

---

# Structured Temporal Causality for Interpretable Multivariate Time Series Anomaly Detection

---

Dongchan Cho\*    Jiho Han    Keumyeong Kang    Minsang Kim

Honggyu Ryu    Namsoon Jung

Industrial AI Lab, SimPlatform Co. Ltd. Affiliate Research Institute  
{cdc0213, jihohan, kykang, mskim, hgryu, nsj}@simplatform.com

## Abstract

Real-world multivariate time series anomalies are rare and often unlabeled. Additionally, prevailing methods rely on increasingly complex architectures tuned to benchmarks, detecting only fragments of anomalous segments and overstating performance. In this paper, we introduce OracleAD, a simple and interpretable unsupervised framework for multivariate time series anomaly detection. OracleAD encodes each variable’s past sequence into a single causal embedding to jointly predict the present time point and reconstruct the input window, effectively modeling temporal dynamics. These embeddings then undergo self-attention mechanism to project them into a shared latent space and capture spatial relationships. These relationships are not static, since they are modeled by a property that emerges from each variable’s temporal dynamics. The projected embeddings are aligned to a Stable Latent Structure (SLS) representing normal-state relationships. Anomalies are identified using a dual scoring mechanism based on prediction error and deviation from the SLS, enabling fine-grained anomaly diagnosis at each time point and across individual variables. Since any noticeable SLS deviation originates from embeddings that violate the learned temporal causality of normal data, OracleAD directly pinpoints the root-cause variables at the embedding level. OracleAD achieves state-of-the-art results across multiple real-world datasets and evaluation protocols, while remaining interpretable through SLS.

## 1 Introduction

Unsupervised anomaly detection in multivariate time series (MTSAD) [3, 46] underpins reliability and safety in domains such as industrial control systems [28, 31], healthcare monitoring [5, 13], and cyber-physical security [1]. In these settings, anomalies are rare, unlabeled, and context-dependent, posing challenges to conventional methods and requiring actionable insights to maintain operational reliability. Models must therefore not only flag irregularities but also explain their causal origins. Thus, effective MTSAD methods should (i) define anomalies within a multivariate temporal framework, (ii) model the temporal and spatial signals of anomaly formation, and (iii) integrate these signals for both detection and root-cause diagnosis.

Although intrinsic temporal and structural changes occur when anomalies emerge, most MTSAD methods overlook these signals. They often rely on reconstruction errors computed over fixed "normal" and "abnormal" windows, on superficial feature-level differences after arbitrary model transformations, or on frequency-domain representations divorced from causal time dynamics. More complex deep-model pipelines (e.g., dual-axis attention) further amplify this issue by creating artificial

---

\*Corresponding Author.

separations between normal and anomalous data through their complex forward-propagation outputs, rather than by explicitly modeling how anomalies form and spread. Such complexity rarely translates into practical gains, as these methods frequently fail to outperform simpler baselines or to justify their architectural overhead empirically or theoretically [33, 47]. Transformer-based approaches exemplify this misalignment by using large temporal contexts and bidirectional attention, which ignore the unidirectional, irreversible nature of time and compromise real-time deployment. Furthermore, recent studies [10, 47] demonstrate that compact causal architectures with lightweight designs can match or exceed Transformer performance across diverse forecasting benchmarks while greatly reducing computational costs.

At the core of these issues lies a fundamental misunderstanding regarding what naturally constitutes an anomaly in multivariate time series. We argue that anomalies inherently manifest through two interconnected signals. First, anomalies begin as prediction errors when a variable’s current state deviates from the expectations derived from its past states, signifying a breakdown in temporal causality within that variable. Second, temporal disruptions propagate into structural deviations by breaking the stable inter-variable relationships that are typically preserved under normal conditions. Thus, we explicitly define anomalies in multivariate time series as processes originating from a loss of temporal causality in specific variables, which subsequently alters the inter-variable relational structure.

Motivated by these perspectives, we propose **OracleAD**, an interpretable framework that captures temporal causality and leverages the natural signals inherent in multivariate anomalies. OracleAD processes the time series data with a sliding window of length  $L$ . For each variable  $i$ , a per-variable LSTM encoder processes its past  $L - 1$  observations to produce hidden states  $\{h_i^1, \dots, h_i^{L-1}\}$ . We then apply attention pooling over these hidden states to yield a single *causal embedding*  $c_i$ , which summarizes the temporal information necessary to predict the final timepoint  $x_i^L$  in the window. To instill temporal causality in the encoder, we jointly train it with two objectives: one-step point prediction of  $x_i^L$  and reconstruction of the window. The learned causal embeddings are projected into a shared latent space via a self-attention mechanism, where attention weights quantify dynamic interaction strengths between variables at each time point.

During training, OracleAD computes pairwise dissimilarities between projected causal embeddings across variables at each time step, forming a dissimilarity matrix that explicitly reflects the current spatial organization. At the end of each training epoch, these matrices are aggregated into a *Stable Latent Structure (SLS)*, which serves as a stable, statistically derived reference structure representing consistent inter-variable relationships under normal conditions. At inference time, OracleAD employs a dual scoring system. The *prediction score* quantifies the error between the actual and predicted states at each time step, signaling temporal irregularities that violate past-informed expectations. The *deviation score* captures structural inconsistencies by measuring the deviation between the current latent dissimilarity matrix and the SLS, flagging shifts in inter-variable relationships that deviate from learned causal structures. By integrating these complementary signals, OracleAD not only detects anomalies accurately but also transparently pinpoints root-cause variables at the embedding level, addressing critical limitations of existing methods and providing causal explanations of anomaly formation and propagation.

OracleAD demonstrates state-of-the-art anomaly detection performance across multiple real-world datasets. Experimental results show consistent improvements in detection accuracy, anomaly localization, and robustness under diverse evaluation protocols. Recognizing and explicitly modeling natural anomaly signals enables OracleAD to detect anomalies more efficiently using compact temporal windows. This approach is supported by recent findings [44], which show that anomalies strongly correlate with adjacent time points. OracleAD provides interpretability through a dual scoring system based on prediction error and structural deviation, enabling users to identify anomalies and trace their root causes at both the temporal and variable levels. By grounding anomaly detection in temporal causality and stable latent representations, OracleAD addresses core limitations of existing methods and offers a principled, practical direction for advancing multivariate time series anomaly detection.

## 2 Background

### 2.1 Unsupervised Multivariate Time Series Anomaly Detection

Early approaches to multivariate time-series anomaly detection primarily relied on classical outlier detection methods [4, 23, 36]. However, these early methods treated each time point as an independent multivariate sample and did not leverage the rich temporal dependencies inherent in sequential data. Subsequently, research gradually shifted toward deep sequence modeling based on reconstruction or prediction. EncDec-AD [27], OmniAnomaly [37], and recurrent ensemble variants [18] train autoencoder or variational models to reproduce normal behavior, flagging large reconstruction or prediction errors as anomalies. Although these approaches are conceptually simple and robust, they treat each channel independently, failing to capture inter-variable dependencies critical for complex fault patterns. MAD-GAN [20] extends this line by applying adversarial training via LSTM-based generator-discriminator networks, enabling joint modeling of temporal dynamics and inter-variable structures. However, unstable GAN optimization often limits robustness and interpretability.

To address this, explicit spatio-temporal modeling techniques emerged. MSCRED [48] uses ConvLSTM [35] to reconstruct correlation matrices over sliding windows, and InterFusion [21] employs hierarchical VAEs to disentangle temporal and spatial factors. Transformer-based architectures inject attention mechanisms into MTSAD. Anomaly Transformer [44] characterizes anomalies as discrepancies between learned attention weights and prior temporal associations, whereas TranAD [39] combines predictive and adversarial objectives to sharpen sensitivity to subtle deviations. Prior to this, Gangopadhyay et al. [12] proposed a spatio-temporal attention mechanism for interpretive forecasting in MTS, showing that attention can offer explanatory cues across time and variable dimensions. PatchTST [29] and SparseTSF [22] further improve efficiency by processing fixed-size patches or periodic samples, but at the expense of strict causal consistency and fine-grained localization. Graph-wise and dimension-wise attention methods capture variable interactions more flexibly [50]. GDN [8] learns a data-driven adjacency matrix to model dependencies (albeit static at inference), while Crossformer [49] and iTransformer [25] deploy multi-axis attention to fuse temporal and cross-variable signals, though large receptive fields may dilute local anomalies. Recently, Zheng et al. [51] introduced a correlation-aware spatial-temporal graph learning model. Contrastive and spectral approaches offer orthogonal gains: DCdetector [45] forgoes reconstruction entirely, using contrastive grouping of normal versus perturbed views; SARAD [6] adds regularization across adjacent subsequences; CATCH [43] isolates spectral anomalies via frequency-domain patching.

### 2.2 Limitations of Existing Benchmarks

Several widely used MTSAD benchmarks, such as SWaT [28], SMAP, and MSL [17], are frequently regarded as fully multivariate datasets. However, in practice, most real-world attacks and faults affect only a limited number of variables. For example, many scenarios in the SWaT dataset involve perturbations to only a small subset of sensors. Similarly, in the SMAP and MSL datasets, many sensor streams exhibit behavior that closely resembles step functions. A recent study by Liu et al. [24] demonstrates that converting these datasets into univariate formats results in comparable detection performance. This observation raises concerns regarding the actual necessity of using complex models that are specifically designed to capture cross-variable dependencies.

In addition to dataset limitations, the evaluation protocols used in MTSAD research exhibit several inconsistencies [19, 42]. The point-adjusted F1 score, for instance, tends to give excessive credit for any detection that falls within an anomaly window, even if the localization is inaccurate. Metrics such as AUC-ROC and AUC-PR are threshold-independent and widely used, but they often produce misleading results in highly imbalanced settings where true negatives dominate. The Affiliation F1 score [16] attempts to compensate for localization errors through post-processing, but it does not fundamentally resolve the problem of poor anomaly localization. More recently, researchers have proposed volume-under-surface metrics such as VUS-ROC and VUS-PR [24, 30], which jointly measure detection consistency and temporal precision across a range of thresholds.

Considering these limitations, it is difficult to regard strong performance on a single benchmark or metric as a reliable indicator of general robustness [2]. In order to provide an accurate assessment of MTSAD methods, it is necessary to conduct comprehensive evaluations that include multiple anomaly types, realistic variable correlations, and diverse metrics such as VUS-based measures.

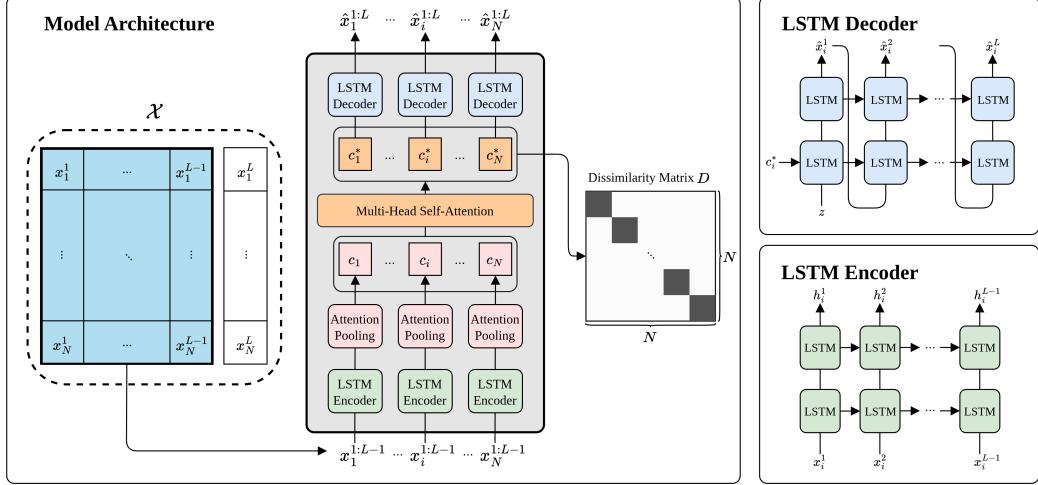


Figure 1: Overview of the OracleAD pipeline. Each variable’s past window is encoded by an LSTM and pooled via attention. The resulting embeddings are refined through multi-head self-attention to capture inter-variable dependencies. These embeddings are then fed into an LSTM decoder for reconstruction and next-step prediction. Finally, the pairwise distance matrix of the embeddings is compared against the Stable Latent Structure (SLS) to detect anomalies. The decoder uses an all-zero vector  $z$  as its initial input.

### 3 Method

Let  $\mathcal{X} \in \mathbb{R}^{N \times L}$  denote a fixed-length input window of multivariate observations, defined as

$$\mathcal{X} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^L\}, \quad \mathbf{x}^\ell \in \mathbb{R}^N, \quad (1)$$

where each window element is written as

$$\mathbf{x}^\ell = (x_1^\ell, x_2^\ell, \dots, x_N^\ell)^\top, \quad 1 \leq \ell \leq L, \quad (2)$$

and  $\mathbf{x}^L$  is the most recent observation vector. OracleAD determines whether  $\mathbf{x}^L$  is anomalous based on (i) its prediction error relative to temporal patterns learned within the window, and (ii) its deviation from the stable inter-variable structure observed under normal conditions.

#### 3.1 Temporal Causality Modeling

OracleAD models each variable independently using a per-variable encoder–decoder architecture. This contrasts with prior approaches [41, 44, 48] that process the multivariate sequence jointly, often with a shared model. However, in real-world settings, different variables often follow heterogeneous dynamics, and anomalies may affect only a subset of them. Shared architectures can entangle unrelated temporal patterns, hindering both detection sensitivity and interpretability. To address this, we adopt a modular structure where each variable learns its own causal dynamics.

As shown in Figure 1, for each variable segment  $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^{L-1})^\top \in \mathbb{R}^{L-1}$ , an LSTM [27] encoder  $\text{Enc}_i$  processes the input window and outputs a sequence of hidden states:

$$\{h_i^1, h_i^2, \dots, h_i^{L-1}\}, \quad h_i^l \in \mathbb{R}^d, \quad (3)$$

where  $d$  is the hidden dimension. To distill salient temporal information and suppress noise, we apply a learnable attention mechanism over the hidden states:

$$f(h_i^l) = w^\top h_i^l + b, \quad w \in \mathbb{R}^d, b \in \mathbb{R}, \quad (4)$$

followed by a softmax normalization over the temporal axis ( $l = 1, \dots, L - 1$ ):

$$\alpha_i^l = \text{softmax}(f(h_i^l)), \quad c_i = \sum_{l=1}^{L-1} \alpha_i^l h_i^l. \quad (5)$$

The resulting vector  $c_i \in \mathbb{R}^d$ , termed the causal embedding of variable  $i$ , aggregates its past hidden states weighted by their relevance in predicting  $x_i^L$ , thereby encoding the temporal causality inherent in that variable's history.

To incorporate inter-variable interactions, we apply Multi-Head Self-Attention (MHSA) [40] to the stacked embeddings

$$C = [c_1, \dots, c_N]^\top \in \mathbb{R}^{N \times d}, \quad (6)$$

yielding the attention-refined output

$$C^* = [c_1^*, \dots, c_N^*] \in \mathbb{R}^{N \times d}. \quad (7)$$

Each  $c_i^*$  is computed as

$$c_i^* = \sum_{h=1}^H W_h^O \text{softmax}\left(\frac{(W_h^Q c_i)(W_h^K C)^{\top}}{\sqrt{d_h}}\right) W_h^V C, \quad (8)$$

where  $H$  denotes the number of attention heads, and  $W_h^Q, W_h^K, W_h^V, W_h^O$  are learnable projection matrices for query  $Q$ , key  $K$ , value  $V$ , and output in head  $h$ . The per-variable embeddings  $\{c_1, \dots, c_N\}$  are first projected into a shared latent space via head-specific linear transformations. Subsequent multi-head self-attention mixes these projections, so each  $c_i^*$  absorbs contextual cues from all other variables and captures soft, dynamic dependencies without a predefined static graph.

These context-aware embeddings  $c_i^*$  are then passed into their corresponding LSTM decoders  $\text{Dec}_i$ , each trained to both reconstruct the past window and predict the next value:

$$\hat{x}_i^{1:L-1}, \hat{x}_i^L = \text{Dec}_i(c_i^*). \quad (9)$$

The outputs of each decoder are used to compute the reconstruction loss  $\|x_i^{1:L-1} - \hat{x}_i^{1:L-1}\|^2$  and the prediction loss  $\|x_i^L - \hat{x}_i^L\|^2$ , which together ensure that each causal embedding  $c_i$  faithfully captures its variable's temporal dynamics under a causal, windowed regime.

The final multivariate prediction at time  $L$  is assembled as

$$\hat{\mathbf{x}}^L = [\hat{x}_1^L, \hat{x}_2^L, \dots, \hat{x}_N^L] \in \mathbb{R}^N. \quad (10)$$

This per-variable framework guarantees robust modeling of the normal data distribution by (i) condensing all  $L - 1$  past steps into each causal embedding  $c_i$ , thereby covering the full spectrum of normal temporal patterns rather than overfitting to local snippets, (ii) employing independent encoder-decoder branches so that each  $c_i$  adapts to its variable's unique dynamics, (iii) fusing these embeddings via MHSA into context-aware vectors  $c_i^*$ , which learn the joint normal manifold and soft correlations across variables, and (iv) suppressing the influence of rare spikes through attention pooling and Stable Latent Structure (SLS)-based structural regularization.

### 3.2 Stable Latent Structure (SLS)

To detect structural anomalies and promote latent consistency during training, OracleAD maintains a reference structure called the **Stable Latent Structure (SLS)**. The SLS summarizes stable inter-variable relationships observed in the latent space during normal behavior. Unlike static graphs or manually defined priors, it is constructed directly from data using the attention-refined latent representations of individual variables.

Let  $c_i^* \in \mathbb{R}^d$  denote the latent vector of variable  $i$  ( $i = 1, \dots, N$ ) after temporal encoding and relational attention. For each training window  $k$ , we form a pairwise dissimilarity matrix

$$D_{ij}^{(k)} = \|c_i^{*(k)} - c_j^{*(k)}\|_2, \quad D^{(k)} \in \mathbb{R}^{N \times N}, \quad (11)$$

where we adopt the L2 distance because it accounts for both directional mismatch and latent energy differences in the embeddings. Empirically, L2 also provided more stable and discriminative results than alternatives such as cosine similarity or L1 distance (see Appendix E.4).

Aggregating over the  $M$  windows that appear in one training epoch yields the stable template

$$\mathbf{SLS} = \frac{1}{M} \sum_{k=1}^M D^{(k)}, \quad (12)$$

which provides a compact baseline of canonical inter-variable structure. Since the first epoch has no prior SLS to build upon, the updating process begins at its end, and the resulting SLS is then used from the second epoch onward. During training, it acts as a regularizer that encourages the encoders to preserve this relational pattern. At inference time, for each time step  $t$  in the entire input sequence, OracleAD computes the deviation matrix

$$\mathcal{D}_{\text{matrix}}^t = |D^t - \mathbf{SLS}|, \quad (13)$$

which measures element-wise deviations from the normal inter-variable structure. High-magnitude rows or columns in  $\mathcal{D}_{\text{matrix}}^t$  indicate variables that disrupt relational stability, enabling root-cause identification.

### 3.3 Training Objective

OracleAD is trained with a composite loss function that enforces temporal causality, reconstructs normal patterns, and regularizes inter-variable structure. The total loss is defined as

$$\mathcal{L} = \underbrace{\|\mathbf{x}^L - \hat{\mathbf{x}}^L\|^2}_{\text{prediction loss}} + \lambda_{\text{recon}} \cdot \underbrace{\|\mathbf{x}^{1:L-1} - \hat{\mathbf{x}}^{1:L-1}\|^2}_{\text{reconstruction loss}} + \lambda_{\text{dev}} \cdot \underbrace{\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (D_{ij} - \mathbf{SLS}_{ij})^2}_{\text{deviation loss}}, \quad (14)$$

where  $\mathbf{x}^\ell \in \mathbb{R}^N$  denotes the  $\ell$ -th element of the window and  $D$  is the dissimilarity matrix defined in Eq. 11.

The first term encourages each decoder to forecast the most recent observation  $\mathbf{x}^L$  using only the preceding window elements, thereby enforcing temporal causality. The second term guides the model to reconstruct the past window  $\mathbf{x}^{1:L-1}$ , preserving contextual memory and improving robustness to noise. This reconstruction term prevents the model from relying on shortcut predictions based only on local trends, forcing it instead to leverage the full temporal context within the window. The third term penalizes discrepancies between the current dissimilarity matrix  $D$  and the reference SLS from Eq. 12, ensuring that latent representations remain aligned with relational patterns observed under normal conditions.

The hyperparameters  $\lambda_{\text{recon}}$  and  $\lambda_{\text{dev}}$  (set to 0.1 and 3 by default, respectively) balance the reconstruction and structural regularization terms against the prediction objective, promoting both predictive accuracy and structural coherence. In the first epoch, the deviation loss is omitted because no SLS is available; it is included from the second epoch onward once the initial SLS has been constructed.

### 3.4 Anomaly Scoring

During inference, OracleAD computes two complementary scores for each time step  $t$  in the entire input sequence. The prediction score is defined as

$$\mathcal{P}_{\text{score}}^t = \frac{1}{N} \sum_{i=1}^N |x_i^t - \hat{x}_i^t|, \quad (15)$$

which measures the average absolute error between the observed variables and their reconstructions.

The deviation score is defined as

$$\mathcal{D}_{\text{score}}^t = \|D^t - \mathbf{SLS}\|_F, \quad (16)$$

where  $D^t$  is the pairwise dissimilarity matrix of the causal–context embeddings at time  $t$ , and  $\mathbf{SLS}$  is the stable relational template from Eq. 12. Here  $\|\cdot\|_F$  denotes the Frobenius norm.

The final anomaly score is then defined as

$$\mathcal{A}_{\text{score}}^t = \mathcal{P}_{\text{score}}^t \cdot \mathcal{D}_{\text{score}}^t, \quad (17)$$

which jointly captures both temporal and spatial anomalies in multivariate time series. By grounding detection in both temporal prediction and relational stability in latent space, OracleAD achieves enhanced sensitivity and interpretability compared to methods relying solely on univariate errors or static dependency graphs. The diagnostic role of each component is illustrated with real-world examples in Section 4.3.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We evaluate OracleAD on three widely adopted benchmark datasets: **SMD** [37], **PSM** [1], and **SWaT** [28]. These datasets span diverse industrial scenarios, including cloud server telemetry (SMD, 38 variables), real-world industrial sensors (PSM, 25 variables), and a water treatment testbed (SWaT, 51 variables). Each dataset contains labeled anomaly segments with varying temporal and spatial characteristics. Full details are provided in Appendix A.

**Evaluation Metrics.** To comprehensively evaluate model performance, we report results using seven metrics: standard F1 (not point-adjusted), range-based F1 (R-F1) [38], affiliation F1 (Aff-F1) [16], AUC-ROC (A-ROC) [11], AUC-PR (A-PR) [7], VUS-ROC (V-ROC), and VUS-PR (V-PR) [30]. These metrics capture different aspects of anomaly detection, including temporal accuracy, interval-level coherence, robustness to class imbalance, and threshold stability. More detailed descriptions of each metric are provided in Appendix B.

**Baselines.** We benchmark OracleAD against a diverse set of classical and deep learning baselines. For reconstruction-based method, **AutoEncoder (A.E)** [32] and **OmniAnomaly (Omni)** [37] are considered. Spatially-structured and attention-based deep models include **AnomalyTransformer (A.T)** [44], **DCdetector (DC)** [45], **SARAD** [6] and **CATCH** [43]. Forecasting-oriented approaches such as **PatchTST (Patch)** [29], **TimesNet (TsNet)** [41], **DLinear (DLin)** [47], **NLinear (NLin)** [47], **iTransformer (iTrans)** [25] and **ModernTCN (Modern)** [9] are also considered, due to their high performance in general time-series tasks. All baselines are run using publicly available implementations and reimplementations based on the original paper’s configurations. The implementation details of baselines are provided in Appendix C.

**Implementation Details** OracleAD is implemented in a fully unsupervised setting using only normal training data. We use a sliding window of length  $L = 10$  for all datasets, and exclude the first  $L - 1$  time steps of the test data from evaluation. Following the recommendations from recent benchmark studies [15, 24, 34], we consider thresholding to be orthogonal to the quality of the model itself. Therefore, we either report threshold-independent metrics or evaluate detection performance under an optimal threshold to better isolate the effectiveness of the anomaly scoring function. Further details, sensitivity analysis of hyperparameters, and computational cost are provided in Appendix D.

Table 1: Performance (%) of deep anomaly detection models across seven evaluation metrics on three benchmark datasets (PSM, SMD, SWaT). All results are averaged over five random seeds. **Boldface** indicates the best score and underline indicates the second-best score.

Dataset	Metric	A.E	Omni	A.T	Patch	TsNet	DLin	NLin	DC	iTrans	Modern	SARAD	CATCH	OracleAD	
PSM	F1	47.55	<u>45.90</u>	43.45	43.45	43.45	43.45	43.45	43.45	43.45	45.75	44.33	<b>65.85</b>		
	R-F1	45.93	<u>40.58</u>	12.17	34.75	37.49	34.74	32.10	21.35	35.91	35.48	13.54	<u>54.19</u>	<b>54.66</b>	
	Aff-F1	73.70	73.49	69.68	69.43	69.43	69.43	69.59	69.53	69.43	77.12	<b>79.16</b>	78.07		
	A-ROC	<u>66.79</u>	63.95	38.35	58.68	59.09	58.02	58.56	49.86	59.22	59.21	62.86	64.75	<b>84.78</b>	
	A-PR	47.27	45.00	24.28	37.91	38.95	37.17	37.64	27.78	38.27	38.59	41.67	43.40	<b>68.11</b>	
	V-ROC	<u>68.20</u>	61.23	50.96	50.62	50.67	50.53	50.69	50.11	51.23	50.68	57.63	57.00	<b>84.24</b>	
SMD	V-PR	49.66	<u>52.49</u>	49.76	50.31	49.86	49.93	46.71	27.89	44.54	50.55	38.64	45.95	<b>68.17</b>	
	F1	25.78	<u>32.16</u>	7.98	7.98	7.98	7.98	8.00	7.98	7.98	25.92	7.98	<b>43.03</b>		
	R-F1	28.09	<u>35.81</u>	5.84	13.81	9.43	13.37	13.51	8.23	9.02	14.43	10.35	2.90	<b>38.88</b>	
	Aff-F1	83.34	<u>82.57</u>	67.43	67.43	67.43	67.43	67.49	67.43	67.43	75.87	67.43	<b>84.73</b>		
	A-ROC	76.80	71.25	50.40	73.85	59.09	72.73	73.83	49.97	74.57	72.13	72.84	<u>80.96</u>	<b>83.56</b>	
	A-PR	19.40	27.73	4.57	14.82	13.66	13.88	14.10	4.16	14.57	13.01	25.87	17.09	<b>44.83</b>	
SWaT	V-ROC	<b>74.38</b>	<u>73.26</u>	50.34	51.66	51.10	51.52	51.76	49.92	51.76	51.55	65.57	50.95	69.57	
	V-PR	22.50	31.18	36.86	<u>41.60</u>	41.21	41.26	39.15	4.27	32.85	41.54	19.33	35.25	<b>47.52</b>	
	F1	74.46	<u>75.40</u>	21.65	21.65	7.98	12.14	21.65	21.65	21.65	57.30	21.65	<b>76.50</b>		
	R-F1	36.63	<b>38.72</b>	15.48	12.23	20.80	5.13	12.06	15.54	11.89	12.88	21.37	14.74	28.15	
	Aff-F1	<b>75.30</b>	72.09	70.06	69.51	73.47	67.43	69.68	69.25	69.25	72.08	<u>74.88</u>	73.68	71.97	
	A-ROC	81.97	82.15	42.58	24.36	28.97	50.48	23.31	49.89	24.57	24.50	<b>85.40</b>	33.31	<b>82.71</b>	
SWaT	A-PR	67.51	<b>72.73</b>	11.93	8.40	10.71	4.53	8.00	12.14	8.33	9.00	64.77	13.39	<u>72.39</u>	
	V-ROC	81.89	81.98	50.72	49.84	20.80	50.36	49.02	50.00	49.90	50.15	<b>86.30</b>	51.89	82.42	
	V-PR	65.89	64.42	17.00	11.92	51.20	28.47	10.55	11.92	12.05	13.05	62.72	18.70	<b>74.16</b>	

## 4.2 Main Results

As shown in Table 1, OracleAD achieves the highest scores across most of the reported metrics on SMD (average over all subsets), PSM, and SWaT. This demonstrates strong robustness against diverse anomaly types and dimension of data.

Among the various metrics, we emphasize the standard point-wise F1 score, which directly reflects alignment between detected anomalies and ground truth. OracleAD consistently outperforms all baselines with large margins in F1 (PSM:+19.95%pt, SMD:+10.87%pt, SWaT:+0.9%pt), confirming its ability to clearly separate anomalous from normal points. These high scores are often accompanied by strong performance in threshold-independent metrics (AUC, VUS). Recent benchmarks [24] identify VUS-PR as a key indicator of detection consistency and localization quality. In this metric, OracleAD significantly outperforms all baselines (PSM:+15.68%pt, SMD:+5.92%pt, SWaT:+8.27%pt), which underscores its superior anomaly localization.

In contrast, Affiliation F1 tends to yield high scores across most models regardless of their performance on other metrics. This reveals a key limitation of the metric, Affiliation F1 rewards predictions made near the anomaly segment, even if they do not directly overlap with the ground truth. Experiments in [24] demonstrate that models can score highly on affiliation even with random predictions or by detecting only a small part of an anomaly segment. This explains why some baselines (e.g., A.T, CATCH) exhibit high affiliation F1 despite poor F1, AUC-PR, and VUS-PR, indicating their failure to fully capture anomaly regions. Overall, these results confirm OracleAD’s dual-score design delivers both sharp detection and robust localization across diverse scenarios.

## 4.3 Visual Interpretation and Anomaly Diagnosis

A key interpretability component is the deviation matrix  $\mathcal{D}_{\text{matrix}}^t = |D^t - \mathbf{SLS}|$ , which highlights each variable’s disruption of the learned structure. Figure 3 visualizes deviation matrices at representative anomaly timestamps from the SMD dataset. For example, variables 32, 33 in Figure 3(a) and 10, 15, 18 in Figure 3(d) show dominant deviations. Their corresponding raw signals and scoring trends in Figure 2 validate these candidates, as each shows strong alignment with prediction peaks and anomaly score elevations, confirming OracleAD’s ability to localize root causes through unsupervised latent analysis. A practical heuristic is that variables with multiple highlighted rows or columns are more likely to be root causes, as their relational patterns remain consistently perturbed across time. When arranged sequentially, these matrices reveal how the structural disruptions of key variables evolve over the course of an anomaly segment. Quantitative analysis based on this row-wise aggregation and more examples are provided in Appendix G.

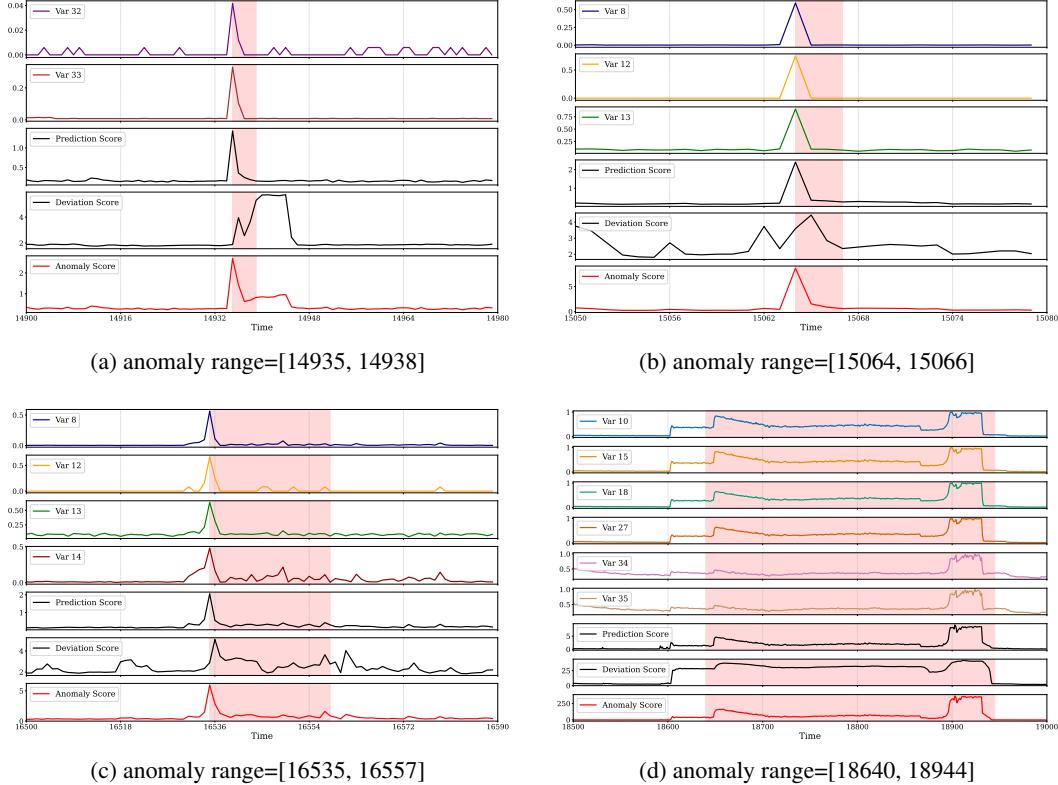


Figure 2: Raw signals of the identified root-cause variables and anomaly score visualization for selected intervals in the SMD dataset. Shaded red regions indicate ground-truth anomalies.

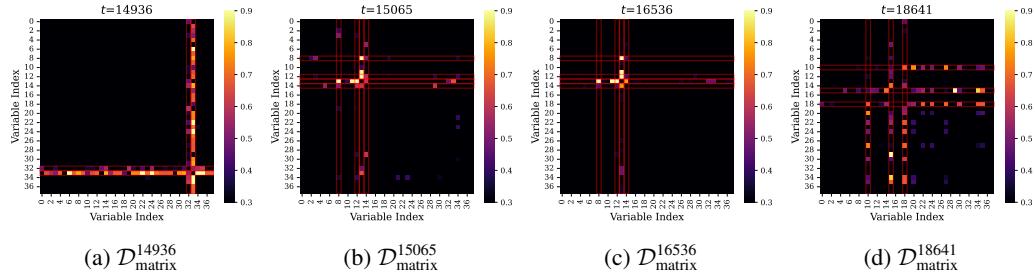


Figure 3: Visualization for Deviation Matrices  $D_{\text{matrix}}^t$  at anomalous time points in SMD dataset. Each matrix corresponds respectively to the subfigures in Figure 2. The color bar of each heatmap represents the magnitude of pairwise dissimilarity between variables. A bigger row over column value means a higher chance of an anomaly in that variable. For visualization, each matrix is independently min–max normalized and values are clipped to the range [0.3, 0.9] to enhance contrast.

As shown in Figure 2, OracleAD combines prediction and deviation scores into a final anomaly score. The prediction score measures point-level discrepancies between observation and forecast, responding sharply to sudden deviations and achieving high temporal precision. However, its response is typically brief and localized to the moment of disruption, potentially missing prolonged structural anomalies. In contrast, the deviation score compares the current dissimilarity matrix to the Stable Latent Structure (SLS) and captures longer-term relational disturbances (see Figure 2(d)) that persist even after the raw signal stabilizes. This property aligns with real-world anomaly labeling, where fault intervals span the full duration of a system disturbance rather than just its initiation. Consequently, the deviation score remains elevated across an entire fault segment, reflecting sustained relational instability.

Nonetheless, the deviation score may exhibit a slight temporal lag because structural disruption can only be captured after anomalous patterns have entered the input window. This means the effect of a recent anomaly is only reflected after it enters the window, and the score can also be

influenced by residual effects from recent anomalies (see Figure 2(a)), potentially leading to false positives in recovery periods. The multiplicative combination of these two scores enables OracleAD to balance their respective strengths and mitigate their individual weaknesses. While the prediction score anchors detection to the precise moment of temporal disruption, the deviation score extends sensitivity across the whole anomaly segment. This joint scoring mechanism also suppresses false positives and recovers false negatives. High deviation scores that arise in non-anomalous contexts are suppressed by low prediction error, while false negatives from prediction-only detection are compensated by the sustained elevation of structural deviations. Together, they offer both sharpness and stability, enabling OracleAD to detect anomalies with improved precision, temporal coverage, and interpretability across diverse anomaly types and durations. More case studies are provided in appendix F.

#### 4.4 Ablation Study

Table 2: Ablation study results (%) on the PSM, SMD, and SWaT datasets. All results are averaged over five random seeds.

Component	Variant	PSM			SMD			SWaT		
		F1	V-ROC	V-PR	F1	V-ROC	V-PR	F1	V-ROC	V-PR
Loss Function(Sec 3.3)	w/o Reconstruction Loss	58.03	74.56	54.40	56.47	75.59	54.29	76.61	82.92	71.95
Anomaly Score(Sec 3.4)	$\mathcal{A}_{\text{score}}^t = \mathcal{D}_{\text{score}}^t$	59.06	75.78	56.11	47.32	77.50	37.02	76.92	81.39	70.77
	$\mathcal{A}_{\text{score}}^t = \mathcal{P}_{\text{score}}^t$	55.33	60.99	60.99	58.98	87.39	53.71	70.49	82.68	68.50
<b>OracleAD (Full Model)</b>		65.85	84.24	68.17	60.19	80.81	56.63	76.50	82.42	74.16

We conduct an ablation study to evaluate the contributions of reconstruction loss and anomaly scoring strategies. Results on PSM, SWaT, and SMD (average of 4 subsets) are summarized in Table 2. Per-subset results of SMD are provided in Appendix E.

**Loss Function.** The full model achieves the highest performance across most of the datasets and metrics. Removing the reconstruction loss results in moderate degradation, especially in PSM (F1:-7.82%pt, V-ROC:-9.68%pt, V-PR:-13.77%pt) and SMD (F1:-3.72%pt, V-ROC:-5.22%pt, V-PR:-2.34%pt), indicating that reconstruction helps stabilize temporal representations and improve robustness.

**Anomaly Score.** While deviation-only scoring ( $\mathcal{A}_{\text{score}}^t = \mathcal{D}_{\text{score}}^t$ ) performs well on SWaT, it underperforms on PSM and SMD, likely because it lacks temporal context. Prediction-only scoring ( $\mathcal{A}_{\text{score}}^t = \mathcal{P}_{\text{score}}^t$ ) can detect sharp temporal deviations but fails to capture relational disruptions, leading to weak performance on SWaT (F1:-6.01%pt, V-PR:-6.66%pt). As discussed in Section 2.2, SWaT anomalies often affect only a small subset of variables, which can limit the effectiveness of prediction-based scoring. However, structural disruptions caused by these localized anomalies still alter inter-variable relationships, allowing deviation-based scoring to retain strong performance. The combined scoring ( $\mathcal{A}_{\text{score}}^t = \mathcal{P}_{\text{score}}^t \cdot \mathcal{D}_{\text{score}}^t$ ) consistently performs best overall, confirming the complementarity of temporal and structural dimensions in OracleAD’s design.

## 5 Conclusion

We introduced **OracleAD**, an unsupervised framework that jointly models temporal causality and inter-variable structure through a unified latent representation. OracleAD derives a compact causal embedding for each variable, aligns them via attention into a shared latent space, and detects anomalies through deviations from a learned Stable Latent Structure (SLS).

OracleAD achieves state-of-the-art performance across multiple real-world datasets and evaluation protocols, demonstrating strong robustness, localization accuracy, and interpretability. Its dual scoring mechanism captures both prediction errors and structural inconsistencies, enabling precise detection and fine-grained root-cause analysis even without labels. We also found limitations in our framework when applied to complex or multimodal systems, where globally consistent inter-variable relationships and continuous input assumptions may not hold.

Our paper highlights the importance of grounding anomaly detection in temporal causality and latent relational consistency, establishing a principled foundation for generalizable and interpretable multivariate time series analysis.

## References

- [1] Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical Approach to Asynchronous Multivariate Time Series Anomaly Detection and Localization. In *Proceedings of the 27th ACM SIGKDD Conference on KDD*, pages 2485–2494. ACM, August 2021.
- [2] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Do deep neural networks contribute to multivariate time series anomaly detection? *Pattern Recognition*, 132:108945, 2022.
- [3] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, 54(3):1–33, 2021.
- [4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.
- [5] Sucheta Chauhan and Lovekesh Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. In *2015 IEEE international conference on DSAA*, pages 1–7. IEEE, 2015.
- [6] Zhihao Dai, Ligang He, Shuanghua Yang, and Matthew Leeke. SARAD: Spatial Association-Aware Anomaly Detection and Diagnosis for Multivariate Time Series. In *The 38th Annual Conference on NeurIPS*, November 2024.
- [7] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *Proc. of the 23rd ICML*, pages 233–240. ACM, 2006.
- [8] Ailin Deng and Bryan Hooi. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. *Proc. of the AAAI Conference on Artificial Intelligence*, 35(5):4027–4035, May 2021.
- [9] Luo Donghao and Wang Xue. ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. In *Proc. of ICLR*, October 2023.
- [10] Vijay Ekambaram et al. Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series. In *The 38th Annual Conference on NeurIPS*, November 2024.
- [11] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.
- [12] Tryambak Gangopadhyay, Sin Yong Tan, Zhanhong Jiang, Rui Meng, and Soumik Sarkar. Spatiotemporal attention for multivariate time series prediction and interpretation. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 3560–3564. IEEE, 2021.
- [13] Ary L. Goldberger et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220, June 2000.
- [14] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.
- [15] Mononito Goswami, Cristian Ignacio Challu, Laurent Callot, Lenon Minorics, and Andrey Kan. Unsupervised Model Selection for Time Series Anomaly Detection. In *The 11th ICLR*, September 2022.
- [16] Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local Evaluation of Time Series Anomaly Detection Algorithms. In *Proc. of the 28th ACM SIGKDD Conference on KDD*, pages 635–645. ACM, August 2022.
- [17] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. In *Proc. of the 24th ACM SIGKDD International Conference on KDD*, pages 387–395. ACM, July 2018.

- [18] Tung Kieu, Bin Yang, Chenjuan Guo, and Christian S Jensen. Outlier detection for time series with recurrent autoencoder ensembles. In *IJCAI*, pages 2725–2732, 2019.
- [19] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7194–7201, 2022.
- [20] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks*, pages 703–716. Springer, 2019.
- [21] Zhihan Li et al. Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding. In *Proc. of the 27th ACM SIGKDD Conference on KDD*, pages 3220–3230. ACM, August 2021.
- [22] Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. SparseTSF: Modeling Long-term Time Series Forecasting with 1k Parameters. In *Proc. of Machine Learning Research*, volume 202, June 2024.
- [23] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 8th IEEE ICDM*, pages 413–422. IEEE, December 2008.
- [24] Qinghua Liu and John Paparrizos. The Elephant in the Room: Towards A Reliable Time-Series Anomaly Detection Benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in NeurIPS*, volume 37, pages 108231–108261. Curran Associates, Inc., 2024.
- [25] Yong Liu et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *Proc. of ICLR*, October 2023.
- [26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [27] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. In *ICML2016ADW*, 2016.
- [28] Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: A Water Treatment Testbed for Research and Training on ICS Security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pages 31–36. IEEE, April 2016.
- [29] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *Proc. of ICLR*, September 2022.
- [30] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. *Proc. of the VLDB Endow.*, 15(11):2774–2787, July 2022.
- [31] Cory A. Rieth, Ben D. Amsel, Randy Tran, and Maia B. Cook. Issues and Advances in Anomaly Detection Evaluation for Joint Human-Automated Systems. In Jessie Chen, editor, *AISC*, volume 595, pages 52–63. Springer International Publishing, 2018.
- [32] Mayu Sakurada and Takehisa Yairi. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proc. of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11. ACM, December 2014. ISBN 978-1-4503-3159-3. doi: 10.1145/2689746.2689747.
- [33] M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, and Marios Koulakis. Position: Quo Vadis, Unsupervised Time Series Anomaly Detection? In *Proc. of the 41st ICML*, pages 43461–43476. PMLR, July 2024.
- [34] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly Detection in Time Series: A Comprehensive Evaluation. *Proc. of the VLDB Endow.*, 15(9):1779–1797, May 2022.

- [35] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
- [36] 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 foundations and new directions of data mining workshop*, pages 172–179. IEEE Press Piscataway, NJ, USA, 2003.
- [37] 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 *Proc. of the 25th ACM SIGKDD International Conference on KDD*, KDD ’19, pages 2828–2837. Association for Computing Machinery, July 2019.
- [38] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. Precision and Recall for Time Series. In *Advances in NeurIPS*, volume 31. Curran Associates, Inc., 2018.
- [39] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proc. of the VLDB Endow.*, 15(6):1201–1214, February 2022.
- [40] Ashish Vaswani et al. Attention is All you Need. In *Advances in NeurIPS*, volume 30. Curran Associates, Inc., 2017.
- [41] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *Proc. of ICLR*, September 2022.
- [42] Renjie Wu and Eamonn J Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE transactions on knowledge and data engineering*, 35(3):2421–2429, 2021.
- [43] Xingjian Wu et al. CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching. In *Proc. of ICLR*, February 2025.
- [44] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *Proc. of ICLR*, October 2021.
- [45] Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection. In *Proc. of the 29th ACM SIGKDD Conference on KDD*, pages 3033–3045. ACM, August 2023.
- [46] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024.
- [47] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting? *Proc. of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, June 2023.
- [48] Chuxu Zhang et al. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. *Proc. of the AAAI Conference on Artificial Intelligence*, 33(01):1409–1416, July 2019.
- [49] Yunhao Zhang and Junchi Yan. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *Proc. of ICLR*, September 2022.
- [50] 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 *2020 IEEE international conference on data mining (ICDM)*, pages 841–850. IEEE, 2020.
- [51] Yu Zheng, Huan Yee Koh, Ming Jin, Lianhua Chi, Khoa T Phan, Shirui Pan, Yi-Ping Phoebe Chen, and Wei Xiang. Correlation-aware spatial–temporal graph learning for multivariate time-series anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):11802–11816, 2023.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Our research domain, the problem setting within that domain, and our perspective and approach to tackling the problem are clearly articulated in both the abstract and the introduction. The main claims made are consistent with the contributions demonstrated throughout the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of the work are discussed in Section 5 and Appendix I. We address assumptions made by our method, potential edge cases, and its scope of applicability in real-world scenarios.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results such as theorems or formal proofs. The focus of the work is on architectural design, empirical validation, and interpretability of an unsupervised anomaly detection method.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide full details necessary for reproducibility in Appendix A, C and D, including training settings, hyperparameters, and dataset configurations used for our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All datasets used in the paper are publicly available, and full details for accessing and preprocessing them are provided. While our code is not yet released due to internal procedures, we plan to release the official implementation to support reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.1 and Appendix D provide all relevant training and evaluation details, including data splits, model hyperparameters, optimizer settings, and selection strategies, which are sufficient to understand and reproduce the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experiments were conducted using five different random seeds. We report the mean performance across runs, and use standard deviation as a measure of variability. This is documented in Appendix E.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix D.4 and Appendix D.5 provides details on the computational resources used, including GPU type and memory specifications.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research complies fully with the NeurIPS Code of Ethics. It does not involve human subjects, personal data, or high-risk models, and adheres to ethical standards for reproducibility and transparency.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix H includes a discussion of both potential positive impacts of our work, including its applicability to industrial systems.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve models or datasets with high risk for misuse. All datasets used are publicly available benchmarks, and the proposed method is not designed for generative or dual-use purposes.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We provide proper citations and baseline code implementations used in our work. These details are included in Appendix A and C

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce new datasets, models, or other assets that require documentation. It builds upon existing publicly available datasets and frameworks.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve any research with human subjects and therefore does not require IRB approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core methodology does not involve large language models (LLMs) as an essential or original component. Any LLM usage, if any, was limited to non-substantive tasks such as editing or formatting.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

---

# Appendix for "Structured Temporal Causality for Interpretable Multivariate Time Series Anomaly Detection"

---

## A Datasets

### A.1 Dataset Statistics

Table S1: Statistics of the benchmark datasets used for multivariate time-series anomaly detection.

Dataset	Subsets	Features	Avg Set Size		Anomalies		Segment Lengths			Sampling
			Train	Test	Count	Ratio	Min	Med	Max	
PSM	1	25	132,481	87,841	71	27.73%	1	5	8,861	1 min
SMD	28	38	25,300	25,301	327	4.16%	2	11	3,161	1 min
SWaT	1	51	496,800	449,919	34	12.02%	101	447	35,900	1 sec

Table S1 summarizes the key characteristics of the benchmark datasets used in our experiments. These datasets differ substantially in terms of dimensionality, time series length, anomaly density, and segment structure.

**PSM** features relatively few but lengthy anomaly segments, resulting in a high anomaly ratio despite a limited number of anomalous events. In contrast, **SMD** comprises numerous short anomaly segments with a low overall anomaly density, making it sensitive to brief temporal disruptions. **SWaT** falls between these extremes, containing a moderate number of anomalies with segment lengths ranging from under 200 to tens of thousands of timestamps.

Together, these datasets span a spectrum of dense vs. sparse anomalies, short vs. long-term disruptions, and uni-source vs. multi-entity dynamics. This diversity allows comprehensive evaluation of model robustness under heterogeneous anomaly scenarios.

### A.2 Dataset Descriptions

**PSM (Pooled Server Metrics) [1]** The PSM dataset comprises server-side metrics collected from production servers at eBay. It includes 25 variables recorded over approximately 21 weeks, totaling 132,481 training points and 87,841 testing points. The test set contains both short-duration spikes and prolonged system-level anomalies, with an anomaly ratio of 27.73%. The combination of high density and mixed-length segments makes PSM a challenging benchmark.

**SMD (Server Machine Dataset) [37]** The SMD comprises univariate telemetry from 28 individual machines, each monitored with 38 variables. Every machine's data spans 10 days, with the first 5 days containing only normal behavior (training set) and the last 5 days including injected anomalies (testing set), resulting in an average anomaly rate of 4.16%.

Following the original authors' recommendation, we treat each machine as an independent subset and train a separate model per subset. This approach respects the natural isolation of system dynamics and failure modes across machines. For ablation studies and hyperparameter tuning, we exclusively utilize a few selected subsets, without overlap with those used for final evaluation.

**SWaT (Secure Water Treatment) [28]** The SWaT dataset captures sensor and actuator data from a realistic water treatment plant testbed. It contains 51 variables over 11 days, with the first 7 days considered normal and the remaining 4 days including 41 attack scenarios. Around 12% of the test points are labeled anomalous. SWaT's closed-loop dynamics and complex multivariate dependencies

make it especially suitable for evaluating the causal reasoning and interpretability capabilities of detection models.

### A.3 Dataset Sources

We use three widely adopted, publicly available MTSAD datasets. Each dataset is accessible for academic research, subject to the terms listed below:

- **PSM (Pooled Server Metrics):** Released by eBay and available at <https://github.com/netflix/psm-dataset>. Although no explicit license is provided, the dataset is publicly hosted and widely referenced in research.
- **SMD (Server Machine Dataset):** First introduced in the OmniAnomaly paper [37], the dataset is hosted at <https://github.com/NetManAIOps/OmniAnomaly> under the MIT License.
- **SWaT (Secure Water Treatment):** Developed by iTrust, SUTD. The dataset requires a formal access request via <https://itrust.sutd.edu.sg/>. Use is permitted under a non-commercial academic research agreement.

## B Metrics

Table S2: Comparison of evaluation metrics for time-series anomaly detection. ‘Partial’ denotes limited or conditional support depending on implementation details (e.g., decay functions, overlap thresholds).

Metric	Granularity	Threshold-Free	Temporal-Aware	Partial Credit	Imbalance-Robust
Standard F1	Point-wise	No	No	No	Partial
Range-based F1	Segment-wise	No	Partial	Yes	Partial
Affiliation F1	Segment-wise	No	Yes	Yes	Partial
AUC-ROC	Point-wise	Yes	No	No	No
AUC-PR	Point-wise	Yes	No	No	Yes
VUS-ROC	Point+Time	Yes	Yes	Yes	Yes
VUS-PR	Point+Time	Yes	Yes	Yes	Yes

Table S2 summarizes the core properties of evaluation metrics commonly used in time-series anomaly detection. Each metric captures different facets of detection quality, including temporal alignment, threshold sensitivity, and tolerance to misclassification. Notably, no single metric is universally optimal: metrics vary in granularity (point- vs. segment-level), and in their capacity to reflect delay, calibration, or imbalance.

In our view, using multiple complementary metrics is essential for a fair and comprehensive assessment. Different datasets, anomaly types, and operational goals may favor different evaluation criteria. For instance, segment-based metrics may benefit coarse anomaly annotations, while ranking-based metrics offer insight into score separation. However, overreliance on permissive or post hoc-adjusted metrics may mask modeling deficiencies, especially in time-sensitive applications.

We assume that multivariate time-series anomaly detection (MTSAD) demands stricter evaluation. Point-wise and delay-sensitive metrics, such as Standard F1 and VUS, impose higher requirements for detection precision, temporal consistency, and calibration. While these metrics may be unforgiving, they serve as catalysts for improving the model itself, encouraging detection systems that are not merely explainable or reactive, but fundamentally precise.

### B.1 Standard F1 Score (Point-wise)

The standard F1 score is defined as the harmonic mean of point-level precision and recall:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad (\text{S1})$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (\text{S2})$$

Despite its well-known limitations, F1 remains one of the most rigorous indicators of model capability in pointwise anomaly detection. Evaluating each time step independently avoids assumptions about anomaly structure and instead forces the model to separate normal and anomalous scores clearly. This property can be harsh for weak models, but acts as a fair and objective measure for well-calibrated ones.

Temporal fragmentation and misalignment are often cited as weaknesses of F1. However, we argue that a strong model should inherently produce temporally coherent anomaly scores, without relying on the metric to compensate for structural noise. In this view, fragmentation bias is not a failure of the metric, but a meaningful signal about a model’s inability to maintain temporal consistency.

F1 does require thresholding, which introduces variability across datasets. However, in controlled evaluation settings, identifying the optimal threshold gives insight into a model’s best achievable distinction and reflects the strength of score contrast between normal and abnormal patterns.

Finally, F1 is partially robust to class imbalance due to its exclusion of true negatives, though its reliability degrades in extremely sparse regimes. Overall, we consider F1 a high-precision, low-tolerance metric that is particularly suitable for isolating model quality in tightly controlled, point-level evaluations.

## B.2 Range-based F1 Score [38]

We adopt a range-aware variant of the F1 score implemented in the TSB-AD benchmark suite [24], which provides a finer-grained assessment of segment-level anomaly detection. Unlike standard range-based F1 that uses binary overlap rules to match predicted and ground-truth segments, this version assigns *partial credit* based on three interpretable components: existence, overlap, and cardinality.

The range-based recall is defined as:

$$\text{Recall}_{\text{range}} = \frac{1}{|\mathcal{R}|} \sum_{i=1}^{|\mathcal{R}|} [\omega(R_i, \mathcal{P}) \cdot \text{CF}(R_i, \mathcal{P})] + \alpha \cdot \text{ER}(\mathcal{R}, \mathcal{P}), \quad (\text{S3})$$

and  $\text{Precision}_{\text{range}}$  is computed symmetrically by swapping the roles of predictions and ground truth (with  $\alpha = 0$ ). Here,  $\mathcal{P} = \{p_1, \dots, p_m\}$  denotes the set of predicted anomaly time steps, and  $\mathcal{R} = \{R_1, \dots, R_{|\mathcal{R}|}\}$  the set of ground-truth anomaly ranges.

For each ground-truth range  $R_i$ , a weighted overlap score  $\omega(R_i, \mathcal{P})$  measures the fraction of the range covered by predictions using a bias-adjusted coverage function. A *cardinality factor* CF penalizes cases where multiple predicted segments correspond to a single ground-truth anomaly, while an *existence reward* ER grants partial credit if any prediction intersects the range. The parameter  $\alpha \in [0, 1]$  balances the overlap and existence terms.

The final score is computed as the harmonic mean:

$$\text{F1}_{\text{range}} = \frac{2 \cdot \text{Precision}_{\text{range}} \cdot \text{Recall}_{\text{range}}}{\text{Precision}_{\text{range}} + \text{Recall}_{\text{range}}}. \quad (\text{S4})$$

While this formulation enables smoother credit assignment and interpretable decomposition of detection quality, its reliance on hyperparameters (e.g., bias function,  $\alpha$ ) and the aggregation procedure makes it sensitive to implementation details. Therefore, standardized parameter settings are essential for fair cross-method comparisons.

## B.3 Affiliation F1 Score [16]

The Affiliation F1 score evaluates prediction quality by assigning partial credit based on temporal proximity between predictions and ground-truth anomaly segments. Using soft decay functions  $f(\cdot)$ , it rewards predictions that are “close enough,” even if not perfectly aligned.

The score is computed as:

$$\text{Precision}_{\text{aff}} = \frac{1}{m} \sum_{j=1}^m f_p \left( \min_{R_i \in \mathcal{R}} d(p_j, R_i) \right), \quad \text{Recall}_{\text{aff}} = \frac{1}{|\mathcal{R}|} \sum_{i=1}^{|\mathcal{R}|} f_r \left( \min_{p_j \in \mathcal{P}} d(R_i, p_j) \right), \quad (\text{S5})$$

$$\text{F1}_{\text{aff}} = \frac{2 \cdot \text{Precision}_{\text{aff}} \cdot \text{Recall}_{\text{aff}}}{\text{Precision}_{\text{aff}} + \text{Recall}_{\text{aff}}}, \quad (\text{S6})$$

where function  $d(p_j, R_i)$  computes the temporal distance between prediction  $p_j$  and range  $R_i$ .

Though often described as temporally sensitive, Affiliation F1 tends to produce inflated scores even when predictions are loosely aligned with anomaly segments. The decay function  $f(\cdot)$ , which grants high credit to any prediction within a short temporal window, effectively smooths over temporal errors. As a result, predicting only a small portion of a segment or merely grazing its edge may yield disproportionately high F1.

This behavior is analogous to post hoc adjustments like point- or range-based scoring. It does not penalize imprecise or fragmented detection, and may mask deficiencies in a model’s ability to produce temporally coherent scores. While this soft matching is useful when ground-truth labels are coarse or ambiguous, we argue that it should be treated as a model-independent alignment compensator, not a primary measure of detection capability.

In our view, a high-performing model should not rely on the metric to explain away timing errors. Affiliation F1 may serve as a secondary diagnostic tool to assess tolerance to misalignment, but we do not consider it sufficient for rigorous performance evaluation.

## B.4 AUC-ROC [11] and AUC-PR [7]

AUC-ROC and AUC-PR measure the threshold-independent ranking quality of anomaly scores. For a continuous anomaly score function, they integrate model performance across all thresholds.

The AUC-ROC computes the area under the TPR–FPR curve:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}, \quad (\text{S7})$$

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t) d(\text{FPR}(t)), \quad (\text{S8})$$

while AUC-PR evaluates the precision–recall trade-off:

$$\text{AUC-PR} = \int_0^1 \text{Precision}(r) d(\text{Recall}(r)). \quad (\text{S9})$$

Although widely adopted, these metrics ignore the critical temporal structure in time-series anomaly detection (TSAD). In particular, they are insensitive to detection continuity, localization lag, and whether anomalies are detected early, partially, or at all. As emphasized by Liu et al. [24], a model that correctly ranks only a single point in a long anomaly segment can still achieve a high AUC, even if it fails to detect the anomaly holistically.

Additionally, AUC-ROC is known to be unreliable in highly imbalanced settings, since the false positive rate denominator includes the large pool of true negatives. This allows naive models to achieve deceptively high AUC-ROC scores by merely avoiding false positives, without demonstrating meaningful detection. AUC-PR partially addresses this by focusing only on the positive class, making it more appropriate for sparse anomaly regimes.

Nonetheless, both AUC metrics are temporally agnostic and may obscure model behavior over time. We report them for completeness, but caution against over-reliance in TSAD tasks where the timing and continuity of detection are central.

## B.5 VUS-ROC and VUS-PR [30]

Volume Under the Surface (VUS) metrics evaluate the consistency and timeliness of anomaly score rankings by integrating performance over both threshold levels and temporal tolerance windows. Given a set of allowable lags  $\Omega = \{\omega_1, \omega_2, \dots\}$ , where each  $\omega$  defines a symmetric temporal buffer, the predicted points are matched to ground-truth anomalies if they fall within  $\pm\omega$  time steps. For each  $\omega$ , we compute a corresponding AUC score:

$$\text{VUS-ROC} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \text{AUC-ROC}_\omega, \quad (\text{S10})$$

$$\text{VUS-PR} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \text{AUC-PR}_\omega. \quad (\text{S11})$$

VUS metrics address two limitations of classical AUC: their insensitivity to detection delay and their lack of temporal context. By explicitly modeling detection within variable tolerance windows, VUS captures how consistently a model produces timely, high-ranking anomaly scores across a range of thresholds. This makes it a strong tool for evaluating delay-aware detection performance, especially in streaming or real-time settings.

Unlike segment-level metrics such as Affiliation F1, which apply smooth decay functions to soften the penalty for temporal misalignment, VUS does not interpolate between correct and incorrect predictions. It avoids heuristic credit assignment and instead tests whether the model can repeatedly detect anomalies within strict or relaxed time boundaries, without requiring threshold or decay tuning. This makes it a better measure of intrinsic anomaly scoring quality.

We find that Affiliation F1 and VUS serve complementary purposes: Affiliation F1 is suitable for analyzing misaligned segment predictions in the presence of noisy labels or lag-prone detection. At the same time, VUS offers a more principled, high-resolution assessment of detection precision and consistency across time and score thresholds.

## C Baselines

### C.1 Implementation Details

All baseline models were implemented using official or widely adopted public repositories, with minor formatting adjustments to ensure evaluation consistency. Hyperparameters were either taken from the original papers or selected via lightweight tuning. All input features were normalized using a `StandardScaler` fitted on the training set (zero mean and unit variance), and no separate validation set was used. Each model was trained and evaluated using only the provided training/test splits. To ensure robustness, we repeated all experiments with five independent random seeds and report the average performance. As described in Section 4.1, we apply optimal threshold selection for each method to decouple detection performance from threshold sensitivity. The threshold search procedure is described in Appendix D.3, and standard deviations across seeds are reported in Appendix E.3.

### C.2 Codebase of Implementation

Most baseline models, excluding OmniAnomaly and SARAD, are implemented using the unified codebase of CATCH:

- CATCH (ICLR2025): [github.com/decisionintelligence/CATCH](https://github.com/decisionintelligence/CATCH)

Below, we list the official implementation links for all baseline models used in our study:

- OmniAnomaly (KDD 2019): [github.com/NetManAIOps/OmniAnomaly](https://github.com/NetManAIOps/OmniAnomaly)
- AnomalyTransformer (ICLR 2021): [github.com/thumt/Anomaly-Transformer](https://github.com/thumt/Anomaly-Transformer)
- PatchTST (ICLR 2022): [github.com/yuqinie98/PatchTST](https://github.com/yuqinie98/PatchTST)
- TimesNet (ICLR 2022): [github.com/thumt/TimesNet](https://github.com/thumt/TimesNet)
- DLinear (AAAI 2023): [github.com/honeywell121/DLinear](https://github.com/honeywell121/DLinear)
- NLinear (AAAI 2023): [github.com/honeywell121/DLinear](https://github.com/honeywell121/DLinear)
- DCdetector (KDD 2023): [github.com/DAMO-DI-ML/KDD2023-DCdetector](https://github.com/DAMO-DI-ML/KDD2023-DCdetector)
- iTransformer (ICLR 2023): [github.com/thumt/iTransformer](https://github.com/thumt/iTransformer)
- ModernTCN (ICLR 2023): [github.com/luodhhh/ModernTCN](https://github.com/luodhhh/ModernTCN)
- SARAD (NeurIPS 2024): [github.com/daidahao/SARAD](https://github.com/daidahao/SARAD)

## D More Details of Experiment

### D.1 Implementation

We trained all models using the AdamW [26] optimizer with default hyperparameters, as it provided stable convergence across datasets. To account for dataset-specific scale differences, we set the learning rate to  $5e-5$  for PSM, while a higher value of  $5e-4$  was used for both SMD and SWaT. Prior to training, all input features were standardized using a `StandardScaler`<sup>1</sup> fitted on the training set, ensuring zero mean and unit variance. Since our evaluation protocol does not rely on validation-based model selection, we did not use a separate validation set. Instead, each model was trained and evaluated solely on the provided training/test splits.

### D.2 Hyperparameter Sensitivity

We evaluated the sensitivity of OracleAD to four key hyperparameters: the window length, the batch size, the deviation weighting factor  $\lambda_{\text{dev}}$ , and the number of encoder and decoder layers. The analysis was performed on three benchmark datasets, and the results are summarized in Figures S1 through S4, using F1, VUS-ROC, and VUS-PR as evaluation metrics.

For the SMD dataset, we followed the protocol from OmniAnomaly<sup>1</sup>, treating each machine as an independent subset and selecting representative subsets for this analysis to avoid overlap with the evaluation set. The reported performance is averaged over these selected subsets.

The default configuration used across all datasets fixes the window length to 10, the batch size to 1024, the deviation weight to 3, and the number of layers to 2. While this setting does not always achieve the highest performance, it offers a stable balance between accuracy and training efficiency.

**Window length.** As shown in Figure S1, OracleAD achieves its best performance with window lengths of 10 or 20 across most datasets, including PSM, SMD, and SWaT. Larger windows such as 40 occasionally improve VUS-ROC, but often lead to reduced precision in VUS-PR, likely due to the dilution of localized anomaly patterns. On the other hand, smaller windows such as 5 may fail to capture sufficient temporal context. Overall, OracleAD performs robustly within the range of 10 to 20, suggesting that this moderate window length is well suited for capturing both short-term dynamics and broader structural dependencies.

**Batch Size.** Figure S2 shows that as batch size increases, performance improves in general. For instance, in the case of PSM, F1 score rises from 45.77 to 65.85 as batch size increases from 64 to 1024. This trend suggests that OracleAD benefits from larger batch statistics during training, potentially due to more stable gradient estimates.

**Deviation Weight  $\lambda_{\text{dev}}$ .** As shown in Figure S3, increasing the structural regularization weight  $\lambda_{\text{dev}}$  consistently improves performance across metrics. This effect is most prominent in PSM and SMD, where larger  $\lambda$  values enhance VUS-ROC and VUS-PR score. The results confirm that deviation scoring based on the Stable Latent Structure plays a key role in anomaly localization and benefits from stronger guidance. In our main experiments, we set  $\lambda_{\text{dev}} = 3$ , as it offers the best balance between prediction and structural alignment.

**Model Depth.** Figure S4 illustrates the impact of varying the number of encoder and decoder layers. Using two layers generally leads to the best overall performance across datasets. This setting provides sufficient model capacity to capture temporal and structural dependencies without overfitting. Increasing the number of layers to three does not consistently improve performance and, in some cases such as PSM and SMD, leads to degradation in F1 score. These results suggest that additional depth may introduce unnecessary complexity or optimization instability, particularly in datasets with limited variability or high noise. Therefore, a moderate depth of two layers strikes a good balance between expressiveness and generalization.

<sup>1</sup><https://github.com/NetManAIOps/OmniAnomaly>

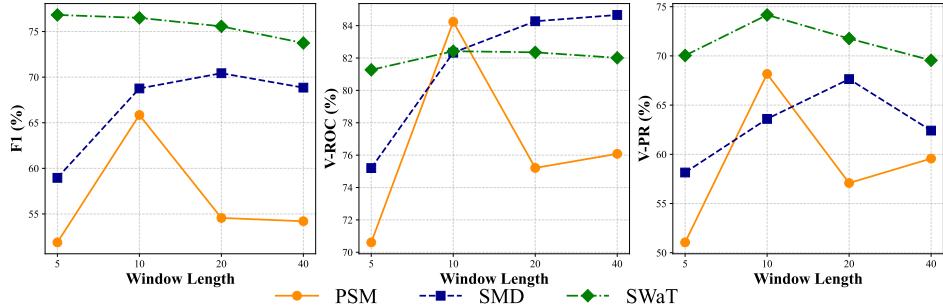


Figure S1: Effect of input window length on OracleAD performance. Each subplot shows a different evaluation metric. All results are averaged over five random seeds.

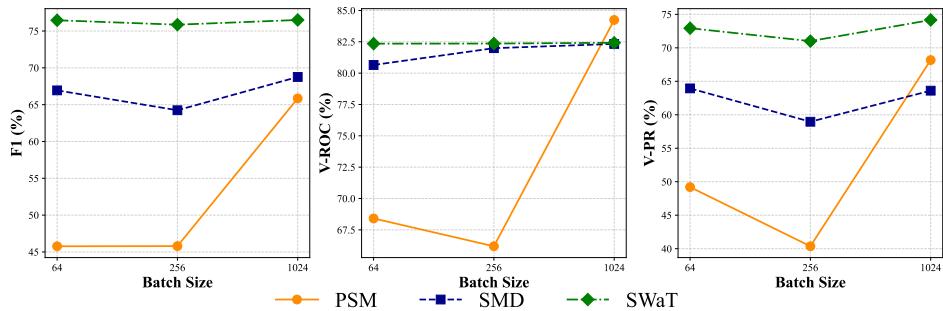


Figure S2: Effect of training batch size on OracleAD performance. Each subplot shows a different evaluation metric. All results are averaged over five random seeds.

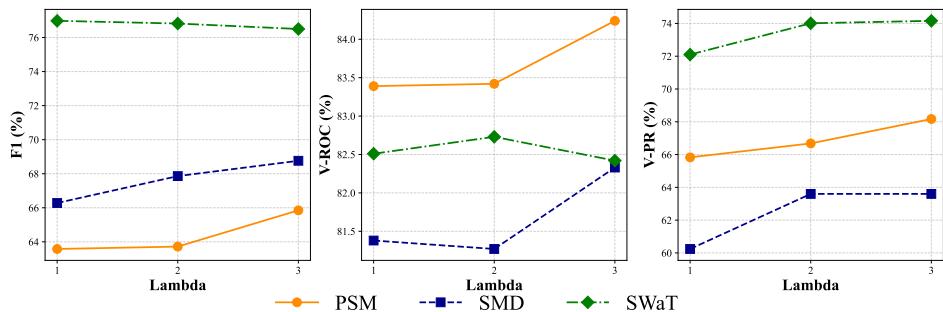


Figure S3: Effect of the deviation regularization weight  $\lambda_{\text{dev}}$  on OracleAD performance. Each subplot shows a different evaluation metric. All results are averaged over five random seeds.

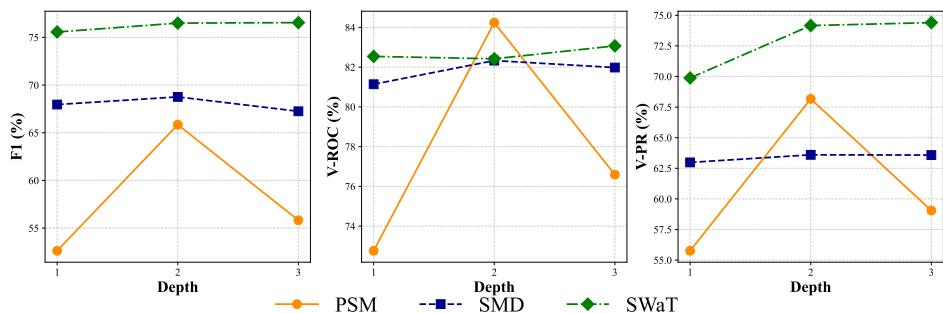


Figure S4: Effect of the encoder and decoder depth on OracleAD performance. Each subplot shows a different evaluation metric. All results are averaged over five random seeds.

### D.3 Thresholding Strategy

We adopt a score calibration procedure that selects the optimal threshold based on a target evaluation metric to convert continuous anomaly scores into binary predictions.

Given a sequence of anomaly scores  $\mathbf{s} = \{s_t\}_{t=1}^T$  and corresponding binary ground-truth labels  $\mathbf{y} = \{y_t\}_{t=1}^T$ , the optimal threshold  $\tau^*$  is obtained by maximizing an evaluation metric  $\mathcal{M}$  over a discrete candidate set  $\mathcal{T}$ :

$$\tau^* = \arg \max_{\tau \in \mathcal{T}} \mathcal{M}(\hat{\mathbf{y}}^\tau, \mathbf{y}), \quad (\text{S12})$$

where  $\hat{y}_t^\tau = \mathbb{1}[s_t \geq \tau]$  denotes the binary prediction at time  $t$  based on threshold  $\tau$ .

The candidate set  $\mathcal{T}$  is constructed by uniformly partitioning the score range into  $N$  bins:

$$\mathcal{T} = \left\{ \tau_i \mid \tau_i = \min(\mathbf{s}) + \frac{i}{N-1} (\max(\mathbf{s}) - \min(\mathbf{s})), i = 0, 1, \dots, N-1 \right\}. \quad (\text{S13})$$

We set  $N = 200$  in all experiments, following the protocol used by Liu et al. [24] and adopted in the official TSB-UAD benchmark suite<sup>2</sup>. This value offers a good balance between threshold granularity and computational cost, and has been shown to yield stable evaluation results in large-scale studies. The threshold search is performed independently for each model and dataset to ensure fair and metric-sensitive calibration.

### D.4 System Configuration

All experiments were conducted on a single workstation equipped with an NVIDIA GeForce RTX 5090 GPU (32GB) and an Intel Core Ultra 7 265K CPU (20 cores). The system had 96GB of DDR5 RAM and ran Ubuntu 24.04 LTS (64-bit). The software environment included Python 3.12, PyTorch 2.7.0, and CUDA Toolkit 12.8.

---

<sup>2</sup><https://github.com/TheDatumOrg/TSB-AD>

## D.5 Computational Cost

OracleAD employs a separate LSTM encoder–decoder for each of the  $N$  variables, giving a per-window temporal cost of  $O(N \cdot T)$ . Multi-head self-attention requires  $O(N^2)$  pairwise comparisons, leading to an overall per-window complexity of  $O(N \cdot T + N^2)$ . Although this scales superlinearly with  $N$ , the quadratic cost is comparable to basic correlation computations and is a common requirement in models that capture explicit inter-variable structure.

We empirically profiled training and inference efficiency across real-world (PSM, SMD, SWaT) and synthetic datasets (up to  $N = 400$  variables). Training results (Table S3) show that while memory usage increases with dimensionality, throughput remains practical. Inference throughput (Table S4) demonstrates that OracleAD remains feasible for real-time monitoring in systems with hundreds of variables.

Table S3: Training costs across datasets and batch sizes.

Dataset (N)	Batch Size	Memory Usage	Training Speed (it/s)
PSM (25)	64	1.5 GB	12.6
	1024	8.5 GB	9.6
SMD (38)	64	1.9 GB	8.0
	1024	13.0 GB	6.8
SWaT (51)	64	2.4 GB	6.0
	1024	18.5 GB	5.0
Syn-PSM (100)	64	4.7 GB	3.1
	512	23.3 GB	2.9
Syn-PSM (400)	32	20.4 GB	1.35

Table S4: Inference throughput by dataset dimensionality.

Dataset (N)	Inference Speed (time steps/s)
PSM (25)	54.0
SMD (38)	37.0
SWaT (51)	28.0
Syn-PSM (100)	13.5
Syn-PSM (400)	3.15

## E Additional Experiments

### E.1 Full Evaluation Results with Classical and Deep Baselines

Table S5: Performance (%) of anomaly detection models on the PSM, SMD, and SWaT datasets. **Boldface** indicates the best score and underline indicates the second-best score.

Dataset	Metric	PCA	HBOS	IF	A.E	Omni	A.T	Patch	TsNet	DLin	NLin	DC	iTrans	Modern	SARAD	CATCH	OracleAD
PSM	F1	46.45	45.98	43.45	<b>47.55</b>	45.90	43.45	43.45	43.45	43.45	43.45	43.45	43.45	45.75	44.33	<b>65.85</b>	
	R-F1	46.18	47.74	45.78	45.93	40.58	12.17	34.75	37.49	34.74	32.10	21.35	35.91	35.48	13.54	<u>54.19</u>	
	Aff-F1	73.90	72.77	<b>81.98</b>	73.70	73.49	69.68	69.43	69.43	69.43	69.59	69.53	69.43	77.12	<u>79.16</u>	78.07	
	A-ROC	64.82	61.98	54.23	<u>66.79</u>	63.95	38.35	58.68	59.09	58.02	58.56	49.86	59.22	59.21	62.86	<u>64.75</u>	<b>84.78</b>
	A-PR	46.75	39.34	33.41	<u>47.27</u>	45.00	24.28	37.91	38.95	37.17	37.64	27.78	38.27	38.59	41.67	43.40	<b>68.11</b>
	V-ROC	66.59	63.20	56.47	<u>68.20</u>	61.23	50.96	50.62	50.67	50.53	50.69	50.11	51.23	50.68	57.63	57.00	<b>84.24</b>
	V-PR	49.13	41.18	37.78	<u>49.66</u>	<u>52.49</u>	49.76	50.31	49.86	49.93	46.71	27.89	44.54	50.55	38.64	45.95	<b>68.17</b>
SMD	F1	21.97	17.66	14.91	25.78	<u>32.16</u>	7.98	7.98	7.98	8.00	7.98	7.98	7.98	25.92	7.98	<b>43.03</b>	
	R-F1	27.74	14.98	21.74	28.08	<u>35.81</u>	5.84	13.81	9.43	13.37	13.51	8.23	9.02	14.43	10.35	2.90	<b>38.88</b>
	Aff-F1	72.35	69.28	81.50	<u>83.34</u>	82.57	67.43	67.43	67.43	67.43	67.43	67.43	67.43	75.87	67.43	<u>84.73</u>	
	A-ROC	67.88	62.57	66.44	76.80	<u>71.25</u>	50.40	73.85	59.09	72.73	73.83	49.97	74.57	72.13	72.84	<u>80.96</u>	<b>83.56</b>
	A-PR	12.75	14.53	12.17	19.40	<u>27.73</u>	4.57	14.82	13.66	13.88	14.10	4.16	14.57	13.01	25.87	17.09	<b>44.83</b>
	V-ROC	66.50	65.24	68.58	<u>74.38</u>	73.26	50.34	51.66	51.10	51.52	51.76	49.92	51.76	51.55	65.57	50.95	69.57
	V-PR	15.02	16.09	16.02	22.50	31.18	36.86	41.60	41.21	41.26	39.15	42.7	32.85	41.54	19.33	35.25	<b>47.52</b>
SWaT	F1	75.36	<u>76.04</u>	21.65	74.46	75.40	21.65	21.65	7.98	12.14	21.65	21.65	21.65	57.30	21.65	<b>76.50</b>	
	R-F1	29.84	20.18	14.77	<u>36.63</u>	<b>38.72</b>	15.48	12.23	20.80	5.13	12.06	15.54	11.89	12.88	21.37	14.74	28.15
	Aff-F1	72.18	70.63	72.04	<b>75.30</b>	72.09	70.06	69.51	73.47	67.43	69.68	69.25	69.25	72.08	<u>74.88</u>	73.68	71.97
	A-ROC	81.79	<u>83.52</u>	36.35	81.97	82.15	42.58	24.36	28.97	50.48	23.31	49.89	24.57	24.50	<b>85.40</b>	33.31	82.71
	A-PR	72.47	<b>73.99</b>	10.27	67.51	<u>72.73</u>	11.93	8.40	10.71	4.53	8.00	12.14	8.33	9.00	64.77	13.39	72.39
	V-ROC	81.75	<u>83.49</u>	41.37	81.89	81.98	50.72	49.84	20.80	50.36	49.02	50.00	49.90	50.15	<b>86.30</b>	51.89	82.42
	V-PR	52.31	<u>73.90</u>	10.17	65.89	64.42	17.00	11.92	51.20	28.47	10.55	11.92	12.05	13.05	62.72	18.70	<b>74.16</b>

This section presents the complete evaluation results of OracleAD alongside both deep learning and classical anomaly detection baselines on the PSM, SMD, and SWaT datasets. While deep learning models were trained and evaluated over five random seeds to capture stochastic variation, classical baselines such as Principal Component Analysis (PCA) [36], Histogram-Based Outlier Score (HBOS) [14], and Isolation Forest (IF) [23] were evaluated under a fixed random seed. This is because they are either fully deterministic or only marginally affected by random initialization, and thus produce effectively consistent outputs.

Interestingly, as shown in Table S5, several classical baselines outperform many deep learning models across a range of metrics. This reflects a broader phenomenon we also highlighted in Section 1, complex or novel architectures do not necessarily lead to better performance. We observe that many deep models achieve their highest F1 scores when recall is maximized, meaning recall is close to one. This implies that thresholding tends to favor high precision, often at the cost of missing many true anomalies, especially those that fall within labeled anomaly segments. Consequently, these models often struggle on strict evaluation metrics such as VUS-PR and VUS-ROC, which emphasize consistent detection quality across thresholds and accurate alignment with ground-truth structure.

In contrast, classical models tend to produce relatively strong AUC-ROC and VUS-ROC scores. This indicates that while they have high overall detection rates, they often lack precision stability, which leads to weaker performance on precision-sensitive metrics like AUC-PR. The combination of high recall and low precision reveals that these models are often overly sensitive, generating many false positives.

Taken together, these findings clarify OracleAD’s distinctive advantage. It is explicitly designed to address both detection coverage and accuracy through a principled dual-scoring mechanism that combines prediction error with structural deviation. Unlike many deep models that overfit to a single scoring heuristic, OracleAD integrates temporal and inter-variable signals, resulting in stable performance across a broad range of evaluation criteria. This is particularly evident in metrics such as VUS-PR and VUS-ROC, where OracleAD consistently outperforms both classical and deep learning baselines.

## E.2 Ablation study on the SMD dataset

Table S6: Ablation results (%) on selected SMD subsets. All results are averaged over five random seeds.

Component	Variant	machine 1-5			machine 2-3			machine 2-7			machine 3-6		
		F1	V-ROC	V-PR									
Loss Function	w/o Recon Loss	59.61	81.82	49.24	51.94	73.70	55.60	80.07	81.33	76.94	34.25	65.51	35.37
Anomaly Score	w/o $\mathcal{P}_{\text{score}}^t$	23.31	72.35	11.13	50.29	80.26	51.84	81.50	79.14	62.17	34.18	78.23	22.95
	w/o $\mathcal{D}_{\text{score}}^t$	54.32	86.25	49.60	52.89	86.60	42.86	82.66	86.94	71.66	46.03	89.76	50.70
<b>OracleAD (Full Model)</b>		62.03	82.44	52.15	53.99	81.71	54.83	80.35	81.75	76.42	44.39	77.32	43.11

The SMD dataset is characterized by its diversity in anomaly types, ranging from abrupt point anomalies to gradually developing structural failures. This heterogeneity results in variable performance across scoring mechanisms depending on the subset. For instance, the prediction-based score  $\mathcal{P}_{\text{score}}^t$  often excels when anomalies manifest as sharp temporal deviations, whereas the deviation-based score  $\mathcal{D}_{\text{score}}^t$  is more effective when anomalies involve shifts in inter-variable structure. Table S6 shows that each score alone captures certain types of anomalies well, but fails to generalize across the full spectrum. In contrast, combining both scores consistently yields more stable performance across all subsets, striking a balance between temporal precision and structural sensitivity.

The role of the reconstruction loss also varies across subsets. In machine 2-7, excluding the reconstruction objective has little impact on performance, suggesting that prediction alone suffices to model normal dynamics. However, in subsets like machine 2-3 and 3-6, removing reconstruction objective significantly degrades structural metrics such as VUS-PR and causes instability in overall detection performance. This indicates that the reconstruction loss is a form of regularization that stabilizes the latent space and improves robustness, particularly when the data contains noise or complex normal variation.

These findings confirm that OracleAD’s complete model design offers the most consistent and generalizable performance across the varied anomaly profiles present in SMD.

## E.3 Statistical Significance

Table S7: Standard deviation (%) of anomaly detection performance across seven evaluation metrics on three benchmark datasets (PSM, SMD, and SWaT). All values are computed over five random seeds for each model. Lower values indicate higher stability and consistency across runs. Metrics include F1, R-F1, Aff-F1, AUC-ROC, AUC-PR, VUS-ROC, and VUS-PR.

Dataset	Metric	A.E	Omni	A.T	Patch	TsNet	DLin	NLin	DC	iTrans	Modern	SARAD	CATCH	OracleAD
PSM	F1	0.44	1.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	1.59	2.98	
	R-F1	0.48	2.49	7.50	0.44	0.61	0.61	0.11	24.03	1.67	0.38	7.54	5.46	2.93
	Aff-F1	0.24	0.65	0.55	0.00	0.00	0.00	0.00	0.36	0.23	0.00	4.30	4.76	0.94
	A-ROC	1.07	1.89	2.10	0.06	0.56	0.07	0.05	0.29	0.11	0.14	1.52	0.78	2.49
	A-PR	0.59	2.34	0.88	0.10	0.46	0.05	0.05	0.24	0.11	0.18	1.26	0.75	4.30
	V-ROC	0.96	0.56	0.95	0.05	0.03	0.08	0.01	0.56	0.24	0.03	4.98	5.14	2.47
	V-PR	0.56	0.52	13.09	0.76	1.40	0.80	0.51	0.82	1.02	0.46	1.75	0.87	4.31
SMD	F1	1.60	0.11	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.35	0.00	0.92
	R-F1	2.48	0.26	3.69	0.44	0.84	0.48	0.15	5.40	1.21	0.18	2.28	0.14	0.93
	Aff-F1	1.28	0.29	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.35	0.00	0.78
	A-ROC	0.98	0.07	0.27	0.24	0.42	0.11	0.04	0.25	0.11	0.15	0.32	0.23	0.47
	A-PR	1.78	0.11	0.10	0.13	0.32	0.06	0.01	0.25	0.18	0.06	2.53	0.22	0.90
	V-ROC	2.69	0.09	0.21	0.08	0.04	0.07	0.04	0.22	0.18	0.03	0.75	0.08	0.46
	V-PR	2.20	0.10	4.35	0.60	0.52	0.47	0.26	0.07	0.45	0.34	0.37	0.89	1.34
SWaT	F1	2.34	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.29	0.00	0.31
	R-F1	1.67	3.48	14.46	0.28	0.40	0.12	0.16	0.40	0.11	0.60	9.58	0.84	2.72
	Aff-F1	0.77	0.05	0.96	0.40	0.64	0.00	0.08	0.00	0.00	0.72	1.51	0.94	1.80
	A-ROC	0.64	0.07	4.39	0.07	0.32	0.12	3.54	0.29	0.16	0.08	0.91	0.27	1.22
	A-PR	6.48	0.02	0.41	0.02	0.09	0.01	1.10	0.03	0.01	0.18	12.61	0.35	1.99
	V-ROC	0.67	0.11	0.12	0.02	0.10	0.06	0.00	0.10	0.08	0.08	0.56	0.16	0.90
	V-PR	7.38	3.58	0.42	0.03	0.76	0.04	0.00	0.31	0.20	0.19	9.02	0.73	4.38

#### E.4 Distance Metrics for Dissimilarity Matrix

Table S8: Performance comparison of cosine similarity, L1, and L2 distances for latent vector dissimilarity on the PSM dataset. All results are averaged over five random seeds. **Boldface** indicates the best score.

Method	F1	R-F1	Aff-F1	A-ROC	A-PR	V-ROC	V-PR
Cosine Sim.	46.81	43.68	78.18	66.46	48.59	65.83	48.82
L1	57.72	51.54	75.87	76.33	54.29	73.33	53.52
L2	<b>65.85</b>	<b>54.66</b>	<b>78.07</b>	<b>84.78</b>	<b>68.11</b>	<b>84.24</b>	<b>68.17</b>

In Eq. 11, we define the pairwise dissimilarity matrix using the L2 distance. We choose L2 over other metrics because it captures both directional mismatch and magnitude differences (latent energy variation) in the embeddings. To validate this choice, we empirically compared L2 against cosine similarity and L1 distance across multiple evaluation metrics. As shown in Table S8, L2 consistently outperforms the alternatives, yielding higher and more stable performance across five random seeds.

#### E.5 Variance Dynamics of Dissimilarity Matrices

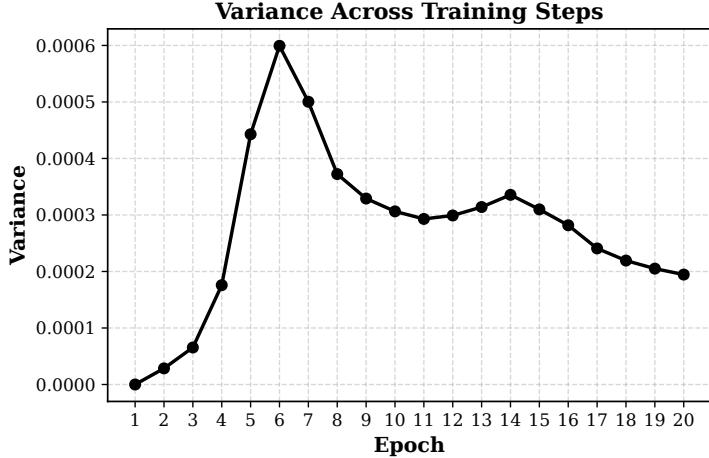


Figure S5: Variance of dissimilarity matrices across epochs on the PSM dataset. All results are averaged over five random seeds.

To further analyze the stability of the Stable Latent Structure (SLS), we track the variance of the dissimilarity matrices  $D^{(k)}$  across training epochs. This variance reflects how much the learned inter-variable relationships fluctuate from one sliding window to another within the same epoch.

Figure S5 shows the variance trend on the PSM dataset. The trajectory can be divided into two phases: in the early epochs, variance increases rapidly as the model explores latent relational geometry from sliding windows, reflecting active adaptation consistent with the reviewer's observation that rising variance indicates meaningful structural learning; in later epochs, variance gradually decreases and converges, showing that the latent representations stabilize toward a consistent inter-variable template. This stabilization confirms that the learned SLS serves as a robust baseline for normal relational structure.

## E.6 Comparison of Anomaly Score Fusion Strategies

Table S9: Performance comparison between multiplicative and additive anomaly score fusion on PSM and SWaT datasets. Results are averaged over five random seeds.

Dataset	Anomaly Score	F1	R-F1	Aff-F1	A-ROC	A-PR	V-ROC	V-PR
PSM	$\mathcal{P}_{\text{score}}^t \cdot \mathcal{D}_{\text{score}}^t$	65.85	54.66	78.07	84.78	68.11	84.24	68.17
	$\mathcal{P}_{\text{score}}^t + \mathcal{D}_{\text{score}}^t$	68.36	55.48	78.52	85.03	67.19	84.82	67.43
SWaT	$\mathcal{P}_{\text{score}}^t \cdot \mathcal{D}_{\text{score}}^t$	76.50	28.15	71.97	82.71	72.39	82.42	74.16
	$\mathcal{P}_{\text{score}}^t + \mathcal{D}_{\text{score}}^t$	76.91	37.75	73.43	81.73	71.73	81.42	71.07

To find an effective anomaly score construction, we compare two fusion strategies that combine the temporal prediction score( $\mathcal{P}_{\text{score}}^t$ ) and the structural deviation score( $\mathcal{D}_{\text{score}}^t$ ). Their fusion determines how OracleAD balances temporal precision and structural sensitivity.

Table S9 presents results for multiplicative and additive fusion on PSM and SWaT. While additive fusion also achieves strong performance, it requires dataset-specific tuning to balance the relative influence of  $\mathcal{P}_{\text{score}}^t$  and  $\mathcal{D}_{\text{score}}^t$ , as their scales and sensitivities can differ across domains. In contrast, multiplicative fusion inherently balances the two by emphasizing co-occurring signals, leading to sharper and more consistent anomaly responses. Furthermore, higher A-PR and V-PR scores indicate that the multiplicative form is more effective in reducing false positives, making it better suited for real-world applications where reliability and precision are critical.

Overall, both methods achieve competitive performance, but the multiplicative design better aligns with OracleAD’s objective, providing improved stability and interpretability in multivariate detection.

## F Case Studies for Interpretability

### F.1 Complex Multivariate Anomaly

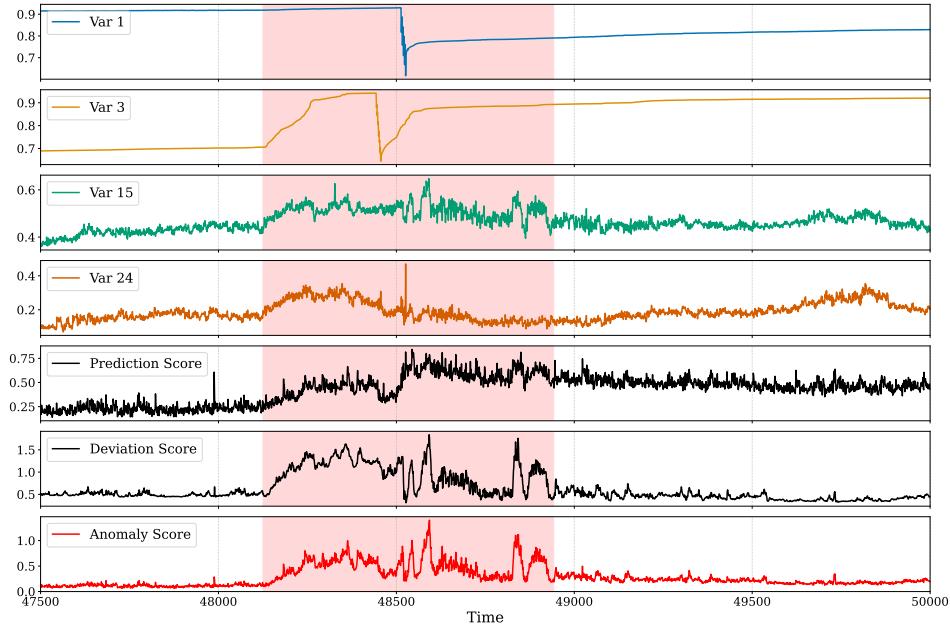


Figure S6: Raw signals of the identified root-cause variables and anomaly score visualization for selected intervals in the PSM dataset. Shaded red regions indicate ground-truth anomalies.

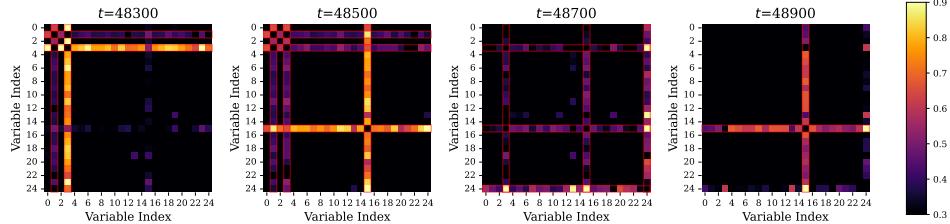


Figure S7: Visualization for Deviation Matrices  $\mathcal{D}_{\text{matrix}}^t$  at anomalous time points in PSM dataset. Each matrix corresponds respectively to the subfigures in Figure S6. Each matrix is independently min–max normalized and values are clipped to the range [0.3, 0.9] to enhance contrast.

Figure S6 presents a case from the PSM dataset in which a single anomaly segment contains multiple overlapping anomaly signals across different variables. While the majority of variables in the PSM follow relatively stable patterns, several variables (e.g., 1, 3, 15, 24) exhibit subtle yet complex dynamics that makes it difficult to detect through prediction alone. In the highlighted anomaly segment, these variables show anomalous behavior either concurrently or sequentially, creating a challenging detection scenario.

The prediction score captures local deviations in the latter part of the segment but remains relatively low or fragmented in the earlier phase. In contrast, the deviation score maintains a more consistent profile, reflecting sustained disruptions in the latent structure. This contrast demonstrates how the deviation score is more effective in representing the internal disorganization of the system, especially when multiple variables interact in a non-trivial way.

Overall, this case illustrates that OracleAD’s dual-scoring mechanism enables anomaly detection that is both sensitive and robust across a variety of complex conditions. The prediction score is responsive to short-term deviations in individual variable behavior, while the deviation score captures structural inconsistencies across variables over time. Together, these two signals allow OracleAD to handle heterogeneous and overlapping anomaly patterns effectively.

## F.2 Long-Duration Anomaly

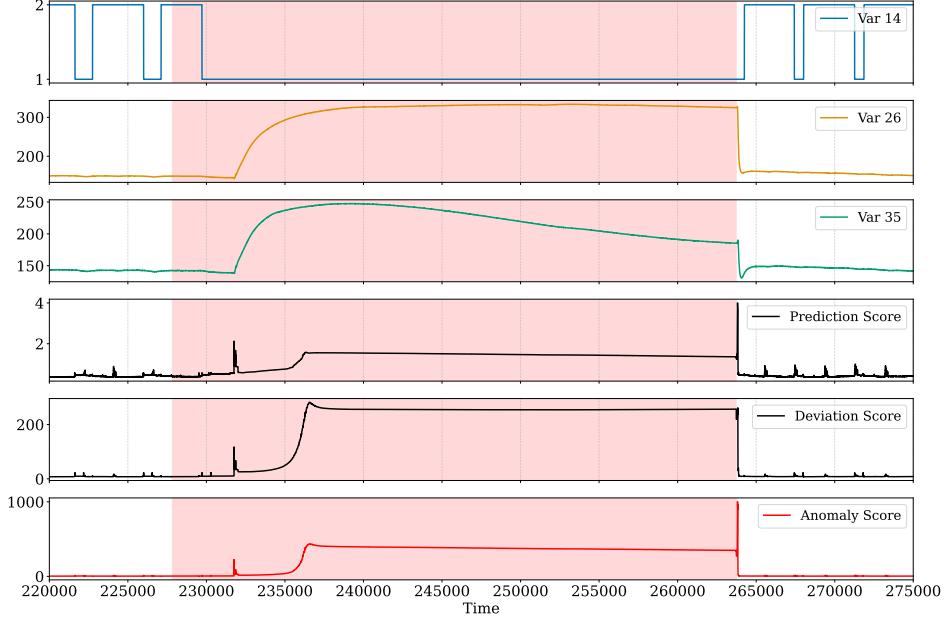


Figure S8: Raw signals of the identified root-cause variables and anomaly score visualization for selected intervals in the SWaT dataset. Shaded red regions indicate ground-truth anomalies.

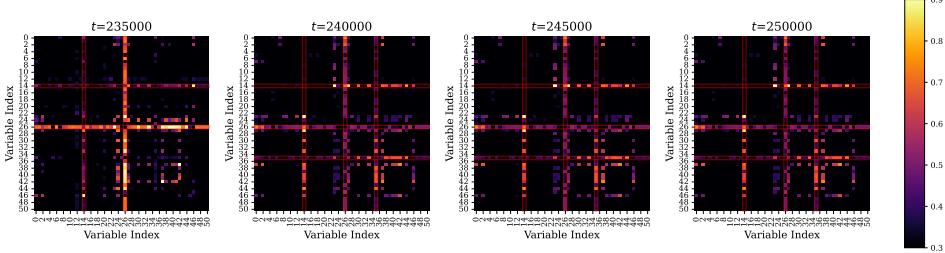


Figure S9: Visualization for Deviation Matrices  $\mathcal{D}_{\text{matrix}}^t$  at anomalous time points in SWaT dataset. Each matrix corresponds respectively to the subfigures in Figure S8. Each matrix is independently min–max normalized and values are clipped to the range [0.3, 0.9] to enhance contrast.

Figure S8 illustrates a long anomaly segment in the SWaT dataset, notable for its persistent and structurally coherent anomaly behavior. Among all available sequences, this segment contains the longest continuous anomaly window, allowing us to evaluate OracleAD’s interpretability in a sustained abnormal context.

As shown in Figure S8, variable 26 deviates sharply and persistently from its normal operating range throughout the entire anomaly segment. This deviation is visually evident and positions variable 26 as the dominant root cause of the anomaly. Both the prediction and the deviation scores rise significantly within the anomaly interval, the deviation score plays a pivotal role in this case.

The deviation matrices, visualized in Figure S9, further articulate the identification of root-cause variable. Across multiple time stamps within the anomaly segment, variable 26 consistently dominates the matrix in terms of deviation strength. This pattern indicates its sustained role as the structural root cause.

This case demonstrates how OracleAD leverages deviation scoring to detect sustained and structure-driven anomalies that may not exhibit dynamic variation within individual variables. By modeling the relational structure among variables, OracleAD maintains sensitivity even when traditional prediction-based detectors may fail due to temporal flatness.

## G Quantitative Analysis of Interpretability

Table S10: Top-3 predicted variables ranked by aggregate deviation scores compared with ground-truth causal variables across representative anomaly timestamps in the SMD dataset. Each case is drawn from a distinct anomaly segment.

	$t=2106$	$t=4986$	$t=6032$	$t=7667$	$t=14936$	$t=15065$	$t=16536$	$t=18641$
	Var. Score	Var. Score	Var. Score					
Top-3 Predicted	13 9.00	13 9.49	32 15.00	29 15.99	33 12.84	13 9.39	13 13.55	15 55.97
	12 5.46	8 7.36	33 10.16	33 4.48	32 6.11	8 7.50	12 6.62	10 49.98
	8 3.91	12 5.79	31 3.49	34 2.95	35 3.42	12 5.57	8 5.56	18 45.22
Ground Truth	12, 13	8, 12, 13	32, 33	29	32, 33	8, 12, 13, 14	8, 12, 13, 14	10, 15, 18

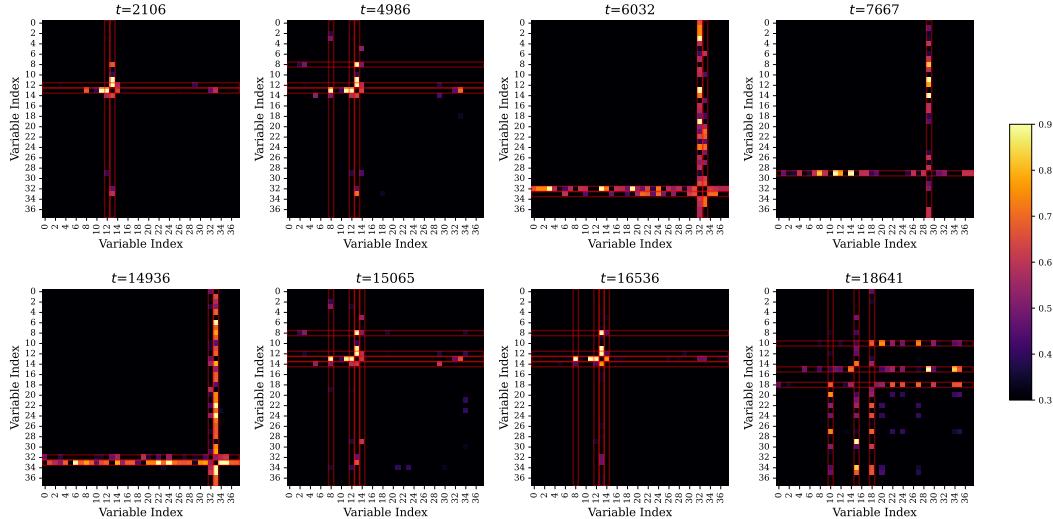


Figure S10: Deviation matrices correspond to a deviation matrix visualization in Table S10. Row and column activations highlight the variables most responsible for abnormal behavior, and red boxes mark the ground-truth root-cause variables. For visualization, each matrix is independently min–max normalized and values are clipped to the range [0.3, 0.9] to enhance contrast.

A distinctive strength of OracleAD lies in its ability to localize root-cause variables through the deviation matrix  $\mathcal{D}_{\text{matrix}}^t = |D^t - \text{SLS}|$ . Each entry measures the extent to which a variable pair's relationship deviates from the stable latent structure, and row-wise aggregation highlights which variables most strongly disrupt normal interactions.

Figure S10 shows deviation matrices at representative anomaly timestamps from the SMD dataset. Each anomaly produces a characteristic disruption pattern. As discussed in Section 4.3, variables associated with multiple high-magnitude rows or columns are more likely to correspond to root causes. This enables not only root-cause analysis but also a clear visualization of how structural disruptions in key variables evolve over the course of an anomaly segment.

For precise localization, however, we rely on the aggregate deviation score obtained by summing each row (or column). This yields a variable-level anomaly score that can be ranked to produce candidate root causes. As shown in Table S10, most cases quantitatively validate this approach. The top-3 ranked variables coincide with the ground-truth causal set, and even for multivariable anomalies, OracleAD reliably recovers the majority of causal variables.

## H Broader Impacts

Anomaly detection in multivariate time series data has wide-reaching applications across domains such as industrial monitoring, cybersecurity, healthcare, and environmental systems. The ability to detect abnormal behavior in a timely and interpretable manner can improve operational safety, prevent financial or physical damage, and assist in early diagnostics. OracleAD aims to address key limitations of existing models by offering interpretable, structure-aware, and segment-sensitive anomaly detection.

However, as with all data-driven systems, there are potential risks. If deployed in high-stakes domains such as healthcare or autonomous control systems, undetected anomalies or false alarms may lead to significant consequences. Furthermore, the model's performance relies heavily on the quality and representativeness of the training data. If the training data lacks diversity or contains unannotated anomalies, the model may reinforce biases or exhibit degraded performance in underrepresented conditions.

To mitigate such risks, practitioners should carefully evaluate the model under realistic operational conditions and complement automated decisions with human oversight, particularly in critical applications. We believe that OracleAD can contribute positively by promoting reliable and interpretable anomaly detection, but responsible deployment and continuous evaluation are necessary to ensure its benefits are realized without unintended harms.

## I Limitations and Future Works

While OracleAD demonstrates strong generalization and interpretability across diverse benchmarks, several limitations remain. First, the Stable Latent Structure (SLS) assumes globally consistent inter-variable relationships learned from continuous-valued inputs. This assumption may not hold in complex or multimodal systems such as SWaT, where heterogeneous subsystems or mixed variable types (e.g., continuous and discrete sensors) coexist. Under such conditions, the deviation matrix may lose discriminative sharpness, and the anomaly scoring can become less reliable when one component signal (prediction or deviation) weakens.

To address these challenges, future work will focus on developing (1) an anomaly detection framework that is agnostic to variable types, seamlessly handling continuous, categorical, and discrete signals; and (2) a dynamically updated SLS mechanism that adapts over time rather than remaining static during inference. These extensions would enhance OracleAD's applicability to large-scale, dynamic, and heterogeneous industrial systems.