CHIME: CONTEXTUAL HIERARCHICAL INCONSISTENCY MODELING FOR MULTIVARIATE TIME SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction-based anomaly detection assumes that anomalous observations are harder to reconstruct than normal ones. In multivariate time series, this assumption can fail when anomalies disrupt temporal or inter-variable relationships while remaining accurately reconstructable. This paper presents CHIME, a framework for Contextual Hierarchical Inconsistency Modeling that incorporates historical context into the assessment of cross-scale structural discrepancies. Preceding windows are encoded into context tokens that guide bidirectional cross-attention among fine-, middle-, and coarse-scale representations. Coarse-to-fine attention relates local patterns to broader trends, while fine-to-coarse attention updates coarse representations with local evidence. A variable-relation branch captures dependencies at each resolution. The anomaly score combines reconstruction error with two forms of hierarchical inconsistency: discrepancies between aligned temporal representations and discrepancies between variable-relation representations across scales. Experiments on ten public benchmarks show that CHIME achieves a mean point-adjusted F1 score of 95.54%, ranking first on eight datasets among the compared methods.

1 INTRODUCTION

Multivariate time-series anomaly detection (MTSAD) is widely used in industrial monitoring, cloud-server operations, spacecraft telemetry, and cyber-physical systems Liu & Paparrizos (2024); Dai et al. (2024). Compared with univariate settings, MTSAD is more challenging because anomalies may arise not only from abnormal values, but also from disruptions in temporal dynamics and inter-variable dependencies Dai et al. (2024); Shimillas et al. (2026); Jeong & Yoon (2026). Real-world multivariate time series are also often high-dimensional, non-stationary, and heterogeneous across temporal resolutions.

Reconstruction-based methods commonly assume that normal samples are easier to reconstruct than anomalous ones. However, expressive models may over-generalize and reconstruct anomalous observations with errors comparable to normal samples Song et al. (2023); Sun et al. (2025); Zhu et al. (2026); Kim et al. (2026). Moreover, some anomalies remain numerically plausible while violating temporal or inter-variable structures. Recent studies therefore explore association discrepancy, spatial dependencies, and relational structures as complementary anomaly evidence Xu et al. (2022); Dai et al. (2024); Jeong & Yoon (2026).

Such structural relationships naturally exist across temporal resolutions. Fine-grained representations capture local fluctuations, whereas coarse representations describe broader trends and long-term states. Multi-scale modeling has shown strong capability in representing such hierarchical temporal patterns Wang et al. (2025); Liu & Paparrizos (2024). Meanwhile, variable dependencies also provide important structural information Dai et al. (2024); Wu et al. (2025). However, most existing methods use different scales or variable relations mainly as additional representations, rather than explicitly measuring whether their structures remain consistent. Furthermore, the interpretation of a local pattern depends on recent operating context, and fixed-window modeling may miss anomalies caused by changes relative to preceding system states Li et al. (2025); He et al. (2026).

Motivated by these observations, we propose **CHIME** (**C**ontextual **H**ierarchical **I**nconsistency **M**odeling) for multivariate time-series anomaly detection. CHIME encodes preceding windows as

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

contextual representations to condition multi-resolution temporal features and employs bidirectional cross-scale interaction to connect coarse trends with fine-grained local patterns. A scale-specific variable-relation branch further captures dependencies among variables. Rather than relying solely on reconstruction error, CHIME detects two complementary forms of hierarchical inconsistency: temporal inconsistency across aligned resolutions and variable-relation inconsistency across scales. These signals are combined with reconstruction discrepancy to produce the final anomaly score.

Our main contributions are summarized as follows:

- We introduce contextual hierarchical inconsistency modeling, which characterizes anomalies as violations of context-dependent structural relationships beyond reconstruction difficulty.
- We develop a context-conditioned bidirectional cross-scale mechanism together with scale-specific variable-relation modeling to capture temporal and inter-variable structures across resolutions.
- We combine reconstruction discrepancy with temporal and variable-relation hierarchical inconsistency for anomaly scoring. Experiments on ten public benchmarks show that CHIME achieves a mean point-adjusted F1-score of 95.54% and ranks first on eight datasets among the compared methods.

2 Related Work

2.1 Reconstruction- and Prediction-Based Detection

Reconstruction and forecasting are two dominant paradigms for multivariate time-series anomaly detection. Reconstruction-based methods train autoencoders, variational autoencoders, recurrent networks, or Transformers to recover normal windows and use deviations from the input as anomaly scores. Forecasting-based methods learn normal evolution from historical observations and identify anomalies when predictions deviate substantially from future observations. Representative models include OmniAnomaly (Su et al., 2019), USAD (Audibert et al., 2020), TranAD (Tuli et al., 2022), and Anomaly Transformer (Xu et al., 2022).

Despite their effectiveness, these approaches may reconstruct anomalous segments well when model capacity is high. Their errors also mainly describe numerical deviations and may overlook anomalies that appear as inconsistencies between local variations and long-range trends or as changes in inter-variable dependencies. Recent work has therefore incorporated association discrepancies (Xu et al., 2022; Yang et al., 2023), spatial or channel relations (Dai et al., 2024; Wu et al., 2025), and multi-scale representations (Xie et al., 2025; Li et al., 2025). CHIME follows this structural perspective while jointly assessing reconstruction, cross-scale consistency, and variable-relation consistency.

2.2 Association- and Relation-Based Anomaly Detection

Beyond numerical reconstruction or prediction errors, recent MTSAD methods exploit structural relationships as complementary anomaly evidence. Anomaly Transformer introduces association discrepancy to distinguish abnormal temporal associations from normal patterns Xu et al. (2022), while DCdetector learns contrasting representations through dual-attention structures Yang et al. (2023). For multivariate dependencies, SARAD models spatial associations among variables Dai et al. (2024), and CATCH captures channel correlations under frequency-aware representations Wu et al. (2025). More recent approaches further exploit dynamic inter-variable relationships for anomaly detection and localization Shimillas et al. (2026); Jeong & Yoon (2026). These studies demonstrate that structural relationships can reveal anomalies that remain difficult to distinguish using numerical errors alone. However, most existing approaches characterize associations within a particular representation or resolution, leaving the consistency of variable-relation structures across temporal resolutions relatively underexplored. CHIME addresses this issue by constructing scale-specific relation representations and measuring their cross-resolution inconsistency.

2.3 Multi-Scale and Context-Aware Temporal Modeling

Temporal patterns in real-world time series occur at different resolutions. Fine-grained representations capture abrupt changes and local fluctuations, whereas coarse representations characterize

broader trends, periodicity, and long-term states. Multi-resolution architectures such as TimeMixer++ demonstrate the effectiveness of hierarchical temporal representations for general time-series modeling Wang et al. (2025). In anomaly detection, CrossAD explicitly models cross-scale associations and further incorporates cross-window information to provide broader temporal context Li et al. (2025). Patch-based approaches such as PaAno additionally emphasize the importance of local temporal structures Park & Kang (2026), while recent methods such as TimeRadar explore contextual deviation beyond global reconstruction signals He et al. (2026). Despite these advances, multi-scale representations are often used primarily to enrich features or improve reconstruction, rather than to explicitly test whether representations at different resolutions remain structurally compatible under the current operating context. CHIME instead conditions multi-resolution representations on historical context and models bidirectional interactions between adjacent resolutions, enabling anomalies to be identified through violations of context-dependent hierarchical temporal consistency.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERALL ARCHITECTURE

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ be a multivariate time series with $\mathbf{x}_t \in \mathbb{R}^N$. A window ending at time t is $\mathbf{X}_t = [\mathbf{x}_{t-W+1}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{W \times N}$. We train on normal data and estimate anomaly evidence for individual positions in test windows. Reconstruction alone may miss observations that preserve numerical appearance while violating temporal or inter-variable dependencies. CHIME therefore assesses whether the current window agrees with the hierarchical structure expected from its recent history.

As illustrated in Figure 1, CHIME consists of historical context modeling, hierarchical cross-scale temporal modeling, and variable-relation scoring. Historical context conditions multi-resolution temporal representations, which subsequently interact through bidirectional cross-attention between adjacent resolutions. A progressive decoder reconstructs the input across scales, while a relation branch models scale-specific inter-variable dependencies. Reconstruction, temporal inconsistency, and variable-relation inconsistency are combined for final anomaly scoring.

3.2 HISTORICAL CONTEXT MODELING

Given K historical windows separated by stride r, we write $\mathbf{C}_t = (\mathbf{X}_{t-Kr}, \ldots, \mathbf{X}_{t-r})$. With $r = W$ by default, complete historical windows precede the current window except near a sequence boundary, where unavailable starts are clamped to the first window. A shared temporal convolution extracts features $\mathbf{F}_{t,k} \in \mathbb{R}^{W \times d_h}$ from each historical window. Attention over the K windows aggregates features at relative position u:

$$\omega_{t,k,u} = \mathrm{softmax}_k\left(f_w(\mathbf{F}_{t,k,u})\right), \qquad \mathbf{H}_{t,u} = \mathrm{LN}\left(\mathbf{W}_h \sum_{k=1}^{K} \omega_{t,k,u} \mathbf{F}_{t,k,u}\right). \tag{1}$$

The resulting context sequence $\mathbf{H}_t \in \mathbb{R}^{W \times d}$ provides a local operating-state reference and conditions the multi-resolution representations before bidirectional cross-scale interaction, rather than producing a separate anomaly score.

3.3 HIERARCHICAL CROSS-SCALE TEMPORAL MODELING

Multi-resolution tokens. We apply average pooling directly to $\mathbf{X}_t$ using factors $k_0 > \cdots > k_{M-1}$, from coarse to fine:

$$\mathbf{X}_t^{(s)} = D_{k_s}(\mathbf{X}_t), \qquad T_s = \lceil W/k_s \rceil, \qquad s = 0, \ldots, M-1. \tag{2}$$

Right-side replication handles incomplete pooling blocks. The default factors $[16, 8, 4, 2]$ yield four encoded resolutions; coarse, intermediate, and fine refer to resolution ranges, not exactly three discrete scales. Each variable is embedded separately using patches of length and stride p, producing $\mathbf{P}_t^{(s)} \in \mathbb{R}^{L_s \times N \times d}$ with $L_s = \lceil T_s/p \rceil$. Patch-aligned context tokens $\mathbf{H}_t^{(s)} \in \mathbb{R}^{L_s \times d}$ are obtained by pooling and patch averaging $\mathbf{H}_t$.

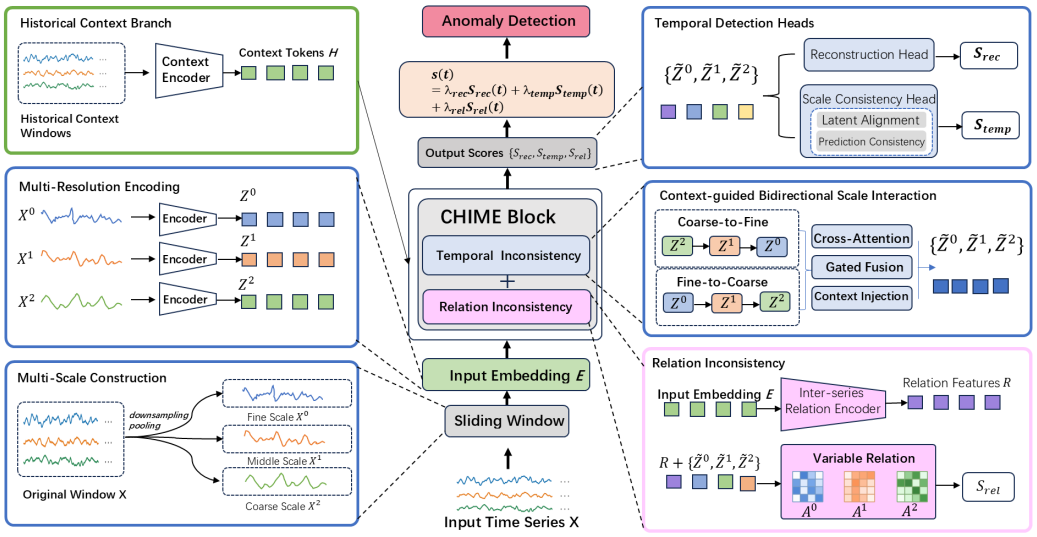

Figure 1: Overall architecture of CHIME. Historical context conditions multi-resolution temporal representations, while bidirectional cross-scale interactions model hierarchical temporal dependencies. Temporal inconsistency and variable-relation inconsistency are combined with reconstruction error for anomaly detection.

In the default local-patch configuration, historical context conditions the scale tokens through a gated residual:

$$\bar{\mathbf{P}}_t^{(s)} = \mathbf{P}_t^{(s)} + \eta_{\text{ctx}}\, \sigma\Big(f_g([\mathbf{P}_t^{(s)}; \mathbf{H}_t^{(s)}])\Big) \odot \mathbf{H}_t^{(s)}. \tag{3}$$

Context tokens are broadcast across variables. A shared, scale-masked Transformer encoder then maps $\bar{\mathbf{P}}_t^{(s)}$ to $\mathbf{Z}_t^{(s)} \in \mathbb{R}^{L_s \times N \times d}$. Its mask restricts encoder self-attention to tokens at the same resolution. Thus historical context conditions the tokens *before* cross-scale interaction; it is not an additional context-to-scale cross-attention operation in this configuration.

Context-guided bidirectional cross-scale interaction. For adjacent coarse and fine token sequences $(\mathbf{C}, \mathbf{F})$, let $\mathcal{R}_{a\to b}$ linearly interpolate a token sequence of length a to length b along its temporal axis. Coarse-to-fine attention updates local representations first:

$$\mathbf{F}' = \mathrm{LN}(\mathbf{F} + \eta_{\text{cf}}\, \mathrm{CA}(\mathbf{F}, \mathcal{R}(\mathbf{C}), \mathcal{R}(\mathbf{C}))). \tag{4}$$

Fine-to-coarse attention then incorporates the updated local evidence:

$$\mathbf{C}' = \mathrm{LN}(\mathbf{C} + \eta_{\text{fc}}\, \mathrm{CA}(\mathbf{C}, \mathcal{R}(\mathbf{F}'), \mathcal{R}(\mathbf{F}'))). \tag{5}$$

Here $\mathrm{CA}(Q, K, V)$ is multi-head cross-attention and interpolation aligns keys and values with the query length. The learnable scalar residual coefficients are initialized to small values. Applying this update once to each adjacent pair in coarse-to-fine order produces enhanced tokens $\{\tilde{\mathbf{Z}}_t^{(s)}\}_{s=0}^{M-1}$. A middle resolution carries its previous update into the interaction with its finer neighbor. Since context has already modified the input tokens, both directions are conditioned on recent history.

Progressive reconstruction. A shared decoder upsamples the enhanced tokens and attends to the encoder output to reconstruct targets $\mathbf{Y}_t^{(j)} = \mathbf{X}_t^{(j+1)}$ for $j < M-1$ and $\mathbf{Y}_t^{(M-1)} = \mathbf{X}_t$. Let $\hat{\mathbf{Y}}_t^{(j)}$ be the corresponding outputs with lengths U_j. For the default four encoder resolutions, the decoder targets have pooling factors $[8, 4, 2, 1]$. The full-resolution reconstruction is $\hat{\mathbf{X}}_t = \hat{\mathbf{Y}}_t^{(M-1)}$.

3.4 HIERARCHICAL INCONSISTENCY SCORING

Temporal inconsistency. We compare adjacent aligned latent representations and decoder outputs. For scale s and relative position u, define

$$e^{(s)}_{\text{lat},t}(u,n) = \frac{1}{d}\left\|\tilde{\mathbf{Z}}^{(s)}_t(u,n,:) - \text{sg}\Big[\mathcal{R}_{L_{s+1}\to L_s}(\tilde{\mathbf{Z}}^{(s+1)}_t)(u,n,:)\Big]\right\|_2^2, \tag{6}$$

$$e^{(s)}_{\text{dec},t}(u,n) = \left|\hat{\mathbf{Y}}^{(s)}_t(u,n) - \mathcal{R}_{U_{s+1}\to U_s}(\hat{\mathbf{Y}}^{(s+1)}_t)(u,n)\right|. \tag{7}$$

Here n indexes variables and sg stops gradients through the finer latent target during training. At inference, the temporal inconsistency score adds the two errors after interpolation to the original window length:

$$S_{\text{temp},t}(u) = \frac{1}{N}\sum_{n=1}^{N}\sum_{s=0}^{M-2}\left[\mathcal{R}_{L_s\to W}(e^{(s)}_{\text{lat},t})_{u,n} + \mathcal{R}_{U_s\to W}(e^{(s)}_{\text{dec},t})_{u,n}\right]. \tag{8}$$

Thus, the subsequent bidirectional cross-attention operates on context-conditioned representations.

Variable-relation inconsistency. An inter-series encoder produces $\mathbf{R}_t \in \mathbb{R}^{W\times d_r}$. Define $\mathbf{B}_t = \mathbf{X}_t + \rho\ \text{Linear}_r(\mathbf{R}_t)$ and the scale-specific variable signal

$$\mathbf{V}^{(s)}_t = \text{Linear}_z(\tilde{\mathbf{Z}}^{(s)}_t) + \rho\,\mathcal{R}_{W\to L_s}(\mathbf{B}_t) \in \mathbb{R}^{L_s\times N}. \tag{9}$$

Here ρ is a shared learnable residual coefficient used in both fusion stages, initialized to 0.1. Consequently, the projected inter-series features enter $\mathbf{V}^{(s)}_t$ with coefficient ρ^2. Projecting each scalar $V^{(s)}_t(u,n)$ into query $\mathbf{q}_{u,n}$ and key $\mathbf{k}_{u,n}$ vectors of dimension d_a gives the row-normalized variable graph

$$A^{(s)}_t(u,n,m) = \text{softmax}_m\Big(\mathbf{q}^{\mathsf{T}}_{u,n}\mathbf{k}_{u,m}/\sqrt{d_a}\Big), \qquad \mathbf{A}^{(s)}_t \in \mathbb{R}^{L_s\times N\times N}. \tag{10}$$

We compare adjacent graphs only after aligning the temporal dimension:

$$e^{(s)}_{\text{rel},t}(u,n) = \frac{1}{N}\sum_{m=1}^{N}\left|A^{(s)}_t(u,n,m) - \mathcal{R}_{L_{s+1}\to L_s}(\mathbf{A}^{(s+1)}_t)(u,n,m)\right|. \tag{11}$$

The variable-relation score is $S_{\text{rel},t}(u) = N^{-1}\sum_n\sum_{s=0}^{M-2}\mathcal{R}_{L_s\to W}(e^{(s)}_{\text{rel},t})_{u,n}$. This score measures changes in cross-resolution variable dependencies without requiring relation graphs at different resolutions to be identical.

3.5 TRAINING OBJECTIVE

Write $\langle\cdot\rangle$ for the mean over all entries, including the batch during training. Progressive reconstruction is optimized by

$$\mathcal{L}_{\text{rec}} = \frac{\sum_{j=0}^{M-1} U_j\left\langle(\hat{\mathbf{Y}}^{(j)} - \mathbf{Y}^{(j)})^2\right\rangle}{\sum_{j=0}^{M-1} U_j}. \tag{12}$$

The full objective combines temporal and relation consistency with reconstruction:

$$\mathcal{L}_{\text{temp}} = \sum_{s=0}^{M-2}\big(\langle e^{(s)}_{\text{lat}}\rangle + \langle e^{(s)}_{\text{dec}}\rangle\big), \tag{13}$$

$$\mathcal{L}_{\text{rel}} = \sum_{s=0}^{M-2}\langle e^{(s)}_{\text{rel}}\rangle, \tag{14}$$

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha_{\text{temp}}\mathcal{L}_{\text{temp}} + \alpha_{\text{rel}}\mathcal{L}_{\text{rel}}. \tag{15}$$

These auxiliary objectives encourage agreement under normal training data. The default training weights are $\alpha_{\text{temp}} = 0.02$ and $\alpha_{\text{rel}} = 0.005$; they are distinct from inference fusion weights.

3.6 Anomaly Score Fusion

The per-position reconstruction score adds squared errors across decoder resolutions, upsampled to length W and averaged across variables:

$$S_{\mathrm{rec},t}(u) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=0}^{M-1}\left[\mathcal{R}_{U_j\to W}\big((\hat{\mathbf{Y}}_t^{(j)} - \mathbf{Y}_t^{(j)})^2\big)\right]_{u,n}. \tag{16}$$

For each $c \in \{\mathrm{rec}, \mathrm{temp}, \mathrm{rel}\}$, compute the median m_c and median absolute deviation a_c of training-loader position scores. CHIME uses

$$\bar{S}_{c,t}(u) = \frac{S_{c,t}(u) - m_c}{\max(a_c, 10^{-6})}, \qquad S_t(u) = \lambda_{\mathrm{rec}}\bar{S}_{\mathrm{rec},t}(u) + \lambda_{\mathrm{temp}}\bar{S}_{\mathrm{temp},t}(u) + \lambda_{\mathrm{rel}}\bar{S}_{\mathrm{rel},t}(u). \tag{17}$$

We use $(\lambda_{\mathrm{rec}}, \lambda_{\mathrm{temp}}, \lambda_{\mathrm{rel}}) = (0.8, 0.1, 0.1)$ by default. For a window position u corresponding to time $i = t - W + u$, an anomaly is predicted when its fused score exceeds a threshold τ: $\hat{y}_i = \mathbb{I}[S(i) > \tau]$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CHIME on ten public benchmarks covering spacecraft telemetry, server monitoring, industrial control, environmental sensing, and large-scale system traces. SMAP and MSL contain NASA spacecraft telemetry (Hundman et al., 2018); SMD and PSM contain server-monitoring metrics (Su et al., 2019; Abdulaal et al., 2021); SWaT and WADI capture water-treatment and water-distribution processes (Mathur & Tippenhauer, 2016; Ahmed et al., 2017); NIPS-TS-Water (derived from the GECCO water-quality benchmark) contains drinking-water sensor measurements (Moritz et al., 2018); TAO contains oceanic and atmospheric observations with strong seasonality (Crout et al., 2012); Exathlon contains high-dimensional traces from Apache Spark applications (Jacob et al., 2021); and HAI21 contains normal and attack scenarios from a hardware-in-the-loop industrial control testbed (Shin et al., 2021). We follow the public benchmark definitions and splits used in reliable TSAD evaluation (Liu & Paparrizos, 2024). Dataset statistics are provided in Appendix A.

Evaluation Metrics. We report precision (PC), recall (RC), F1-score, VUS-ROC, VUS-PR, area under the ROC curve (AUC-ROC), and affiliation-based F1 (AF-F1). Precision measures false-alarm control, recall measures the proportion of detected anomalies, and F1-score balances the two. VUS-ROC and VUS-PR jointly evaluate score thresholds and temporal tolerance ranges, reducing sensitivity to a single threshold and strict point-wise alignment (Paparrizos et al., 2022). AF-F1 evaluates the temporal proximity between predicted and ground-truth anomaly events (Huet et al., 2022). Following recommendations for reliable TSAD evaluation (Liu & Paparrizos, 2024), we report complementary range- and event-aware metrics in addition to point-adjusted F1. Detailed preprocessing, implementation, optimization, and evaluation settings are provided in Appendices A and B.

4.2 Comparison with Baselines

We compare CHIME with ten representative methods covering reconstruction-based detection (Omni-Anomaly (Su et al., 2019) and TranAD (Tuli et al., 2022)), attention- and association-based detection (Anomaly Transformer (Xu et al., 2022), DCdetector (Yang et al., 2023), SARAD (Dai et al., 2024), and ALoRa (Shimillas et al., 2026)), multi-scale modeling (CATCH (Wu et al., 2025), CrossAD (Li et al., 2025), and TimeRadar (He et al., 2026)), and patch-based representation learning (PaAno (Park & Kang, 2026)).

Table 1 reports point-adjusted F1-scores on all ten benchmarks. CHIME achieves the highest F1-score on eight datasets and the second-highest result on TAO and Exathlon, ranking among the top two methods on every benchmark. Its average F1-score reaches 95.54%, indicating consistent performance across datasets with substantially different dimensions, anomaly ratios, and temporal characteristics.

Table 1: Point-adjusted F1-scores (%). Best results are bold and second-best results are underlined.

Method	SMD	MSL	SMAP	SWaT	PSM	NIPS-TS-Water	WADI	TAO	HAI21	Exathlon
PaAno	51.33	55.91	68.04	84.51	91.97	40.93	65.40	56.41	39.05	81.98
TimeRadar	47.83	56.71	68.67	85.79	82.53	48.22	53.31	72.45	84.56	58.35
ALoRa	76.37	88.22	69.65	88.42	93.97	52.63	27.36	91.08	89.42	73.10
DCdetector	70.24	<u>91.52</u>	78.35	<u>97.18</u>	97.40	39.72	<u>93.33</u>	55.88	48.60	**99.46**
OmniAnomaly	81.96	88.85	70.73	88.44	93.67	49.42	31.98	91.44	<u>92.03</u>	73.13
TranAD	78.46	88.83	69.82	84.86	91.34	52.83	31.70	91.35	88.50	73.12
Anomaly Transformer	<u>85.36</u>	91.17	<u>78.84</u>	91.89	<u>97.53</u>	51.40	52.66	**96.45**	91.66	95.84
SARAD	60.76	59.84	66.20	85.28	95.31	49.53	44.10	65.08	74.09	55.54
CATCH	73.47	66.03	69.01	91.51	96.90	70.26	42.18	90.82	85.73	85.22
CrossAD	79.40	88.51	74.47	84.17	95.62	<u>70.56</u>	35.65	92.65	88.85	58.84
CHIME	**94.23**	**94.85**	**97.91**	**98.16**	**98.46**	**91.02**	**94.67**	<u>94.43</u>	**93.54**	<u>98.09</u>

Table 2 further compares a representative subset of methods using complementary evaluation criteria. Among the methods reported in the table, CHIME achieves the highest F1, AF-F1, VUS-PR, and AUC-ROC on SMAP, SWaT, WADI, and Exathlon. On TAO, CHIME obtains the highest F1-score, whereas CrossAD achieves higher AF-F1, VUS-PR, and AUC-ROC. These results indicate that CHIME remains competitive under range- and event-aware criteria that are less dependent on point adjustment.

Table 2: Detailed results (%) on five representative datasets. In each cell, the first row reports F1 / AF-F1 and the second row reports VUS-PR / AUC-ROC.

Method	SMAP	SWaT	WADI	Exathlon	TAO
	68.04/41.33	84.51/56.65	65.40/54.20	81.98/45.21	56.41/44.76
PaAno	41.28/82.34	72.28/97.77	37.54/86.90	77.57/96.47	42.03/93.86
	68.67/56.71	85.79/15.48	53.31/54.37	58.35/16.44	72.45/66.08
TimeRadar	43.14/79.80	75.88/97.90	26.39/66.09	46.08/87.31	77.76/99.02
	69.65/51.10	88.42/23.19	27.36/34.74	73.10/15.18	91.08/89.50
ALoRa	50.50/83.99	62.76/93.50	11.24/64.10	61.44/92.35	91.31/99.78
	69.82/51.48	84.86/4.35	31.70/45.54	73.12/15.43	91.35/88.36
TranAD	57.94/91.21	47.30/89.83	12.75/64.57	59.23/91.33	94.03/99.86
	66.20/46.71	85.28/42.12	44.10/47.56	55.54/19.10	65.08/54.87
SARAD	42.58/78.59	54.41/95.28	18.20/65.43	37.66/78.71	64.93/98.40
	74.47/64.58	84.17/5.38	35.65/46.03	58.84/15.66	92.65/91.51
CrossAD	67.47/93.26	48.58/91.15	13.12/64.81	42.54/82.72	95.68/99.91
	97.91/67.58	**98.16/69.02**	**94.67/70.39**	**98.09/68.26**	**94.43**/75.18
CHIME	**93.21/99.59**	**96.85/99.92**	**86.03/99.16**	**95.16/99.74**	74.01/97.75

4.3 Ablation Study

We evaluate CHIME on NIPS-TS-Water, SMD, SWaT, and TAO from two perspectives: the contribution of the structural modeling components and the effectiveness of hierarchical inconsistency scores.

Module ablation. We remove the historical context module and bidirectional scale interaction (BSI) from the full model. As shown in Table 3, both variants degrade performance. The average F1-score decreases from 0.9446 for the full model to 0.9171 without BSI and 0.9218 without context. The drop is especially clear on TAO, where F1 falls from 0.9443 to 0.8784 and 0.8892, respectively, indicating that historical state information and cross-scale interaction are important for modeling complex temporal structures.

Score-component ablation. We further compare reconstruction only (S_{rec}), reconstruction with temporal inconsistency ($S_{\mathrm{rec}} + S_{\mathrm{temp}}$), reconstruction with variable-relation inconsistency ($S_{\mathrm{rec}} + S_{\mathrm{rel}}$), and the full fusion. Table 4 shows that the average F1-score increases from 0.9000

Table 3: Module ablation results. PC and RC denote precision and recall, respectively.

Variant	Dataset	F1	PC	RC	VUS-ROC	VUS-PR
w/o BSI	NIPS-TS-Water	0.8735	0.8548	0.8942	0.7230	0.4583
	SMD	0.9365	0.9225	0.9511	0.8271	0.7850
	SWaT	0.9798	0.9625	0.9978	0.9838	0.9751
	TAO	0.8784	0.8845	0.8723	0.7768	0.6953
w/o Context	NIPS-TS-Water	0.8796	0.8809	0.8849	0.7263	0.4555
	SMD	0.9391	0.9242	0.9545	0.8349	0.7869
	SWaT	0.9792	0.9647	0.9942	0.9841	0.9757
	TAO	0.8892	0.9114	0.8731	0.7658	0.6856
Full	NIPS-TS-Water	**0.9102**	0.8849	0.9370	0.7301	0.4874
	SMD	**0.9423**	0.9217	0.9639	0.8311	0.7837
	SWaT	**0.9816**	0.9701	0.9933	0.9836	0.9788
	TAO	**0.9443**	0.8711	0.9760	0.8074	0.7324

with reconstruction alone to 0.9384 and 0.9342 after adding temporal and relation inconsistency, respectively, and reaches 0.9446 with the full model.

Table 4: Score-component ablation in terms of point-adjusted F1-score.

Score Composition	NIPS-TS-Water	SMD	SWaT	TAO	Avg.
S_{rec}	0.7921	0.9153	0.9719	0.9206	0.9000
$S_{\mathrm{rec}} + S_{\mathrm{temp}}$	0.9003	0.9371	0.9786	0.9377	0.9384
$S_{\mathrm{rec}} + S_{\mathrm{rel}}$	0.8968	0.9339	0.9761	0.9301	0.9342
Full CHIME	**0.9102**	**0.9423**	**0.9816**	**0.9443**	**0.9446**

The largest gain appears on NIPS-TS-Water, where F1 improves from 0.7921 with reconstruction alone to 0.9102 with full fusion. Overall, both inconsistency terms consistently improve over S_{rec}, while their combination performs best on all four datasets. These results support the use of temporal and variable-relation inconsistency as complementary anomaly evidence beyond reconstruction error.

4.4 Parameter Sensitivity

We analyze the sensitivity of CHIME to three groups of hyperparameters: temporal patch length, multi-scale pooling configuration, and anomaly-score fusion mode. Figure 2 reports the point-adjusted F1-scores on six representative datasets.

As shown in Figure 2(a), CHIME is relatively stable across patch lengths of 4, 6, and 8 on most datasets. A patch length of 6 provides a favorable overall trade-off, while increasing the patch length to 8 yields a noticeable improvement on NIPS-TS-Water. This suggests that most datasets are not highly sensitive to moderate changes in local temporal granularity, whereas datasets with stronger long-range or periodic patterns may benefit from a larger patch size.

Figure 2(b) evaluates three multi-scale configurations, namely $[12, 6, 3, 2]$, $[16, 8, 4, 2]$, and $[20, 10, 5, 2]$. The performance of MSL, PSM, SMAP, SWaT, and WADI remains relatively stable across these settings, whereas NIPS-TS-Water benefits more clearly from larger temporal scales. This result indicates that the choice of temporal resolutions has a dataset-dependent effect, particularly for time series with pronounced seasonal or multi-resolution structures. We use $[16, 8, 4, 2]$ as the default configuration because it provides a balanced representation across fine- and coarse-grained temporal patterns.

Finally, Figure 2(c) compares the default and fine-dominant anomaly-score fusion strategies. The two settings produce similar results on most datasets, indicating that CHIME is relatively robust to moderate changes in score-fusion weights. The fine-dominant strategy provides a larger improvement on NIPS-TS-Water, suggesting that emphasizing fine-grained anomaly evidence can be beneficial for datasets containing localized or strongly periodic deviations. Overall, CHIME shows stable

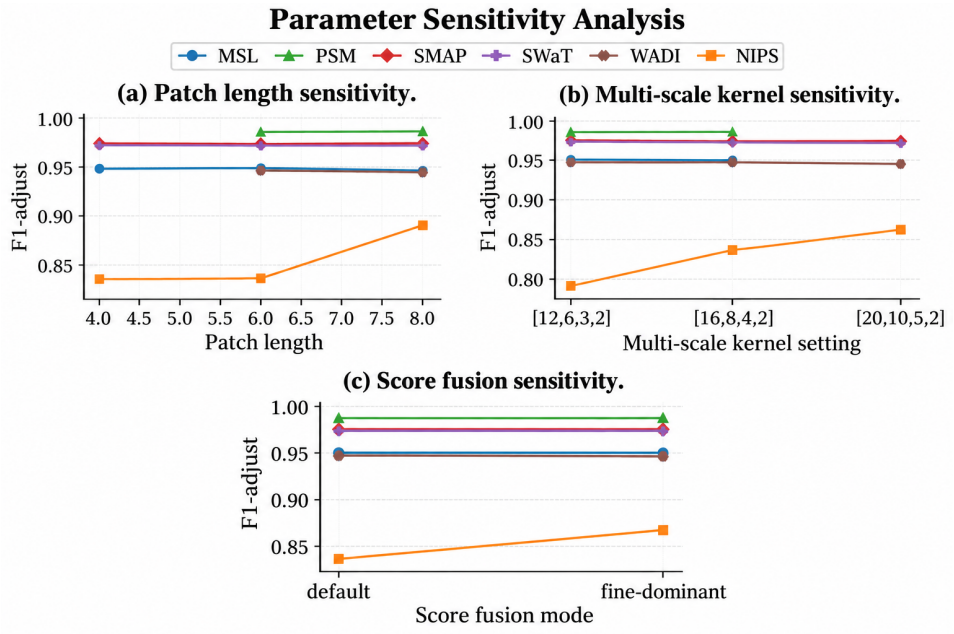

Figure 2: Parameter sensitivity of CHIME on six representative datasets: (a) temporal patch length, (b) multi-scale pooling configuration, and (c) anomaly-score fusion mode. NIPS denotes NIPS-TS-Water.

performance across the evaluated hyperparameter settings, while the optimal temporal granularity may vary with dataset characteristics.

5 Conclusion

We presented CHIME, a framework that augments reconstruction-based anomaly detection with context-conditioned hierarchical inconsistency modeling for multivariate time series. Historical context is used to condition multi-resolution representations, while discrepancies across temporal representations, reconstruction outputs, and variable-relation structures provide complementary anomaly evidence beyond reconstruction error alone. Experiments on ten public benchmarks demonstrate strong point-adjusted F1 performance, while complementary event- and range-aware metrics further support the effectiveness of CHIME across several datasets. Ablation results also indicate that temporal and variable-relation inconsistency provide complementary contributions to anomaly detection. These findings suggest that hierarchical inconsistency can capture structural abnormalities that may not be sufficiently reflected by reconstruction error. Future work will investigate the conditions under which such inconsistency signals are most beneficial, improve robustness to changing operating states and contaminated historical context, and develop more efficient variable-relation modeling for high-dimensional time series.

AI Use Statement

Generative AI tools were used to assist with Chinese-to-English translation, language editing, and LaTeX formatting. All AI-assisted content was reviewed by the authors, who take full responsibility for the technical claims, equations, experimental results, and conclusions.

Reproducibility Statement

Section 3 specifies the model components, cross-scale interactions, and anomaly-score construction. Section 4 describes the datasets, comparison methods, and evaluation metrics, while Appendix A reports the dataset statistics. All methods are compared using a common evaluation pipeline, and

complementary range- and event-aware metrics are reported to reduce dependence on point adjustment.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2485–2494. Association for Computing Machinery, 2021. ISBN 9781450383325. doi: 10.1145/3447548.3467174. URL https://doi.org/10.1145/3447548.3467174.

Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P. Mathur. Wadi: a water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, CySWATER '17, pp. 25–28. Association for Computing Machinery, 2017. ISBN 9781450349758. doi: 10.1145/3055366.3055375. URL https://doi.org/10.1145/3055366.3055375.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 3395–3404. Association for Computing Machinery, 2020. ISBN 9781450379984. doi: 10.1145/3394486.3403392. URL https://doi.org/10.1145/3394486.3403392.

Richard Crout, Dawn Petraitis, and NOAA National Data Buoy Center. Physical and meteorological data from the tropical atmosphere ocean (TAO) array in the tropical pacific ocean. Dataset, 2012. Tropical Atmosphere Ocean (TAO) Array dataset.

Zhihao Dai, Ligang He, Shuang-Hua Yang, and Matthew Leeke. Sarad: Spatial association-aware anomaly detection and diagnosis for multivariate time series. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 48371–48410. Curran Associates, Inc., 2024. doi: 10.52202/079017-1533. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/56ad264ac7448239145606cf4106042f-Paper-Conference.pdf.

Hui He, Hezhe Qiao, Yutong Chen, Kun Yi, and Guansong Pang. Timeradar: A domain-rotatable foundation model for time series anomaly detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 1602–1613. Association for Computing Machinery, 2026. ISBN 9798400722592. doi: 10.1145/3770855.3818062. URL https://doi.org/10.1145/3770855.3818062.

Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 635–645. Association for Computing Machinery, 2022. ISBN 9781450393850. doi: 10.1145/3534678.3539339. URL https://doi.org/10.1145/3534678.3539339.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 387–395. Association for Computing Machinery, 2018. ISBN 9781450355520. doi: 10.1145/3219819.3219845. URL https://doi.org/10.1145/3219819.3219845.

Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. Exathlon: a benchmark for explainable anomaly detection over time series. *Proc. VLDB Endow.*, 14(11): 2613–2626, July 2021. ISSN 2150-8097. doi: 10.14778/3476249.3476307. URL https://doi.org/10.14778/3476249.3476307.

Jahoon Jeong and Hyunsoo Yoon. Pgrf-net: A prototype-guided relational fusion network for diagnostic multivariate time-series anomaly detection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 119368–119391,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c204d12afa0175285e5aac65188808b4-Paper-Conference.pdf.

Jongwon Kim, Samuel Yoon, Young Ko, Yerin Kim, Sung Kim, and JAEUNG TAE. Complexity- and statistics-guided anomaly detection in time series foundation models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 128285–128306, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d06768735ce7d11ade5baef7099b4e05-Paper-Conference.pdf.

Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. Crossad: Time series anomaly detection with cross-scale associations and cross-window modeling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 137959–137985. Curran Associates, Inc., 2025. doi: 10.52202/085713-4605. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c9658e8c20879632cb1cfca91d80ceb7-Paper-Conference.pdf.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 108231–108261. Curran Associates, Inc., 2024. doi: 10.52202/079017-3437. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c3f3c690b7a99fba16d0efd35cb83b2c-Paper-Datasets_and_Benchmarks_Track.pdf.

Aditya P. Mathur and Nils Ole Tippenhauer. Swat: a water treatment testbed for research and training on ics security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pp. 31–36, 2016. doi: 10.1109/CySWater.2016.7469060.

Steffen Moritz, Frederik Rehbach, Sowmya Chandrasekaran, Margarita Rebolledo, and Thomas Bartz-Beielstein. GECCO industrial challenge 2018 dataset: A water quality dataset for the "internet of things: Online anomaly detection for drinking water quality" competition at the genetic and evolutionary computation conference 2018, kyoto, japan, 2018. Dataset.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *Proc. VLDB Endow.*, 15(11):2774–2787, July 2022. ISSN 2150-8097. doi: 10.14778/3551793.3551830. URL https://doi.org/10.14778/3551793.3551830.

Jinju Park and Seokho Kang. Paano: Patch-based representation learning for time-series anomaly detection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 54464–54491, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5991785875cf5eee9c29e2ed7d26d1c0-Paper-Conference.pdf.

Charalampos Shimillas, Kleanthis Malialis, Konstantinos Fokianos, and Marios Polycarpou. Low rank transformer for multivariate time series anomaly detection and localization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 59972–60005, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/61c00c07e6d27285e4b952e96cc65666-Paper-Conference.pdf.

Hyeok-Ki Shin, Woomyo Lee, Jeong-Han Yun, and Byung-Gil Min. Two ics security datasets and anomaly detection contest on the hil-based augmented ics testbed. In *Proceedings of the 14th Cyber Security Experimentation and Test Workshop*, CSET '21, pp. 36–40. Association for Computing Machinery, 2021. ISBN 9781450390651. doi: 10.1145/3474718.3474719. URL https://doi.org/10.1145/3474718.3474719.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: Memory-guided transformer for multivariate time series anomaly detection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 57947–57963. Curran Associates, Inc., 2023. doi: 10.52202/

075280-2525. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b4c898eb1fb556b8d871fbe9ead92256-Paper-Conference.pdf.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2828–2837. Association for Computing Machinery, 2019. ISBN 9781450362016. doi: 10.1145/3292500.3330672. URL https://doi.org/10.1145/3292500.3330672.

Xin Sun, Heng Zhou, and Chao Li. Multivariate time series anomaly detection with idempotent reconstruction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 160900–160958. Curran Associates, Inc., 2025. doi: 10.52202/085713-5374. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ebba26f4dbcfe7318e4a263c54e9cf42-Paper-Conference.pdf.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endow.*, 15(6):1201–1214, February 2022. ISSN 2150-8097. doi: 10.14778/3514061.3514067. URL https://doi.org/10.14778/3514061.3514067.

Shiyu Wang, Jiawei LI, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. Timemixer++: A general time series pattern machine for universal predictive analysis. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 16980–17016, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2b187165e28fdfdc0ffb34d1bfff2b0c-Paper-Conference.pdf.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 17017–17045, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2b25c39788e5cf11d3541de433ebf4c0-Paper-Conference.pdf.

Yongzheng Xie, Hongyu Zhang, and Muhammad Ali Babar. Multivariate time series anomaly detection by capturing coarse-grained intra- and inter-variate dependencies. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 697–705. Association for Computing Machinery, 2025. ISBN 9798400712746. doi: 10.1145/3696410.3714941. URL https://doi.org/10.1145/3696410.3714941.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=LzQQ89U1qm_.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3033–3045. Association for Computing Machinery, 2023. ISBN 9798400701030. doi: 10.1145/3580305.3599295. URL https://doi.org/10.1145/3580305.3599295.

Xiaokun Zhu, Louis Carpentier, and Mathias Verbeke. When foundation models are one-liners: Limitations and future directions for time series anomaly detection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 127659–127688, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cf70320e93c08b39b1b29a348097a376-Paper-Conference.pdf.