Conditional Attribution for Root Cause Analysis in Time-Series Anomaly Detection

Shashank Mishra[1] (✉), Karan Patil[1,3], Cedric Schockaert[2], Didier Stricker[1,3], and Jason Rambach[1]

[1] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany {shashank.mishra, karan_sanjay.patil, didier.stricker, jason.rambach}@dfki.de
[2] Paul Wurth S.A, Luxembourg cedric.schockaert@sms-group.com
[3] Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Germany

Abstract. Root cause analysis (RCA) for time-series anomaly detection is critical for the reliable operation of complex real-world systems. Existing explanation methods often rely on unrealistic feature perturbations and ignore temporal and cross-feature dependencies, leading to unreliable attributions. We propose a conditional attribution framework that explains anomalies relative to contextually similar normal system states. Instead of using marginal or randomly sampled baselines, our method retrieves representative normal instances conditioned on the anomalous observation, enabling dependency-preserving and operationally meaningful explanations. To support high-dimensional time-series data, contextual retrieval is performed in learned low-dimensional representations using both variational autoencoder latent spaces and UMAP manifold embeddings. By grounding the retrieval process in the system's learned manifold, this strategy avoids out-of-distribution artifacts and ensures attribution fidelity while maintaining computational efficiency. We further introduce confidence-aware and temporal evaluation metrics for assessing explanation reliability and responsiveness. Experiments on the SWaT and MSDS benchmarks demonstrate that the proposed approach consistently improves root-cause identification accuracy, temporal localization, and robustness across multiple anomaly detection models. These results highlight the practical utility of conditional attribution for explainable anomaly diagnosis in complex time-series systems. Code and models are available at: GitHub repository.

Keywords: Explainable AI · Generative Models · Anomaly Detection.

1 Introduction

Modern anomaly detection systems are increasingly deployed in safety-critical and large-scale environments such as industrial control systems, cyber-physical infrastructure, and cloud platforms [12,24,1]. While recent advances in deep learning have significantly improved anomaly detection accuracy, practical deployment requires more than detection alone: operators must understand *why* an

anomaly occurred in order to diagnose faults, take corrective action, and prevent recurrence. This has positioned root cause analysis (RCA) as a central challenge in multivariate time-series monitoring.

Existing RCA methods often adapt feature attribution techniques [15,20] that assume feature independence. In multivariate time-series, strong temporal and cross-sensor correlations cause these methods to rely on unrealistic, out-of-distribution (OOD) perturbations [21]. Recent attempts to address these limitations through structured or group-wise attributions [10,3] still rely on fixed background datasets or heuristic sampling strategies, which fail to preserve the conditional structure of complex physical processes, scale poorly in high-dimensional settings, and offer limited guarantees regarding the fidelity of the explanations. Consequently, current explanations are often operationally implausible, creating a gap between detection and actionable diagnostic insight.

In this work, we bridge this gap with a **conditional attribution framework** that explains anomalies relative to contextually similar normal system states. To handle high-dimensionality and noise, we perform contextual retrieval in learned manifold representations using both Variational Autoencoders (VAE) [11] and UMAP [6]. By grounding attribution in representative normal instances that reflect realistic operating conditions, our framework preserves structural dependencies and ensures that explanations are both model-agnostic and physically consistent.

Beyond feature-level identification, we address the critical need for *temporal localization* and *principled evaluation* in time-series RCA. Our framework enables precise onset localization within long-sequence anomalies by performing attribution over structured temporal windows, capturing the evolution of abnormal behavior. To bridge the gap between heuristic rankings and operational utility, we introduce two novel metrics: (i) **CW-RCS**, a confidence-aware measure that incorporates attribution strength to penalize diffuse explanations, and (ii) **TemporalHM**, which quantifies the responsiveness and stability of root-cause identification following anomaly onset. Together, these contributions provide a more rigorous assessment of explanation fidelity and diagnostic reliability than standard retrieval-based measures.

We evaluate our framework on the **SWaT** [24] industrial benchmark and the high-dimensional **MSDS** [16] dataset, alongside a real-world case study on industrial blast furnace monitoring data. Our results demonstrate superior localization accuracy and operational utility in complex, dependent systems.

Our contributions are threefold:

1. **Conditional Attribution Framework:** A novel RCA approach that explains anomalies relative to contextually similar normal states, preserving temporal and cross-feature dependencies to avoid out-of-distribution artifacts.
2. **Manifold-Guided Contextual Retrieval:** A scalable, model-agnostic strategy using VAE and UMAP embeddings to retrieve representative baselines in high-dimensional latent spaces.

3. **Rigorous Evaluation Metrics:** We introduce **CW-RCS** (Confidence-Weighted Root Cause Score) and **TemporalHM** to quantify explanation reliability and temporal responsiveness, bridging the gap between detection and actionable diagnosis.

This work provides a principled foundation for high-fidelity, dependency-preserving explanations in complex time-series systems.

2 Related Work

2.1 Root Cause Analysis (RCA) in Multivariate Time-Series

Existing RCA paradigms generally fall into three categories: (i) Statistical methods (e.g., EXstream [26], SHAP [15,10]) which rank features by deviation intensity but often fall into the "symptom-as-cause" trap; (ii) Causal Discovery (e.g., MicroRCA [23], CloudRanger [22]) which utilize dependency graphs but struggle with the high-dimensionality and non-linearity of industrial data; and (iii) Reconstruction-based methods (e.g., OmniAnomaly [19], Interfusion [14]) which use generative models to identify root causes via reconstruction error or counterfactuals . Despite their utility, these methods frequently rely on marginal perturbations or static backgrounds that fail to preserve the complex conditional dependencies inherent in time-series systems.

2.2 Feature Attribution and Explainability

Feature attribution methods aim to quantify the contribution of individual input variables to model predictions. Model-agnostic approaches such as LIME [17] and SHAP [15] approximate local decision boundaries or decompose predictions into additive feature contributions. Gradient-based techniques, including Integrated Gradients [20], extend attribution to deep neural networks. While these methods have been widely adopted for tabular and vision tasks, their direct application to multivariate time-series anomaly detection remains challenging. In particular, many attribution techniques rely on marginal perturbations or independence assumptions that fail to preserve temporal and cross-feature dependencies, limiting their reliability for root cause analysis in complex monitoring systems.

3 Problem Formulation

We study root cause analysis (RCA) for anomalies in multivariate time-series. The key challenge is to produce explanations that are *faithful* to the anomaly detector while respecting the strong temporal and cross-sensor dependencies present in real systems.

3.1 Multivariate Time-Series Anomaly Detection

Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the d-dimensional sensor measurement at time $t \in \{1, \ldots, T\}$, and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ denote the full multivariate time-series. We define the sliding window of length w starting at time t as

$$\mathbf{W}_t := (\mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+w-1}) \in \mathbb{R}^{w \times d}. \tag{1}$$

An anomaly detector is a function $f : \mathbb{R}^{w \times d} \to \mathbb{R}$, which maps a window $\mathbf{W}_t$ to an anomaly score $s_t := f(\mathbf{W}_t)$.

3.2 Root Cause Attribution Objective

The objective of RCA is to estimate an attribution tensor $\boldsymbol{\Phi}_t \in \mathbb{R}^{w \times d}$, where each element $\Phi_t(\tau, j)$ quantifies the contribution of sensor j at relative time τ. The total sensor-level attribution, which we denote as $\phi_{t,j}$ to differentiate it from the local attribution values, is obtained by aggregating over the window:

$$\phi_{t,j} = \sum_{\tau=1}^{w} \Phi_t(\tau, j). \tag{2}$$

3.3 Conditional Attribution Formulation

Let $\mathbf{W}_t^{(-j)}$ denote the window context excluding the j-th sensor's trajectory. To quantify the contribution of sensor j relative to normal system behavior, let $p(\cdot)$ denote the distribution of normal (non-anomalous) system windows, and let $\mathbf{W}'_j$ denote a replacement trajectory for sensor j sampled from the corresponding normal distribution. We define the **marginal attribution** and the **conditional attribution** as:

$$\phi_{t,j}^{\text{marg}} = \mathbb{E}_{\mathbf{W}'_j \sim p(\mathbf{W}_j)} \left[f(\mathbf{W}_t) - f(\mathbf{W}_t^{(-j)}, \mathbf{W}'_j) \right] \tag{3}$$

$$\phi_{t,j}^{\text{cond}} = \mathbb{E}_{\mathbf{W}'_j \sim p(\mathbf{W}_j | \mathbf{W}_t^{(-j)})} \left[f(\mathbf{W}_t) - f(\mathbf{W}_t^{(-j)}, \mathbf{W}'_j) \right]. \tag{4}$$

In practice, direct sampling from the conditional distribution of normal behavior is intractable. We approximate it using a contextual neighborhood $\mathcal{N}(\mathbf{W}_t)$, defined as the set of K normal windows $\mathbf{W}'$ that minimize the contextual distance $d(\mathbf{W}_t^{(-j)}, \mathbf{W}'^{(-j)}) = \|\text{vec}(\mathbf{W}_t^{(-j)}) - \text{vec}(\mathbf{W}'^{(-j)})\|_2$ in the context space. This yields the empirical estimator:

$$\hat{\phi}_{t,j} = \frac{1}{K} \sum_{\mathbf{W}' \in \mathcal{N}(\mathbf{W}_t)} \left[f(\mathbf{W}_t) - f(\mathbf{W}_t^{(-j)}, \mathbf{W}'_j) \right]. \tag{5}$$

Proposition 1 (Dependency-Preserving Attribution). *Assume $f(\cdot)$ is L-Lipschitz continuous with respect to the j-th sensor trajectory, i.e., $|f(\mathbf{W}_t^{(-j)}, \mathbf{u}) -$*

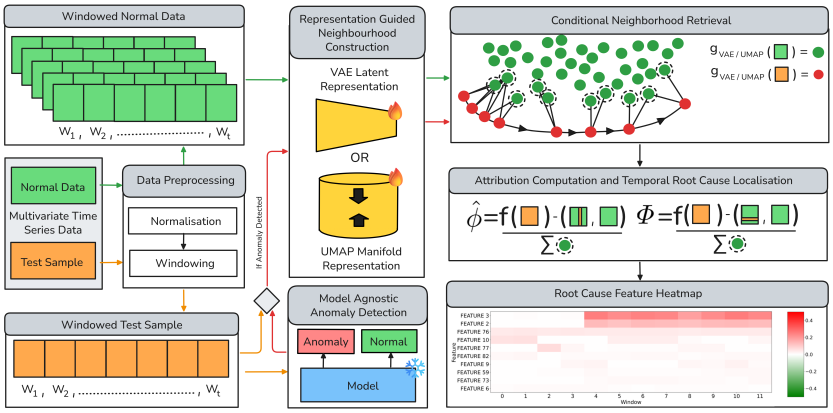


Fig. 1: Overview of the proposed model-agnostic RCA pipeline. Steps include latent-space neighborhood construction, conditional attribution computation, and temporal root cause localization for multivariate industrial sensors.

$f(\mathbf{W}_t^{(-j)}, \mathbf{v})| \leq L\|\mathbf{u} - \mathbf{v}\|_2$ *for any trajectories* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^w$. *The attribution bias induced by marginal perturbation is bounded by:*

$$|\phi_{t,j}^{cond} - \phi_{t,j}^{marg}| \leq L \cdot W_1 \left(p(\mathbf{W}_j \mid \mathbf{W}_t^{(-j)}), p(\mathbf{W}_j) \right), \tag{6}$$

where W_1 *is the Wasserstein-1 distance computed over trajectories in* $\mathbb{R}^w$ *using the* ℓ_2*-norm as the ground metric.*

Proof. Let $g(\mathbf{W}'_j) = f(\mathbf{W}_t^{(-j)}, \mathbf{W}'_j)$. By the linearity of expectation:

$$|\phi_{t,j}^{\text{cond}} - \phi_{t,j}^{\text{marg}}| = \left| \mathbb{E}_{p(\mathbf{W}_j|\mathbf{W}_t^{(-j)})}[g(\mathbf{W}'_j)] - \mathbb{E}_{p(\mathbf{W}_j)}[g(\mathbf{W}'_j)] \right|.$$

By the Lipschitz assumption, g is an L-Lipschitz function. According to the Kantorovich-Rubinstein duality, the following holds:

$$|\mathbb{E}_P[g] - \mathbb{E}_Q[g]| \leq L \cdot W_1(P, Q).$$

By substituting $P = p(\mathbf{W}_j \mid \mathbf{W}_t^{(-j)})$ and $Q = p(\mathbf{W}_j)$, we obtain the bound. The bias represents the systematic error introduced by evaluating the detector on out-of-distribution counterfactuals that violate learned sensor dependencies.

4 Conditional Attribution Framework

We operationalize the conditional attribution for RCA through a scalable, model-agnostic framework illustrated in Figure 1. Since direct sampling from the conditional distribution $P(\mathbf{x}_j \mid \mathbf{x}_{-j})$ is intractable in high-dimensional time-series,

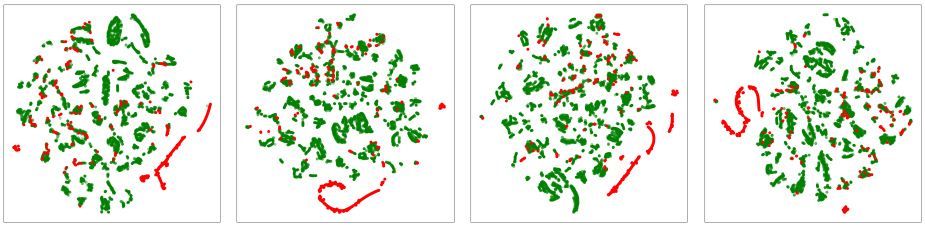


Fig. 2: SWaT Manifold Topology ($d \in 4, 8, 16, 32$). Normal (green) and anomalous (red) regions overlap significantly, indicating anomalies can be locally indistinguishable from normal operating modes. This highlights the need for a conditional framework.

we approximate it via **contextual neighborhood retrieval**. Representative normal windows are selected based on their proximity to the anomalous context in a learned representation space. Attribution is then computed through dependency-preserving counterfactual replacement using these retrieved neighbors, ensuring explanations remain grounded in the system's manifold and avoid out-of-distribution artifacts.

4.1 Representation-Guided Neighborhood Construction

Direct neighborhood construction in the input space becomes unreliable in high-dimensional multivariate time-series due to the curse of dimensionality and noise sensitivity. To obtain semantically meaningful conditional neighborhoods, we perform similarity retrieval in learned low-dimensional representations that preserve system dependencies.

Let $g : \mathbb{R}^{w \times d} \to \mathbb{R}^k$ denote an embedding function that maps windows to a k-dimensional representation space. The conditional neighborhood is then defined as

$$\mathcal{N}(\mathbf{W}_t) = \underset{\{\mathbf{W}^{(i)}\} \subset \mathcal{D}_{\text{norm}}}{\overset{K}{\arg\min}} \left\| g(\mathbf{W}_t^{(-j)}) - g(\mathbf{W}^{(i)(-j)}) \right\|_2 . \tag{7}$$

By performing retrieval in representation space, the neighborhood better reflects latent system structure and nonlinear dependencies, yielding more faithful approximations of the conditional distribution.

VAE Latent Representation We first instantiate $g(\cdot)$ using the encoder of a variational autoencoder [11] trained on normal system windows. Let the encoder map each window to a latent variable $\mathbf{z} \sim q_\theta(\mathbf{z} \mid \mathbf{W})$. We define the embedding as the posterior mean:

$$g_{\text{VAE}}(\mathbf{W}) = \mathbb{E}[\mathbf{z} \mid \mathbf{W}]. \tag{8}$$

Neighborhood retrieval in latent space captures nonlinear system dependencies while reducing dimensionality, improving both computational efficiency and conditional fidelity.

UMAP Manifold Representation As an alternative representation, we employ Uniform Manifold Approximation and Projection (UMAP) to learn a low-dimensional embedding that preserves local neighborhood topology. Let

$$g_{\text{UMAP}} : \mathbb{R}^{w \times d} \to \mathbb{R}^k \tag{9}$$

denote the learned manifold embedding. Conditional neighborhoods constructed in this space capture geometric similarity between system states, enabling non-parametric retrieval that preserves local manifold connectivity without the reconstruction bias of generative models.

4.2 Conditional Neighborhood Retrieval

To approximate $p(\mathbf{W}_j \mid \mathbf{W}_t^{(-j)})$, we construct a sensor-specific neighborhood $\mathcal{N}_j(\mathbf{W}_t)$ from a reference set of normal windows $\mathcal{D}_{\text{norm}}$. For an anomalous window $\mathbf{W}_t$, we retrieve the K nearest neighbors conditioned on the context $\mathbf{W}_t^{(-j)}$:

$$\mathcal{N}_j(\mathbf{W}_t) = \text{KNN}_K\left(\mathbf{W}_t^{(-j)}, \mathcal{D}_{\text{norm}}\right), \tag{10}$$

where $\text{KNN}_K(\cdot)$ returns the K nearest neighbors under the Frobenius distance $\|\cdot\|_F$. This conditioning ensures that retrieved states preserve system dependencies, enabling counterfactual replacements for sensor j that remain within the normal data manifold.

4.3 Attribution Computation

We operationalize the framework by estimating the empirical conditional attribution $\hat{\phi}_{t,j}$ as defined in Equation (5). This involves averaging the model's response across the retrieved neighborhood $\mathcal{N}(\mathbf{W}_t)$ while keeping the non-target sensor context $\mathbf{W}_t^{(-j)}$ fixed. By grounding the counterfactual replacement in representative normal instances rather than global averages, the computation ensures that the resulting importance scores reflect realistic system deviations. This model-agnostic approach allows the framework to scale efficiently across high-dimensional sensor suites without making restrictive assumptions about the underlying anomaly detector's architecture.

4.4 Temporal Root Cause Localization

Beyond identifying contributing sensors, root cause analysis in time-series requires localizing when anomalous behavior emerges. To achieve this, we extend conditional attribution to the temporal dimension by computing contributions over localized segments within the window.

For each sensor j and relative time index τ, we construct time-localized counterfactuals by replacing the value $x_{t+\tau-1,j}$ (or short temporal segments centered at τ) with corresponding values drawn from neighborhood windows while preserving the remaining context. The temporal attribution is then estimated as

$$\Phi_t(\tau, j) = \frac{1}{K} \sum_{\mathbf{W}' \in \mathcal{N}(\mathbf{W}_t)} \left[f(\mathbf{W}_t) - f(\mathbf{W}_t^{(-(\tau,j))}, \mathbf{W}'_{(\tau,j)}) \right], \tag{11}$$

where $\mathbf{W}_t^{(-(\tau,j))}$ denotes the window with the value at (τ, j) replaced, and $\mathbf{W}'_{(\tau,j)}$ denotes the corresponding value from the neighborhood window.

Aggregating $\Phi_t(\tau, j)$ across τ recovers sensor-level attribution, while temporal aggregation across j enables anomaly onset localization, yielding a structured sensor–time explanation map.

4.5 Computational Complexity Analysis

Let N denote the number of normal windows, w the window length, and d the number of sensors. Input-space neighborhood retrieval requires computing distances in $\mathbb{R}^{w \times d}$, resulting in a complexity of $\mathcal{O}(Nwd)$ per query.

When retrieval is performed in a learned representation space $g(\cdot) \in \mathbb{R}^k$ with $k \ll wd$, the cost reduces to $\mathcal{O}(Nk)$. Attribution requires K counterfactual evaluations per sensor, leading to a total cost of $\mathcal{O}(dK)$ model evaluations.

Thus, representation-guided retrieval improves scalability while preserving conditional neighborhood fidelity.

5 Evaluation Metrics

We evaluate RCA using standard retrieval metrics and introduce two novel measures for explanation reliability and temporal responsiveness. Unlike ranking metrics that ignore attribution strength and detection delay, our metrics capture attribution concentration and detection timeliness, enabling a more rigorous evaluation of operational utility in multivariate time series.

5.1 Retrieval-Based Metrics

Following prior work [25,5], we evaluate root cause identification using the **Top-K Recall** ($R@K$) metric. Let $\mathcal{S}_t \subseteq \{1, \ldots, d\}$ denote the set of ground-truth root cause sensors for anomaly window W_t, and $\widehat{\mathcal{R}}_t^{(K)}$ be the set of top-K sensors ranked by the aggregated attribution scores $\phi_{t,j}$. We assign a uniform weight $w_{t,j} = 1/|\mathcal{S}_t|$ to each causal sensor. The recall for anomaly t is:

$$\mathrm{R@}K(t) = \sum_{j \in \mathcal{S}_t} w_{t,j}\, \mathbf{1}(j \in \widehat{\mathcal{R}}_t^{(K)})$$

The dataset-level score $\mathrm{R@}K$ is the average of $\mathrm{R@}K(t)$ over all N anomalous windows. We report scores for $K \in \{3, 5, 10\}$, reflecting realistic settings where operators inspect only the top candidate sensors.

5.2 Confidence-Aware Metrics

Traditional RCA metrics evaluate whether ground-truth sensors appear in the top rankings but ignore the relative *magnitude* of attribution. To reward explanations that assign dominant attribution mass to the correct causes, we propose the **Confidence-Weighted Root Cause Score (CW-RCS)**.

Definition. Let $\phi_{t,j}$ be the attribution for sensor j at time t. We define the normalized confidence weight as $\tilde{\phi}_{t,j} = |\phi_{t,j}| / \sum_{\ell=1}^{d} |\phi_{t,\ell}|$. Let $\mathcal{S}_t$ denote the set of ground-truth root cause sensors and $\widehat{\mathcal{R}}_t^{(K)}$ be the set of top-K ranked sensors. The CW-RCS for an anomaly at time t is:

$$\text{CW-RCS}(t) = \frac{1}{|\mathcal{S}_t|} \sum_{j \in \mathcal{S}_t \cap \widehat{\mathcal{R}}_t^{(K)}} \tilde{\phi}_{t,j}. \tag{12}$$

The dataset-level score is the mean over N anomalous windows:

$$\text{CW-RCS@}K = \frac{1}{N} \sum_{t=1}^{N} \text{CW-RCS}(t). \tag{13}$$

Unlike rank-based metrics, CW-RCS rewards the model for maximizing the attribution gap between true causes and noise.

Property. The proposed metric is bounded by the standard Recall@K:

$$\text{CW-RCS}(t) \leq \text{Recall@}K(t). \tag{14}$$

This holds because for any sensor j, the normalized attribution $\tilde{\phi}_{t,j} \leq 1$. Consequently, the sum of weights for correctly retrieved sensors in $\mathcal{S}_t \cap \widehat{\mathcal{R}}_t^{(K)}$ is naturally upper-bounded by the cardinality of that intersection. CW-RCS thus penalizes cases where the detector identifies the correct sensor but fails to distinguish it clearly from spurious attributions.

5.3 Temporal Identification Metrics

Explanation quality in time-series RCA depends on both the *latency* of identification and its *consistency* throughout the anomaly. We evaluate these via two complementary measures:

Components: Early Identification (E) and Persistence (A). Let t_0 be the anomaly onset, T_a its duration, and t^* the first timestamp where any ground-truth sensor $j \in \mathcal{S}$ appears in the Top-K set $\widehat{\mathcal{R}}_t^{(K)}$. We define these metrics as:

$$E = \max\left(0, 1 - \frac{t^* - t_0}{T_a}\right), \qquad A = \frac{1}{T_a} \sum_{t=t_0}^{t_0+T_a-1} \mathbf{1}(\mathcal{S} \cap \widehat{\mathcal{R}}_t^{(K)} \neq \emptyset). \tag{15}$$

Harmonic Combined Temporal Score (TemporalHM). To ensure responsiveness and stability are treated as mutually necessary, we propose the **TemporalHM**, generalized via the F_β form:

$$\text{TemporalHM}_\beta = \frac{(1+\beta^2)EA}{\beta^2 E + A + \varepsilon} \tag{16}$$

where ε provides numerical stability and β allows prioritizing latency ($\beta > 1$) or persistence ($\beta < 1$). This formulation heavily penalizes methods that achieve high early scores but fail to maintain attribution consistency, or vice-versa.

6 Experiments

We evaluate the proposed RCA framework to answer the following questions:

- **RQ1 (RCA Accuracy):** To what extent can the framework accurately identify the ground-truth root cause sensors (measured by Top@KR)?
- **RQ2 (Attribution Confidence):** Does the method effectively concentrate attribution mass on true causal variables, reducing noise in the explanation (measured by CW-RCS@K)?
- **RQ3 (Temporal Dynamics):** How early and consistently does the framework localize root causes throughout the duration of an anomaly (measured by TemporalHM)?
- **RQ4 (Model Agnosticism):** Is the framework robust and effective when integrated with diverse underlying anomaly detection architectures?

The following subsections describe the datasets, models and baselines, and implementation details.

6.1 Datasets

We evaluate our framework on two widely-used benchmarks: SWaT (Secure Water Treatment) [24], containing 51 sensors from a continuous industrial process with 11 days of operation, and MSDS [16] (Multi-Source Distributed System), a high-dimensional dataset capturing 10 metrics across 12 nodes in a distributed computing environment. In our experiments we utilize system metrics modality.

6.2 Models and Baselines

Backbone Models: To demonstrate the detector-agnostic nature of the proposed RCA framework,we employ a diverse suite of anomaly detection architectures, including Autoencoder (AE) [7], Variational Autoencoder (VAE) [11], LSTM-based detectors [9], TCN [2], and Transformer-based [25] models. This selection ensures the attribution framework is tested against varying inductive biases and internal representations.

RCA Baseline: We compare our approach against representative RCA baselines spanning both causal inference and attribution-based methods: ϵ-Diagnosis [18], RCD [8], CIRCA [13], and AERCA [5] for causal root cause localization, as well as KernelSHAP [15] and ShaTS [4] for feature attribution-based explanations.

Proposed Variants: We evaluate two framework variants:**CondAttr-VAE** uses a learned VAE latent space for neighborhood search, while **CondAttr-UMAP** leverages a UMAP manifold embedding. This design ensures the framework remains model-agnostic and decoupled from the specific anomaly detection backbone. Refer to Appendix A for implementation and hyperparameter details.

6.3 Implementation Details

All sensor variables are normalized using standard z-score normalization. Sliding windows of length $w = 50$ are constructed to capture temporal system dynamics. Anomaly detection models are trained using only normal operation data.

For contextual retrieval, the neighborhood size is fixed to nearest contextual neighbor ($K = 3$). In the VAE variant, the latent representation is learned using a bottleneck dimension of $z = 8$, while the UMAP variant constructs a low-dimensional manifold embedding of the input windows for neighbor search.

Experiments are conducted in PyTorch on an NVIDIA RTX A6000 GPU; see Appendix B for hyperparameter studies.

7 Results and Analysis

7.1 Root Cause Identification Performance

Table 1: **Root Cause Identification Performance**. Top-K Recall reported per dataset. **Best** and second-best results are highlighted.

Method	**SWaT**			**MSDS**		
	Top@3R	**Top@5R**	**Top@10R**	**Top@3R**	**Top@5R**	**Top@10R**
ϵ-Diagnosis	0.125	0.125	0.375	0.266	0.452	1.000
RCD	0.000	0.000	0.300	0.573	0.984	1.000
CIRCA	0.000	0.000	0.300	0.860	0.917	1.000
AERCA	0.290	0.330	0.455	0.908	0.974	1.000
KernelSHAP	0.055	0.064	0.138	0.311	0.467	1.000
ShaTS	0.393	0.513	0.601	0.915	0.986	1.000
CondAttr-VAE	**0.537**	**0.601**	**0.694**	0.948	**1.000**	1.000
CondAttr-UMAP	0.481	0.523	0.638	**0.956**	**1.000**	1.000

As shown in Table 1, **CondAttr-UMAP** and **CondAttr-VAE** consistently outperform all baselines across both benchmarks. On SWaT, our manifold-guided approaches provide a significant lift over the strongest baseline (ShaTS),

with improvements of up to **36.64%** in Top-3 Recall. Similar gains on the high-dimensional MSDS dataset underscore the scalability of our contextual retrieval strategy. These results empirically validate that conditioning attributions on the learned system manifold, rather than using marginal perturbations, better preserves inter-sensor dependencies and reduces out-of-distribution artifacts, leading to more precise root-cause localization.

Additional comparisons between conditional and unconditional retrieval strategies are provided in Appendix B.6, further illustrating the benefits of conditioning the reference set on the anomalous context.

7.2 Confidence-Aware Evaluation

Table 2 evaluates attribution quality via the proposed *CW-RCS@K* metric, denoted as *CW@K* in the table for brevity. **CondAttr-UMAP** consistently outperforms the strongest baseline, [4], nearly doubling the scores across all K on SWaT [24]. While absolute scores are lower than standard Recall@K due to the strictness of confidence weighting, the substantial performance gap confirms that our conditional framework effectively concentrates attribution mass on ground-truth root causes. Unlike marginal baselines that disperse importance across spurious correlations, our manifold-guided approach enhances the **separability** of root causes from background noise, providing more decisive and operationally actionable explanations.

Table 2: **Confidence-Aware Evaluation**. CW-RCS@K per dataset. **Best** and second-best results are highlighted.

Method	**SWaT**			**MSDS**		
	CW@3	**CW@5**	**CW@10**	**CW@3**	**CW@5**	**CW@10**
KernelSHAP	0.004	0.004	0.009	0.050	0.093	0.144
ShaTS	0.122	0.134	0.141	0.462	0.489	0.517
CondAttr-VAE	**0.245**	**0.253**	0.257	0.551	0.607	0.664
CondAttr-UMAP	0.243	0.250	**0.258**	**0.569**	**0.624**	**0.668**

7.3 Cross-Model Explanation Consistency

Table 3 evaluates the robustness of our framework across diverse anomaly detection backbones. While marginal attribution methods exhibit high variance and poor confidence scores, **CondAttr-UMAP** maintains superior performance across all detectors. Notably, on high-capacity models like the Transformer and TCN, our method achieves nearly double the CW@3 scores compared to ShaTS [4]. This consistency suggests that by decoupling the explanation logic from the detector's internal parameters and grounding it in the system's normal manifold,

Table 3: **Cross-model Consistency**. Top-3 Recall and confidence-weighted recall (CW-RCS@3) across anomaly detection backbones. **Best** and second-best results are highlighted.

Method	VAE		AE		LSTM		TCN		Transformer	
	Top@3R	CW@3	Top@3R	CW@3	Top@3R	CW@3	Top@3R	CW@3	Top@3R	CW@3
KernelSHAP	0.055	0.004	0.097	0.038	0.041	0.006	0.046	0.017	0.092	0.011
ShaTS	0.393	0.122	0.332	0.158	0.407	0.126	0.462	0.138	0.475	0.196
CondAttr-VAE	**0.537**	**0.245**	**0.425**	0.216	**0.490**	**0.229**	0.555	0.284	**0.564**	**0.325**
CondAttr-UMAP	0.481	0.238	0.402	**0.222**	0.462	0.213	**0.569**	**0.342**	0.518	0.303

we mitigate model-specific biases. Our approach thus provides a model-agnostic layer of reliability, ensuring that root cause attributions remain stable regardless of the underlying detection architecture.

7.4 Temporal Localization Analysis

As shown in Table 4, our manifold-guided strategies lead in **TemporalHM** (denoted *TempHM* in tables), with the UMAP variant achieving the highest fidelity. The substantial margin over ShaTS [4], particularly at $K = 3$, underscores our framework's capacity to minimize localization latency while ensuring persistence. Unlike marginal perturbations that often yield transient or "flickering" attributions, our conditional approach stabilizes explanations by grounding them in the system's learned normal manifold. This ensures root causes are identified promptly at onset and tracked reliably as the anomaly evolves, providing the stability required for real-time industrial monitoring.

Table 4: **Temporal Localization Analysis**. **Best** and second-best results are highlighted.

Dataset	Method	TempHM@3	TempHM@5	TempHM@10
SWaT	KernelSHAP	0.064	0.078	0.171
	ShaTS	0.422	0.519	0.562
	CondAttr-VAE	0.503	0.562	**0.650**
	CondAttr-UMAP	**0.504**	**0.568**	0.629

8 Industrial Case Study: Blast Furnace Monitoring

To assess the practical utility of the proposed framework, we conduct an industrial case study on real-world blast furnace monitoring data. The study evaluates root cause attribution performance in a complex thermo-chemical manufacturing environment under domain-expert supervision.

8.1 System Overview and Data Description

We validate our framework in collaboration with **Paul Wurth** using data from an operational blast furnace. This complex thermo-chemical system is monitored by over 100 heterogeneous sensors (temperature, gas composition, pressure) across multiple structural levels. The dataset contains 2 million observations over several production cycles. While industrial confidentiality precludes public release, the data underwent rigorous curation, including sensor validation, noise filtering, and synchronization, in close coordination with domain experts to ensure physically consistent inputs for root cause analysis.

8.2 Modeling Pipeline

Following standard preprocessing (normalization and alignment), we constructed sliding windows with lengths selected in consultation with domain experts to match furnace reaction and material transit times. We trained reconstruction-based detectors, specifically Autoencoder (AE) and VAE models, on normal operating data. Both architectures achieved stable reconstruction and the necessary sensitivity for downstream RCA, serving as the model-agnostic backbones for our attribution framework.

8.3 Operational Requirements at Paul Wurth

For Paul Wurth, anomaly detection is only actionable if it identifies the specific causal sensors driving a deviation. Given that blast furnace thermo-chemical cycles span several hours, late-stage discovery leads to irreversible material loss and significant energy waste. Our framework addresses this by isolating sensors deviating from their *context-dependent* patterns. This provides operators with precise localization needed for targeted intervention during the critical early phases of a process deviation, transforming a binary alert into a diagnostic insight.

8.4 Validation and Stress Testing

We validated our framework against expert-annotated historical anomalies and process logs. To further quantify reliability, we conducted controlled stress tests by injecting synthetic anomalies into individual sensors and correlated groups. Across both scenarios, the framework consistently recovered the ground-truth drivers, demonstrating high sensitivity to deviations at their earliest stages. As illustrated by the attribution heatmaps in Figure 3, our approach provides the interpretability required for industrial diagnosis, accurately isolating causal sensors from normal system noise. Appendix C provides additional anomaly cases and experimental results.

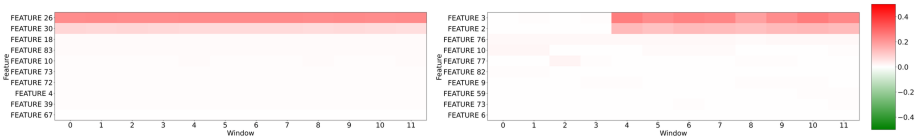


Fig. 3: **Root cause feature heatmaps for representative Paul Wurth samples.** The left panel shows a real test sample with `feature26` identified as the root cause, while the right panel shows a synthetic anomaly with perturbations injected in `feature2` and `feature3` after a given time point.

8.5 Industrial Impact and Portability

The Paul Wurth case study validates the practical utility of conditional attribution in mitigating downtime and resource loss through early diagnostic intervention. Beyond blast furnace monitoring, the framework's detector-agnostic design ensures seamless integration with diverse anomaly detection architectures. This versatility supports broad transferability to other complex multivariate industrial systems, providing a scalable solution for high-fidelity root cause diagnosis where system dependencies are critical.

9 Conclusion

In this paper, we addressed the fundamental limitations of marginal attribution in time-series RCA, which often lead to unreliable, out-of-distribution explanations. We proposed a conditional attribution framework grounded in the system's learned manifold, utilizing VAE and UMAP representations to retrieve contextually relevant normal states for counterfactual construction. To rigorously assess the utility of these explanations, we introduced two novel metrics: the confidence-weighted *CW-RCS* and the stability-focused *TemporalHM*. Our experiments on the SWaT and MSDS benchmarks, along with a case study on industrial blast furnace data from **Paul Wurth**, demonstrate that our approach consistently improves root-cause identification accuracy and temporal responsiveness. Qualitative evaluation by domain experts further confirms the practical utility of our approach, as the identified root causes showed strong alignment with known physical process failures. By providing high-fidelity, dependency-preserving explanations, our framework bridges the gap between complex anomaly detection models and actionable industrial diagnostics. Future work will explore the extension of this manifold-guided strategy to online, incremental learning settings where system dynamics evolve over time.

10 Acknowledgements

This work was partially funded by the German Federal Ministry of Research, Technology and Space (BMFTR) under Grant Agreement No. 16IW24009 (COPPER).

References

1. Abshari, D., Sridhar, M.: A survey of anomaly detection in cyber-physical systems. arXiv preprint arXiv:2502.13256 (2025)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
3. Covert, I., Lundberg, S., Lee, S.I.: Understanding global feature contributions with additive importance measures (2020), https://arxiv.org/abs/2004.00668
4. De La Peña, M.F., Gómez, Á.L.P., Maimó, L.F.: Shats: A shapley-based explainability method for time series artificial intelligence models. Future Generation Computer Systems p. 108178 (2025)
5. Han, X., Absar, S., Zhang, L., Yuan, S.: Root cause analysis of anomalies in multivariate time series through granger causal discovery. In: The Thirteenth International Conference on Learning Representations (2025)
6. Healy, J., McInnes, L.: Uniform manifold approximation and projection. Nature Reviews Methods Primers **4**(1), 82 (2024)
7. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006)
8. Ikram, A., Chakraborty, S., Mitra, S., Saini, S., Bagchi, S., Kocaoglu, M.: Root cause analysis of failures in microservices through causal discovery. Advances in Neural Information Processing Systems **35**, 31158–31170 (2022)
9. Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., Tatbul, N.: Exathlon: A benchmark for explainable anomaly detection over time series. arXiv preprint arXiv:2010.05073 (2020)
10. Jullum, M., Redelmeier, A., Aas, K.: groupshapley: Efficient prediction explanation with shapley values for feature groups (2021), https://arxiv.org/abs/2106.12228
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Kumar, P., Pandi, S.S., Kumar, L.B., Karthick, R.: Anomaly detection in industrial control systems using machine learning. In: 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS). pp. 1–6. IEEE (2025)
13. Li, M., Li, Z., Yin, K., Nie, X., Zhang, W., Sui, K., Pei, D.: Causal inference-based root cause analysis for online service systems with intervention recognition. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 3230–3240 (2022)
14. Li, Z., Zhao, Y., Han, J., Su, Y., Jiao, R., Wen, X., Pei, D.: Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 3220–3230 (2021)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
16. Nedelkoski, S., Bogatinovski, J., Mandapati, A.K., Becker, S., Cardoso, J., Kao, O.: Multi-source distributed system data for ai-powered analytics. In: European conference on service-oriented and cloud computing. pp. 161–176. Springer (2020)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

18. Shan, H., Chen, Y., Liu, H., Zhang, Y., Xiao, X., He, X., Li, M., Ding, W.: ?-diagnosis: Unsupervised and real-time diagnosis of small-window long-tail latency in large-scale microservice platforms. In: The World Wide Web Conference. pp. 3215–3222 (2019)
19. Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D.: Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2828–2837 (2019)
20. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
21. Vlassopoulos, G., van Erven, T., Brighton, H., Menkovski, V.: Explaining predictions by approximating the local decision boundary. arXiv preprint arXiv:2006.07985 (2020)
22. Wang, P., Xu, J., Ma, M., Lin, W., Pan, D., Wang, Y., Chen, P.: Cloudranger: Root cause identification for cloud native systems. In: 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). pp. 492–502. IEEE (2018)
23. Wu, L., Tordsson, J., Elmroth, E., Kao, O.: Microrca: Root cause localization of performance issues in microservices. In: IEEE/IFIP Network Operations and Management Symposium (NOMS) (2020)
24. Xie, X., Wang, B., Wan, T., Tang, W.: Multivariate abnormal detection for industrial control systems using 1d cnn and gru. Ieee Access **8**, 88348–88359 (2020)
25. Xu, J., Wu, H., Wang, J., Long, M., Wang, J.: Anomaly transformer: Time series anomaly detection with association discrepancy. In: International Conference on Learning Representations (2022)
26. Zhang, H., Diao, Y., Meliou, A.: Exstream: Explaining anomalies in event stream monitoring. In: Proceedings of the 20th international conference on extending database technology (EDBT) (2017)

Appendix: Conditional Attribution for Root Cause Analysis in Time-Series Anomaly Detection

A Additional Model Details

A.1 Model Details

VAE Architecture and Hyperparameters The variational autoencoder (VAE) employed in *CondAttr-VAE* acts as the underlying latent generative model for conditional attribution. Given an input time-series window $\mathbf{X} \in \mathbb{R}^{T \times D}$, the encoder maps the input to the parameters of a latent Gaussian distribution, namely the mean vector $\boldsymbol{\mu}$ and log-variance vector $\log \boldsymbol{\sigma}^2$. A latent representation is then sampled via the reparameterization trick and passed through the decoder to obtain the reconstruction $\hat{\mathbf{X}}$.

The VAE is trained using a combined objective consisting of a reconstruction loss, a KL regularization term, and a time-axial consistency loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{time}}\mathcal{L}_{\text{time}}, \tag{1}$$

where the reconstruction loss is defined as

$$\mathcal{L}_{\text{rec}} = \sum_{t=1}^{T}\sum_{d=1}^{D}\left(X_{t,d} - \hat{X}_{t,d}\right)^2, \tag{2}$$

and the KL divergence term is

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2}\sum_{j=1}^{L}\left(1 + \log\sigma_j^2 - \mu_j^2 - \exp(\log\sigma_j^2)\right), \tag{3}$$

where L denotes the latent dimension. To further encourage temporal consistency in the reconstruction, we include a time-axial loss defined as

$$\mathcal{L}_{\text{time}} = \sum_{t=1}^{T}\left(\frac{1}{D}\sum_{d=1}^{D}X_{t,d} - \frac{1}{D}\sum_{d=1}^{D}\hat{X}_{t,d}\right)^2. \tag{4}$$

Table 1 summarizes the architectural and training hyperparameters of the VAE used in our experiments.

Table 1: **VAE architecture and training hyperparameters.** Hyperparameter configuration of the timeVAE model used in our experiments.

Hyperparameter	**Value**
Input dimension	50
Latent dimension	8
Hidden layer sizes	[50, 100, 200]
Reconstruction weight	3.0
KL-Divergence weight	1.0
Time-Axial weight	1.0
Batch size	512
Maximum epochs	300

UMAP Architecture and Hyperparameters The UMAP-based variant, referred to as *CondAttr-UMAP*, employs Uniform Manifold Approximation and Projection (UMAP) as the nonlinear dimensionality reduction module underlying conditional attribution. Given a multivariate time-series window of length T with D features, each input sample is reshaped into a flattened vector of dimension $T \times D$ before being projected into a lower-dimensional embedding space. In our implementation, UMAP is configured with `n_neighbors = 30`, `n_components = 10`, `min_dist = 0.1`, and `random_state = 42`. All remaining parameters are retained at their default settings, including the Euclidean distance metric, spectral initialization, learning rate of 1.0, spread of 1.0, and automatic selection of the number of optimization epochs.

Table 2: **UMAP hyperparameters.** Hyperparameter configuration of the UMAP model used in our experiments.

Hyperparameter	**Value**
Output dimension	10
Number of neighbors	30
Minimum distance	0.1
Distance metric	Euclidean
Initialization	Spectral
Learning rate	1.0
Spread	1.0
Optimization epochs	Automatic
Random state	42

A.2 Anomaly Detection Scores

Table 3 reports the anomaly detection performance of different backbone models on the SWaT dataset.

Table 3: **Anomaly detection performance on the SWaT dataset** across different backbone models.

Model	Precision	Recall	F1-Score	ROC-AUC
VAE	0.952	0.901	0.926	0.954
AE	0.960	0.910	0.934	0.962
LSTM	0.979	0.923	0.948	0.980
TCN	0.989	0.961	0.975	0.990
Transformer	0.969	0.949	0.959	0.983

B Ablation

B.1 Additional Dataset Details

We evaluate our method on two real-world multivariate time-series benchmarks:

SWaT (Secure Water Treatment) is derived from a scaled-down but high-fidelity six-stage water treatment testbed. The SWaT.A1_Dec 2015 release consists of 11 days of continuous operation, of which 7 days correspond to normal operation and 4 days contain attack scenarios. The dataset provides labeled measurements from 51 sensors and actuators, and includes 41 attacks during the abnormal period.

MSDS (Multi-Source Distributed System) is collected from a complex distributed OpenStack environment and is designed to facilitate AIOps tasks, including anomaly detection and root cause analysis. It comprises multi-source observability data, including metrics, logs, and distributed traces, along with workload and fault scripts that provide ground truth. In our experiments, we use the system metrics modality and adopt the benchmark setting with 10 variables.

Table 4: **Summary of dataset statistics.** Number of features corresponds to the number of columns in each multivariate time-series dataset.

Dataset	Features	Timesteps
SWaT	51	49,500
MSDS	10	29,268

B.2 Inference Time

Table 5 presents a comparison of explanation methods in terms of both computational efficiency and localization performance. Specifically, for each dataset, we report the average inference time (in seconds) required to explain a single anomalous window, along with the corresponding Top-K Recall. While inference

time reflects the practical scalability of an explanation method, Top-K Recall indicates its ability to recover the true anomalous variables among the highest-ranked explanations.

Table 5: **Comparison of inference latency and Top@3R across explanation methods.** We report the mean inference time (seconds per window) and Top@3R accuracy for both benchmarks.

Method	SWAT		MSDS	
	Time (s)	Top@3R	Time (s)	Top@3R
KernelSHAP	18.653	0.055	0.402	0.311
ShaTS	21.822	0.393	0.441	0.915
CondAttr-VAE	21.952	0.537	0.461	0.948
CondAttr-UMAP	22.011	0.481	0.475	0.956

B.3 Window Size Impact

Table 6 reports the sensitivity of the proposed conditional attribution methods to the choice of input window size. Specifically, we evaluate CondAttr-VAE and CondAttr-UMAP using Top@3R, CW-RCS@3, and TemporalHM@3 over different temporal window lengths. The results indicate that explanation performance is generally higher for smaller or intermediate window sizes, whereas larger windows tend to degrade root cause localization quality. This behavior suggests a trade-off between capturing sufficient temporal context and preserving attribution specificity. Based on this trade-off, we choose a window size of 50 for reporting the main results in the paper, as it represents a moderate setting that balances temporal context and localization performance. The remaining ablation results for shorter and longer window sizes are provided here for completeness.

Table 6: **Sensitivity analysis of RCA performance across varying window sizes.**

Window Size	CondAttr-VAE			CondAttr-UMAP		
	Top@3R	CW@3	TempHM@3	Top@3R	CW@3	TempHM@3
5	0.592	0.318	0.436	0.523	0.312	0.413
10	0.574	0.320	0.446	0.564	0.327	0.481
20	0.509	0.275	0.437	0.513	0.318	0.465
50	0.537	0.245	0.503	0.481	0.243	0.504
100	0.435	0.196	0.397	0.453	0.236	0.479
200	0.287	0.193	0.317	0.393	0.189	0.412
500	0.287	0.177	0.279	0.425	0.239	0.410

B.4 Attribution Sizes Impact

Table 7 presents the sensitivity of CondAttr-VAE and CondAttr-UMAP to the attribution sample size. We report Top@3R, CW-RCS@3, and TemporalHM@3 for varying numbers of attribution samples. The results show that increasing the attribution size from very small values leads to initial improvements, particularly for CondAttr-VAE. Beyond this range, however, the gains become marginal, indicating a clear performance plateau for larger attribution sizes. This suggests that moderate attribution sizes are sufficient for obtaining reliable explanations, while further increases offer limited additional benefit.

Table 7: **Impact of attribution size on RCA performance.** Results for CondAttr-VAE and CondAttr-UMAP across different attribution sizes, evaluated using Top@3R, CW@3, and TempHM@3. Performance generally stabilizes for larger attribution sizes.

Attribution Size	**CondAttr-VAE**			**CondAttr-UMAP**		
	Top@3R	CW@3	TempHM@3	Top@3R	CW@3	TempHM@3
1	0.472	0.235	0.446	0.468	0.237	0.484
2	0.513	0.246	0.489	0.476	0.241	0.496
3	0.537	0.245	0.503	0.481	0.243	0.504
4	0.537	0.241	0.504	0.482	0.244	0.504
5	0.537	0.239	0.499	0.493	0.244	0.501
10	0.550	0.243	0.467	0.508	0.246	0.486
20	0.550	0.238	0.470	0.509	0.246	0.478
50	0.550	0.241	0.461	0.511	0.245	0.468

B.5 Input vs Representation-Space Retrieval

In high-dimensional sensor data, nearest-neighbor retrieval in the **raw input space** is unreliable because distance metrics (e.g., Euclidean) treat all features uniformly, causing irrelevant sensor fluctuations to distort similarity relationships and produce noisy neighborhoods.

To address this, we perform retrieval in **learned representation spaces**. A Variational Autoencoder (VAE) maps the input $x \in \mathbb{R}^d$ to a lower-dimensional latent representation $z \in \mathbb{R}^k$, capturing dominant system factors while suppressing noise. We further apply UMAP to preserve the **local manifold structure** of the latent space.

Retrieval quality is evaluated using *CW-RCS@3*. As shown in Fig. 1, representation space retrieval (VAE latent and UMAP manifold) significantly outperforms raw-input retrieval, indicating that learned representations better capture the *intrinsic system manifold* and yield more consistent nearest neighbors for anomaly analysis.

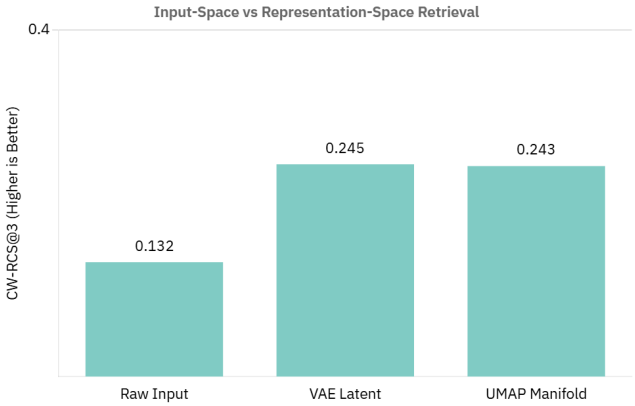


Fig. 1: Comparison of retrieval performance across spaces. Representation-space retrieval (VAE latent and UMAP manifold) achieves higher CW-RCS than raw input space.

B.6 Unconditional vs Conditional Retrieval

We study the impact of retrieval context on anomaly attribution. **Unconditional retrieval** constructs the reference set using K randomly sampled normal windows (global baseline), ignoring the current operating regime. This mixes heterogeneous system states and leads to noisy attribution.

Conditional retrieval instead selects K reference samples similar to the *non-anomalous sensors* of the current sample, conditioning the baseline on the current system state.

As shown in Fig. 2, unconditional retrieval produces diffuse heatmaps with multiple false positives, whereas conditional retrieval yields a cleaner attribution with a localized root-cause sensor.

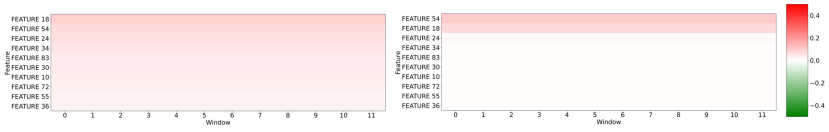
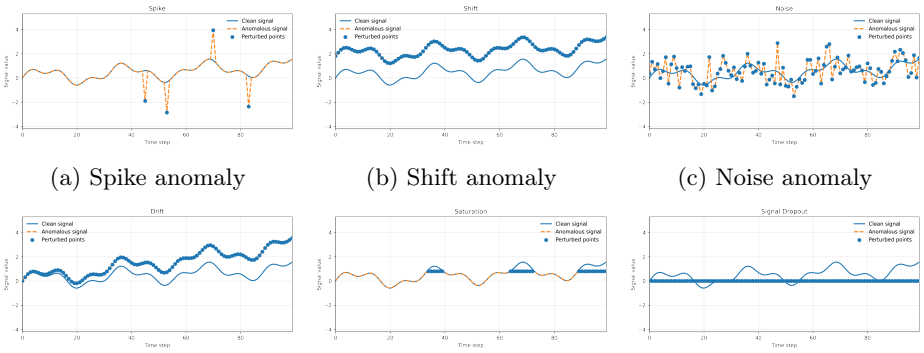


Fig. 2: **Unconditional versus conditional retrieval.** Conditional retrieval produces cleaner and more localized attributions by focusing on more relevant reference patterns.



(a) Spike anomaly (b) Shift anomaly (c) Noise anomaly

(d) Drift anomaly (e) Saturation anomaly (f) Signal Dropout anomaly

Fig. 3: Examples of synthetic anomaly types used in our experiments: spike, shift, noise, drift, saturation, and signal dropout. Each subplot compares the clean signal and the anomalous signal, while highlighting the perturbed points.

C Industrial Dataset Details and Evaluation

This section describes the evaluation setting for the Paul Wurth industrial dataset. To assess the proposed method comprehensively, we consider both controlled and real anomaly scenarios. We first introduce synthetic anomaly injections used to stress-test the method under varying perturbation conditions with known affected sensors. In addition, expert evaluation confirms the diagnostic utility of the proposed approach. Process engineers verified that the attribution heatmaps accurately captured the progression of thermal instabilities and identified the correct sensor groups in nearly all evaluated anomalies. This qualitative assessment highlights the practical relevance of the framework and supports its suitability for real-world deployment in complex metallurgical monitoring environments.

C.1 Synthetic Anomaly Injection

To rigorously evaluate the proposed explainable anomaly detection method, we inject synthetic anomalies into selected sensors of the Paul Wurth industrial time-series data. These perturbations are introduced at different time steps and with varying intensities, enabling a controlled stress test of the method under diverse anomaly conditions. Since the injected anomalies have known locations and affected sensors, they provide ground-truth root causes for assessing whether the proposed approach can accurately localize the underlying sources of anomalous behavior. The considered perturbations are designed to reflect a range of realistic failure patterns commonly encountered in industrial monitoring systems. Alongside the quantitative evaluation, we further provide representative feature plots as qualitative examples, with one plot for each synthetic anomaly type, to

visualize the corresponding perturbation patterns. Specifically, we consider the following anomaly types, illustrated in Fig. 3:

- **Spike** anomaly introduces large abnormal values at randomly selected time steps within a window, simulating sudden transient disturbances, impulsive faults, or brief sensor surges.
- **Shift** anomaly adds a constant offset to the entire window, thereby altering the baseline level of the signal throughout the affected segment. This reflects persistent calibration errors or operating-point shifts.
- **Noise** anomaly perturbs the full window with random fluctuations, modeling measurement corruption or stochastic disturbances affecting the signal over the entire interval.
- **Drift** anomaly evolves gradually over time, producing a slow increase or decrease in the signal baseline. This represents sensor aging, calibration drift, or slowly changing process conditions.
- **Signal Dropout** anomaly abruptly sets the entire window to zero, representing sudden sensor failure, communication loss, or complete signal interruption.
- **Saturation** anomaly clips the signal at an upper or lower bound, modeling actuator limits, sensor saturation, or hardware clipping.

Representative feature plots illustrating the different synthetic anomaly types are shown in Figures 4–9, complementing the quantitative evaluation with qualitative analysis.

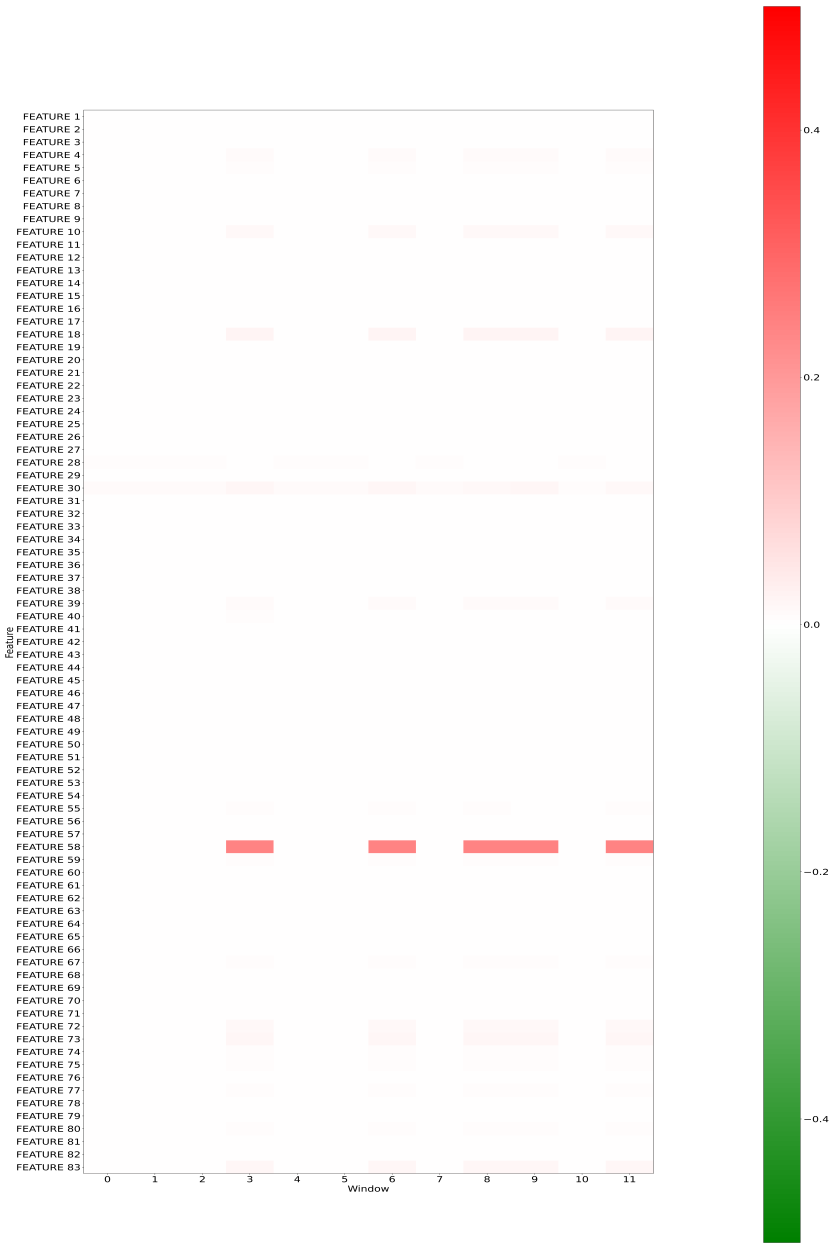


Fig. 4: **Spike anomaly on Feature 58.** The figure shows a representative synthetic spike anomaly, where Feature 58 exhibits large abnormal values at randomly selected time steps, simulating a sudden transient disturbance or brief sensor surge.

Fig. 5: **Shift anomaly on Feature 26.** Feature 26 is affected by a synthetic shift anomaly, where a constant offset is added across the window, altering the baseline level of the signal.

Fig. 6: **Noise anomaly on Feature 22.** Feature 22 is perturbed by a mild synthetic noise anomaly, and the corresponding attribution pattern shows that the anomalous signal is still correctly detected despite the relatively small disturbance.

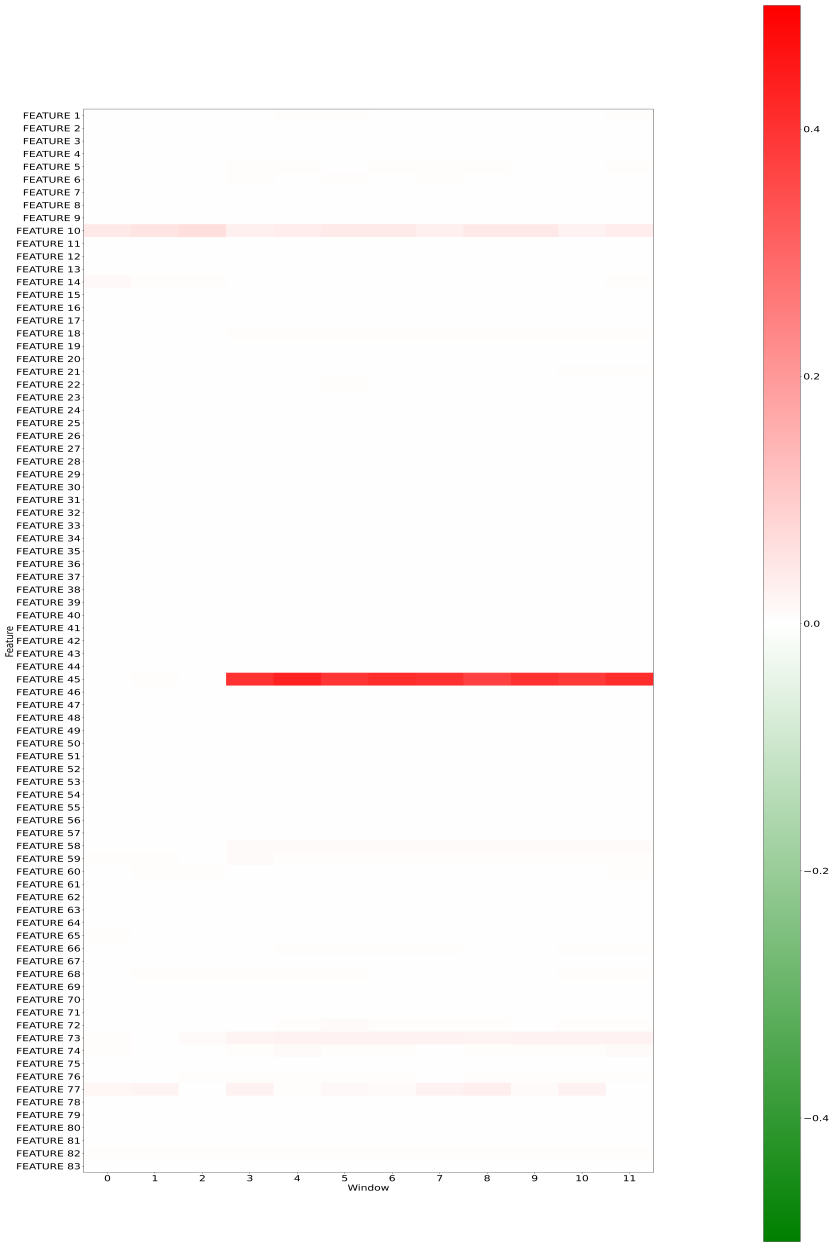


Fig. 7: **Drift anomaly on Feature 45.** Feature 45 exhibits a gradual temporal change corresponding to a synthetic drift anomaly, which is correctly identified by the attribution method.

Fig. 8: **Saturation anomaly on Feature 3.** Feature 3 exhibits a synthetic saturation anomaly, in which the signal is clipped at a limit, and the anomaly is correctly identified.

Fig. 9: **Signal dropout anomaly on Feature 11.** Feature 11 is affected by a synthetic signal dropout anomaly, where the signal abruptly drops to zero over the affected interval, and the attribution pattern correctly highlights the anomalous feature.

D Per-Attack Results on SWaT

This section reports the complete per-attack results on the SWaT dataset for the considered ablation studies. Tables 8–10 present the results for CondAttr-VAE, while Tables 11–13 present the corresponding results for CondAttr-UMAP.

Table 8: **Per-attack performance of CondAttr-VAE on the SWaT dataset.** For each attack instance, we report Top@3R, CW-RCS@3, and TemporalHM@3, alongside the ground-truth sensors correctly identified within the Top-3.

Attack #	Top@3R	CW@3	TempHM@3	FeatureID
1	1.000	0.145	0.104	MV101
2	1.000	0.121	0.545	P102
3	0.000	0.000	0.307	—
4	0.000	0.000	0.000	—
6	1.000	0.439	1.000	AIT202
7	1.000	0.888	1.000	LIT301
8	1.000	0.451	0.882	DPIT301
10	1.000	0.890	1.000	FIT401
11	1.000	0.104	0.705	FIT401
13	0.000	0.000	0.000	—
14	0.000	0.000	0.193	—
16	1.000	0.256	0.774	LIT301
17	0.000	0.000	0.000	—
19	0.000	0.000	0.000	—
20	1.000	0.925	1.000	AIT504
21	0.500	0.136	0.617	LIT101
22	0.333	0.031	0.363	P501
23	0.000	0.000	0.049	—
24	0.000	0.000	0.153	—
25	0.500	0.386	0.628	LIT401
26	0.000	0.000	0.101	—
27	0.500	0.000	0.435	LIT401
28	1.000	0.093	0.711	P302
29	0.000	0.000	0.000	—
30	0.000	0.000	0.337	—
31	1.000	0.407	0.444	LIT401
32	1.000	0.790	1.000	LIT301
33	1.000	0.127	0.722	LIT101
34	1.000	0.286	0.666	P101
35	0.000	0.000	0.105	—
36	0.000	0.000	0.716	—
37	0.000	0.000	0.105	—
38	1.000	0.688	1.000	AIT502, AIT402
39	0.500	0.487	0.666	FIT401
40	1.000	0.223	0.888	FIT401
41	1.000	0.946	0.876	LIT301

Table 9: **Per-attack performance of CondAttr-VAE on the SWaT dataset.** For each attack instance, we report Top@5R, CW-RCS@5, and TemporalHM@5, alongside the ground-truth sensors correctly identified within the Top-5.

Attack #	**Top@5R**	**CW@5**	**TempHM@5**	**FeatureID**
1	1.000	0.145	0.104	MV101
2	1.000	0.121	0.666	P102
3	0.000	0.000	0.566	—
4	0.000	0.000	0.000	—
6	1.000	0.439	1.000	AIT202
7	1.000	0.888	1.000	LIT301
8	1.000	0.451	1.000	DPIT301
10	1.000	0.890	1.000	FIT401
11	1.000	0.104	0.777	FIT401
13	0.000	0.000	0.000	—
14	0.000	0.000	0.193	—
16	1.000	0.256	1.000	LIT301
17	0.000	0.000	0.000	—
19	0.000	0.000	0.000	—
20	1.000	0.925	1.000	AIT504
21	0.500	0.136	0.666	LIT101
22	0.667	0.123	0.500	P501, UV401
23	0.333	0.002	0.049	—
24	0.500	0.037	0.153	P203
25	0.500	0.386	0.628	LIT401
26	0.000	0.000	0.301	—
27	0.500	0.000	0.462	LIT401
28	1.000	0.093	0.915	P302
29	0.000	0.000	0.000	—
30	0.667	0.152	0.660	P101, MV201
31	1.000	0.407	0.444	LIT401
32	1.000	0.790	1.000	LIT301
33	1.000	0.127	0.866	LIT101
34	1.000	0.286	0.666	P101
35	0.500	0.013	0.105	P101
36	0.000	0.000	0.783	—
37	0.000	0.000	0.200	—
38	1.000	0.688	1.000	AIT502, AIT402
39	0.500	0.487	0.666	FIT401
40	1.000	0.223	1.000	FIT401
41	1.000	0.946	0.876	LIT301

Table 10: **Per-attack performance of CondAttr-VAE on the SWaT dataset.** For each attack instance, we report Top@10R, CW-RCS@10, and TemporalHM@10, alongside the ground-truth sensors correctly identified within the Top-10.

Attack #	**Top@10R**	**CW@10**	**TempHM@10**	**FeatureID**
1	1.000	0.145	0.104	MV101
2	1.000	0.121	0.769	P102
3	1.000	0.027	0.566	—
4	0.000	0.000	0.000	—
6	1.000	0.439	1.000	AIT202
7	1.000	0.888	1.000	LIT301
8	1.000	0.451	1.000	DPIT301
10	1.000	0.890	1.000	FIT401
11	1.000	0.104	1.000	FIT401
13	0.000	0.000	0.000	—
14	0.000	0.000	0.193	—
16	1.000	0.256	1.000	LIT301
19	1.000	0.010	1.000	—
20	1.000	0.925	1.000	AIT504
21	0.500	0.136	0.666	LIT101
22	0.667	0.123	0.540	P501, UV401
23	0.333	0.002	0.049	—
24	0.500	0.037	0.400	P203
25	0.500	0.386	0.628	LIT401
26	0.500	0.036	0.564	—
27	0.500	0.000	0.462	LIT401
28	1.000	0.093	0.976	P302
29	0.000	0.000	0.000	—
30	1.000	0.240	0.946	P101, MV201
31	1.000	0.407	0.444	LIT401
32	1.000	0.790	1.000	LIT301
33	1.000	0.127	0.866	LIT101
34	1.000	0.286	1.000	P101
35	0.500	0.013	0.285	P101
36	0.000	0.000	0.783	—
37	0.500	0.011	0.285	P501
38	1.000	0.688	1.000	AIT502, AIT402
39	0.500	0.487	0.769	FIT401
40	1.000	0.223	1.000	FIT401
41	1.000	0.946	0.876	LIT301

Table 11: **Per-attack performance of CondAttr-UMAP on the SWaT dataset.** For each attack instance, we report Top@3R, CW-RCS@3, and TemporalHM@3, alongside the ground-truth sensors correctly identified within the Top-3.

Attack #	**Top@3R**	**CW@3**	**TempHM@3**	**FeatureID**
1	0.000	0.000	0.000	—
2	1.000	0.300	0.933	P102
3	0.000	0.000	0.600	—
4	0.000	0.000	0.000	—
6	1.000	0.443	1.000	AIT202
7	1.000	0.911	1.000	LIT301
8	1.000	0.375	0.882	DPIT301
10	1.000	0.851	1.000	FIT401
11	1.000	0.118	0.533	FIT401
13	0.000	0.000	0.000	—
14	0.000	0.000	0.000	—
16	0.000	0.000	0.774	—
17	0.000	0.000	0.000	—
19	0.000	0.000	0.000	—
20	1.000	0.999	1.000	AIT504
21	0.500	0.151	0.666	LIT101
22	0.667	0.217	0.540	UV401, P501
23	0.333	0.169	0.500	MV302
24	0.500	0.068	0.153	P205
25	0.500	0.088	0.545	LIT401
26	0.500	0.019	0.403	LIT301
27	0.500	0.000	0.420	LIT401
28	0.000	0.000	0.000	—
29	0.000	0.000	0.000	—
30	0.333	0.002	0.180	P101
31	1.000	0.407	0.923	LIT401
32	1.000	0.931	0.912	LIT301
33	0.000	0.000	0.452	—
34	1.000	0.342	0.666	P101
35	0.000	0.000	0.105	—
36	0.000	0.000	0.716	—
37	0.000	0.000	0.000	—
38	1.000	0.803	1.000	AIT502, AIT402
39	0.500	0.439	0.720	FIT401
40	1.000	0.167	0.571	FIT401
41	1.000	0.956	0.907	LIT301

Table 12: **Per-attack performance of CondAttr-UMAP on the SWaT dataset.** For each attack instance, we report Top@5R, CW-RCS@5, and TemporalHM@5, alongside the ground-truth sensors correctly identified within the Top-5.

Attack #	**Top@5R**	**CW@5**	**TempHM@5**	**FeatureID**
1	0.000	0.000	0.000	—
2	1.000	0.300	1.000	P102
3	0.000	0.000	0.600	—
4	0.000	0.000	0.000	—
6	1.000	0.443	1.000	AIT202
7	1.000	0.911	1.000	LIT301
8	1.000	0.375	1.000	DPIT301
10	1.000	0.851	1.000	FIT401
11	1.000	0.118	0.900	FIT401
13	0.000	0.000	0.000	—
14	0.000	0.000	0.000	—
16	1.000	0.063	0.774	LIT301
17	0.000	0.000	0.000	MV303
19	0.000	0.000	0.000	AIT504
20	1.000	0.999	1.000	AIT504
21	0.500	0.151	0.666	LIT101
22	0.667	0.217	0.744	UV401, P501
23	0.333	0.169	0.500	MV302, DPIT301
24	1.000	0.245	0.285	P205, P203
25	0.500	0.088	0.545	LIT401
26	0.500	0.019	0.501	LIT301
27	0.500	0.000	0.462	LIT401
28	0.000	0.000	0.000	—
29	0.000	0.000	0.000	—
30	0.333	0.002	0.266	P101, LIT101
31	1.000	0.407	0.923	LIT401
32	1.000	0.931	1.000	LIT301
33	0.000	0.000	0.731	—
34	1.000	0.342	1.000	P101
35	0.000	0.000	0.105	—
36	0.000	0.000	0.716	—
37	0.000	0.000	0.104	—
38	1.000	0.803	1.000	AIT502, AIT402
39	0.500	0.439	0.720	FIT401
40	1.000	0.167	1.000	FIT401
41	1.000	0.956	0.907	LIT301

Table 13: **Per-attack performance of CondAttr-UMAP on the SWaT dataset.** For each attack instance, we report Top@10R, CW-RCS@10, and TemporalHM@10, alongside the ground-truth sensors correctly identified within the Top-10.

Attack #	Top@10R	CW@10	TempHM@10	FeatureID
1	0.000	0.000	0.000	—
2	1.000	0.300	1.000	P102
3	0.000	0.000	0.600	—
4	0.000	0.000	0.000	—
6	1.000	0.443	1.000	AIT202
7	1.000	0.911	1.000	LIT301
8	1.000	0.375	1.000	DPIT301
10	1.000	0.851	1.000	FIT401
11	1.000	0.118	1.000	FIT401
13	0.000	0.000	0.000	—
14	0.000	0.000	0.000	—
16	1.000	0.063	1.000	LIT301
17	1.000	0.000	0.000	—
19	1.000	0.008	0.774	—
20	1.000	0.999	1.000	AIT504
21	0.500	0.151	0.666	LIT101
22	0.667	0.217	0.875	UV401, P501
23	0.667	0.339	0.528	MV302
24	1.000	0.245	0.285	P205, P203
25	0.500	0.088	0.628	LIT401
26	0.500	0.019	0.634	LIT301
27	0.500	0.000	0.462	LIT401
28	0.000	0.000	0.002	—
29	0.000	0.000	0.000	—
30	0.667	0.006	0.458	P101
31	1.000	0.407	0.923	LIT401
32	1.000	0.931	1.000	LIT301
33	1.000	0.073	1.000	—
34	1.000	0.342	1.000	P101
35	0.500	0.011	0.285	P101
36	0.000	0.000	0.716	—
37	0.000	0.000	0.200	—
38	1.000	0.803	1.000	AIT502, AIT402
39	0.500	0.439	0.720	FIT401
40	1.000	0.167	1.000	FIT401
41	1.000	0.956	0.907	LIT301