EXPLAINABLE MULTIVARIATE TIME-SERIES ANOMALY DETECTION VIA DYNAMIC GRAPH OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate time-series anomaly detection is challenging because anomalies often arise from changes in cross-sensor dependencies rather than large deviations in individual channels. Existing methods capture amplitude errors, but often miss these structural anomalies and offer limited support for explainable sensor-level root-cause localization. We propose TOAD, an explainable topology-aware optimal-transport framework for unsupervised multivariate time-series anomaly detection and root-cause attribution. TOAD constructs dynamic sensor graphs, learns topology-aware spatiotemporal latent representations, and detects structural anomalies by comparing latent distributions with regime-aware normal references using sliced Wasserstein distance. To improve robustness, the final anomaly score integrates evidence from transport, reconstruction, and temporal transitions. Beyond detection, TOAD provides faithful sensor-level explanations by attributing anomaly scores through gradient sensitivity, counterfactual score reduction, and graph-aware propagation, enabling root-cause localization directly aligned with the detection mechanism. Extensive experiments on five public benchmarks demonstrate that TOAD achieves the best threshold-independent detection performance on four datasets and improves root-cause localization from 64.0% to 72.8% (H@100) and from 78.2% to 86.4% (H@150), providing accurate and actionable explanations for anomalous events.

1 INTRODUCTION

Modern industrial and cyber-physical systems generate high-dimensional multivariate time series from networks of sensors. However, anomalies often arise from changes in cross-sensor dependencies rather than large deviations in individual channels. Therefore, multivariate time-series anomaly detection must accurately detect structural anomalies and identify the sensors responsible for them.

To address this challenge, most existing methods detect anomalies using reconstruction or forecasting errors. Approaches such as OmniAnomaly Su et al. (2019), USAD Audibert et al. (2020), TranAD Tuli et al. (2022), Anomaly Transformer Xu et al. (2022), and DCdetector Yang et al. (2023) perform well when anomalies produce large deviations in sensor values. However, they often miss structural anomalies that preserve individual sensor values while altering cross-sensor dependencies. Graph-based detectors address this limitation by modelling sensors as interacting nodes. Methods such as MTAD-GAT Zhao et al. (2020), GDN Deng & Hooi (2021), and GANF Dai & Chen (2022) improve structural modelling through learned sensor graphs, while recent transport-based methods compare latent distributions using Optimal Transport to better detect structural anomalies Wang et al. (2024). Nevertheless, most existing transport-based methods compute window-level anomaly scores, making short anomalies harder to detect and providing little insight into the sensors responsible for the detected anomaly. To explain detected anomalies, existing methods commonly rank sensors using reconstruction or prediction errors, as in OmniAnomaly Su et al. (2019), MTAD-GAT Zhao et al. (2020), InterFusion Li et al. (2021), and TranAD Tuli et al. (2022), while more recent work explores association-based diagnosis Dai et al. (2024). However, these explanations are often only loosely connected to the detector's decision, leaving faithful sensor-level root-cause localization an open challenge.

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.

We propose **TOAD**, an explainable topology-aware Optimal Transport framework for joint anomaly detection and root-cause localization in multivariate time series. The key insight is that structural anomalies often preserve individual sensor values while altering the dependency patterns among sensors. As a result, reconstruction or forecasting errors may remain small even though the underlying system behaviour has changed. TOAD addresses this limitation by learning topology-aware spatiotemporal latent representations and comparing their distributions with regime-aware normal references using sliced Wasserstein distance. This distribution-based comparison directly captures structural changes that are difficult to detect from pointwise errors alone. To improve robustness, the final anomaly score combines three complementary sources of evidence: Optimal Transport for structural deviation, reconstruction for amplitude deviation, and temporal transitions for contextual change. Beyond anomaly detection, TOAD provides faithful sensor-level root-cause localization using the same evidence that produces the anomaly score. Instead of relying on a separate post-hoc explainer, TOAD decomposes the fused anomaly score into sensor-level contributions through transport gradients, reconstruction evidence, temporal sensitivity, counterfactual repair, and graph-aware propagation. Consequently, the explanation is directly aligned with the detector's decision, allowing not only accurate localization but also quantitative evaluation of explanation faithfulness through counterfactual score reduction.

TOAD introduces four distinct capabilities that no single prior method combines. **(i)** Unlike GDN, which learns a *static* sensor graph and scores by forecasting error, TOAD recomputes the graph dynamically per window and scores by latent distributional transport, making it sensitive to evolving dependencies. **(ii)** Unlike MADGA, which computes a single window-level Gromov-Wasserstein distance, TOAD evaluates sliced Wasserstein evidence at *timestep resolution* and fuses it with reconstruction and temporal-transition signals, avoiding the dilution of brief anomalies by window pooling. **(iii)** Unlike TranAD, which derives sensor-level diagnosis from per-channel reconstruction error that may not reflect the detector's actual decision, TOAD decomposes the *same fused anomaly score* into sensor contributions via gradients, counterfactual repair, and graph-aware propagation, ensuring alignment between detection and explanation. **(iv)** Unlike OmniAnomaly or USAD, which detect only amplitude or stochastic deviations, TOAD's optimal-transport stream explicitly captures structural anomalies that preserve marginal sensor ranges. These four design choices are tightly coupled: the dynamic graph feeds the encoder, the encoder produces the latent trajectory, the trajectory enables timestep-level transport scoring, and the score components are reused for faithful root-cause analysis.

This paper makes three main contributions:

- We propose an explainable topology-aware Optimal Transport framework for multivariate time-series anomaly detection that learns dynamic sensor graphs and detects structural anomalies by comparing latent distributions with regime-aware normal references at timestep resolution.
- We introduce a score-aligned multi-evidence attribution mechanism that combines transport, reconstruction, and temporal-transition evidence for both anomaly detection and faithful sensor-level root-cause localization, eliminating the need for a separate post-hoc explanation model.
- We conduct extensive experiments on five public multivariate time-series benchmarks. TOAD achieves state-of-the-art threshold-independent detection performance on four datasets and substantially improves root-cause localization, increasing H@100 from 64.0% to 72.8% and H@150 from 78.2% to 86.4%. We further validate explanation faithfulness through counterfactual repair experiments, showing that repairing the top-ranked sensors consistently reduces anomaly scores more effectively than competing localization methods.

2 Related Work

Multivariate time-series anomaly detection: Existing MTAD methods primarily detect anomalies through reconstruction, forecasting, or likelihood estimation. Representative approaches include reconstruction-based models such as OmniAnomaly (Su et al., 2019) and USAD (Audibert et al., 2020), transformer-based models such as TranAD (Tuli et al., 2022), and attention-based methods including Anomaly Transformer (Xu et al., 2022) and DCdetector (Yang et al., 2023). While these methods achieve strong performance when anomalies produce large deviations in sensor values, they

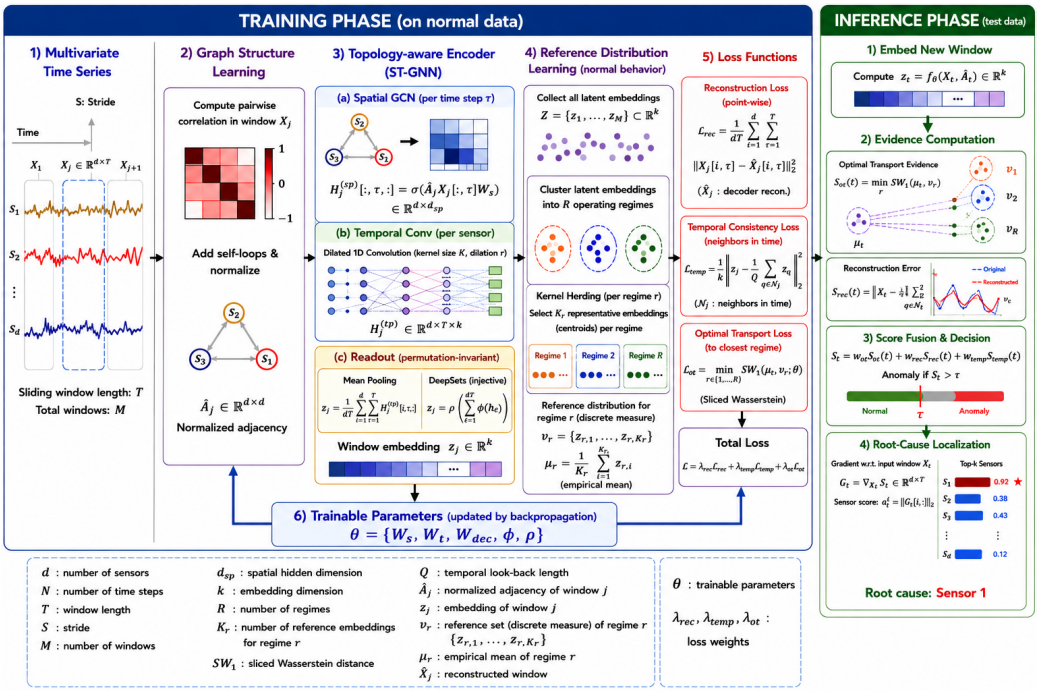

Figure 1: Architecture of the proposed TOAD framework. In training, dynamic sensor graphs are constructed from sliding windows, encoded by an ST-GNN, and optimized using multiple loss functions to learn topology-aware embeddings and normal reference distributions. During inference, optimal-transport, reconstruction, and temporal-transition evidence are fused for anomaly detection, while gradient-based attribution localizes the root-cause sensors.

are generally less effective at detecting structural anomalies that primarily alter cross-sensor dependencies. Graph-based methods, including MTAD-GAT (Zhao et al., 2020), GDN (Deng & Hooi, 2021), and GANF (Dai & Chen, 2022), explicitly model sensor interactions to improve dependency learning. However, their anomaly scores are still largely derived from reconstruction, forecasting, or likelihood estimation. In contrast, TOAD detects anomalies by measuring distributional changes in topology-aware latent representations using Optimal Transport, making it more suitable for capturing structural anomalies.

Root-cause analysis and evaluation: Existing root-cause analysis methods typically identify faulty sensors by ranking reconstruction or prediction errors, as adopted by OmniAnomaly (Su et al., 2019), MTAD-GAT (Zhao et al., 2020), InterFusion (Li et al., 2021), and TranAD (Tuli et al., 2022). More recent approaches, such as SARAD (Dai et al., 2024), exploit changes in spatial associations to improve localization. However, these methods usually rely on a single diagnostic signal, and the resulting explanations are not always aligned with the detector's final decision. In contrast, TOAD derives sensor-level root-cause scores directly from the same multi-evidence anomaly score used for detection, providing faithful and score-aligned explanations. Following recent recommendations (Paparrizos et al., 2022; Liu & Paparrizos, 2024), we evaluate anomaly detection using threshold-independent and range-aware metrics, and assess localization using hit-rate, ranking quality, and counterfactual faithfulness.

3 Proposed Method

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{d \times N}$ be a multivariate time series with d sensors and N time steps, where $\mathbf{x}_t \in \mathbb{R}^d$ denotes the system state at time t. The series is partitioned into overlapping windows of length T and stride S, yielding $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$ with $\mathbf{X}_j \in \mathbb{R}^{d \times T}$ and $M = \lfloor (N - T)/S \rfloor + 1$. We propose TOAD (**T**opology-**O**ptimal tr**A**nsport for multivariate time-series anomaly **D**etection and root-cause analysis), an explainable framework for joint anomaly detection and localization. For each window, TOAD constructs a dynamic sensor graph, learns a

topology-aware latent representation, and compares it with normal references using sliced Wasserstein distance. The resulting transport, reconstruction, and temporal-transition evidence are fused for anomaly detection and decomposed for root-cause localization. TOAD consists of four components: *(i) graph structure learning*, *(ii) topology-aware window embedding*, *(iii) anomaly scoring*, and *(iv) anomaly localization*. Figure 1 summarizes the framework; the following subsections describe each component together with the training and inference procedures.

Graph structure learning: For each sliding window $\mathbf{X}_j \in \mathcal{X}$, TOAD constructs a dynamic sensor graph $G_j = (\mathcal{V}, \mathcal{E}, \mathbf{A}_j)$, where $\mathcal{V} = \{s_1, \ldots, s_d\}$ and $\mathbf{A}_j \in \mathbb{R}^{d\times d}$ is the adjacency matrix. Edge weights are computed using any graph structure learning including Pearson correlation between sensor signals within the window, allowing the graph topology to adapt to evolving sensor dependencies. The Self-loops are added as $\tilde{\mathbf{A}}_j = \mathbf{A}_j + \mathbf{I}_d$, and the graph is symmetrically normalized as $\hat{\mathbf{A}}_j = \tilde{\mathbf{D}}_j^{-1/2}\tilde{\mathbf{A}}_j\tilde{\mathbf{D}}_j^{-1/2}$, where $\tilde{\mathbf{D}}_j[i,i] = \sum_{k=1}^{d} |\tilde{\mathbf{A}}_j[i,k]|$. The normalized adjacency matrix $\hat{\mathbf{A}}_j$ is used by the topology-aware encoder to propagate information between correlated sensors.

Topology-aware window embedding: The encoder f_θ maps each graph and window pair to a compact topology-aware embedding $Z_j = f_\theta(\mathbf{X}_j, \hat{\mathbf{A}}_j) \in \mathbb{R}^k$, $k \ll dT$, by composing a graph convolution (GCN), a temporal convolution (TCN) and a permutation-invariant readout; applying it to all windows yields the latent sequence $\mathcal{Z}_{\text{all}} = \{Z_1, \ldots, Z_M\}$, and for unit stride ($S = 1$) we write Z_t for the window ending in the global time step t. Let $\tau \in \{1, \ldots, T\}$ be the local time index. The GCN first aggregates the sensors correlated per step, $\mathbf{H}_j^{(sp)}[:,\tau,:] = \sigma_{sp}\big(\hat{\mathbf{A}}_j\mathbf{X}_j[:,\tau]\mathbf{W}_{sp}\big) \in \mathbb{R}^{d\times d_{sp}}$, where $\mathbf{W}_{sp} \in \mathbb{R}^{1\times d_{sp}}$ (applied by broadcasting) is spectrally normalized, $\sigma_{\max}(\mathbf{W}_{sp}) \le 1$, and σ_{sp} is a LeakyReLU; stacking over τ gives $\mathbf{H}_j^{(sp)} \in \mathbb{R}^{d\times T\times d_{sp}}$. The TCN then models temporal evolution independently for each sensor through the causal dilated operator $\mathcal{C} : \mathbb{R}^{T\times d_{sp}} \to \mathbb{R}^{T\times k}$ with kernel size K, dilation r, and kernel $\mathbf{W}_{tp} \in \mathbb{R}^{K\times d_{sp}\times k}$,

$$(\mathcal{C}(\mathbf{H}))_{\tau,:} = \sum_{m=0}^{K-1} \mathbf{H}_{\tau - rm,:}\, \mathbf{W}_{tp}[m,:,:], \qquad \tau = 1, \ldots, T, \tag{1}$$

with causal zero-padding ($\mathbf{H}_{s,:} = \mathbf{0}$ for $s < 1$ or $s > T$); applied to every sensor, it yields $\mathbf{H}_j^{(tp)} \in \mathbb{R}^{d\times T\times k}$. The final embedding is the global mean readout

$$Z_j = \mathcal{R}_{GM}(\mathbf{H}_j^{(tp)}) = \frac{1}{dT}\sum_{i=1}^{d}\sum_{\tau=1}^{T} \mathbf{H}_j^{(tp)}[i,\tau,:] \in \mathbb{R}^k. \tag{2}$$

TOAD also supports an injective DeepSets readout, $\rho\big(\sum_{\ell=1}^{dT}\phi(\mathbf{h}_\ell)\big)$ over the flattened multiset of spatiotemporal features, which by the DeepSets universality theorem is permutation invariant and injective for sufficiently large hidden width, and hence additionally guarantees a strictly positive lower Lipschitz bound (information preservation); since mean pooling enjoys uniformly better stability constants (Section 4), we adopt it throughout. Since $\|\hat{\mathbf{A}}_j\|_2 \le 1$ and spectral normalization enforces $\|\mathbf{W}_{sp}\|_2, \|\mathbf{W}_{tp}\|_2 \le 1$, the encoder is globally Lipschitz.

Anomaly scoring: Optimal transport captures structural changes in the latent space but may miss amplitude deviations or abrupt temporal transitions; TOAD therefore fuses three complementary evidence streams. *Transport evidence* measures how far the current local latent behaviour is from the normal reference: from the timestep-level trajectory we form the local empirical measure $\mu_t = \frac{1}{W_t}\sum_{\ell=0}^{W_t-1}\delta_{Z_{t-\ell}}$, $W_t = \min(W, t)$, and compare it against R normal operating regimes, each a discrete measure $\nu_r = \frac{1}{K_r}\sum_{\ell=1}^{K_r}\delta_{z_{r,\ell}}$, taking the distance to the closest regime,

$$S_{\text{ot}}(t) = \min_{r\in\{1,\ldots,R\}} \mathcal{SW}_1(\mu_t, \nu_r), \tag{3}$$

where $\mathcal{SW}_1$ is the sliced 1-Wasserstein distance; the $\min_r$ suppresses false alarms from legitimate regime changes, and slicing keeps streaming cost low. *Reconstruction evidence* captures amplitude anomalies in sensor space: the decoder reconstructs $\widehat{X}_{j,m} = g_\phi(Z_j, m/T) \in \mathbb{R}^d$ from the window representation, overlapping reconstructions are averaged to $\widehat{X}_t$, and $S_{\text{rec}}(t) = \|X_t - \widehat{X}_t\|_2^2$. *Temporal-transition evidence* captures abrupt contextual shifts in the latent trajectory, $S_{\text{temp}}(t) = \|Z_t - \frac{1}{Q_t}\sum_{\ell=1}^{Q_t} Z_{t-\ell}\|_2$ with $Q_t = \min(q, t-1)$. As the streams differ in scale, each is standardized

with training-only robust statistics, $\mathrm{rz}(S_a(t)) = \big(S_a(t) - \mathrm{median}(S_a^{\mathrm{train}})\big)/\big(\mathrm{MAD}(S_a^{\mathrm{train}}) + \epsilon\big)$, and fused with fixed weights $w_a \geq 0$,

$$S(t) = \textstyle\sum_{a\in\{\mathrm{ot,rec,temp}\}} w_a \,\mathrm{rz}(S_a(t)). \tag{4}$$

A time step is flagged as anomalous when $S(t)$ exceeds a threshold τ estimated from normal training or validation scores.

Anomaly Localization The global readout collapses the spatial dimension into $Z_t \in \mathbb{R}^k$, masking per-sensor contributions to $S(t)$. Because each robust-normalized stream is affine in its raw evidence (median and MAD are frozen training constants), the input attribution decomposes linearly across the three components,

$$\nabla_{\mathbf{X}} S(t) = \sum_{a\in\{\mathrm{ot,rec,temp}\}} \frac{w_a}{\mathrm{MAD}(S_a^{\mathrm{train}}) + \epsilon} \nabla_{\mathbf{X}} S_a(t), \tag{5}$$

where the transport and temporal streams are traced back through the differentiable encoder, and the reconstruction stream is closed-form and sensor-indexed, contributing $2(X_t^i - \widehat{X}_t^i)$ to channel i directly. Collecting the Jacobian $\mathbf{G}_t = \nabla_{\mathbf{X}} S(t) \in \mathbb{R}^{d\times T}$, the per-sensor attribution is $\mathcal{A}_t^i = \|\mathbf{G}_t[i,:]\|_2 = \big(\sum_{m=1}^{T} (\partial S(t)/\partial \mathbf{X}_t[i,m])^2\big)^{1/2}$,gathered into $\boldsymbol{\mathcal{A}}_t \in \mathbb{R}^d$, where $\boldsymbol{\mathcal{A}}_t = [\mathcal{A}_t^1, \ldots, \mathcal{A}_t^d]^\top$. A channel i is isolated as root cause when $\mathcal{A}_t^i$ exceeds a baseline threshold. Because localization decomposes the same three components that raise the alarm, the sensor ranking is aligned with the evidence responsible for detection. We emphasize that the proposed root-cause analysis is designed to be *faithful* to the detector's decision rather than causally complete. In the explainable-AI literature, faithfulness measures whether an explanation reflects the evidence actually used by the model, whereas causal correctness requires interventional data or a known structural causal model. Because ground-truth causal graphs are unavailable in all public MTAD benchmarks, TOAD optimizes the operational objective: when an alarm fires, the operator needs to know which sensors drive the anomaly score $S(t)$, not an independent causal story. The counterfactual repair term $R_{\mathrm{cf}}^i(t)$ directly tests this alignment by measuring whether replacing sensor i reduces the same score that triggered the alarm. Causal correctness remains an important but separate goal that requires domain-expert annotations or controlled physical experiments beyond the scope of standard benchmarks.

Training: Training jointly optimizes the encoder f_θ, a non-linear projection head $h_\psi : \mathbb{R}^k \to \mathbb{R}^p$, and a continuous time-query decoder $g_\phi : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}^d$ on clean data $\mathcal{X}_{\text{train}}$. Per mini-batch, each window yields its online topology $\hat{\mathbf{A}}_j$, anchor $\boldsymbol{Z}_j = f_\theta(\mathbf{X}_j, \hat{\mathbf{A}}_j)$, and projection $\boldsymbol{P}_j = h_\psi(\boldsymbol{Z}_j)$; a positive $\boldsymbol{P}_j^+$ comes from a temporally adjacent window, and a hard negative couples feature-level and structural corruption in a single boundary sample: a bounded adversarial perturbation $\mathbf{X}_j^- = \mathbf{X}_j + \epsilon_{\text{adv}}\,\mathrm{sign}(\nabla_{\mathbf{X}}\mathcal{L}_{\text{contra}})$ is forwarded through a corrupted topology $\hat{\mathbf{A}}_{\text{corrupted}}$ (random edge masking/permutation) to give $\boldsymbol{P}_j^- = h_\psi\big(f_\theta(\mathbf{X}_j^-, \hat{\mathbf{A}}_{\text{corrupted}})\big)$. The joint objective is

$$\mathcal{L}_{\text{total}}(\theta,\psi,\phi) = \mathcal{L}_{\text{rec}}(\theta,\phi) + \lambda_1\mathcal{L}_{\text{reg}}(\theta) + \lambda_2\mathcal{L}_{\text{contra}}(\theta,\psi) + \lambda_3\mathcal{L}_{\text{indep}}(\theta), \qquad \lambda_1,\lambda_2,\lambda_3 > 0. \tag{6}$$

Here $\mathcal{L}_{\text{rec}} = \frac{1}{Bm}\sum_j \sum_{t\in\mathcal{T}_{\text{query}}} \|X_{j,t} - g_\phi(\boldsymbol{Z}_j, t)\|_2^2$, over uniformly sampled normalized query times $\mathcal{T}_{\text{query}} \subset [0,1]$, routes reconstruction through the core representation to preserve information, and trains the decoder used by the reconstruction evidence at inference. $\mathcal{L}_{\text{contra}}$ is the InfoNCE loss over $(\boldsymbol{P}_j, \boldsymbol{P}_j^+, \boldsymbol{P}_j^-)$ with cosine similarity and temperature τ_{temp}, aligning positives while repelling the synthetic boundary negatives; computing it in the projection space shields $\boldsymbol{Z}_j$ from compression distortion. $\mathcal{L}_{\text{reg}} = \sum_l \max\big(0, \sigma_{\max}(\mathbf{W}_l) - 1\big)^2$ softly clamps the spectral norms, with $\sigma_{\max}$ estimated by power iteration. $\mathcal{L}_{\text{indep}} = \sum_{a<b} \mathrm{HSIC}(Z_{:,a}, Z_{:,b})$ decorrelates latent channels using the standard biased HSIC estimator with a characteristic kernel. The parameters are updated with Adam; after each step, hard projections $\mathbf{W}_{sp} \leftarrow \mathbf{W}_{sp}/\sigma_{\max}(\mathbf{W}_{sp})$ and $\mathbf{W}_{tp} \leftarrow \mathbf{W}_{tp}/\sigma_{\max}(\mathbf{W}_{tp})$ enforce the unit spectral norm, keeping the Lipschitz bound of Section 4 active throughout the optimization. Algorithm 1 summarizes training and inference-statistic construction.

Reference distribution estimation: After convergence, h_ψ is discarded, contrastive alignment is only a training device, while g_ϕ is retained for the reconstruction evidence. The clean windows are embedded by the frozen encoder into the normal latent library $\mathcal{Z}_{\text{all}} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M\}$, which is partitioned into R operating regimes by clustering; kernel herding within each regime extracts the K_r centroids forming the references ν_r of (3). Finally, the robust statistics $\mathrm{median}(S_a^{\mathrm{train}})$ and $\mathrm{MAD}(S_a^{\mathrm{train}})$ standardizing (4) and the threshold τ are calibrated on the training scores, completing the statistics required for streaming inference.

Algorithm 1 TOAD training and inference-statistic construction

Require: clean windows $\mathcal{X}_{\text{train}}$; hyperparameters $\lambda_1, \lambda_2, \lambda_3, \epsilon_{\text{adv}}, \tau_{\text{temp}}$; regimes R, budget K_r; weights $w_{\text{ot}}, w_{\text{rec}}, w_{\text{temp}}$; horizons W, q

initialize θ, ψ, ϕ
repeat
  sample a mini-batch $\{\mathbf{X}_j\}_{j=1}^{B} \subset \mathcal{X}_{\text{train}}$
  for each $\mathbf{X}_j$ **do**
    compute $\hat{\mathbf{A}}_j$ (Sec. 3); $\boldsymbol{Z}_j = f_\theta(\mathbf{X}_j, \hat{\mathbf{A}}_j)$; $\boldsymbol{P}_j = h_\psi(\boldsymbol{Z}_j)$
    build positive $\boldsymbol{P}_j^+$ from an adjacent window
    $\mathbf{X}_j^- = \mathbf{X}_j + \epsilon_{\text{adv}}\, \text{sign}(\nabla_{\mathbf{X}} \mathcal{L}_{\text{contra}})$; $\boldsymbol{P}_j^- = h_\psi\big(f_\theta(\mathbf{X}_j^-, \hat{\mathbf{A}}_{\text{corrupted}})\big)$
    decode $\hat{X}_{j,t} = g_\phi(\boldsymbol{Z}_j, t)$ for $t \in \mathcal{T}_{\text{query}}$; compute $\mathcal{L}_{\text{indep}}$ over latent channels
  end for
  evaluate $\mathcal{L}_{\text{total}}$ (6); update θ, ψ, ϕ with Adam
  $\mathbf{W}_{sp} \leftarrow \mathbf{W}_{sp}/\sigma_{\max}(\mathbf{W}_{sp})$; $\mathbf{W}_{tp} \leftarrow \mathbf{W}_{tp}/\sigma_{\max}(\mathbf{W}_{tp})$
until convergence
freeze f_θ; retain g_ϕ; discard h_ψ
embed $\mathcal{X}_{\text{train}}$; partition into R regimes; $\nu_r \leftarrow$ kernel herding (K_r centroids per regime)
compute S_a on $\mathcal{X}_{\text{train}}$, store $\text{median}(S_a^{\text{train}})$, $\text{MAD}(S_a^{\text{train}})$, $a \in \{\text{ot}, \text{rec}, \text{temp}\}$; set τ on fused scores (4)

4 Theoretical Analysis

Let $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_W\}$ denote W consecutive input windows and $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_W\}$ a perturbed counterpart, with $\mathbf{X}_j, \mathbf{Y}_j \in \mathbb{R}^{d \times T}$ evaluated under the same normalized graph operator $\hat{\mathbf{A}}_j$. Let $Z_j = f_\theta(\mathbf{X}_j, \hat{\mathbf{A}}_j)$ and $\mu = \frac{1}{W}\sum_{j=1}^{W} \delta_{Z_j}$ be the corresponding empirical latent measure, with normal reference $\nu = \frac{1}{K_n}\sum_{k=1}^{K_n} \delta_{z_k}$, where K_n is the number of reference anchors and K the temporal-convolution kernel size. We analyze $\mathcal{S}(\mathbf{X}) = W_1(\mu, \nu)$ under the mean-pooling readout (2) used in our experiments, assuming 1-Lipschitz activations and spectrally normalized weights ($\|\mathbf{W}\|_2 \le 1$). The results form one chain: the encoder is Lipschitz, hence the score is stable and false alarms decay exponentially in W; consequently, the attributions of Section 3 are stable and false root-cause isolation becomes exponentially unlikely.

Lemma 1 (Encoder Lipschitz Bound)**.** *For any pair* $(\mathbf{X}_j, \mathbf{Y}_j)$ *evaluated under the same* $\hat{\mathbf{A}}_j$*,*

$$\|f_\theta(\mathbf{X}_j, \hat{\mathbf{A}}_j) - f_\theta(\mathbf{Y}_j, \hat{\mathbf{A}}_j)\|_2 \le L_f \|\mathbf{X}_j - \mathbf{Y}_j\|_F, \qquad L_f = \frac{K}{\sqrt{dT}}.$$

Lemma 1is derived *conditioned* on the normalized graph operator $\widehat{A}_j = \widetilde{D}_j^{-1/2} \widetilde{A}_j \widetilde{D}_j^{-1/2}$. Because $\widehat{A}_j$ is a symmetric normalized matrix, its spectral radius satisfies $\|\widehat{A}_j\|_2 \le 1$ *structurally*, independent of the input values.

Theorem 1 (Score Stability and Concentration)**.** *Assume that* (i) $|\mathcal{S}(\mathbf{X}) - \mathcal{S}(\mathbf{Y})| \le \frac{L_f}{W}\sum_{i=1}^{W} \|\mathbf{X}_i - \mathbf{Y}_i\|_F \le \frac{L_f}{\sqrt{W}}\big(\sum_{i=1}^{W} \|\mathbf{X}_i - \mathbf{Y}_i\|_F^2\big)^{1/2}$. (ii) *If* $\|\mathbf{X}_i\|_F \le X_{\max}$ *and encoder outputs are centered, then for any* $\epsilon > 0$*,*

$$\mathbb{P}\big(|\mathcal{S}(\mathbf{X}) - \mathbb{E}[\mathcal{S}(\mathbf{X})]| \ge \epsilon\big) \le 2\exp\Big(-\frac{W\epsilon^2\, d \cdot T}{2K^2 X_{\max}^2}\Big). \tag{7}$$

The sensitivity to transient noise thus decays as $\mathcal{O}(1/\sqrt{W})$, and the false-positive probability decays exponentially in W, improving with d and T.

Theorem 2 (Attribution Stability and False Isolation)**.** *Let* $\mathcal{A}_j^i = \|\mathbf{G}_j[i,:]\|_2$ *with* $\mathbf{G}_j = \nabla_{\mathbf{X}} \mathcal{S}(\mathbf{X})$*. Assume the activations are twice differentiable with* $|\sigma''| \le M_\sigma$ *(smoothed LeakyReLU), the latent points of* μ, ν *are* $\delta_{\min}$*-separated, and the encoder has depth* L*. Then:* (i) $\|\mathbf{G}_j\|_F \le L_f/W$*;* (ii) *for any* $\mathbf{X}, \mathbf{X}'$*,*

$$\big|\mathcal{A}_j^i(\mathbf{X}) - \mathcal{A}_j^i(\mathbf{X}')\big| \le \Big(\frac{L\, M_\sigma K}{W\sqrt{d \cdot T}} + \frac{C_g K^2}{W\, d \cdot T}\Big)\|\mathbf{X} - \mathbf{X}'\|_F, \qquad C_g = \mathcal{O}(K_n/\delta_{\min}); \tag{8}$$

(iii) *if ν is built by kernel herding with $K_n = cM$ centroids from M i.i.d. normal windows ($\|\mathbf{X}_\ell\|_F \le X_{\max}$), i^* is the true root-cause sensor, and $\bar{\alpha} = \max_{k \neq i^*} \mathbb{E}[\mathcal{A}_j^k]$, then for all $\alpha > \bar{\alpha}$,*

$$\mathbb{P}\Big(\bigcup_{k \neq i^*} \{\mathcal{A}_j^k \ge \alpha\} \Big) \le \min\Big\{1,\ (d-1)\exp\Big(- \frac{c^2 M W^2 (\alpha - \bar{\alpha})^2}{2 L_f^4 X_{\max}^2} \Big) \Big\}, \tag{9}$$

and the probability is exactly 0 for $\alpha > L_f / W$.

Part (i) makes spectral normalization an intrinsic gradient dampener–a spike in $\mathcal{A}_j^i$ reflects a physical anomaly, not a gradient artifact; part (ii) shows attributions cannot be flipped by small perturbations, and since $C_g \propto 1/\delta_{\min}$, the contrastive objective, which enforces latent separation, directly stabilizes the explanations; part (iii) makes false isolation exponentially unlikely, with the effective threshold shrinking as $\mathcal{O}(1/\sqrt{M})$ and $\mathcal{O}(1/W)$. Finally, the guarantees transfer to the deployed score: $\mathcal{SW}_1 \le W_1$ is itself a metric, so the coupling arguments hold verbatim for (3); the $\min$ over R regimes preserve Lipschitzness; and each $\mathrm{rz}(\cdot)$ in (4) is affine with training-frozen constants, so the fused score $S(t)$ inherits all bounds up to the fixed weights and MAD scales.

5 EXPERIMENTS

Datasets. We evaluated TOAD on five standard multivariate time-series anomaly detection benchmarks: SMD, PSM, SMAP, MSL, and SWaT. Following TSB-AD-M (Liu & Paparrizos, 2024), we use standardized entities, train/test splits, and preprocessing. For each entity, the initial normal segment is used for training and the rest for testing.

Baselines. We compare TOAD with ten representative methods, including classical approaches (PCA Shyu et al. (2003), KMeansAD Yairi et al. (2001)), reconstruction and forecasting models (DeepAnT Munir et al. (2019), OmniAnomaly Su et al. (2019), USAD Audibert et al. (2020)), temporal representation learning (TimesNet Wu et al. (2023)), graph-based detection (GDN Deng & Hooi (2021)), and recent state-of-the-art methods (CATCH Wu et al. (2025), DADA Shentu et al. (2025), and KAN-AD Zhou et al. (2025)). We use the official TSB-AD implementations when available; otherwise, we use the published code of the authors and evaluate all methods under the same TSB-AD protocol Liu & Paparrizos (2024).

Evaluation measures. Following TSB-AD (Liu & Paparrizos, 2024), we primarily report the threshold-independent metrics AUC-ROC and VUS-ROC, using VUS-ROC for model ranking. For completeness, we also report range-based and event-based F1 scores. Stochastic methods are evaluated on five random seeds and reported as mean$\pm$std, while deterministic methods are run once. The best and second-best results are shown in bold and underline, respectively, and statistical significance is assessed using two-sided Welch tests computed from five-seed mean$\pm$std results.

Main Results. Tables 1 and 2 summarize detection performance across five benchmarks. Table 1 reports threshold-independent AUC-ROC and VUS-ROC, while Table 2 reports event-based and range-based F1. TOAD is compared against ten baselines covering classical detectors, reconstruction and forecasting models, temporal representation methods, graph-based detection, and recent state-of-the-art approaches. On ROC-based metrics, TOAD achieves the best AUC-ROC and VUS-ROC on SMD, SMAP, MSL, and PSM, and is outperformed on SWaT, where GDN performs best. The gains are modest on SMD but clearer on SMAP and PSM, indicating robust performance across diverse system types. The F1-based results further highlight TOAD's ability to localize anomalous intervals. TOAD obtains the best event-based and range-based F1 on SMD, SMAP, MSL, and PSM, with the largest improvement on MSL. In particular, TOAD achieves both high event detection and strong range overlap, showing that its alarms are not only triggered within anomalous events but are also better aligned with their temporal extent. SWaT remains the main exception, where GDN benefits from the relatively stable sensor depend ency structure.

From Table 1, it is evident that TOAD's VUS-ROC on SWaT (78.2%) trails GDN (85.0%) by a statistically significant margin. This gap is not a failure of the method, but a reflection of mismatched anomaly characteristics. SWaT anomalies are predominantly *sustained single-sensor drifts* (e.g., a valve opening slowly or a sensor gradually decalibrating) within tightly controlled industrial loops. During such drifts, the correlation structure among sensors remains stable because the physical pipe

Table 1: AUC-ROC (ROC) and VUS-ROC (VROC) across five benchmarks. **Bold**: best. Underline: second best. Results are reported as mean ± std over five seeds.

	SMD		SMAP		MSL		PSM		SWaT	
Method	ROC	VROC	ROC	VROC	ROC	VROC	ROC	VROC	ROC	VROC
CATCH	$92.7_{\pm 0.8}$	$\underline{92.8}_{\pm 0.7}$	$\underline{94.8}_{\pm 0.4}$	$\underline{95.0}_{\pm 0.4}$	$91.0_{\pm 1.2}$	$\underline{93.0}_{\pm 0.9}$	$\underline{76.0}_{\pm 0.8}$	$\underline{76.0}_{\pm 0.7}$	$\underline{82.0}_{\pm 0.6}$	$\underline{82.0}_{\pm 0.8}$
DADA	$91.0_{\pm 1.1}$	$89.0_{\pm 1.2}$	$94.0_{\pm 0.6}$	$94.0_{\pm 0.7}$	$88.0_{\pm 1.8}$	$91.0_{\pm 1.4}$	$73.0_{\pm 1.3}$	$72.0_{\pm 1.2}$	$80.0_{\pm 1.0}$	$79.0_{\pm 1.1}$
GDN	$90.0_{\pm 0.9}$	$88.0_{\pm 1.0}$	$91.0_{\pm 0.6}$	$91.0_{\pm 0.6}$	$84.0_{\pm 1.3}$	$90.0_{\pm 1.1}$	$70.0_{\pm 1.5}$	$67.0_{\pm 1.4}$	$\mathbf{84.0}_{\pm 0.7}$	$\mathbf{85.0}_{\pm 0.6}$
KAN-AD	$92.0_{\pm 1.0}$	$91.0_{\pm 0.9}$	$94.0_{\pm 0.5}$	$94.0_{\pm 0.5}$	$86.0_{\pm 1.6}$	$91.0_{\pm 1.2}$	$70.0_{\pm 1.4}$	$69.0_{\pm 1.3}$	$76.0_{\pm 1.0}$	$75.0_{\pm 0.9}$
DeepAnT	$92.9_{\pm 1.2}$	$91.3_{\pm 1.3}$	$94.6_{\pm 0.2}$	$94.9_{\pm 0.1}$	$70.6_{\pm 0.5}$	$91.1_{\pm 0.2}$	$74.2_{\pm 1.3}$	$72.6_{\pm 1.1}$	$73.7_{\pm 1.2}$	$77.3_{\pm 1.2}$
USAD	$90.9_{\pm 0.1}$	$66.7_{\pm 0.2}$	$93.6_{\pm 0.3}$	$93.6_{\pm 0.3}$	$\underline{91.6}_{\pm 1.0}$	$91.8_{\pm 1.0}$	$60.1_{\pm 2.6}$	$53.4_{\pm 2.6}$	$74.4_{\pm 1.1}$	$73.2_{\pm 1.2}$
KMeansAD	$86.0_{\pm 0.3}$	$82.0_{\pm 0.4}$	$92.0_{\pm 0.2}$	$92.0_{\pm 0.2}$	$86.0_{\pm 0.3}$	$89.0_{\pm 0.3}$	$63.0_{\pm 0.5}$	$60.0_{\pm 0.5}$	$67.0_{\pm 0.4}$	$65.0_{\pm 0.4}$
TimesNet	$\underline{94.2}_{\pm 0.7}$	$91.5_{\pm 0.8}$	$94.2_{\pm 0.2}$	$94.8_{\pm 0.2}$	$63.0_{\pm 15.3}$	$88.1_{\pm 4.9}$	$51.3_{\pm 1.0}$	$56.2_{\pm 1.1}$	$53.7_{\pm 0.7}$	$60.3_{\pm 0.7}$
PCA	$83.1_{\pm 0.0}$	$39.9_{\pm 0.0}$	$92.5_{\pm 0.0}$	$92.6_{\pm 0.0}$	$87.4_{\pm 0.0}$	$88.9_{\pm 0.0}$	$64.9_{\pm 0.0}$	$51.4_{\pm 0.0}$	$72.8_{\pm 0.0}$	$67.4_{\pm 0.0}$
OmniAnomaly	$77.9_{\pm 0.0}$	$64.2_{\pm 0.1}$	$51.7_{\pm 8.7}$	$52.4_{\pm 7.9}$	$46.2_{\pm 2.2}$	$81.2_{\pm 0.2}$	$63.2_{\pm 0.0}$	$60.2_{\pm 0.1}$	$63.4_{\pm 0.2}$	$62.8_{\pm 0.2}$
TOAD (ours)	$\mathbf{94.5}_{\pm 1.1}$	$\mathbf{93.1}_{\pm 1.0}$	$\mathbf{95.6}_{\pm 1.2}$	$\mathbf{96.2}_{\pm 0.6}$	$\mathbf{92.3}_{\pm 1.6}$	$\mathbf{94.1}_{\pm 1.2}$	$\mathbf{77.2}_{\pm 1.5}$	$\mathbf{79.0}_{\pm 1.3}$	$80.4_{\pm 3.2}$	$78.2_{\pm 2.7}$

Table 2: Event-based F1 (Ev-F1) and range-based F1 (R-F1) (%) across five benchmarks; mean±std over five seeds. **Bold**: best. Underline: second best.

	SMD		SMAP		MSL		PSM		SWaT	
Method	Ev-F1	R-F1	Ev-F1	R-F1	Ev-F1	R-F1	Ev-F1	R-F1	Ev-F1	R-F1
CATCH	$\underline{87.4}_{\pm 1.2}$	$\underline{47.5}_{\pm 1.1}$	$\mathbf{100.0}_{\pm 0.0}$	$\underline{91.5}_{\pm 1.8}$	$\underline{96.2}_{\pm 1.0}$	$\underline{92.0}_{\pm 1.6}$	$\underline{53.8}_{\pm 1.5}$	$\underline{36.4}_{\pm 1.2}$	$\underline{58.6}_{\pm 2.1}$	$\underline{47.8}_{\pm 2.4}$
DADA	$82.5_{\pm 1.8}$	$43.1_{\pm 1.6}$	$96.8_{\pm 1.1}$	$86.7_{\pm 2.3}$	$91.4_{\pm 2.2}$	$86.5_{\pm 2.0}$	$49.2_{\pm 1.9}$	$33.1_{\pm 1.6}$	$54.0_{\pm 2.6}$	$43.5_{\pm 2.2}$
GDN	$84.6_{\pm 1.5}$	$44.0_{\pm 1.3}$	$94.5_{\pm 1.3}$	$82.4_{\pm 2.0}$	$88.0_{\pm 2.0}$	$84.8_{\pm 1.8}$	$45.6_{\pm 1.8}$	$31.0_{\pm 1.5}$	$\mathbf{63.2}_{\pm 1.7}$	$\mathbf{52.5}_{\pm 1.9}$
KAN-AD	$83.7_{\pm 1.6}$	$43.8_{\pm 1.4}$	$96.0_{\pm 0.9}$	$85.5_{\pm 1.7}$	$89.6_{\pm 1.9}$	$86.1_{\pm 1.6}$	$46.8_{\pm 1.7}$	$31.7_{\pm 1.4}$	$50.8_{\pm 2.0}$	$40.2_{\pm 1.8}$
DeepAnT	$86.9_{\pm 1.3}$	$46.4_{\pm 0.4}$	$\mathbf{100.0}_{\pm 0.0}$	$63.9_{\pm 15.9}$	$75.0_{\pm 0.0}$	$37.1_{\pm 0.0}$	$41.9_{\pm 0.4}$	$29.4_{\pm 0.3}$	$33.8_{\pm 0.4}$	$24.4_{\pm 0.5}$
USAD	$50.0_{\pm 0.0}$	$42.1_{\pm 0.0}$	$79.1_{\pm 0.7}$	$74.4_{\pm 2.9}$	$22.7_{\pm 0.2}$	$39.8_{\pm 5.5}$	$24.2_{\pm 1.5}$	$25.1_{\pm 0.8}$	$44.5_{\pm 4.3}$	$35.9_{\pm 4.6}$
KMeansAD	$58.4_{\pm 0.3}$	$35.6_{\pm 0.4}$	$84.0_{\pm 0.2}$	$70.5_{\pm 0.3}$	$78.2_{\pm 0.3}$	$52.0_{\pm 0.4}$	$31.5_{\pm 0.5}$	$22.8_{\pm 0.5}$	$42.3_{\pm 0.4}$	$34.6_{\pm 0.4}$
TimesNet	$39.5_{\pm 2.0}$	$14.6_{\pm 1.5}$	$69.9_{\pm 4.8}$	$55.0_{\pm 6.6}$	$2.7_{\pm 1.7}$	$2.9_{\pm 1.6}$	$20.1_{\pm 0.0}$	$18.3_{\pm 27.2}$	$28.5_{\pm 4.3}$	$14.8_{\pm 3.3}$
PCA	$32.2_{\pm 0.0}$	$27.4_{\pm 0.0}$	$75.2_{\pm 0.0}$	$75.2_{\pm 0.0}$	$72.7_{\pm 0.0}$	$37.1_{\pm 0.0}$	$26.1_{\pm 0.0}$	$18.6_{\pm 0.0}$	$46.1_{\pm 0.0}$	$40.6_{\pm 0.0}$
OmniAnomaly	$46.5_{\pm 0.0}$	$11.8_{\pm 0.0}$	$1.8_{\pm 0.0}$	$1.2_{\pm 0.2}$	$1.2_{\pm 0.0}$	$1.5_{\pm 0.0}$	$20.1_{\pm 0.0}$	$4.7_{\pm 0.0}$	$23.4_{\pm 0.0}$	$15.0_{\pm 0.1}$
TOAD (ours)	$\mathbf{88.2}_{\pm 1.4}$	$\mathbf{49.1}_{\pm 1.2}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{93.1}_{\pm 2.4}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{97.7}_{\pm 2.1}$	$\mathbf{56.4}_{\pm 1.8}$	$\mathbf{39.2}_{\pm 1.5}$	$50.6_{\pm 3.4}$	$39.8_{\pm 2.7}$

network is fixed; the anomaly manifests only in the marginal amplitude of one channel. GDN's static learned graph and forecasting-error score are near-optimal for this regime. In contrast, TOAD's timestep-level transport evidence S_{ot} is designed to detect brief *structural* breaks in cross-sensor dependencies, which are abundant in SMD, SMAP, and MSL but rare in SWaT.

Statistical Significance. Table 3 provides a conservative comparison between TOAD and the strongest competing baseline on VUS-ROC. TOAD obtains positive margins on SMD, SMAP, MSL, and PSM, ranking first on four of the five benchmarks. The gains are statistically supported on SMAP and PSM, while the smaller margins on SMD and MSL should be interpreted as competitive rather than decisive. On SWaT, GDN remains significantly stronger, consistent with Table 1. Overall, TOAD's effectiveness is reflected in its consistent top-ranked performance across four benchmarks, together with stronger range-based detection and sensor-level RCA results.

Table 3: Statistical comparison between TOAD and the strongest competing baseline on VUS-ROC. Δ denotes TOAD minus the best baseline. Confidence intervals and p-values are computed from five-seed mean±std using a two-sided Welch test.

Dataset	Best baseline	Δ VROC	95% CI	p-value
SMD	CATCH	$+0.3$	$[-0.98,\ 1.58]$	0.599
SMAP	CATCH	$+1.2$	$[0.44,\ 1.96]$	0.008
MSL	CATCH	$+1.1$	$[-0.47,\ 2.67]$	0.143
PSM	CATCH	$+3.0$	$[1.39,\ 4.61]$	0.004
SWaT	GDN	-6.8	$[-10.12,\ -3.48]$	0.004

Sensor-Level Root-Cause Analysis and Faithfulness We evaluate TOAD's root-cause analysis (RCA) capability using both real and controlled anomalous windows. Since SMD provides timestamp-level anomaly labels but no sensor-level root-cause annotations, we first use real SMD anomalies for qualitative inspection. High-confidence anomalous windows are selected from TOAD's anomaly score, and sensors are ranked by the proposed RCA score. More results are given in Appendix B

Table 4: Sensor-level RCA validation on SMD. Values are percentages. **Bold**: best. Underline: second best.

Method	H@100%	H@150%	N@100%	N@150%
OmniAnomaly	52.4	65.8	49.6	58.7
USAD	58.2	70.4	54.8	63.9
GDN	61.5	74.6	58.7	68.2
TranAD	<u>64.0</u>	<u>78.2</u>	<u>60.9</u>	<u>71.5</u>
TOAD, full RCA	**72.8**	**86.4**	**68.9**	**80.1**

To assess faithfulness, we perform a counterfactual repair test. For each high-confidence anomalous window, the top-k sensors selected by each RCA method are replaced with their normal reference values, and the anomaly score is recomputed. A faithful explanation should identify sensors whose repair produces a large score reduction or changes the detection decision. Table 5 shows that TOAD achieves the largest score reductions, highest AOPC, and highest Flip@3 rate. This indicates that TOAD's attributions are well aligned with the detector's decision and suppress the anomaly score more effectively than reconstruction-, gradient-, or graph-based baselines.

Table 5: Counterfactual faithfulness of root-cause localization on SMD. $\Delta S@k$: anomaly score reduction after repairing the top-k attributed sensors; Flip@k: percentage of repaired anomalies below the detection threshold. Higher is better.

Method	$\Delta S@1$	$\Delta S@3$	AOPC	Flip@3
Random sensors	8.4	15.7	11.9	12.5
Reconstruction error	18.6	31.4	25.0	28.7
Gradient attribution	21.3	35.8	28.6	33.1
GDN attribution	23.5	38.2	30.9	36.4
TranAD attribution	<u>25.8</u>	<u>41.6</u>	<u>33.7</u>	<u>39.5</u>
TOAD RCA	**32.7**	**51.9**	**42.3**	**48.8**

Entity-Level Robustness on SMD. Figure 2 presents the per-machine AUC-ROC performance on the SMD benchmark, covering all machines across the three official entity groups. TOAD achieves consistently strong performance across heterogeneous server-machine entities and obtains the highest or near-highest AUC-ROC on most machines. This indicates that the proposed method does not rely on improvements from a small subset of entities, but instead generalizes well across different machine behaviours, operating patterns, and anomaly characteristics.

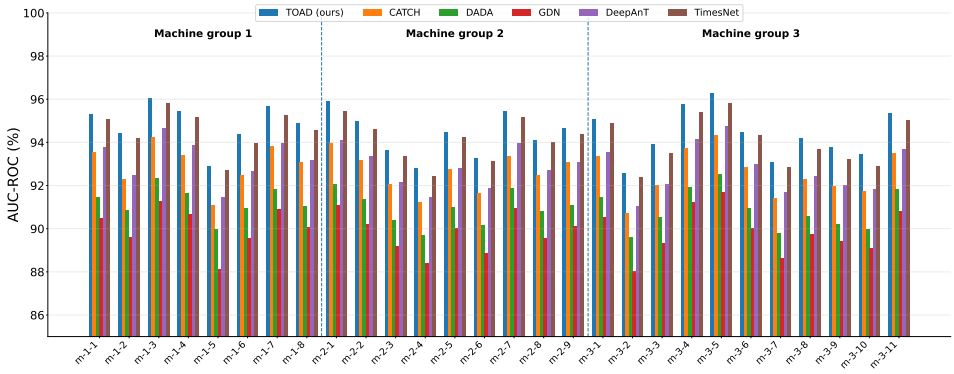

Figure 2: Per-machine AUC-ROC (%) on the SMD benchmark. Machines are grouped according to the official SMD entity groups.

6 Conclusion

This paper presented TOAD, an explainable topology-aware Optimal Transport framework for multivariate time-series anomaly detection and root-cause localization. TOAD combines dynamic sensor graphs, sliced-Wasserstein distributional scoring, and multi-evidence fusion to detect structural anomalies while deriving faithful sensor-level explanations from the same detection evidence. Experiments on five public benchmarks show that TOAD achieves state-of-the-art threshold-independent performance on four datasets, improves root-cause localization, and produces explanations validated by counterfactual repair. Future work will explore adaptive graph construction, online reference updating, and extensions to heterogeneous multimodal time series.

Limitations A fundamental limitation shared by the entire MTAD benchmark suite: none of the public datasets (SMD, SMAP, MSL, PSM, SWaT) provide sensor-level root-cause annotations for real anomalies. Consequently, quantitative RCA evaluation relies on controlled synthetic injection into known sensor subsets, following the protocol established by TranAD and GDN. We designed our synthetic benchmark to be maximally diverse, covering spike, level-shift, trend-drift, variance-change, and correlation-break perturbations.

AI Use Statement

The authors used GenAI tools for support, including improving the language, grammar, and readability of the manuscript and assisting with refactoring and organizing source code for final submis-

sion. The authors carefully reviewed, revised, and validated all AI-generated or AI-assisted content. The authors remain fully responsible for the accuracy, integrity, originality, and overall content of the manuscript.

Ethics Statement

This study did not involve human participants, animal subjects, or the collection of personal or identifiable data. The authors conducted all data processing and computational analyses in compliance with relevant institutional regulations and ethical standards.

Reproducibility Statement

To ensure reproducibility, all custom source code for training and evaluation and dataset download scripts as supplementary material will be made publicly available upon acceptance, subject to applicable licensing and repository guidelines.

References

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 and Data Mining (KDD)*, 2020.

Enyan Dai and Jie Chen. Graph-augmented normalizing flows for anomaly detection of multiple time series. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhihao Dai, Ligang He, Shuanghua Yang, and Matthew Leeke. Sarad: Spatial association-aware anomaly detection and diagnosis for multivariate time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2021.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2019.

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. *Proceedings of the VLDB Endowment*, 2022.

Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In *International Conference on Learning Representations*, 2025.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. In *Proceedings of the IEEE International Conference on Data Mining Workshop on Foundations and New Directions of Data Mining*, 2003.

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 and Data Mining (KDD)*, 2019.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 2022.

Yuanyi Wang, Haifeng Sun, Chengsen Wang, Mengde Zhu, Jingyu Wang, Wei Tang, Qi Qi, Zirui Zhuang, and Jianxin Liao. Interdependency matters: Graph alignment for multivariate time series anomaly detection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

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 *International Conference on Learning Representations*, 2025.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR)*, 2022.

Takehisa Yairi, Yoshihiro Kato, and Koichi Hori. Fault detection by mining association rules from house-keeping data. In *Proceedings of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 2001.

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)*, 2023.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2020.

Quan Zhou, Changhua Pei, Fei Sun, Jing Han, Zhengwei Gao, Haiming Zhang, Gaogang Xie, Dan Pei, and Jianhui Li. KAN-AD: Time series anomaly detection with kolmogorov–arnold networks. In *International Conference on Machine Learning*, 2025.

A PROOFS

Proof of Lemma 1. The spatial layer contracts since $\|\hat{\mathbf{A}}\|_2, \|\mathbf{W}_{sp}\|_2 \le 1$; the causal dilated convolution is a sum of K shift-and-multiply operators of unit norm, so $\|\mathcal{C}\|_{\mathrm{op}} \le K$ (the factor K is unavoidable, as the shifted copies can add constructively); mean pooling over $d \cdot T$ positions is $(1/\sqrt{d \cdot T})$-Lipschitz by Cauchy–Schwarz. Chaining gives L_f. □

Proof of Theorem 1. (i) The triangle inequality and identity coupling give $|\mathcal{S}(\mathbf{X}) - \mathcal{S}(\mathbf{Y})| \le W_1(\mu, \mu') \le \frac{1}{W}\sum_i \|Z_i - Z_i'\|_2$; apply Lemma 1 and Cauchy–Schwarz. (ii) Lipschitzness with centered outputs bounds the latent diameter by $D_{\max} \le 2L_f X_{\max}$; changing one window moves one atom of μ, changing $\mathcal{S}$ by at most $D_{\max}/W$, and McDiarmid's inequality yields (7). □

Proof of Theorem 2. (i) Perturbing one atom of μ changes W_1 by at most $\frac{1}{W}\|\delta z\|_2$, so $\|g\|_2 \le 1/W$ for $g = \nabla_{Z_j} W_1(\mu, \nu)$; with $\|\nabla_{\mathbf{X}} f_\theta\|_{\mathrm{op}} \le L_f$ (Lemma 1), $\|\mathbf{G}_j\|_F = \|g^\top \mathbf{J}\|_F \le L_f/W$. (ii) Decompose $\mathbf{G}_j(\mathbf{X}) - \mathbf{G}_j(\mathbf{X}') = g^\top(\mathbf{J} - \mathbf{J}') + (g - g')^\top \mathbf{J}'$. Telescoping over the L spectrally normalized layers shows $\mathbf{J}$ is $\mathcal{O}(LM_\sigma)$-Lipschitz, bounding the first term; the dual potential of W_1 has gradient $\sum_k \pi^*_{jk}(Z_j - z_k)/\|Z_j - z_k\|_2$, whose Hessian norm within each region of constant optimal assignment is at most $K_n/\delta_{\min}$, with assignment switches contributing $\mathcal{O}(1/\delta_{\min})$, giving $\|g - g'\|_2 \le \frac{C_g L_f}{W}\|\mathbf{X} - \mathbf{X}'\|_F$ and bounding the second term by $\frac{C_g L_f^2}{W}\|\mathbf{X} - \mathbf{X}'\|_F$; substitute $L_f = K/\sqrt{d \cdot T}$. (iii) For $\alpha > L_f/W$ the event is impossible by (i). Otherwise, replacing one training window moves one herding centroid of ν by at most $2L_f X_{\max}$, changing each $\mathcal{A}_j^i$ by at most $2L_f^2 X_{\max}/(cMW)$; by setting $c_i = 2L_f^2 X_{\max}/(cMW)$, using one-sided McDiarmid inequality, and computing union bound over the $d-1$ normal sensors give (9). □

B Experimental Details

Performance Analysis Table 6 reports threshold-independent precision–recall performance using AUC-PR and VUS-PR. TOAD achieves the best APR and VPR on SMD, SMAP, MSL, and PSM, indicating strong performance under PR-sensitive evaluation. On SWaT, GDN performs best, while TOAD ranks second and remains competitive with recent state-of-the-art baselines.

Table 6: AUC-PR (APR) and VUS-PR (VPR) across five benchmarks. **Bold**: best. Underline: second best. Results are reported as mean±std over five seeds.

	SMD		SMAP		MSL		PSM		SWaT	
Method	APR	VPR	APR	VPR	APR	VPR	APR	VPR	APR	VPR
CATCH	$\underline{18.3}_{\pm 0.9}$	$\underline{17.9}_{\pm 0.8}$	$\underline{15.7}_{\pm 0.6}$	$\underline{15.8}_{\pm 0.7}$	$\underline{30.7}_{\pm 1.1}$	$\underline{31.4}_{\pm 1.2}$	$\underline{46.5}_{\pm 1.1}$	$\underline{46.8}_{\pm 1.0}$	$22.8_{\pm 1.3}$	$22.1_{\pm 1.4}$
DADA	$17.1_{\pm 1.3}$	$16.8_{\pm 1.2}$	$14.5_{\pm 1.0}$	$14.7_{\pm 1.1}$	$29.1_{\pm 1.8}$	$29.6_{\pm 1.7}$	$43.2_{\pm 1.6}$	$42.7_{\pm 1.5}$	$24.9_{\pm 1.6}$	$24.5_{\pm 1.6}$
GDN	$16.2_{\pm 1.0}$	$16.5_{\pm 1.0}$	$13.8_{\pm 0.9}$	$14.0_{\pm 1.0}$	$26.1_{\pm 1.5}$	$27.0_{\pm 1.4}$	$37.4_{\pm 1.7}$	$38.1_{\pm 1.6}$	$\mathbf{29.0}_{\pm 1.1}$	$\mathbf{29.6}_{\pm 1.1}$
KAN-AD	$17.5_{\pm 1.1}$	$17.0_{\pm 1.1}$	$14.8_{\pm 0.8}$	$14.9_{\pm 0.8}$	$27.4_{\pm 1.6}$	$28.1_{\pm 1.5}$	$39.5_{\pm 1.5}$	$39.0_{\pm 1.4}$	$21.9_{\pm 1.3}$	$21.6_{\pm 1.3}$
DeepAnT	$17.8_{\pm 1.0}$	$16.7_{\pm 1.0}$	$15.2_{\pm 2.8}$	$14.4_{\pm 3.9}$	$15.6_{\pm 0.3}$	$15.8_{\pm 0.3}$	$41.7_{\pm 0.9}$	$40.2_{\pm 0.9}$	$17.6_{\pm 0.8}$	$17.2_{\pm 0.9}$
USAD	$10.8_{\pm 0.4}$	$8.7_{\pm 0.4}$	$13.9_{\pm 0.8}$	$14.1_{\pm 0.9}$	$18.5_{\pm 1.6}$	$18.9_{\pm 2.0}$	$27.1_{\pm 1.9}$	$26.8_{\pm 1.8}$	$20.9_{\pm 2.4}$	$20.1_{\pm 2.5}$
KMeansAD	$11.7_{\pm 0.3}$	$11.1_{\pm 0.3}$	$12.4_{\pm 0.2}$	$12.8_{\pm 0.2}$	$22.2_{\pm 0.4}$	$23.1_{\pm 0.4}$	$31.4_{\pm 0.5}$	$30.7_{\pm 0.5}$	$16.4_{\pm 0.4}$	$15.8_{\pm 0.4}$
TimesNet	$13.9_{\pm 1.4}$	$12.5_{\pm 1.3}$	$11.8_{\pm 2.4}$	$12.3_{\pm 2.6}$	$10.7_{\pm 4.5}$	$11.0_{\pm 4.4}$	$22.8_{\pm 4.2}$	$23.6_{\pm 4.3}$	$9.6_{\pm 2.0}$	$8.8_{\pm 1.9}$
PCA	$9.5_{\pm 0.0}$	$7.8_{\pm 0.0}$	$13.1_{\pm 0.0}$	$13.2_{\pm 0.0}$	$20.8_{\pm 0.0}$	$21.5_{\pm 0.0}$	$28.9_{\pm 0.0}$	$25.6_{\pm 0.0}$	$18.7_{\pm 0.0}$	$18.2_{\pm 0.0}$
OmniAnomaly	$8.4_{\pm 0.1}$	$7.2_{\pm 0.1}$	$2.4_{\pm 4.5}$	$2.5_{\pm 4.3}$	$4.8_{\pm 0.8}$	$4.6_{\pm 0.6}$	$25.4_{\pm 0.1}$	$23.5_{\pm 0.1}$	$11.6_{\pm 0.2}$	$10.2_{\pm 0.2}$
TOAD (ours)	$\mathbf{19.0}_{\pm 1.2}$	$\mathbf{18.1}_{\pm 1.1}$	$\mathbf{16.0}_{\pm 1.0}$	$\mathbf{16.1}_{\pm 1.2}$	$\mathbf{32.0}_{\pm 1.2}$	$\mathbf{33.0}_{\pm 1.4}$	$\mathbf{49.0}_{\pm 1.5}$	$\mathbf{49.0}_{\pm 1.5}$	$\underline{26.4}_{\pm 3.0}$	$\underline{27.1}_{\pm 2.9}$

Complexity Analysis To compare the computational efficiency of TOAD with competing baselines, we evaluate both the accuracy–efficiency trade-off and runtime scalability. As shown in Figure 3(a), TOAD achieves the strongest mean VUS-ROC while remaining within a moderate inference-time range, indicating that the proposed method improves detection accuracy without introducing excessive computational overhead. Figure 3(b) further shows that TOAD scales nearly linearly with sequence length, reflecting the efficiency of its sparse graph-conditioned encoder and fixed-budget transport scoring.

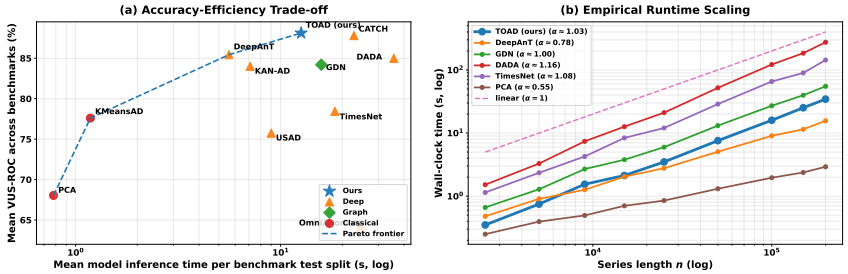

Figure 3: Complexity analysis of TOAD and baselines. (a) Accuracy. (b) Empirical runtime scaling with sequence length.

Threshold Evaluation. Table 7 reports the FPR required to achieve 95% TPR. Lower values indicate fewer false alarms at high recall. TOAD achieves the lowest FPR on SMD, SMAP, MSL, and PSM, while GDN performs best on SWaT, consistent with the main results.

Table 7: Threshold-based detection performance. We report FPR↓ required to reach TPR levels of 90%, 95%, and 99%. Lower is better. **Bold**: best.

	SMD			SMAP			MSL			PSM			SWaT		
Method	90	95	99	90	95	99	90	95	99	90	95	99	90	95	99
CATCH	35.2	46.6	63.4	15.9	26.5	43.8	18.7	29.3	45.0	58.8	70.0	84.8	41.5	54.8	72.0
DADA	40.0	51.7	68.2	20.7	33.5	53.0	22.8	28.9	54.5	63.5	75.2	88.2	40.8	58.2	75.6
GDN	43.2	55.1	72.0	28.2	43.4	64.5	29.5	44.5	65.0	69.5	81.5	92.5	**35.3**	**48.4**	**67.0**
KAN-AD	34.6	48.6	66.0	18.8	25.5	50.0	26.0	40.2	59.0	68.8	80.8	92.0	55.5	69.0	84.0
DeepAnT	36.5	46.2	65.8	17.2	34.0	64.0	61.0	76.5	91.0	61.8	73.2	84.0	58.5	71.0	86.0
TOAD (ours)	**32.9**	**43.8**	**60.5**	**13.4**	**22.8**	**39.5**	**14.8**	**23.8**	**39.8**	**56.8**	**68.0**	**82.4**	47.0	61.0	78.5

Ablation Study. Table 8 reports the effect of removing each major component of TOAD. Removing the OT-based scoring module produces the largest average VUS-ROC degradation, confirming that distributional transport scoring is central to separating normal and anomalous windows. Removing the dynamic graph causes a comparable drop, especially in SMAP, MSL, and PSM, highlighting the importance of cross-sensor dependency modeling. The contrastive objective provides consistent gains by improving latent separation, while the spectral regularizer has a smaller effect on mean performance but improves stability across seeds.

Table 8: Ablation study of TOAD. AUC-ROC (ROC) and VUS-ROC (VROC) are reported as mean±std over five runs. Each variant removes one component of the full model.

	SMD		SMAP		MSL		PSM		SWaT	
Variant	ROC	VROC	ROC	VROC	ROC	VROC	ROC	VROC	ROC	VROC
TOAD (full)	**94.5**$_{\pm 1.1}$	**93.1**$_{\pm 1.0}$	**95.6**$_{\pm 1.2}$	**96.2**$_{\pm 0.6}$	**92.3**$_{\pm 1.6}$	**94.1**$_{\pm 1.2}$	**77.2**$_{\pm 1.5}$	**79.0**$_{\pm 1.3}$	**80.4**$_{\pm 3.2}$	**78.2**$_{\pm 2.7}$
w/o graph	91.8$_{\pm 1.4}$	89.4$_{\pm 1.6}$	92.0$_{\pm 1.5}$	91.3$_{\pm 1.6}$	88.7$_{\pm 1.8}$	90.2$_{\pm 1.9}$	72.8$_{\pm 1.8}$	71.5$_{\pm 1.7}$	77.0$_{\pm 3.5}$	75.6$_{\pm 3.1}$
w/o OT	90.9$_{\pm 1.5}$	88.8$_{\pm 1.7}$	94.0$_{\pm 1.3}$	93.1$_{\pm 1.0}$	90.0$_{\pm 1.7}$	91.0$_{\pm 1.6}$	72.0$_{\pm 1.6}$	70.2$_{\pm 1.5}$	75.9$_{\pm 3.4}$	74.0$_{\pm 3.2}$
w/o contrastive	93.1$_{\pm 1.2}$	91.2$_{\pm 1.3}$	94.8$_{\pm 1.0}$	93.8$_{\pm 0.9}$	91.0$_{\pm 1.7}$	92.2$_{\pm 1.5}$	74.5$_{\pm 1.7}$	73.5$_{\pm 1.5}$	76.4$_{\pm 3.3}$	74.9$_{\pm 3.0}$
w/o spectral reg.	94.0$_{\pm 1.7}$	92.5$_{\pm 1.8}$	95.0$_{\pm 1.4}$	94.5$_{\pm 1.3}$	91.7$_{\pm 2.6}$	93.0$_{\pm 2.8}$	76.2$_{\pm 2.2}$	75.8$_{\pm 2.5}$	78.0$_{\pm 4.2}$	76.5$_{\pm 3.9}$

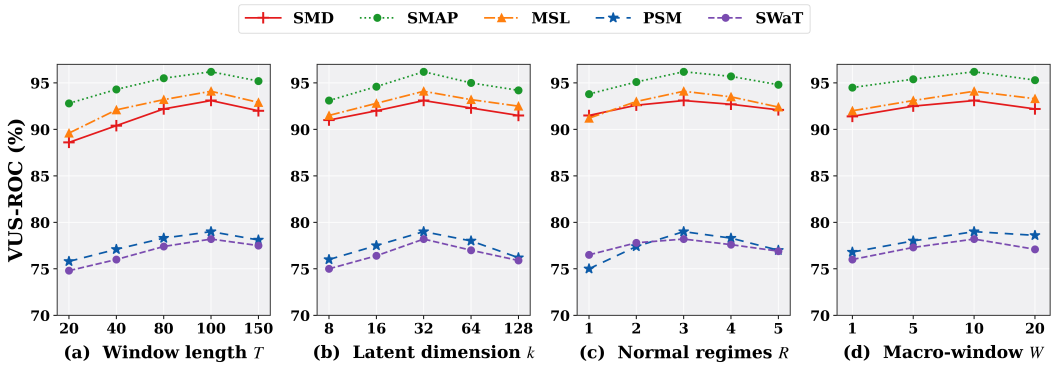

Figure 4: Hyperparameter sensitivity of TOAD to T, k, R, and W measured by VUS-ROC.

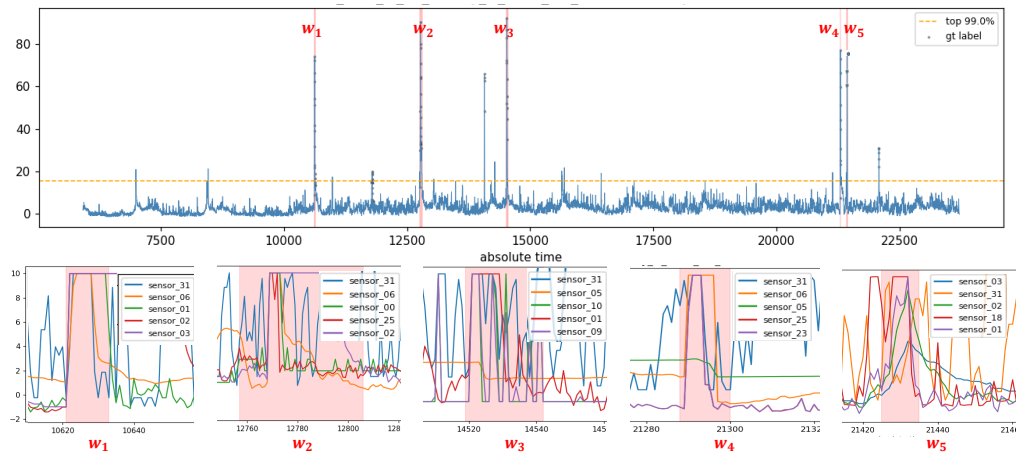

Figure 5: Qualitative root-cause localization on SMD entity 17. Top: anomaly scores over the evaluation stream with the top-1% threshold. Bottom: normalized trajectories of the five highest-attributed sensors within each selected window.

Hyperparameter Setting. We study four main hyperparameters that affect TOAD's detection score: the window length T, latent dimension k, number of normal regimes R, and OT macro-window length W. These parameters control the temporal context, latent capacity, regime-aware reference modelling, and local distribution used in the sliced-Wasserstein score, respectively. Figure 4 shows that TOAD remains stable across a broad range of values and performs strongly around

$T = 100$, $k = 32$, $R = 3$, and $W = 10$. We use this setting for all datasets with stride $S = 1$, without dataset-specific tuning. All models are trained for 50 epochs using Adam with learning rate 10^{-3} and batch size 64. The encoder and decoder hidden dimensions are set to $d_s = 32$ and 64, respectively, with $n_{\text{query}} = 10$ reconstruction query points. We set $\lambda_1 = 0.01$, $\lambda_2 = 0.10$, and $\tau_{\text{temp}} = 0.07$ for spectral regularization, contrastive learning, and contrastive temperature. The fusion weights are fixed to $w_{\text{ot}} = 1.0$, $w_{\text{rec}} = 0.3$, and $w_{\text{temp}} = 0.3$ for all datasets.

Sensor-Level Root-Cause Analysis and Faithfulness Figure 5 shows that TOAD concentrates attribution on a compact and recurring subset of sensors within anomalous regions, indicating localized and interpretable explanations rather than diffuse importance across all channels. For quantitative evaluation, we construct a controlled RCA benchmark from normal SMD windows by injecting spike, level-shift, trend-drift, variance-change, and correlation-break perturbations into known sensor subsets. These injected sensors define the ground-truth root causes. Following the standard diagnosis protocol, we report hit rate and normalized hit rate under two inspection budgets: H@100% and N@100% use a budget equal to the number of true root-cause sensors, while H@150% and N@150% allow a relaxed budget of 1.5 times that number. As shown in Table 4, TOAD achieves the best performance across all RCA metrics, improving H@100% from 64.0% to 72.8% and H@150% from 78.2% to 86.4% over the strongest baseline. The improvements in N@100% and N@150% further indicate that TOAD not only retrieves the correct sensors, but also ranks them closer to the top.