 96$ provides more context for SMD and GECCO but reduces performance on MSL and WADI, whereas $\omega = 64$ achieves the highest average Aff-F1 across datasets by balancing contextual coverage and sensitivity to local anomalies. In Fig. 2(b), $K = 2$ and $K = 3$ are generally close, and adding more bands does not yield a stable gain. This suggests that a limited number of bands may be sufficient to capture the dominant dynamic scales under the tested settings, while finer partitioning may reduce the stability of within-band statistics and relation estimation. Figures 2(c) and 2(d) show that the benefit of attention heads is dataset-dependent. $H_{\mathrm{B}} = 2$ performs well on SMD and MSL and gives the highest average Aff-F1, but increasing the number to eight heads provides no consistent gain; the datasets exhibit different preferences for $H_{\mathrm{x}} = 2$ and $H_{\mathrm{x}} = 4$. Figure 2(e) shows that $\lambda_{\mathrm{div}} = 0.05$ performs best on WADI and achieves the highest four-dataset average Aff-F1 among the tested weights. Removing the constraint degrades WADI performance relative to the default setting, whereas

increasing the weight to 0.1 reduces performance on SMD and GECCO. These results indicate that the effect of the diversity constraint depends on its weight and the dataset. Overall, sensitivity varies across datasets and parameters.

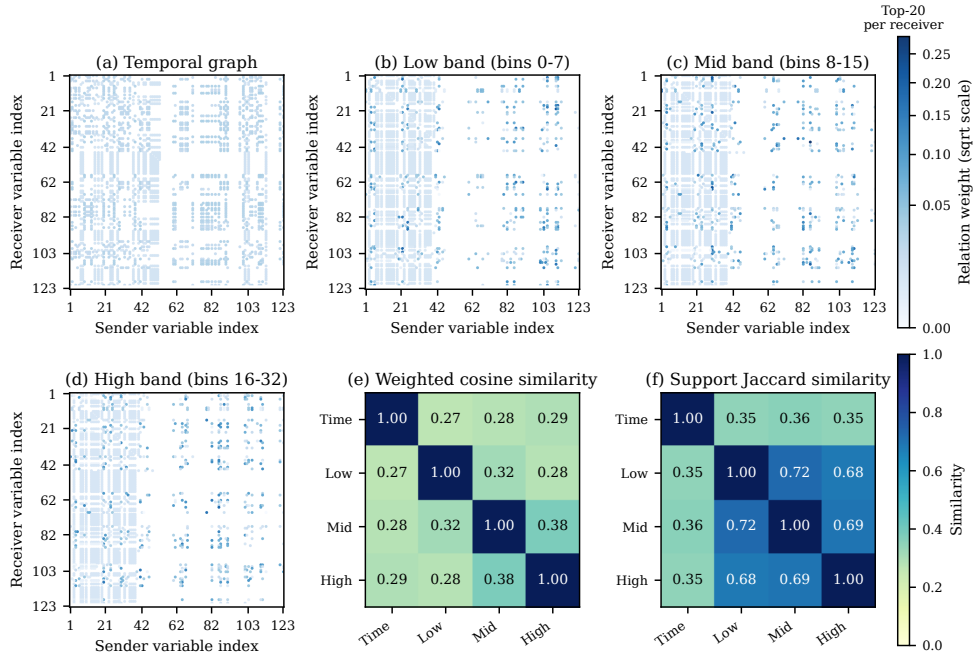

Fig. 3 Multi-view variable relations on a representative anomalous window of WADI: (a) time-domain adjacency matrix; (b)–(d) low-, mid-, and high-frequency adjacency matrices, respectively; (e) weighted cosine similarities; and (f) edge-support Jaccard similarities. Up to 20 highest-weight positive non-self incoming relations per receiving variable are shown for visualization only.

4.5 Relation Interpretability Analysis

To examine whether the variable relations learned by TFMG at different frequency scales provide complementary rather than redundant structural evidence, we conduct a case study on the WADI test set. The selected window spans 11:06:50–11:12:05 on October 10, 2017, and lies entirely within an anomalous interval. Figs. 3(a)–3(d) show visualizations of the temporal and frequency-band adjacency matrices under the same variable ordering and sparsification rule. For visualization clarity, Figs. 3(a)–3(d) display up to 20 highest-weight positive non-self incoming relations for each receiving variable. The four matrices share a square-root color scale to improve the visibility of weak relations while maintaining comparability. The top-20 filtering and square-root color transformation are used only for visualization; all similarity measures, edge densities, and subsequent statistics are computed from the complete learned adjacency matrices before this visualization-only filtering. Entry (i, j) represents the dependence of variable i on variable j.

Figures 3(e) and 3(f) report the weighted cosine similarity and edge-support Jaccard similarity, respectively, among the four relation views. The weighted cosine similarities between the time-domain relation graph and the low-, mid-, and high-frequency relation graphs are 0.268, 0.278, and 0.290, respectively, while the corresponding edge-support Jaccard similarities range from 0.347 to 0.359. These results indicate that the time-domain relation graph and frequency-band relation graphs preserve a certain degree of shared structure while exhibiting clear differences in both connectivity and edge-weight allocation. In contrast, the frequency-band relation graphs show relatively high edge-support overlap (0.680–0.722), whereas their weighted cosine similarities remain low (0.279–0.382). The high support overlap suggests that different frequency bands are still constrained by common process couplings, while the low weight similarity indicates that the importance assigned to the same candidate relations varies across dynamic scales. Thus, the differences among frequency-band relation graphs are mainly reflected in the scale-dependent redistribution of relation strengths rather than in completely independent connectivity patterns. These observations further suggest that frequency-domain relations are not simple replicas of the time-domain relations, but provide scale-discriminative complementary evidence for the current window, supporting independent band-wise propagation and state-aware fusion.

The nonzero edge densities of the time-domain relation graph and the low-, mid-, and high-frequency relation graphs are 40.21%, 43.64%, 42.36%, and 47.66%, respectively. Their relatively small variation indicates that the observed structural differences are not primarily caused by excessive sparsity or density in any particular view. The four views jointly retain 2,839 non-self edges, accounting for 26.25% of the 10,817 edges in their union. This shows that the learned graphs contain both relations shared across views and scale-specific connections that appear only in particular views. Further examination of high-weight edges reveals that different frequency bands do not emphasize exactly the same variable pairs, and that both the endpoints and relative strengths of some relations vary with dynamic scale. This suggests that a relation emphasized at low frequencies may become weaker at mid or high frequencies or be replaced by other relations. It should be emphasized that these edges represent conditional dependencies and information-propagation relations learned from observations within the current anomalous window, rather than physical causal relationships. Overall, the frequency-band relation graphs do not collapse into an identical topology; instead, they preserve shared cross-variable structure while forming scale-distinct relational representations, providing empirical support for the multi-view graph construction and adaptive band-attention fusion used in TFMG.

5 Conclusion

This paper presents TFMG, a time-frequency multi-view graph learning framework designed to capture variable relations at different dynamic scales that are difficult to represent with a single variable graph. The framework jointly learns a window-conditioned time-domain relation graph and multiple frequency-band relation graphs on a shared set of variable nodes. Adaptive band-attention fusion uses

band self-attention and temporal-state-conditioned cross-domain attention for relation interaction and state-aware aggregation. Hierarchical gated relation injection then injects the fused relation representation into the main temporal pathway with variable- and position-specific control. The prediction-step-aware forecasting decoder generates multi-step forecasts and computes anomaly scores from their deviations. Comparisons on four public benchmarks show that TFMG achieves the highest mean Aff-F1 among the compared methods on all four datasets. Ablation studies support the effectiveness of explicit graph propagation, frequency-band relation modeling, adaptive band-attention fusion, and hierarchical gated relation injection. Parameter analysis shows relative stability over the main candidate ranges, and the WADI case study indicates that the time-domain relation graph and frequency-band relation graphs share some connections while exhibiting distinct edge-weight patterns.

The current method uses predefined contiguous frequency bands, and its learned edges reflect window-conditioned dependencies rather than physical causality. This limits the physical interpretability of the learned relations. Future work will investigate data-driven band partitioning and cross-scale graph-structure learning, together with process topologies or domain priors, to improve the physical consistency of relation explanations and adaptation across scenarios.

Statements and Declarations

Author Contributions The authors jointly contributed to the conceptualization, methodology, validation, investigation, and manuscript preparation. All authors read and approved the final manuscript.

Funding No funding was received for conducting this study.

Data Availability The SMD, MSL, WADI, and GECCO datasets analyzed in this study are publicly available. Their original sources are cited in [50–53] and described in Section 4.1.1.

Competing Interests The authors declare no competing interests.

References

[1] Paparrizos J, Boniol P, Liu Q, et al (2026) A comprehensive guide to time-series anomaly detection. In: Proc. ACM Int. Conf. Web Search Data Mining (WSDM), pp 1363–1366, https://doi.org/10.1145/3773966.3777914

[2] Paparrizos J, Boniol P, Liu Q, et al (2025) Advances in time-series anomaly detection: Algorithms, benchmarks, and evaluation measures. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), pp 6151–6161, https://doi.org/10.1145/3711896.3736565

[3] Zhang C, et al (2024) Advancing multivariate time series anomaly detection: A comprehensive benchmark with real-world data from alibaba cloud. In: Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM), pp 5410–5414, https://doi.org/10.1145/3627673.3679128

[4] Liu J, Yang D, Wang J, et al (2026) Adaptive unsupervised anomaly detection for low-quality multivariate time-series data. IEEE Trans Knowl Data Eng 38(8):4851–4867. https://doi.org/10.1109/tkde.2026.3700672

[5] Wu Z, et al (2026) ICAD: Rethinking the role of inference and cues for the anomaly detection of time series in IIoT. IEEE Trans Knowl Data Eng 38(8):5187–5201. https://doi.org/10.1109/tkde.2026.3682751

[6] Tang Q, et al (2026) Bridging classification and reconstruction: Cooperative time series anomaly detection. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), pp 4604–4615, https://doi.org/10.1145/3770855.3818108

[7] Zhang B, et al (2026) An encode-then-decompose approach to unsupervised time series anomaly detection on contaminated training data. In: Proc. IEEE Int. Conf. Data Eng. (ICDE), pp 919–933, https://doi.org/10.1109/icde65706.2026.00074

[8] Xia Z, Ma J, Zheng B, et al (2026) DMRAD: Dynamic decomposition and memory-aware reconstruction for noise-resilient multivariate time series anomaly detection. IEEE Trans Knowl Data Eng 38(3):1665–1678. https://doi.org/10.1109/tkde.2025.3649930

[9] Liu S, et al (2023) Time series anomaly detection with adversarial reconstruction networks. IEEE Trans Knowl Data Eng 35(4):4293–4306. https://doi.org/10.1109/tkde.2021.3140058

[10] Zhao H, Wang Y, Duan J, et al (2020) Multivariate time-series anomaly detection via graph attention network. In: Proc. IEEE Int. Conf. Data Mining (ICDM), pp 841–850, https://doi.org/10.1109/ICDM50108.2020.00093

[11] Deng A, Hooi B (2021) Graph neural network-based anomaly detection in multivariate time series. In: Proc. AAAI Conf. Artif. Intell., pp 4027–4035, https://doi.org/10.1609/aaai.v35i5.16523

[12] Chen Z, Chen D, Zhang X, et al (2022) Learning graph structures with transformer for multivariate time-series anomaly detection in iot. IEEE Internet Things J 9(12):9179–9189. https://doi.org/10.1109/JIOT.2021.3100509

[13] Lim H, Park S, Kim M, et al (2023) Madsgm: Multivariate anomaly detection with score-based generative models. In: Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM), pp 1411–1420, https://doi.org/10.1145/3583780.3614956

[14] Li Y, Peng X, Zhang J, et al (2023) DCT-GAN: Dilated convolutional transformer-based GAN for time series anomaly detection. IEEE Trans Knowl Data Eng 35(4):3632–3644. https://doi.org/10.1109/tkde.2021.3130234

[15] Du B, Sun X, Ye J, et al (2023) GAN-Based anomaly detection for multivariate time series using polluted training set. IEEE Trans Knowl Data Eng

35(12):12208–12219. https://doi.org/10.1109/tkde.2021.3128667

[16] Zheng B, et al (2024) Adversarial graph neural network for multivariate time series anomaly detection. IEEE Trans Knowl Data Eng 36(12):7612–7626. https://doi.org/10.1109/tkde.2024.3419891

[17] Zhang X, et al (2024) Rethinking robust multivariate time series anomaly detection: A hierarchical spatio-temporal variational perspective. IEEE Trans Knowl Data Eng 36(12):9136–9149. https://doi.org/10.1109/tkde.2024.3466291

[18] Zhou Q, He S, Liu H, et al (2024) Label-free multivariate time series anomaly detection. IEEE Trans Knowl Data Eng 36(7):3166–3179. https://doi.org/10.1109/tkde.2024.3349613

[19] Shi D, Zhang T, Duan Y, et al (2026) Time-frequency hybrid decomposition for multivariate time series anomaly detection. Evolving Syst 17(3). https://doi.org/10.1007/s12530-026-09856-4

[20] Nam Y, et al (2024) Breaking the time-frequency granularity discrepancy in time-series anomaly detection. In: Proc. ACM Web Conf. (WWW), pp 4204–4215, https://doi.org/10.1145/3589334.3645556

[21] Wu X, Qiu X, Li Z, et al (2025) CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In: Proc. Int. Conf. Learn. Representations (ICLR), URL https://openreview.net/forum?id=m08aK3xxdJ

[22] Hu S, et al (2026) TAP: Time series anomaly prediction via adaptive period modeling and dual representation learning. IEEE Trans Knowl Data Eng 38(9):6182–6195. https://doi.org/10.1109/tkde.2026.3691368

[23] Wu J, Gao Y, Guan R, et al (2026) FFT-Based frequency-domain guided anomaly transformer: An improved method for multivariate time series anomaly detection. In: Proc. 2026 IEEE Int. Conf. Mechatronics Autom. (ICMA), pp 320–324, https://doi.org/10.1109/ICMA69663.2026.11647709

[24] Xia Z, Xu X, Wu L, et al (2026) TFT-GCN: A time-frequency based model for time series anomaly detection. IEEE Trans Knowl Data Eng 38(9):6196–6207. https://doi.org/10.1109/TKDE.2026.3707947

[25] Ma X, Hong X, Li W, et al (2026) A wave is worth 100 words: Investigating cross-domain transferability in time series. ACM Trans Knowl Discovery Data 20(3):Article 37. https://doi.org/10.1145/3788674

[26] Xu H, Wang Y, Jian S, et al (2024) Calibrated one-class classification for unsupervised time series anomaly detection. IEEE Trans Knowl Data Eng 36(11):5723–5736. https://doi.org/10.1109/tkde.2024.3393996

[27] Liu C, He S, Li S, et al (2025) Detecting both seen and unseen anomalies in time series. ACM Trans Knowl Discovery Data https://doi.org/10.1145/3717071

[28] He Y, et al (2024) Variate associated domain adaptation for unsupervised multivariate time series anomaly detection. ACM Trans Knowl Discovery Data 18(8):1–24. https://doi.org/10.1145/3663573

[29] Darban ZZ, et al (2025) DACAD: Domain adaptation contrastive learning for anomaly detection in multivariate time series. IEEE Trans Knowl Data Eng 37(8):4485–4496. https://doi.org/10.1109/tkde.2025.3569909

[30] Kim H, Kim S, Min S, et al (2024) Contrastive time-series anomaly detection. IEEE Trans Knowl Data Eng 36(10):5053–5065. https://doi.org/10.1109/tkde.2023.3335317

[31] Sun Y, Pang G, Ye G, et al (2024) Unraveling the 'anomaly' in time series anomaly detection: A self-supervised tri-domain solution. In: Proc. IEEE Int. Conf. Data Eng. (ICDE), pp 981–994, https://doi.org/10.1109/icde60146.2024.00080

[32] Wang R, et al (2024) Cutaddpaste: Time series anomaly detection by exploiting abnormal knowledge. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), pp 3176–3187, https://doi.org/10.1145/3637528.3671739

[33] Chen F, Qiu X, Ye Y, et al (2025) KARMAD: KAN-Based adversarial robust model for anomaly detection. In: Proc. IEEE Int. Conf. Data Eng. (ICDE), pp 3425–3438, https://doi.org/10.1109/icde65448.2025.00256

[34] Liu F, et al (2023) Anomaly detection in quasi-periodic time series based on automatic data segmentation and attentional LSTM-CNN (extended abstract). In: Proc. IEEE Int. Conf. Data Eng. (ICDE), pp 3777–3778, https://doi.org/10.1109/icde55515.2023.00315

[35] Tuli S, Casale G, Jennings NR (2022) Tranad: Deep transformer networks for anomaly detection in multivariate time series data. In: Proc. VLDB Endowment, pp 1201–1214, https://doi.org/10.14778/3514061.3514067

[36] Xu J, Wu H, Wang J, et al (2022) Anomaly transformer: Time series anomaly detection with association discrepancy. In: Proc. Int. Conf. Learn. Representations (ICLR), URL https://openreview.net/forum?id=LzQQ89U1qm_

[37] Vaswani A, et al (2017) Attention is all you need. In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp 5998–6008, URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[38] Zeng A, Chen M, Zhang L, et al (2023) Are transformers effective for time series forecasting? In: Proc. AAAI Conf. Artif. Intell., pp 11121–11128, https://doi.org/10.1609/aaai.v37i9.26317

[39] Nie Y, Nguyen NH, Sinthong P, et al (2023) A time series is worth 64 words: Long-term forecasting with transformers. In: Proc. Int. Conf. Learn. Representations (ICLR), URL https://openreview.net/forum?id=Jbdc0vTOcol

[40] Liu Y, Hu T, Zhang H, et al (2024) itransformer: Inverted transformers are effective for time series forecasting. In: Proc. Int. Conf. Learn. Representations (ICLR), URL https://openreview.net/forum?id=JePfAI8fah

[41] Liu Z, Gao M, Jiao P (2025) GCAD: Anomaly detection in multivariate time series from the perspective of granger causality. In: Proc. AAAI Conf. Artif. Intell., pp 19041–19049, https://doi.org/10.1609/AAAI.V39I18.34096

[42] Zeng S, Chao G, Wei J, et al (2026) Lagraph: Laplacian-guided graph learning for time series anomaly detection. IEEE Trans Knowl Data Eng 38(5):2940–2953. https://doi.org/10.1109/tkde.2026.3665696

[43] Chen T, et al (2025) Compatible unsupervised anomaly detection with multi-perspective spatio-temporal learning. In: Proc. IEEE Int. Conf. Data Eng. (ICDE), pp 4066–4078, https://doi.org/10.1109/icde65448.2025.00303

[44] Liu Z, Huang X, Zhang J, et al (2024) Multivariate time-series anomaly detection based on enhancing graph attention networks with topological analysis. In: Proc. ACM Int. Conf. Inf. Knowl. Manage. (CIKM), pp 1555–1564, https://doi.org/10.1145/3627673.3679614

[45] Chen H, Eldardiry H (2024) Graph time-series modeling in deep learning: A survey. ACM Trans Knowl Discovery Data 18(5):1–35. https://doi.org/10.1145/3638534

[46] Veličković P, Cucurull G, Casanova A, et al (2018) Graph attention networks. In: Proc. Int. Conf. Learn. Representations (ICLR), URL https://openreview.net/forum?id=rJXMpikCZ

[47] Zhang G, et al (2023) DBCatcher: A cloud database online anomaly detection system based on indicator correlation. In: Proc. IEEE Int. Conf. Data Eng. (ICDE), pp 1126–1139, https://doi.org/10.1109/icde55515.2023.00091

[48] Idrees R, Maiti A, Garg S (2024) A clustering algorithm for detecting differential deviations in the multivariate time-series iot data based on sensor relationship. Knowl Inf Syst 67(3):2641–2690. https://doi.org/10.1007/s10115-024-02303-3

[49] Martins AFT, Astudillo RF (2016) From softmax to sparsemax: A sparse model of attention and multi-label classification. In: Proc. 33rd Int. Conf. Mach. Learn. (ICML), pp 1614–1623, URL https://proceedings.mlr.press/v48/martins16.html

[50] Su Y, Zhao Y, Niu C, et al (2019) Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proc. ACM SIGKDD Int.

Conf. Knowl. Discovery Data Mining (KDD), pp 2828–2837, https://doi.org/10.1145/3292500.3330672

[51] Hundman K, Constantinou V, Laporte C, et al (2018) Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), pp 387–395, https://doi.org/10.1145/3219819.3219845

[52] Ahmed CM, Palleti VR, Mathur AP (2017) WADI: A water distribution testbed for research in the design of secure cyber physical systems. In: Proc. 3rd Int. Workshop Cyber-Physical Syst. Smart Water Netw. (CySWater), pp 25–28, https://doi.org/10.1145/3055366.3055375

[53] Moritz S, Rehbach F, Chandrasekaran S, et al (2020) GECCO industrial challenge 2018 dataset: A water quality dataset for the internet of things: Online anomaly detection for drinking water quality competition. Zenodo https://doi.org/10.5281/zenodo.3884398

[54] Liu FT, Ting KM, Zhou ZH (2008) Isolation forest. In: Proc. 8th IEEE Int. Conf. Data Mining (ICDM), pp 413–422, https://doi.org/10.1109/ICDM.2008.17

[55] Kim S, Choi K, Choi HS, et al (2022) Towards a rigorous evaluation of time-series anomaly detection. In: Proc. AAAI Conf. Artif. Intell., pp 7194–7201, https://doi.org/10.1609/aaai.v36i7.20680

[56] Boniol P, et al (2025) VUS: effective and efficient accuracy measures for time-series anomaly detection. VLDB J 34(3). https://doi.org/10.1007/s00778-025-00907-x

[57] Huet A, Navarro JM, Rossi D (2022) Local evaluation of time series anomaly detection algorithms. In: Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD), pp 635–645, https://doi.org/10.1145/3534678.3539339