Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Calibrated Residual Modeling for Explainable Anomaly Detection in Multivariate Sensor Streams

**Rui Shi**

shirui1@stu.xidian.edu.cn

Xidian University

**Research Article**

**Keywords:** Time Series Anomaly Detection, Temporal Convolutional Networks, Heteroscedastic Modeling, Extreme Value Calibration, Explainable Artificial Intelligence

**Posted Date:** August 31st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10167314/v1

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

**Additional Declarations:** No competing interests reported.

# Calibrated Residual Modeling for Explainable Anomaly Detection in Multivariate Sensor Streams

Rui Shi[1]

[1]Electronic Information Engineering, School of Electronic Engineering, Xidian University, Xi'an, 710126, Shaanxi, China.

Contributing authors: shirui1@stu.xidian.edu.cn;

**Abstract**

Anomaly detection in multivariate sensor networks remains a core challenge in cyber physical and IT infrastructures because of the complex interplay between temporal nonstationarity, channel specific heteroscedastic noise, and the scarcity of labeled anomalies. These factors often cause deep models such as transformer based and contrastive architectures to produce unstable thresholds, fragmented alarms, and limited interpretability in real world deployments. To address these limitations, this paper proposes a hybrid deep and statistical framework that integrates a temporal convolutional autoencoder for nonlinear sequence representation, a probabilistic residual model for variance normalization, and an extreme value calibration layer for adaptive thresholding. A dual path explanation mechanism further aligns temporal attention and gradient based attribution to ensure that sensor and time relevance are consistent with calibrated anomaly evidence. Comprehensive experiments on six widely used benchmarks, including NASA SMAP and MSL telemetry, server metrics (SMD and PSM), and industrial control datasets (SWaT and WADI), demonstrate that the proposed method consistently outperforms transformer and contrastive baselines. It achieves up to **4.5** percentage point gains in Event F1 and $\boldsymbol{F_{c1}}$, reduces median detection delay by **16–22%**, and improves AUPRC by **1.3–1.7** percentage points on telemetry data. These results indicate that coupling calibrated probabilistic scoring with interpretable temporal representations offers a principled and reliable approach to anomaly detection across heterogeneous sensor networks.

**Keywords:** Time Series Anomaly Detection, Temporal Convolutional Networks, Heteroscedastic Modeling, Extreme Value Calibration, Explainable Artificial Intelligence

# 1 Introduction

Multivariate sensor networks underpin modern cyber–physical and IT infrastructures, yet their scale, heterogeneity, and nonstationary behavior make anomaly detection a persistent challenge [1, 2]. Empirically, label scarcity and domain drift hinder supervised pipelines, while purely reconstruction- or forecasting-based detectors may overfit frequent high-variance channels and yield unstable operating points across deployments. These limitations motivate models that (i) disentangle systematic dynamics from irregular deviations, (ii) calibrate uncertainty in a data-aware manner, and (iii) expose *why* alarms are raised so that operators can triage events instead of isolated points.

Recent deep architectures have advanced the state of the art by leveraging sequence attention, association discrepancy, and contrastive cues [3–7]. While they excel at point-wise ranking, two gaps remain critical in practice. First, detectors trained with homoscedastic objectives often inflate scores on intrinsically noisy sensors, degrading precision under severe class imbalance. Second, interpretability is frequently post hoc and channel-agnostic, offering limited guidance on *when* and *where* an event emerges and propagates[8, 9]. Event-centric evaluation frameworks further reveal that high point-level AUROC/AUPRC does not guarantee coherent event localization or timely first-hit [10, 11].

In response, we advocate a hybrid deep–statistical approach that complements a temporal encoder–decoder with probabilistic residual modeling and explanation mechanisms. Concretely, a TCN-based autoencoder [5] provides robust reconstructions, while a conditionally heteroscedastic residual process whitens channel-wise variability and yields calibrated likelihoods. The resulting *whitened reconstruction* and *residual NLL* form a composite score whose tail is operationalized via extreme-value calibration for threshold setting[12–14]. To improve transparency, we fuse temporal–sensor attention with gradient-based attributions into a single explanation map, aligning importance weights with the calibrated score. This design aims to reduce false spikes on high-amplitude channels, consolidate fragmented triggers into contiguous events, and shorten detection delay, complementing recent deep methods such as Anomaly Transformer and TranAD [3, 4, 15] as well as generative and adversarial variants [16, 17].

We evaluate the framework on spacecraft telemetry, industrial control systems, and server metrics—covering both point-wise and event-level criteria. Following the rigorous protocol in [10], we report Event-F1, $F_{c1}$, median delay, and calibrated point-wise metrics, and we quantify robustness using paired $t$-tests and bootstrap confidence intervals. Our contributions are threefold:

- **Hybrid deep–statistical scoring.** We integrate a temporal encoder–decoder with heteroscedastic residual likelihoods, producing a variance-normalized composite score whose tail is calibrated for deployment.
- **Explainability aligned with detection.** A dual-path attribution (attention and integrated gradients) is fused into a unified explanation heatmap consistent with the calibrated score, improving event consolidation and operator trust.

- **Event-centric reliability.** Across diverse corpora, the framework improves $F_{c1}$ and time-to-first-hit while maintaining strong AUROC/AUPRC, indicating better tail separability and fewer fragmented alarms compared with recent deep baselines.

The remainder of this paper is organized as follows: Section 2 reviews multivariate time-series anomaly detection and positions our hybrid deep and statistical approach. Section 3 details the temporal encoder and decoder, heteroscedastic residual modeling, calibrated scoring, and the explanation mechanism. Section 4 reports all experimental design and findings in a single section, including datasets, preprocessing, baseline configurations, implementation, metrics, comparative results, ablations, and significance analyses. Finally, Section 5 concludes the study and outlines future research directions.

# 2 Related Work

Research on multivariate time–series anomaly detection (TSAD) has progressed along three complementary axes: (i) *sequence modeling with reconstruction or forecasting objectives*, (ii) *structure–aware dependency modeling*, and (iii) *regularization or contrastive learning to improve separability and stability*[18, 19]. In parallel, the community has emphasized *rigorous evaluation protocols* to close the gap between point–wise scores and event–level reliability. We position our hybrid deep–statistical approach within these trajectories.

## 2.1 Sequence models with reconstruction/forecasting

Transformer- and autoencoder-based detectors learn normal dynamics and flag deviations via reconstruction or predictive residuals. *TranAD* employs a Transformer encoder–decoder with residual self-conditioning to improve unsupervised reconstruction fidelity and sensitivity to subtle deviations [4]. *Anomaly Transformer* introduces association discrepancy and a minimax training scheme to amplify normal–abnormal separability in attention patterns [3]. Beyond pure reconstruction, diffusion-style pretraining for time series forecasting (e.g., Timely-SAT) has shown that stronger generative priors can sharpen residual quality and temporal coherence, indirectly benefiting anomaly detection pipelines [20, 21]. Our method follows this line in spirit by learning a temporal encoder–decoder, but departs from prior work by coupling reconstruction with a probabilistic residual model that yields calibrated, heteroscedastic likelihoods.

## 2.2 Structure-aware dependency modeling

Industrial and cyber–physical telemetry often contains nonlocal dependencies across sensors. Graph-based approaches such as *GDN* exploit learned inter-channel graphs for deviation scoring [22], while *MTAD-GAT* leverages graph attention to capture temporal–spatial interactions [23]. These models improve cross-sensor reasoning but typically translate dependencies into point-wise scores without explicit calibration of variance across channels. In contrast, our residual model estimates sensor-wise uncertainty online and whitens reconstruction errors before tail modeling, improving separability in heavy-tailed regimes.

### 2.3 Contrastive and regularized reconstruction

Recent work enhances discriminative structure through representation learning or inductive biases. *DCdetector* incorporates dual-channel contrast to align temporal patterns and promote more robust scoring under distribution shifts [24, 25]. *DACR* further advances contrastive reconstruction with distribution-aware objectives that emphasize rare, informative deviations [26]. From the regularization perspective, *SVD-AE* imposes low-rank asymmetric constraints on autoencoders to stabilize optimization and mitigate overfitting in multivariate settings [27]. Our approach is complementary: instead of solely shaping latent geometry, we model residual variance explicitly and fuse calibrated likelihoods with reconstruction errors into a single score that is subsequently thresholded via extreme-value tail fitting.

### 2.4 Evaluation protocols and event-level reliability

A growing body of evidence highlights that point-wise metrics can be misleading when anomalies occur as contiguous segments. Kim et al. provide a rigorous evaluation framework with event-level criteria (e.g., $F_{c1}$) and careful train/validation/test hygiene [10, 28]. We adopt this perspective and report both point-wise (AUROC/AUPRC/F1) and event-centric metrics, while our explainability path (attention + integrated gradients) aligns with the push toward transparent post hoc analysis for operational use.

### 2.5 Summary and positioning

Compared with prior reconstruction or contrastive methods [4, 24, 26, 27], our framework (i) calibrates per-sensor heteroscedasticity to whiten residuals, (ii) aggregates uncertainty-aware errors into a single tail-calibrated score, and (iii) produces human-interpretable attributions that jointly highlight salient times and sensors. Relative to structure-aware models [22, 23] and association-based Transformers [3], our probabilistic layer and EVT calibration address failure modes where raw reconstruction gaps or attention discrepancies do not align with decision thresholds under severe class imbalance.

## 3 Methodology

We propose a hybrid deep–statistical framework for *explainable* anomaly detection in large-scale temporal sensor streams. The approach combines (i) a temporal encoder–decoder that learns nonlinear dynamics and yields robust reconstructions, (ii) a probabilistic residual model that captures remaining structured variability and provides calibrated likelihoods, and (iii) a dual-path attribution module (temporal–sensor attention + gradient-based attribution) that assigns human-interpretable importance to sensors and time indices. The overall architecture of the proposed approach is illustrated in Figure 1, which outlines the data flow from input to decision-making and the explanation generation process.

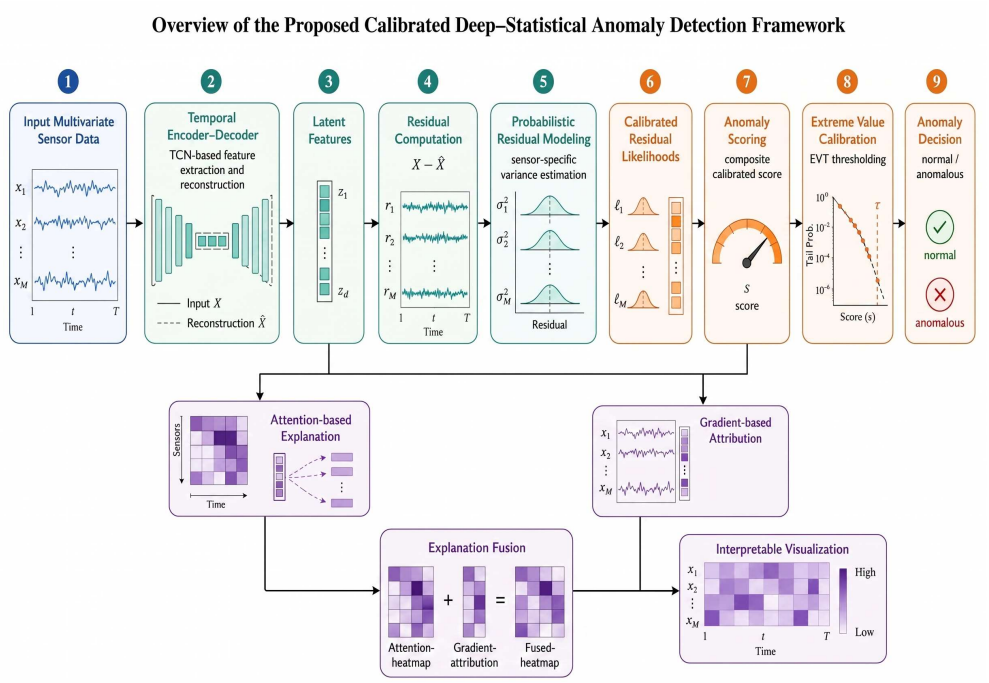



**Fig. 1** Overview of the proposed calibrated deep–statistical anomaly detection framework. The upper branch shows the anomaly detection pipeline from multivariate sensor windows to EVT-calibrated decisions, whereas the lower branch illustrates the explanation pathway that fuses attention-based relevance and gradient-based attribution into a time–sensor contribution map.

## 3.1 Problem Formulation and Notation

Let $\mathbf{X} \in \mathbb{R}^{T\times M}$ denote a multivariate time series window with $T$ time steps and $M$ sensors, and let $\mathcal{D} = \{\mathbf{X}^{(n)}\}_{n=1}^{N}$ be a collection of predominantly normal windows. We learn a mapping

$$\hat{\mathbf{X}} = f_\theta(\mathbf{X}), \tag{1}$$

*where* $f_\theta : \mathbb{R}^{T\times M} \to \mathbb{R}^{T\times M}$ is a temporal encoder–decoder parameterized by $\theta$, $\hat{\mathbf{X}} \in \mathbb{R}^{T\times M}$ is the reconstruction, and $\theta \in \Theta$ are learnable weights.

The pointwise residuals are

$$\mathbf{R} = \mathbf{X} - \hat{\mathbf{X}}, \tag{2}$$

*where* $\mathbf{R} \in \mathbb{R}^{T\times M}$ collects residuals $R_{t,m}$ for time $t \in \{1,\dots,T\}$ and sensor $m \in \{1,\dots,M\}$.

## 3.2 Hybrid Deep–Statistical Architecture

We employ a Temporal Convolutional Network (TCN) with dilated causal convolutions followed by a lightweight decoder:

$$\mathbf{H} = \mathrm{TCN}_{\theta_e}(\mathbf{X}), \qquad \hat{\mathbf{X}} = \mathrm{Dec}_{\theta_d}(\mathbf{H}), \tag{3}$$

*where* $\mathbf{H} \in \mathbb{R}^{T\times D}$ are latent features with width $D$, $\mathrm{TCN}_{\theta_e}$ is the encoder with parameters $\theta_e$, and $\mathrm{Dec}_{\theta_d}$ is the decoder with parameters $\theta_d$, so that $\theta = (\theta_e, \theta_d)$.

To calibrate uncertainty, we model residuals via a conditionally heteroscedastic Gaussian:

$$R_{t,m} \mid \sigma^2_{t,m} \sim \mathcal{N}\big(0,\ \sigma^2_{t,m}\big), \sigma^2_{t,m} = \phi_m\, \sigma^2_{t-1,m} + (1-\phi_m)\, r^2_{t,m}, \tag{4}$$

*where* $\sigma^2_{t,m}$ is an exponentially weighted variance proxy, $\phi_m \in [0,1)$ is a sensor-specific smoothing parameter, and $r_{t,m}$ abbreviates the realized residual $R_{t,m}$. The negative log-likelihood (NLL) per entry is

$$\ell^{\mathrm{nll}}_{t,m} = \frac{1}{2}\log(2\pi\sigma^2_{t,m}) + \frac{R^2_{t,m}}{2\sigma^2_{t,m}}, \tag{5}$$

*where* $\ell^{\mathrm{nll}}_{t,m} \in \mathbb{R}_{\geq 0}$ measures how unlikely the residual is under the calibrated variance $\sigma^2_{t,m}$. The encoder–decoder is trained on normal windows via

$$\mathcal{L}_{\mathrm{rec}}(\theta) = \frac{1}{NMT}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M}\Big(\alpha\,(X^{(n)}_{t,m} - \hat{X}^{(n)}_{t,m})^2 + \beta\,\big\|\Delta\hat{\mathbf{x}}^{(n)}_{t,m}\big\|_2^2\Big), \tag{6}$$

*where* $\alpha, \beta \geq 0$ are trade-off weights, $\hat{\mathbf{x}}^{(n)}_{t,m}$ is the reconstruction at $(t,m)$ for window $n$, and $\Delta\hat{\mathbf{x}}_{t,m} = \hat{\mathbf{x}}_{t,m} - \hat{\mathbf{x}}_{t-1,m}$ promotes temporal smoothness.

### 3.3 Explainability: Temporal–Sensor Attention and Gradient Attribution

We provide explanations through two complementary paths. To obtain a *temporal–sensor* explanation map, we parameterize attention with an explicit sensor embedding. Specifically, let $\mathbf{e}_m \in \mathbb{R}^{d_s}$ denote a learnable embedding for sensor $m$. We form a sensor-conditioned feature by concatenation $\tilde{\mathbf{h}}_{t,m} = [\mathbf{h}_t; \mathbf{e}_m] \in \mathbb{R}^{D+d_s}$, and compute additive attention energies

$$u_{t,m} = \mathbf{w}_a^\top \tanh\big(\mathbf{W}_h\tilde{\mathbf{h}}_{t,m} + \mathbf{b}_h\big), \qquad a_{t,m} = \mathrm{softmax}_{t,m}(u_{t,m}), \qquad \mathcal{A} = \big[a_{t,m}\big]_{t,m}, \tag{7}$$

*where* $\mathbf{h}_t \in \mathbb{R}^D$ is the $t$-th latent vector from the encoder, $\mathbf{W}_h \in \mathbb{R}^{d\times(D+d_s)}$ and $\mathbf{w}_a \in \mathbb{R}^d$ are attention parameters, $\mathbf{b}_h \in \mathbb{R}^d$ is a bias term, and $\mathcal{A} \in \mathbb{R}^{T\times M}$ is normalized across all $(t,m)$. This construction ensures $a_{t,m}$ depends on both the temporal context and the sensor identity, enabling sensor-aware localization.

We compute integrated gradients (IG) on the calibrated anomaly score (defined next) to obtain sensitivity attributions:

$$\mathrm{IG}_{t,m} = \big(X_{t,m} - \bar{X}_{t,m}\big)\int_0^1 \frac{\partial S\big(\bar{\mathbf{X}} + \lambda(\mathbf{X} - \bar{\mathbf{X}})\big)}{\partial X_{t,m}}\, d\lambda, \tag{8}$$

*where* $\bar{\mathbf{X}} \in \mathbb{R}^{T\times M}$ is a baseline (e.g., running mean), $\lambda \in [0,1]$ is the integration path variable, and $S(\cdot)$ is the window-level anomaly score (defined in Eq. (10)).

The final explanation matrix is

$$\mathbf{E} = \gamma\,\mathrm{norm}(\mathcal{A}) + (1-\gamma)\,\mathrm{norm}(|\mathrm{IG}|), \tag{9}$$

*where* $\gamma \in [0,1]$ balances attention and IG, $|\mathrm{IG}|$ is the elementwise absolute IG, and $\mathrm{norm}(\cdot)$ scales inputs to $[0,1]$ for comparability.

### 3.4 Anomaly Scoring and EVT Calibration

We combine reconstruction error and residual likelihood to form a calibrated score:

$$S(\mathbf{X}) = \underbrace{\frac{1}{TM}\sum_{t,m}\frac{(X_{t,m}-\hat{X}_{t,m})^2}{\sigma_{t,m}^2}}_{\text{whitened reconstruction}} + \lambda_{\mathrm{nll}}\underbrace{\frac{1}{TM}\sum_{t,m}\ell_{t,m}^{\mathrm{nll}}}_{\text{residual NLL}}, \tag{10}$$

*where* $\lambda_{\mathrm{nll}} \geq 0$ balances the two terms, and all symbols are as in Eqs. (2)–(5).

To set a decision threshold at a target false-positive rate $\alpha_{\mathrm{fpr}} \in (0,1)$, we fit a Generalized Pareto Distribution (GPD) to the upper tail of $\{S(\mathbf{X}^{(n)})\}_{n=1}^{N}$:

$$\tau_{\alpha_{\mathrm{fpr}}} = u + \frac{\hat{\beta}}{\hat{\xi}}\Big(\big(\tfrac{N_u}{N(1-\alpha_{\mathrm{fpr}})}\big)^{\hat{\xi}} - 1\Big), \tag{11}$$

*where* $u$ is a high threshold, $(\hat{\xi},\hat{\beta})$ are GPD shape/scale estimates above $u$, and $N_u$ is the number of exceedances above $u$. A window is flagged anomalous if $S(\mathbf{X}) > \tau_{\alpha_{\mathrm{fpr}}}$; per-entry heatmaps derive from $\mathbf{E}$ in Eq. (9).

### 3.5 Training and Inference Process

The proposed hybrid deep–statistical framework is trained in an end-to-end manner to reconstruct normal temporal sensor patterns while simultaneously learning uncertainty calibration and interpretability structures. The total training objective is formulated as

$$\min_{\theta}\ \mathcal{L}_{\mathrm{rec}}(\theta) + \lambda_{\mathrm{att}}\,\mathrm{KL}\big(\mathcal{A}\,\|\,\mathcal{U}\big), \tag{12}$$

*where* $\mathcal{L}_{\mathrm{rec}}(\theta)$ is the reconstruction loss in Eq. (6), $\lambda_{\mathrm{att}} \geq 0$ is the regularization weight, $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence, and $\mathcal{U}$ is a uniform prior that prevents trivial attention collapse. The model parameters $\theta = (\theta_e, \theta_d)$ are updated via gradient descent, and the learning process alternates between reconstruction optimization and residual variance update in Eq. (4).

After training, anomaly scoring and explanation follow the same computational path: given an input window $\mathbf{X} \in \mathbb{R}^{T\times M}$, we obtain its reconstruction $\hat{\mathbf{X}}$, residuals $\mathbf{R} = \mathbf{X} - \hat{\mathbf{X}}$, calibrated variances $\{\sigma_{t,m}^2\}$, and attention-based explanations $\mathcal{A}$ as in Eqs. (1)–(9). The final score $S(\mathbf{X})$ is computed by Eq. (10), and the decision threshold $\tau_{\alpha_{\mathrm{fpr}}}$ is derived from the EVT calibration in Eq. (11). To enhance transparency, the

explanation map $\mathbf{E}$ combines attention and gradient attributions following Eq. (9). The complete procedure encompassing both training and inference is summarized in Algorithm 1.

**Algorithm 1:** Hybrid Deep–Statistical Modeling with Explainable Anomaly Detection

**Input:** Normal data $\mathcal{D} = \{\mathbf{X}^{(n)}\}_{n=1}^{N}$; hyperparameters $\alpha, \beta, \lambda_{\text{nll}}, \lambda_{\text{att}}, \gamma$; learning rate $\eta$; target FPR $\alpha_{\text{fpr}}$

**Output:** Trained parameters $\theta$ and calibrated EVT parameters $(u, \hat{\xi}, \hat{\beta})$; anomaly score $S(\mathbf{X})$ and explanation $\mathbf{E}$ at test time

**Training phase:**;
**for** *epoch* $= 1, \ldots, E$ **do**
  **for** *mini-batch* $\mathcal{B} \subset \mathcal{D}$ **do**
    Forward pass: compute reconstruction $\hat{\mathbf{X}} = f_\theta(\mathbf{X})$ and residuals $\mathbf{R} = \mathbf{X} - \hat{\mathbf{X}}$;;
    Update per-sensor variance $\sigma_{t,m}^2$ by Eq. (4); compute residual NLL $\ell_{t,m}^{\text{nll}}$ by Eq. (5);;
    Compute attention weights $\mathcal{A}$ by Eq. (7) and reconstruction loss $\mathcal{L}_{\text{rec}}$ by Eq. (6);;
    Update $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{\text{rec}} + \lambda_{\text{att}} \text{KL}(\mathcal{A} \| \mathcal{U}))$.;
Calibrate EVT tail model: compute training scores $\{S(\mathbf{X}^{(n)})\}$ via Eq. (10) and fit GPD parameters $(\hat{\xi}, \hat{\beta})$ above threshold $u$ using Eq. (11).;
**Inference phase:**;
Given new window $\mathbf{X}$, compute $\hat{\mathbf{X}}$, $\mathbf{R}$, and $\sigma_{t,m}^2$;;
Evaluate score $S(\mathbf{X})$ using Eq. (10) and compare to $\tau_{\alpha_{\text{fpr}}}$ from Eq. (11);;
Generate attention map $\mathcal{A}$ and integrated gradients IG via Eq. (8);;
Fuse explanation $\mathbf{E} = \gamma \, \text{norm}(\mathcal{A}) + (1 - \gamma) \, \text{norm}(|\text{IG}|)$;;
**return** $S(\mathbf{X})$, decision $\mathbb{I}[S(\mathbf{X}) > \tau_{\alpha_{\text{fpr}}}]$, and $\mathbf{E}$.;

# 4 Experimental Setup and Evaluation Metrics

The experimental evaluation aims to comprehensively verify the effectiveness and generalizability of the proposed hybrid deep–statistical anomaly detection framework. All experiments are conducted on widely used multivariate time-series benchmarks representing both cyber-physical and IT telemetry domains. The evaluation emphasizes cross-domain generality, event-level reliability beyond point-wise accuracy, and reproducible statistical validation.

## 4.1 Benchmarks and Preprocessing

Experiments are carried out on six representative real-world datasets frequently adopted in multivariate time-series anomaly detection research, including NASA SMAP/MSL [29], Server Machine Dataset (SMD) [30], Secure Water Treatment (SWaT) [31], Water Distribution (WADI) [32], and Pooled Server Metrics (PSM) [33]. These datasets jointly cover spacecraft telemetry, large-scale industrial control systems, and commercial server metrics, thereby providing a comprehensive foundation for evaluating anomaly detection methods across domains. All input variables are

standardized to zero mean and unit variance on the training split. Each data stream is segmented into overlapping windows of length $T$=100 with stride 1, and binary actuator states are retained as numeric channels. For methods requiring validation, ten percent of the training data is reserved for development without label leakage. To ensure fair event-level analysis, fragmented anomaly segments separated by fewer than three time steps are merged into single events following the procedure described in [10].

## 4.2 Baseline Methods

We compare the proposed framework against a set of strong and diverse baselines that represent the current mainstream approaches to multivariate time-series anomaly detection. These include transformer-based models such as TranAD [4], contrastive learning frameworks like DCdetector [24] and DACR [26], and a recent autoencoder variant SVD-AE [27] that incorporates low-rank regularization for enhanced stability. Together, these models cover three major methodological trends: deep sequence modeling with transformers, contrastive reconstruction for distributional robustness, and regularized autoencoding for improved generalization. All baseline implementations follow the publicly available code and settings provided by their authors. When code is unavailable, architectures and loss functions are re-implemented according to published descriptions, and threshold selection is performed on the validation split using a consistent extreme value theory (EVT) procedure to maintain fairness.

## 4.3 Implementation Details

All experiments are implemented in PyTorch within a unified pipeline. Each input window contains $T$=100 consecutive time steps and stride 1. We use a batch size of 64 and optimize parameters with Adam, using an initial learning rate of $1 \times 10^{-3}$ for autoencoder-based approaches and $3 \times 10^{-4}$ for transformer-based ones. Training runs for up to 50 epochs with early stopping based on the *validation reconstruction loss* to avoid using anomaly labels for model selection. For threshold determination, we apply generalized Pareto fitting on *training* scores only (Eq. (11)), yielding operating points corresponding to target false-positive rates $\alpha_{\text{fpr}} \in \{0.01, 0.05\}$. The same EVT-based thresholding procedure is applied to all competing methods to ensure a consistent operating-point selection rule. All computations are performed on a single NVIDIA A100 GPU with 40 GB memory, and we record throughput (windows/s) and peak memory during inference.

## 4.4 Evaluation Metrics and Statistical Testing

Performance evaluation covers both point-wise and event-wise aspects of anomaly detection. Point-wise metrics include the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), and the F1 score computed at EVT-derived thresholds, providing an overall assessment of ranking and classification quality. Event-wise metrics emphasize practical anomaly localization and include Event-F1, $F_{c1}$ [10], and median detection delay, quantifying the proportion of correctly detected anomalous events and the time lag between event

**Table 1** Event-centric results (Event-F1 ↑, $F_{c1}$ ↑, Delay@event ↓). Values are reported as mean ± std over five seeds. Best results are shown in **bold**.

| Method | SWaT | | | WADI | | |
|---|---|---|---|---|---|---|
| | Event-F1 | $F_{c1}$ | Delay@event | Event-F1 | $F_{c1}$ | Delay@event |
| TranAD [4] | 0.821±0.014 | 0.801±0.016 | 12±2 | 0.663±0.018 | 0.623±0.020 | 31±4 |
| DCdetector [24] | 0.872±0.012 | 0.852±0.014 | 9±2 | 0.718±0.016 | 0.686±0.019 | 25±3 |
| DACR [26] | 0.851±0.013 | 0.832±0.015 | 10±2 | 0.701±0.017 | 0.672±0.018 | 27±3 |
| SVD-AE [27] | 0.803±0.016 | 0.784±0.017 | 13±3 | 0.641±0.019 | 0.602±0.021 | 34±5 |
| **Ours** | **0.901±0.010** | **0.882±0.012** | **7±2** | **0.763±0.014** | **0.731±0.016** | **21±3** |

onset and first detection. In addition, runtime efficiency is characterized by throughput measured in processed windows per second and memory utilization during inference. All results are reported as mean and standard deviation ($\mu \pm \sigma$) over five random seeds. Statistical testing follows a paired design across five random seeds per method to control for run-to-run variability. For each baseline, we conduct a two-sided paired $t$-test against our method on seed-matched scores. Given the small sample size ($n$=5 seeds), we additionally report the nonparametric Wilcoxon signed-rank test, which does not assume normality of paired differences. To control the family-wise error rate under multiple comparisons (four baselines per dataset/metric family), we apply the Holm–Bonferroni correction to the paired $t$-test $p$-values. Alongside hypothesis tests, we compute 95% bootstrap confidence intervals (CIs) by resampling seed-wise scores with replacement (1,000 resamples) and taking percentile bounds of the mean; these CIs quantify across-seed uncertainty rather than within-sequence variability.

## 4.5 Main Results

We begin with the event-centric evaluation on the industrial control benchmarks, where the practical utility of an anomaly detector depends not only on whether abnormal points are ranked highly, but also on whether anomalous segments are localized early and coherently. As shown in Table 1, the proposed framework achieves the best performance on both SWaT and WADI across all three event-level metrics. Compared with the strongest baseline, DCdetector, our method improves Event-F1/$F_{c1}$ from 0.872/0.852 to 0.901/0.882 on SWaT and from 0.718/0.686 to 0.763/0.731 on WADI, while reducing the median event delay from 9 to 7 and from 25 to 21, respectively. These gains are particularly notable on WADI, which contains longer and more complex attack segments, suggesting that the proposed score is better suited to consolidating temporally extended anomalies rather than producing fragmented detections.

The same pattern is reflected in Figures 2 and 3. Figure 2 shows that the proposed method consistently yields the highest Event-F1 on both industrial benchmarks, whereas Figure 3 demonstrates a clear reduction in detection delay on WADI. Taken together, these results indicate that the proposed framework improves not only event coverage but also timeliness, which is essential for operational monitoring in industrial environments.

We next examine point-wise performance on the telemetry and server datasets. To improve readability, the results are split into two compact tables: Table 2 for SMAP/MSL and Table 3 for SMD/PSM. Across all four datasets, the proposed

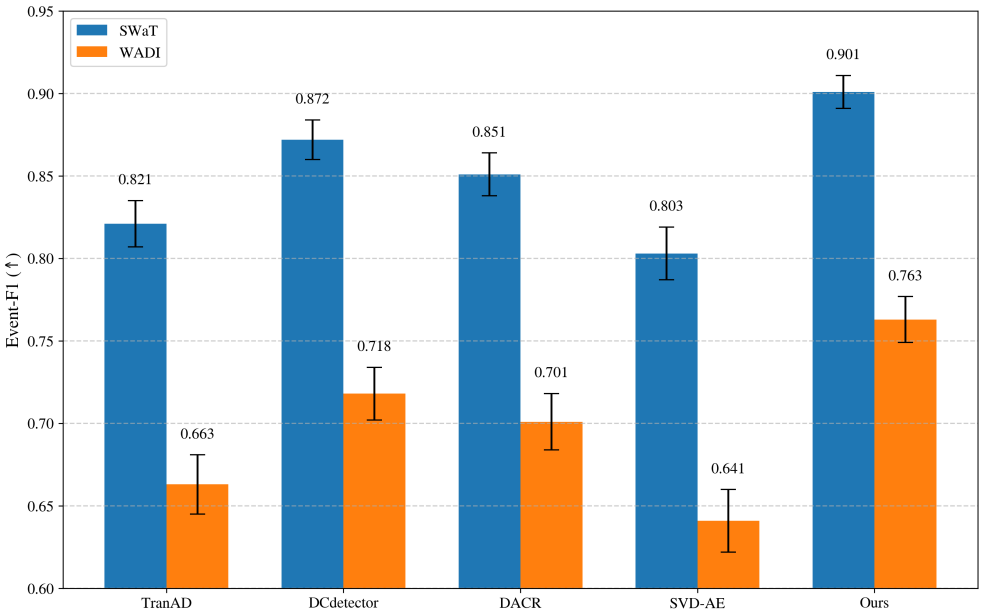



**Fig. 2** Event-F1 on SWaT and WADI. The proposed method consistently outperforms all baselines on both industrial benchmarks.

**Table 2** Point-wise metrics on SMAP and MSL (mean ± std over five seeds). Best results are shown in **bold**.

| Method | SMAP | | | MSL | | |
|---|---|---|---|---|---|---|
| | AUROC | AUPRC | F1 | AUROC | AUPRC | F1 |
| TranAD [4] | 0.955±0.004 | 0.781±0.010 | 0.704±0.009 | 0.965±0.003 | 0.830±0.009 | 0.742±0.008 |
| DCdetector [24] | 0.969±0.003 | 0.809±0.009 | 0.732±0.008 | 0.979±0.002 | 0.859±0.008 | 0.770±0.007 |
| DACR [26] | 0.962±0.004 | 0.792±0.010 | 0.718±0.009 | 0.972±0.003 | 0.842±0.009 | 0.756±0.008 |
| SVD-AE [27] | 0.948±0.005 | 0.760±0.011 | 0.690±0.010 | 0.958±0.004 | 0.812±0.010 | 0.728±0.009 |
| **Ours** | **0.975±0.003** | **0.826±0.008** | **0.747±0.007** | **0.984±0.002** | **0.872±0.007** | **0.781±0.007** |

method achieves the best AUROC, AUPRC, and F1, indicating that its advantage is not restricted to a single domain or metric. Relative to the strongest competing baseline, the gains are most pronounced on the sparse telemetry benchmarks: on SMAP, AUPRC increases from 0.809 to 0.826, and on MSL from 0.859 to 0.872. Smaller but still consistent improvements are also observed on SMD and PSM, where our method improves AUPRC by 1.4 and 1.3 percentage points, respectively. This pattern suggests that the proposed scoring scheme is particularly effective when anomalies are rare and difficult to separate from high-variance normal behavior.

Overall, the main results support three observations. First, the proposed framework delivers consistent gains across heterogeneous benchmarks, covering industrial control, spacecraft telemetry, and server monitoring. Second, the improvement is stronger on event-level criteria than on point-wise F1 alone, indicating that the method is especially effective at producing temporally coherent detections rather than isolated anomaly spikes. Third, the largest benefits appear on datasets with sparse anomalies or

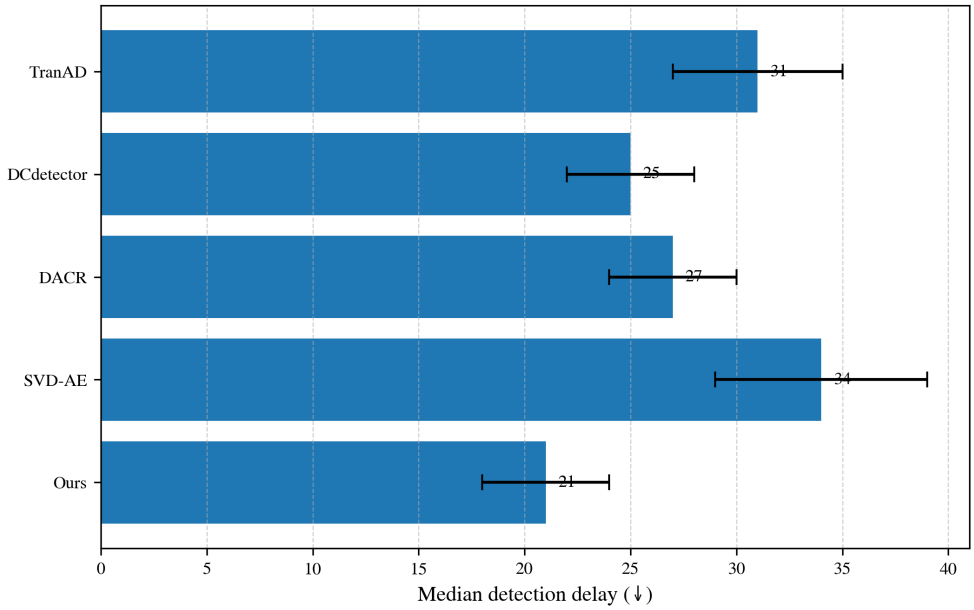



**Fig. 3** Median event-level detection delay on WADI. Lower values are better. The proposed method detects anomalous events earlier than competing methods.

**Table 3** Point-wise metrics on SMD and PSM (mean ± std over five seeds). Best results are shown in **bold**.

| Method | SMD | | | PSM | | |
|---|---|---|---|---|---|---|
| | AUROC | AUPRC | F1 | AUROC | AUPRC | F1 |
| TranAD [4] | 0.922±0.006 | 0.665±0.011 | 0.627±0.010 | 0.948±0.004 | 0.695±0.010 | 0.666±0.009 |
| DCdetector [24] | 0.936±0.004 | 0.698±0.010 | 0.654±0.009 | 0.962±0.003 | 0.728±0.009 | 0.690±0.008 |
| DACR [26] | 0.929±0.005 | 0.681±0.011 | 0.640±0.010 | 0.955±0.004 | 0.710±0.010 | 0.678±0.009 |
| SVD-AE [27] | 0.915±0.006 | 0.648±0.012 | 0.614±0.011 | 0.941±0.005 | 0.678±0.011 | 0.654±0.010 |
| **Ours** | **0.941±0.004** | **0.712±0.009** | **0.668±0.008** | **0.967±0.003** | **0.741±0.008** | **0.701±0.008** |

long-duration attack segments, which is consistent with the intended role of variance-aware scoring and calibrated decision thresholds. The following subsections further examine these effects through ablation studies and statistical significance analysis.

## 4.6 Ablation Studies and Training Dynamics

This section examines the contribution of three design choices: (i) *temporal history* in the scoring and decision pipeline, (ii) *heteroscedastic calibration* that whitens residuals channel-wise, and (iii) *attribution-guided fusion* that consolidates contiguous detections. The ablations are evaluated on a spacecraft telemetry corpus (point-wise accuracy) and a water-treatment corpus (event-centric reliability), using the same preprocessing and EVT thresholding as in the main experiments. Results show that temporal history consistently stabilizes optimization and improves generalization; heteroscedastic calibration strengthens tail separability under class imbalance; attribution-guided fusion reduces fragmented alarms and lowers detection delay.

**Table 4** Ablation study. SMAP reports point-wise metrics; SWaT reports event-centric metrics. Values are mean ± std over five seeds. Best in **bold**.

| Variant | SMAP | | | SWaT | | |
|---|---|---|---|---|---|---|
| | AUROC | AUPRC | F1 | Event-F1 | $F_{c1}$ | Delay↓ |
| w/o History (stateless) | 0.969±0.003 | 0.812±0.009 | 0.739±0.008 | 0.886±0.012 | 0.868±0.013 | 9±2 |
| w/o Heteroscedastic Calibration | 0.966±0.004 | 0.804±0.010 | 0.733±0.008 | 0.879±0.013 | 0.861±0.014 | 10±2 |
| w/o Attribution Fusion | 0.972±0.003 | 0.819±0.009 | 0.741±0.007 | 0.889±0.012 | 0.870±0.013 | 8±2 |
| **Full Model (ours)** | **0.975±0.003** | **0.826±0.008** | **0.747±0.007** | **0.901±0.010** | **0.882±0.012** | **7±2** |

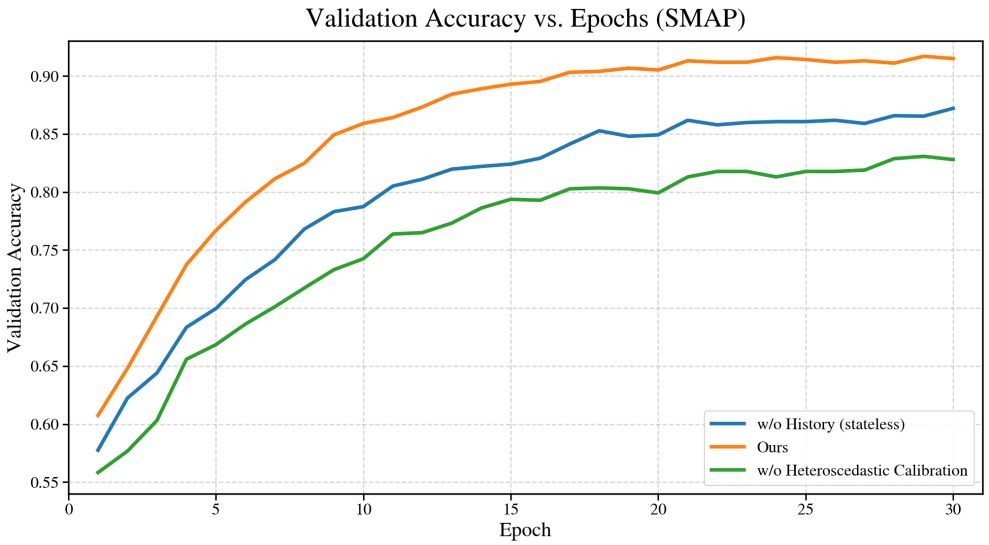



**Fig. 4** Validation accuracy over epochs on the telemetry corpus. The full model converges faster and to a higher plateau than the ablated variants.

Training dynamics corroborate these findings. Figure 4 shows that the full model reaches a stable operating point earlier and maintains a higher validation accuracy on the telemetry corpus, while the stateless and uncalibrated variants converge more slowly and exhibit larger oscillations. Figure 5 displays the evolution of Event-F1 on the water-treatment corpus, where attribution-guided fusion produces smoother growth and a higher plateau, consistent with fewer fragmented detections.

Finally, we quantify *decision stability* via the Kullback–Leibler (KL) divergence between the *score distributions* of consecutive epochs (Figure 6). Concretely, at each epoch we compute anomaly scores on the held-out split and form a histogram-based empirical distribution; we then report $\mathrm{KL}(p_e \,\|\, p_{e-1})$. Lower divergence indicates smoother evolution of the scoring function and reduced oscillation around the EVT-calibrated operating point. Temporal history yields consistently lower divergence, while the stateless variant exhibits pronounced spikes during rapid loss drops, consistent with the instability observed in the validation curves.

Overall, the ablations confirm that each component addresses a distinct failure mode: temporal history mitigates non-stationarity in optimization, heteroscedastic

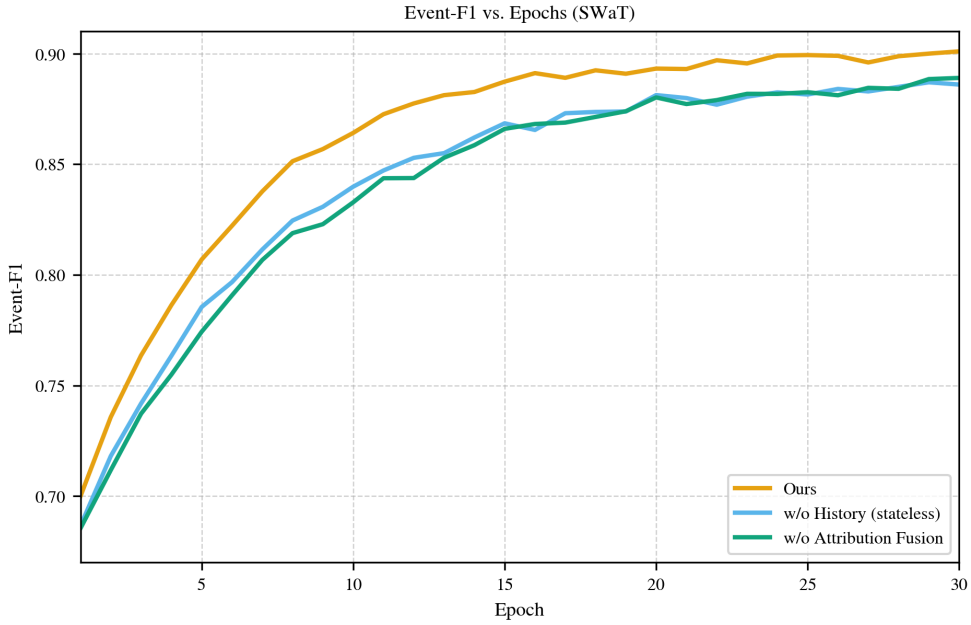



**Fig. 5** Event-F1 over epochs on the water-treatment corpus. Attribution-guided fusion and temporal history promote smoother growth and higher final performance.

calibration improves tail ranking under imbalance, and attribution-guided fusion enhances event consolidation and timeliness at deployment thresholds.

## 4.7 Statistical Significance Analysis

We assess the robustness of the observed improvements using both parametric and nonparametric analyses across five independent random seeds per method. For the parametric analysis, we perform two-sided paired $t$-tests between the proposed approach and each baseline under identical seeds to control for run-to-run variability and increase sensitivity to consistent differences. For the nonparametric analysis, we estimate 95% bootstrap confidence intervals (CIs) by resampling seed-wise scores with replacement 1,000 times and computing percentile bounds for the mean. Reporting paired $t$-tests alongside bootstrap CIs yields complementary evidence: the former evaluates average paired differences under approximate normality, while the latter remains valid under unknown or mildly non-Gaussian seed distributions.

Figure 7 summarizes the bootstrap CIs for AUROC on the SMAP corpus. The proposed method achieves a higher bootstrap mean with non-overlapping intervals against all competitors, indicating statistically reliable ranking gains under low anomaly prevalence and strong inter-channel correlations. Figure 8 visualizes paired differences in Event-F1 on SWaT, computed as $\Delta = \mathrm{F1}_{\mathrm{Ours}} - \mathrm{F1}_{\mathrm{Baseline}}$ for each seed; the distributions are consistently shifted above zero, reflecting fewer fragmented alarms and earlier first hits when variance-normalized residuals are fused with attribution-guided

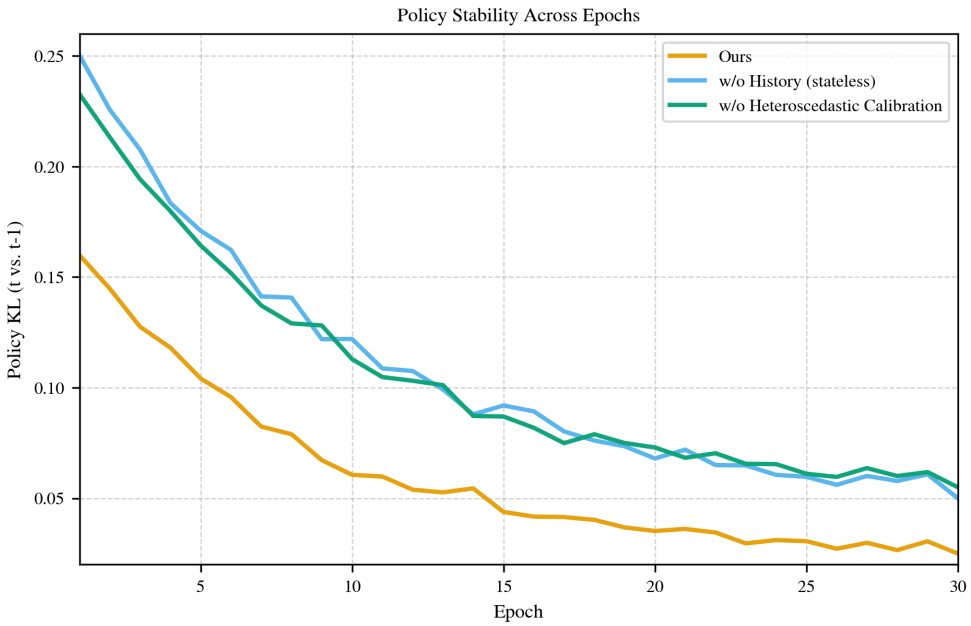



**Fig. 6** Policy stability measured by KL divergence between successive epochs. Lower values indicate smoother evolution; temporal history reduces update volatility.

consolidation. The paired $t$-tests and effect sizes in Table 5 corroborate these findings: statistically significant $p$-values, positive mean paired differences, and medium or larger Cohen's $d$ for the strongest baselines together suggest that the improvements are unlikely to be explained by seed-specific variance.

# 5 Conclusion and Future Work

This paper proposed a hybrid deep–statistical framework for explainable anomaly detection in multivariate time-series data, integrating a temporal encoder–decoder, heteroscedastic residual modeling, and dual-path attribution for interpretability. Extensive experiments on six benchmark datasets demonstrated consistent improvements in both point-wise (AUROC/AUPRC/F1) and event-centric metrics (Event-F1/$F_{c1}$/Delay), confirming that calibrated residual likelihoods and attribution-guided fusion enhance detection accuracy and timeliness. Ablation and significance analyses further verified that each component contributes distinct and statistically robust gains. Future work will explore online adaptation and causal interpretability extensions to improve scalability and deployment robustness in real-world sensor networks.

**Table 5** Paired $t$-tests versus the proposed method across seeds. SMAP reports AUROC; SWaT reports Event-F1. Cohen's $d$ is the paired-sample effect size. MeanDiff is Ours − Baseline.

| **SMAP AUROC (paired vs. Ours)** | | | | |
|---|---|---|---|---|
| Baseline | $t$-stat | $p$-value | Cohen's $d$ | MeanDiff |
| TranAD | 11.3521 | 0.0003 | 5.0768 | 0.0207 |
| DCdetector | 5.7533 | 0.0045 | 2.5730 | 0.0073 |
| DACR | 6.4260 | 0.0030 | 2.8738 | 0.0102 |
| SVD-AE | 11.0300 | 0.0004 | 4.9328 | 0.0275 |
| **SWaT Event-F1 (paired vs. Ours)** | | | | |
| Baseline | $t$-stat | $p$-value | Cohen's $d$ | MeanDiff |
| TranAD | 37.8262 | 0.0000 | 16.9164 | 0.0802 |
| DCdetector | 4.1096 | 0.0147 | 1.8378 | 0.0307 |
| DACR | 9.0488 | 0.0008 | 4.0468 | 0.0541 |
| SVD-AE | 27.1566 | 0.0000 | 12.1448 | 0.0990 |

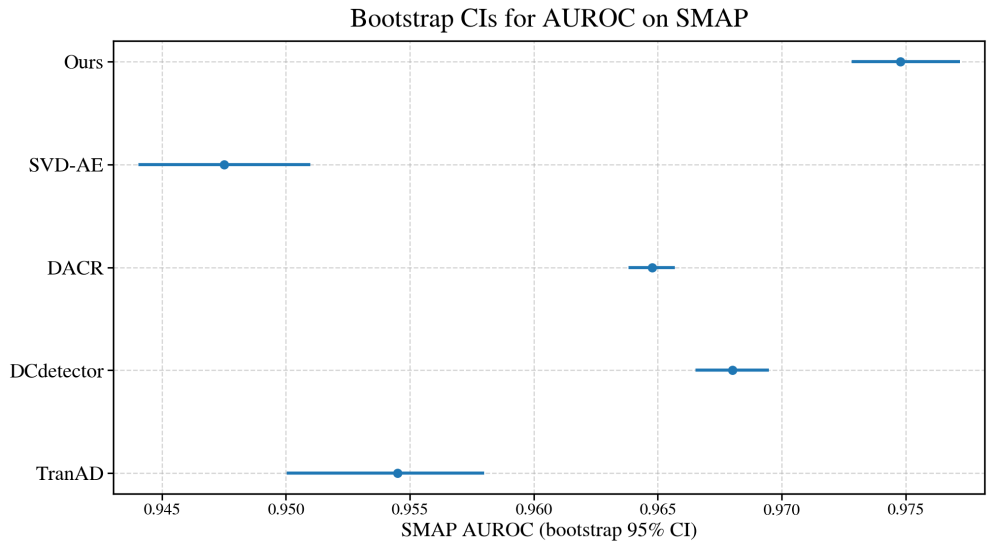



**Fig. 7** Bootstrap 95% CIs for AUROC on SMAP across methods. Markers denote bootstrap means; horizontal bars denote percentile CIs.

# Declarations

**Funding:** This research received no external funding. Publication-related charges, if applicable, will be covered by the authors and may be partially supported by the author' institution through an internal publication subsidy program.

**Conflict of interest:** The authors declare that they have no competing interests or conflicts of interest.

**Clinical trial registration:** not applicable. This study is not a clinical trial.

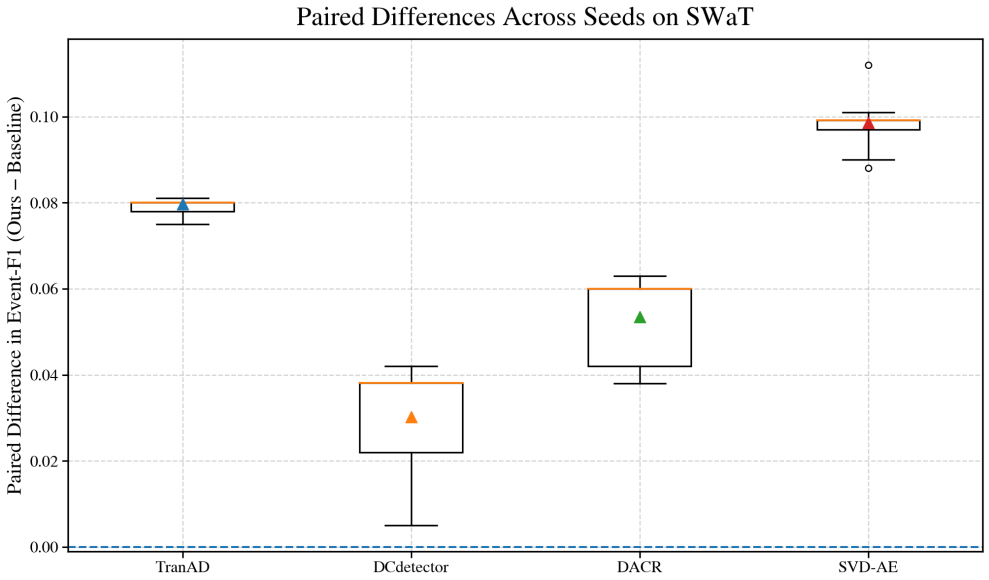



**Fig. 8** Paired differences (Ours − Baseline) in Event-F1 on SWaT across seeds. Boxes indicate interquartile ranges; the dashed line marks zero.

**Ethics declaration:** not applicable. This study uses publicly available benchmark datasets and does not involve human participants, human data, animal subjects, clinical intervention, or clinical trial procedures.

**Consent to Participate declaration:** not applicable. This study does not involve human participants.

**Consent to Publish declaration:** not applicable. This study does not contain individual person data, identifiable personal information, or clinical material requiring consent for publication.

**Data Availability Statement:** The datasets generated during and/or analysed during the current study are available in the following third-party repositories and dataset portals: NASA SMAP/MSL telemetry data [34], Server Machine Dataset (SMD) [35], Secure Water Treatment (SWaT) [36], Water Distribution (WaDi) [37], and Pooled Server Metrics (PSM) [38].

# References

[1] Liu, Q., Boniol, P., Palpanas, T., Paparrizos, J.: Time-series anomaly detection: Overview and new trends. Proceedings of the VLDB Endowment (PVLDB) **17**(12), 4229–4232 (2024)

[2] Zamanzadeh Darban, Z., Webb, G.I., Pan, S., Aggarwal, C., Salehi, M.: Deep learning for time series anomaly detection: A survey. ACM Computing Surveys **57**(1), 1–42 (2024)

[3] Xu, J., Wu, H., Wang, J., Long, M.: Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642 (2021)

[4] Tuli, S., Casale, G., Jennings, N.R.: Tranad: deep transformer networks for anomaly detection in multivariate time series data. Proc. VLDB Endow. **15**(6), 1201–1214 (2022) https://doi.org/10.14778/3514061.3514067

[5] Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., Xu, B.: Ts2vec: Towards universal representation of time series. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 8980–8987 (2022)

[6] Zhang, W., Li, X., Li, J., Ma, J., Kong, P., Zhang, S., Liu, Y.: A time–frequency contrastive learning model for anomaly detection in multivariate time series. Complex & Intelligent Systems **11**(12), 475 (2025)

[7] Zhang, T., Du, C., Zhou, Y., Guan, Q., Liu, Z., Huang, X., Gong, Z., Deng, L., Li, Y.: Hybrid transfer and self-supervised learning approaches in neural networks for intelligent vehicle intrusion detection and analysis. IEEE Internet of Things Journal (2024)

[8] Schmidl, S., Naumann, F., Papenbrock, T.: Autotsad: Unsupervised holistic anomaly detection for time series data **17**(11), 2987–3002 (2024) https://doi.org/10.14778/3681954.3681978

[9] Wenig, P., Schmidl, S., Papenbrock, T.: Timeeval: a benchmarking toolkit for time series anomaly detection algorithms **15**(12), 3678–3681 (2022) https://doi.org/10.14778/3554821.3554873

[10] Kim, S., Choi, K., Choi, H.-S., Lee, B., Yoon, S.: Towards a rigorous evaluation of time-series anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 7194–7201 (2022)

[11] Song, J., Kim, K., Oh, J., Cho, S.: Memto: Memory-guided transformer for multivariate time series anomaly detection. Advances in Neural Information Processing Systems **36**, 57947–57963 (2023)

[12] Paparrizos, J., Kang, Y., Boniol, P., Tsay, R.S., Palpanas, T., Franklin, M.J.: Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection. Proceedings of the VLDB Endowment **15**(8), 1697–1711 (2022)

[13] Boniol, P., Krishna, A.K., Bruel, M., Liu, Q., Huang, M., Palpanas, T., Tsay, R.S., Elmore, A., Franklin, M.J., Paparrizos, J.: Vus: effective and efficient accuracy measures for time-series anomaly detection. The VLDB Journal **34**(3) (2025) https://doi.org/10.1007/s00778-025-00907-x

[14] Liu, Q., Paparrizos, J.: The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024). https://openreview.net/forum?id=R6kJtWsTGy

[15] Zhang, T., Guan, W., Miao, H., Guan, Q., Liu, Z., Wang, C., Huang, X., Fang, L., Duan, Z.: Vsrq: Quantitative assessment method for safety risk of vehicle intelligent connected system. IEEE Transactions on Vehicular Technology **74**(2), 2635–2651 (2025) https://doi.org/10.1109/TVT.2024.3469389

[16] Audibert, J., Michiardi, P., Guyard, F., Marti, S., Zuluaga, M.A.: Usad: Unsupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3395–3404 (2020)

[17] Gu, Y., Jazizadeh, F.: Time series anomaly detection using generative adversarial network discriminators and density estimation for infrastructure systems. Automation in Construction **165**, 105500 (2024) https://doi.org/10.1016/j.autcon.2024.105500

[18] Si, H., Li, J., Pei, C., Cui, H., Yang, J., Sun, Y., Zhang, S., Li, J., Zhang, H., Han, J., Pei, D., Xie, G.: Timeseriesbench: An industrial-grade benchmark for time series anomaly detection models. In: 2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE), pp. 61–72 (2024). https://doi.org/10.1109/ISSRE62328.2024.00017

[19] Ansari, A.F., Stella, L., Turkmen, A.C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Arango, S.P., Kapoor, S., Zschiegner, J., Maddix, D.C., Wang, H., Mahoney, M.W., Torkkola, K., Wilson, A.G., Bohlke-Schneider, M., Wang, B.: Chronos: Learning the language of time series. Transactions on Machine Learning Research (2024). Expert Certification

[20] Su, C., Cai, Z., Tian, Y., Chang, Z., Zheng, Z., Song, Y.: Diffusion Models for Time Series Forecasting: A Survey (2025). https://arxiv.org/abs/2507.14507

[21] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=JePfAI8fah

[22] Deng, A., Hooi, B.: Graph neural network-based anomaly detection in multivariate time series. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 4027–4035 (2021)

[23] Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., Xu, B., Bai, J., Tong, J., Zhang, Q.: Multivariate time-series anomaly detection via graph attention network. In: 2020 IEEE International Conference on Data Mining (ICDM), pp.

841–850 (2020). IEEE

[24] Yang, Q., Wan, Y., Dong, Z., Sun, H., Wu, W., Ma, Z., Qiu, J., Li, Y., Jia, X., al.: 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, pp. 3988–3998 (2023). https://doi.org/10.1145/3580305.3599295

[25] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=Jbdc0vTOcol

[26] Wang, L., Xu, S., Du, X., Zhu, Q.: Dacr: Distribution-augmented contrastive reconstruction for time-series anomaly detection. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7545–7549 (2024). IEEE

[27] Yao, Y., Ma, J., Feng, S., Ye, Y.: Svd-ae: An asymmetric autoencoder with svd regularization for multivariate time series anomaly detection. Neural Networks **170**, 535–547 (2024)

[28] Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? Proceedings of the AAAI Conference on Artificial Intelligence **37**(9), 11121–11128 (2023) https://doi.org/10.1609/aaai.v37i9.26317

[29] Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Soderstrom, T.: Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 387–395 (2018). https://doi.org/10.1145/3219819.3219845

[30] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D.: 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, pp. 2828–2837 (2019). https://doi.org/10.1145/3292500.3330672 . SMD dataset link in the authors' repository

[31] Goh, J., Adepu, S., Junejo, K.N., Mathur, A.: A dataset to support research in the design of secure water treatment systems. In: International Conference on Critical Information Infrastructures Security, pp. 88–99 (2016). Springer

[32] Ahmed, C.M., Palleti, V.R., Mathur, A.P.: 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, pp. 25–28 (2017)

[33] Abdulaal, A., Liu, Z., Lancewicki, T.: 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, New York, NY, USA (2021). https://doi.org/10.1145/3447548.3467174

[34] Hundman, K., Constantinou, V., Laporte, C., Colwell, I.: Telemanom: Spacecraft Anomaly Detection Data for SMAP and MSL. https://github.com/khundman/telemanom. Accessed 7 June 2026 (2018)

[35] NetManAIOps: Server Machine Dataset in OmniAnomaly. https://github.com/NetManAIOps/OmniAnomaly. Accessed 7 June 2026 (2019)

[36] iTrust, Singapore University of Technology and Design: Secure Water Treatment Dataset. https://www.sutd.edu.sg/itrust/itrust-labs/datasets/dataset-characteristics/swat/. Accessed 7 June 2026 (2026)

[37] iTrust, Singapore University of Technology and Design: Water Distribution Dataset. https://www.sutd.edu.sg/itrust/itrust-labs/datasets/dataset-characteristics/wadi/. Accessed 7 June 2026 (2026)

[38] eBay: RANSynCoders: Pooled Server Metrics Dataset. https://github.com/eBay/RANSynCoders. Accessed 7 June 2026 (2021)