LoCo: Local Competitive Learning for Interpretable Multivariate Time Series Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

In streaming data scenarios, multivariate time series anomaly detection is critical for system reliability. Beyond detection, operators require interpretable diagnostics to make informed decisions and repair faults efficiently. However, state-of-the-art detection models are often computationally heavy and opaque, and existing root cause analysis methods are confined to locating anomalous variables at the current timestamp, overlooking the temporal evolution of anomalies along causal chains. Inspired by physical phenomena and real cases, we introduce two design principles—Finite Temporal Horizon and Causal Horizon. These principles and the sparsity of fault-relevant causal connections motivate the design of LoCo, a lightweight interpretable framework based on Local Competitive Learning. LoCo performs online anomaly detection via forecasting residuals and, upon detection, traces root causes along the learned time-lagged causal matrix, from current anomalous variables back to earlier upstream sources. This time-lagged causal matrix not only captures causal influences among variables across different past time steps, but can also be compressed into a global causal graph to accommodate conventional causal discovery tasks. Theoretically, the recovered lagged predictive support coincides with the true structural causal support under identifiability assumptions. Extensive experiments across anomaly detection, root cause analysis, and causal discovery demonstrate that LoCo achieves competitive accuracy and computational efficiency, while providing end-to-end interpretable visualizations suitable for resource-constrained online deployment. The code is available at `https://anonymous.4open.science/r/LoCo-491F/`.

1 Introduction

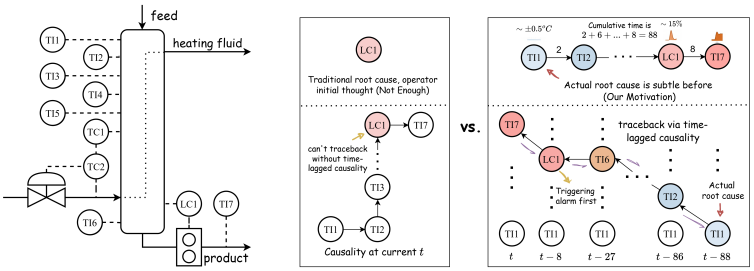
With the rapid growth of streaming data in cyber-physical systems, finance, and industrial monitoring, multivariate time series Anomaly Detection (AD) is vital for ensuring operational reliability, market integrity, and timely decision-making (Wang et al., 2025). Therefore, models must achieve both high accuracy and interpretable diagnostics that enable operators to make informed decisions and repair faults efficiently (Longo et al., 2024; Holly et al., 2022). This demand motivates not only detecting when an anomaly occurs, but also explaining why it occurs and where its root cause lies.

Although advanced detection models have achieved high accuracy, many rely on high-capacity architectures or latent representations whose diagnostic semantics are not explicit (Xu et al., 2022; Tuli et al., 2022; Yang et al., 2023). Some methods focus on Root Cause Analysis (RCA) by reporting root-cause variables at the anomalous timestamp, thereby offering certain diagnostic explanations (Han et al., 2025). Others discover causal relationships among variables to support RCA (Kim et al., 2025b; Liu et al., 2025c). However, both methods overlook that anomalies may propagate to downstream variables along causal relationships, rather than remaining confined to their initial locations. This phenomenon is referred to as the temporal evolution of anomalies. Existing RCA methods are often restricted to locating root causes at the current anomalous step and cannot address such temporal evolution, let alone performance-oriented advanced models.

This temporal evolution is analogous to a light cone in physics. Constrained by the speed of light, what we observe is only a shadow of the past universe, and likewise, severe anomalies that trigger

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

alerts are often shadows of inconspicuous upstream abnormal behaviors. Anomalies are not always instantaneous global mutations, but processes that propagate, accumulate, and diffuse along causal chains over time. A compelling industrial case is shown in Fig. 1 from Bauer & Thornhill (2008). As shown in Fig. 1(a), the process involves a distillation column with multiple temperature, composition, and level measurements (e.g. TI1–TI7, TC1). Fig. 1(b) contrasts traditional RCA and Causal Discovery (CD) with our motivation. In traditional RCA, operators identify LC1, the most deviated variable at the alarm time, as the root cause, and static causality cannot trace back in time. In reality, the disturbance starts at upstream TI1 about **88** sampling intervals earlier with only about $\pm 0.5°\mathrm{C}$ fluctuation, propagates through TI2–TI7 and LC1, and is eventually amplified to roughly 15% at TI7 at t. The alarm-triggering point is therefore not the root cause. Our goal is to trace back along a time-lagged causal graph, from TI7 at t back through LC1 at $t-8$, and finally to TI1 at $t-88$.



(a) Process schematic of the industrial case at Eastman Chemical Company

(b) Traditional root cause analysis and causal discovery vs. our motivation

Figure 1: Industrial case from the Eastman chemical distillation process (Bauer & Thornhill, 2008). (a) Process schematic with sensor and control-loop locations. (b) Comparison between traditional RCA/CD and our motivation. Left: traditional RCA identifies LC1 at the current time step as the root cause, while static causality ignores time lags. Right: the actual disturbance starts at upstream TI1 about 88 sampling intervals earlier, propagates through TI2–TI7 and LC1, and is amplified at TI7 at t. Our method traces back along a time-lagged causal graph from TI7 at t to TI1 at $t-88$.

Grounded in these physical phenomena and realities, we introduce two guiding design principles. (1) **Finite Temporal Horizon**. Causal influence operates within a finite time window, beyond which dependencies drop off and warrant hard truncation, justifying a sliding window input (Granger, 1969). (2) **Causal Horizon**. Conditioned on the direct causal parents of a target variable, indirect causes and non-causal variables become conditionally independent, offering no incremental predictive value (Pearl, 2009). These principles motivate an efficient predictor that meets the accuracy, real-time, and interpretability requirements of streaming data. Our key insight is that fault-relevant direct causal connections are typically sparse relative to all possible variable interactions (Liu et al., 2025b; Blumensath & Davies, 2009), which suggests that a sparse causal topology could emerge by constraining parameter updates. We thus introduce Local Competitive Learning, where backpropagation updates each output variable only on the largest gradient-magnitude connections and zeros out the rest, forcing the model to focus on the most predictive historical nodes.

Building upon these insights, we introduce LoCo (Local Competitive Network), an interpretable and lightweight framework for prediction-based AD. Trained solely on normal data, LoCo employs a single-layer predictor with a gradient mask that updates only the largest few connections, aiming to three tasks. (1) **Online AD.** Single-step forecasting via the learned predictor, triggering real-time alerts based on prediction residuals. Its learned weight matrix naturally serves as a time-lagged causal matrix, encoding how past variables at specific lags influence current targets. (2) **Causal-Driven RCA.** Upon detection, LoCo first identifies the most deviated variable at the current timestamp as the conventional root cause. It then traces backward along this matrix to earlier upstream root causes, yielding fine-grained and quantitative root cause rankings that extend beyond conventional one. (3) **Time-Lagged CD.** The same matrix can be compressed into a global $N \times N$ graph for conventional causal discovery, while also providing a finer-grained $N \times NL$ time-lagged causal graph. Moreover, LoCo offers Linear and Kernel variants, providing transparent white-box weights and capturing nonlinear dynamics via kernel mappings with group-level gradient competition, respectively. Its design ensures low memory overhead, fast inference, and high interpretability, making it suitable for the resource-constrained online deployment. Our contributions are summarized as

- We introduce a Local Competitive Learning strategy for time-series tasks and furthermore design LoCo, a lightweight predictive framework for online anomaly detection. LoCo uses a single-layer predictor with a gradient mask and offers Linear and Kernel variants, achieving minimal parameters, high efficiency, and preserved causal interpretability.
- We utilize the time-lagged causal graph from LoCo, making RCA and CD by-products of AD. Beyond conventional RCA at the current timestamp, LoCo traces backward along this graph to earlier upstream root causes, and the same graph can be compressed into a global $N \times N$ matrix for conventional CD, while also providing a finer-grained $N \times NL$ time-lagged causal graph.
- We theoretically prove that, under suitable assumptions, the introduced sparse gradient competition mechanism aligns with conditional independence and can approximate and recover the true causal topology. Experiments further corroborate these theoretical findings.
- We demonstrate LoCo's competitive accuracy and computational efficiency across three benchmark tasks (i.e., AD, RCA, and CD), supplemented by end-to-end visualizations from detection to traceability to demonstrate the interpretability.

2 Related Work

Multivariate Time Series Anomaly Detection. Anomaly detection has evolved from classical methods such as LOF (Breunig et al., 2000) to deep learning, which mainly follows two paradigms. Reconstruction-based methods learn normal representations and use reconstruction error as the anomaly score (Audibert et al., 2020; Schlegl et al., 2017; Li et al., 2025), but they require the full input window and delay detection. Forecasting-based methods predict future states and treat large residuals as anomalies (Malhotra et al., 2016), enabling immediate assessment and better online deployment. Contrastive learning, pretrained models, and LLMs have also been introduced (Kim et al., 2025a; Yang et al., 2023; Wang et al., 2026a; Goswami et al., 2024; Liu et al., 2025a; Jiang et al., 2025; Lee et al., 2026). Despite this progress, most methods still focus on temporal representation learning and mainly determine whether the current timestamp is anomalous (Holly et al., 2022). In multivariate time series, however, the root cause triggering alarms may involve only one or a few variables, yet their outputs are often not explicitly organized as variable-level causal evidence or time-lagged propagation paths.

CD and RCA in Anomaly Detection. While many models achieve high accuracy, they lack the causal interpretability needed for root cause localization. Recently, causal discovery has been introduced for AD and RCA. Granger causality is attractive for capturing temporal causality (Granger, 1969; Sucar, 2026), and methods such as CAROTS and GCAD integrate causality into contrastive learning or mine nonlinear Granger causality (Kim et al., 2025b; Liu et al., 2025c). For RCA, AERCA treats anomalies as exogenous interventions (Han et al., 2025), and CAAD merges dynamic and structural deviations (Wang et al., 2026b). Despite enhancing interpretability, these methods predominantly restrict their analysis to contemporaneous variables at the current time step. They often fail to fully exploit the discovered causal graphs for historical backtracking or neglect explicit variable–time-lag causal modeling. Consequently, they struggle to pinpoint finer-grained, upstream root causes along the time.

3 Methodology

3.1 Problem Definition

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{N \times T}$ be an observed multivariate time series, where $\mathbf{x}_t \in \mathbb{R}^N$ represents N sensor observations at discrete time step t. We denote the complete sequence of the i-th variable as $\{X_{i,t}\}_{t=1}^{T}$. This paper addresses three tasks:

(1) Anomaly Detection (AD). Given $\mathbf{X}$, determine whether each time step t is anomalous by outputting labels $\mathbf{Y} = \{y_t\}_{t=1}^{T}$, where $y_t \in \{0, 1\}$ (1 for anomaly, 0 for normal).

(2) Root Cause Analysis (RCA). Once an anomaly triggers at t, identify the originating root causes. Rather than restricting RCA to contemporaneous variables at time t, we extend it to fine-grained

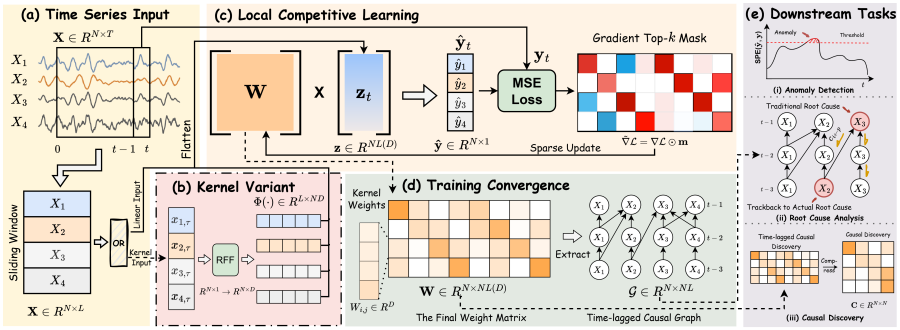


Figure 2: Overview of LoCo. (a) Sliding-window construction from the input multivariate time series. (b) Local competitive learning: a single-layer predictor trained with a Top-k gradient mask. (c) Training convergence and extraction of the time-lagged causal graph $\mathcal{G}$ from the sparse weight matrix. (d) Downstream tasks: (i) anomaly detection via prediction residuals, (ii) root cause analysis by backward tracing along $\mathcal{G}$, and (iii) causal discovery by compressing $\mathcal{G}$ into a $N \times N$ matrix $\mathbf{C}$.

$(variable, time\text{-}lag)$. Formally, a root cause is a tuple $(j, t-\tau)$ such that variable j at time $t-\tau$ triggered the anomaly at t.

(3) Causal Discovery (CD). Discover the global causal relationship matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$, where $C_{i,j}$ quantifies the overall directed causal influence of variable j on variable i. We aim to discover both a fine-grained time-lagged graph $\mathcal{G}$ and its compressed global version $\mathbf{C}$.

Although we extend RCA and CD to the time-lagged setting, existing benchmarks only provide labels for conventional RCA and variable-level causal graphs. We therefore evaluate both tasks under conventional definitions and illustrate the extended outputs via case studies. LoCo follows two principles from Section 1. Principle 1 justifies the sliding window, and Principle 2 motivates gradient competition to isolate direct causal parents, with formal statements in Appendix A.

3.2 LoCo: Local Competitive Network

As shown in Fig. 2(a), the input multivariate time series is first processed by a sliding window to form the historical context vector $\mathbf{z}_t$. Fig. 2(b) shows the optional Kernel variant, which applies variable-wise RFF. Fig. 2(c) illustrates local competitive learning, where only the top-k gradient coordinates are retained at each update. Fig. 2(d) shows the converged sparse weight matrix and the extracted time-lagged causal graph $\mathcal{G}$. Fig. 2(e) summarizes the three downstream tasks.

Let $L \in \mathbb{N}^+$ denote the maximum considered time lag. At time step t, the historical context vector $\mathbf{z}_t$ and the forecast target $\mathbf{y}_t$ are defined as

$$\mathbf{z}_t = [X_{1,t-1}, \ldots, X_{1,t-L}, X_{2,t-1}, \ldots, X_{N,t-L}]^\top \in \mathbb{R}^{NL}, \quad \mathbf{y}_t = \mathbf{x}_t \in \mathbb{R}^N. \tag{1}$$

Linear LoCo adopts a single-layer linear predictor $\hat{\mathbf{y}}_t = \mathbf{W}\mathbf{z}_t$, where $\mathbf{W} \in \mathbb{R}^{N\times NL}$ and $W_{i,(j,\tau)}$ represents the causal strength of lagged variable $X_{j,t-\tau}$ on target $X_{i,t}$. Minimize MSE loss

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N}(\hat{X}_{i,t} - X_{i,t})^2. \tag{2}$$

To enforce Principle 2, LoCo applies a Top-k gradient competition after each backward pass. For each output variable i, only the k connections with the largest gradient magnitudes are retained:

$$\tilde{\nabla}_{W_{i,\cdot}}\mathcal{L} = \nabla_{W_{i,\cdot}}\mathcal{L} \odot \mathbf{m}_i, \tag{3}$$

where $\mathbf{m}_i$ is a row-wise binary mask selecting the top-k elements in $|\nabla_{W_{i,\cdot}}\mathcal{L}|$. This competitive update forces parameters to concentrate on true causal parents, causing a sparse causal topology to naturally emerge during optimization. See Appendix B.1 for algorithm and Appendix 4.1 for proof.

Kernel LoCo, to capture nonlinear dynamics without sacrificing structural interpretability, applies variable-wise Random Fourier Feature (RFF) (Rahimi & Recht, 2007). Most kernel methods project

the entire multivariate vector jointly (i.e. $R^N \to R^D$), which entangles cross-variable features and makes it difficult to disentangle individual causal contributions. To avoid this, Kernel LoCo maps each lagged variable independently. For each lagged variable $X_{j,t-\tau}$, we map its value to a D-dimensional space:

$$\phi_{j,\tau}(x) = \sqrt{\frac{2}{D}} \cos(\boldsymbol{\omega}_{j,\tau} x + \mathbf{b}_{j,\tau}) \in \mathbb{R}^D, \tag{4}$$

where parameters $\boldsymbol{\omega}_{j,\tau}$ and $\mathbf{b}_{j,\tau}$ are randomly sampled and frozen. Concatenating all mapped features yields $\mathbf{\Phi}(\mathbf{z}_t) \in \mathbb{R}^{NLD}$. The weight matrix $\mathbf{W} \in \mathbb{R}^{N \times NLD}$ is partitioned into sub-blocks $\mathbf{W}_{(j,\tau)} \in \mathbb{R}^{N \times D}$. To maintain causal attribution, competition operates at the group level. For each output i, we compute the L_2 norm of the gradient block $v_{i,(j,\tau)} = \|\mathbf{g}_{i,(j,\tau)}\|_2$. Gradient updates are retained strictly for the top-k groups per output, zeroing out the rest. Details are in Appendix B.2.

Time-Lagged Causal Graph Extraction. Upon convergence, the fine-grained time-lagged graph is read directly from the learned parameters without post-hoc testing:

$$\mathcal{G} = \{(X_{j,t-\tau} \to X_{i,t}) \mid s_{i,(j,\tau)} > \epsilon\}, \tag{5}$$

where edge strength $s_{i,(j,\tau)}$ equals $|W_{i,(j,\tau)}|$ for Linear LoCo, and $\|\mathbf{w}^*_{i,(j,\tau)}\|_2$ for Kernel LoCo. The extraction procedure is detailed in Appendix B.1 and Appendix B.2.

3.3 Prediction-Based Anomaly Detection

LoCo trains exclusively on normal data. Online at time t, given input $\mathbf{z}_t$, it performs single-step forecasting $\hat{\mathbf{y}}_t = \mathbf{W}\mathbf{z}_t$ (or $\mathbf{W}\mathbf{\Phi}(\mathbf{z}_t)$ for Kernel LoCo). To accommodate varying variable scales, we define the standardized univariate error and the system-wide Squared Prediction Error (SPE) as

$$e_i(t) = \frac{X_{i,t} - \hat{X}_{i,t}}{\sigma_i}, \quad i = 1, \ldots, N, \qquad \mathrm{SPE}(t) = \sum_{i=1}^{N} e_i(t)^2, \tag{6}$$

where σ_i is the residual standard deviation on training data. An alarm triggers at step t if $\mathrm{SPE}(t) > \theta_{\mathrm{SPE}}$, where θ_{SPE} is the upper α-quantile (e.g., $\alpha = 0.99$) of training SPE scores. Triggering an alarm immediately initiates RCA.

3.4 Causal-Driven Root Cause Analysis

When $\mathrm{SPE}(t) > \theta_{\mathrm{SPE}}$, LoCo decomposes prediction errors backward along $\mathcal{G}$ to isolate originating perturbations. RCA operates on $\mathcal{G}$, while the compressed graph $\mathbf{C}$ is used for CD.

Anomalous Variable Identification. We define individual anomalous variables at time t as $A_t = \{i \mid |e_i(t)| > \theta_i\}$, where θ_i is derived from residual empirical quantiles. The variable exhibiting the highest relative deviation serves as the primary manifestation point $i^* = \arg\max_i(|e_i(t)|/\theta_i)$.

Error Decomposition. For Linear, prediction errors are analytically attributed to parent $p = (j, \tau)$

$$c_{i \leftarrow p} = W_{i,p} \cdot \frac{X_p - \hat{X}_p^{\mathrm{ref}}}{\sigma_i}, \tag{7}$$

where $\hat{X}_p^{\mathrm{ref}}$ is the normal reference value (the model's prediction $\hat{X}_{j,t-\tau}$). For Kernel LoCo, where analytical decomposition is intractable, we perform counterfactual perturbation by evaluating prediction $\hat{X}_{i,t}^{(p)}$ with parent p's input set to its reference value, yielding $c^{\mathrm{K}}_{i \leftarrow p} = (\hat{X}_{i,t} - \hat{X}_{i,t}^{(p)})/\sigma_i$.

Contribution-Guided Traceback. Starting from each $i \in A_t$, LoCo recursively searches upstream along paths with $|c_{i \leftarrow p}| > \zeta$. If a parent node $p = (j, \tau)$ is itself anomalous at $t - \tau$, recursion continues through p. When recursion reaches a node whose parents are normal, this node is identified as a root cause candidate $(j, t - \tau)$ and added to set R.

Quantitative Ranking and Reporting. For each candidate $p^* \in R$, its final attribution score is quantified by the drop in system SPE when replacing p^* with its normal reference:

$$\mathrm{Attr}(p^*) = \mathrm{SPE}(t) - \mathrm{SPE}^{(p^*)}(t). \tag{8}$$

LoCo outputs an actionable report containing: (1) the alarm timestamp t and $\mathrm{SPE}(t)$; (2) the primary variable i^*; (3) ranked root causes $(variable, timestamp)$ sorted by Attr; and (4) propagation paths, e.g., $X_k \xrightarrow{c_1} X_j \xrightarrow{c_2} X_{i^*}$. We provide more details above in Appendix B.3

3.5 Temporal Compression: Global Causal Discovery

Since mainstream CD benchmarks operate at the variable level, we compress the fine-grained graph $\mathcal{G}$ into a global matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$. Reshaping $\mathbf{W}$ into $\mathbf{W}' \in \mathbb{R}^{N\times N\times L}$, where $W'_{i,j,\tau}$ denotes the influence of variable j at lag τ on i, we average over the lag dimension and then apply antisymmetrization to the off-diagonal entries

$$C_{i,j} = \frac{1}{L}\sum_{\tau=1}^{L} W'_{i,j,\tau}, \qquad \mathbf{C} \leftarrow \max(\mathbf{C} - \mathbf{C}^\top, 0), \tag{9}$$

which retains the dominant causal direction while preserving diagonal self-autoregressive terms. A final quantile-based thresholding suppresses residual noise, yielding the global graph $\mathbf{C}$. Kernel weights are first mapped to signed magnitudes, and more details are discussed in Appendix B.4.

4 Theoretical Analysis

We formalize LoCo's local competitive learning as a gradient hard-thresholding (GHT) update, which allows analyzing it within sparse support selection. This section addresses three questions. (1) why gradients carry causal information, (2) why non-causal variables are suppressed, and (3) why GHT can select the sparse predictive support. Phase I characterizes the population-level gradient at the true parameter, while Phase II analyzes the algorithmic behavior of GHT under restricted eigenvalue and gradient-gap conditions. The global assumptions are provided in Appendix C.1.

4.1 Phase I: Gradient as Conditional Independence Test

Theorem 4.1 (Gradient–Residual Equivalence)**.** *Under MSE loss, the expected gradient of a single-layer predictor with respect to a historical input is proportional to the cross-covariance between that input and the model residual:*

$$\mathbb{E}[\nabla_{W_{i,j,\tau}}\mathcal{L}] \propto -\mathrm{Cov}(r_{i,t}, X_{j,t-\tau}), \quad r_{i,t} = X_{i,t} - \hat{X}_{i,t}. \tag{10}$$

This result shows that the expected gradient measures whether the input and residual still exhibit unexplained correlation. The proof is in Appendix C.

Lemma 4.2 (Vanishing Gradients for Non-Causal Variables)**.** *At the true parameter w_i^*, if the model has correctly fitted the true parents, the expected gradient of a non-causal variable is zero:*

$$\mathbb{E}[\nabla_{W_{i,j,\tau}}\mathcal{L}] = -\mathrm{Cov}(\epsilon_{i,t}, X_{j,t-\tau}) = 0, \quad j \notin \mathrm{Pa}_i. \tag{11}$$

This lemma characterizes the population-level screening property at the true parameter. In practice, finite-sample gradients fluctuate, and early stopping together with the contractive dynamics of Phase II mitigates this effect. The proof is in Appendix C.

4.2 Phase II: Gradient Hard-Thresholding for Support Selection

LoCo updates parameters via Gradient Hard-Thresholding (GHT)

$$W^{(m+1)} = W^{(m)} - \eta\,\mathcal{H}_k\big(\nabla_{W^{(m)}}\mathcal{L}\big), \tag{12}$$

where $\mathcal{H}_k(\cdot)$ retains the k largest-magnitude entries and zeros out the rest, η is the learning rate, and m is the iteration index. Let $S_i^* = \mathrm{supp}(w_i^*)$ with $|S_i^*| \le k$.

Theorem 4.3 (Population-Level Support Screening)**.** *Under Assumptions C.1–C.5, for any target variable i with $|S_i^*| \le k$, the population gradient $g_i(w_i^*)$ satisfies*

$$S_i^* \subseteq \mathrm{Top}_k\big(|g_i(w_i^*)|\big). \tag{13}$$

Consequently, at the population level, a single GHT update retains all true lagged parents.

Proof Sketch. By Theorem 4.1, $g_{i,q}(w_i^*) = -\mathrm{Cov}(\epsilon_{i,t}, Z_{t,q})$. By Lemma 4.2, $g_{i,q}(w_i^*) = 0$ for $q \notin S_i^*$. By Assumption C.5, $|g_{i,q}(w_i^*)| \ge \Delta_{\mathrm{pop}} > 0$ for $q \in S_i^*$. Hence, selecting the top-k coordinates in magnitude with $k \ge |S_i^*|$ necessarily includes all of S_i^*. □

Proposition, proofs and more theoretical analysis, including the finite-sample support selection, the comparison with IHT and Granger causality, are provided in Appendix C and Appendix D.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate on three tasks. For **AD**, we use five widely adopted benchmarks (MSL, SMAP, PSM, SMD, SWaT). However, recent studies have pointed out that these widely used benchmarks are flawed (Wu & Keogh, 2023). We therefore additionally adopt four more challenging datasets (SWAN, Creditcard, GECCO, and CICIDS). For **RCA**, we use an annotated SWaT segment and MSDS. For **CD**, we use RiversEastGermany from CausalRivers to construct five subgraph sets with different structures (Stein et al., 2025). More details are provided in Appendix E.

Baselines. We compare LoCo against a comprehensive set of baselines across the three tasks. For AD, this includes classical detectors, deep detectors (including causality-aware detectors), and generative detectors. For RCA, we compare against ϵ-Diagnosis (Shan et al., 2019) and AERCA (Han et al., 2025). For CD, we compare against PCMCI (Runge et al., 2017), VAR (Assaad et al., 2022) and more. The complete list is provided in Appendix F.

Metrics. (1) Anomaly Detection. Since score-based anomaly detection methods are highly sensitive to the threshold in point-wise metrics such as Precision, we primarily use threshold-independent metrics, AUROC and AUPRC. We note that many existing methods select the optimal threshold on the test set. In contrast, LoCo determines its threshold strictly from training prediction errors, which may be less favorable on some datasets but better reflects realilty. And we further report Affiliation metrics (Huet et al., 2022). **(2) Root Cause Analysis.** We report $AC@K$ and $AC^*@K$, together with their averages $Avg@K$ and $Avg^*@K$ (Han et al., 2025). **(3) Causal Discovery.** Following conventional evaluation, we compress the time-lagged graph into a global matrix to get AUROC and Individual AUROC (InROC) for each sampled subgraph. More details are in Appendix G.

5.2 MAIN RESULTS

Anomaly Detection with AD-Specific Baselines. We compare LoCo against anomaly detection-specific baselines on five widely used benchmarks and four additional, more challenging datasets. Since recent studies have pointed out limitations of the widely used benchmarks (Wu & Keogh, 2023), the additional datasets provide a more comprehensive evaluation. Table 1 reports AUROC and AUPRC on all nine datasets. Overall, LoCo achieves the best or second-best performance on most datasets and metrics, consistently outperforming the baselines. Some baselines such as GCAD are also competitive, but they require substantially higher model capacity. This indicates that local competitive learning provides strong anomaly detection performance with a lightweight design.

Table 1: Performance comparison on nine benchmark datasets. The **best** result in each column is highlighted in bold, and the second result is underlined.

Method	MSL		PSM		SMAP		SWaT		SMD		GECCO		SWAN		Creditcard		CICIDS	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Iforest	52.35	11.41	54.24	33.41	48.75	12.17	34.56	9.33	66.44	12.18	61.94	5.16	48.75	38.00	78.70	0.26	71.60	7.36
LOF	55.69	12.02	73.03	46.20	62.59	17.41	64.54	6.70	70.59	**60.21**	32.13	3.85	71.11	62.13	50.98	0.19	42.23	1.05
AnomalyTran	50.04	10.55	50.13	28.03	49.93	12.77	54.35	14.76	50.80	4.44	52.59	1.43	43.26	31.59	53.27	1.90	49.20	0.09
TranAD	46.95	12.60	63.41	49.99	37.35	9.75	81.78	72.77	63.70	10.75	73.15	26.81	59.90	52.75	95.27	10.65	69.36	0.14
DCdetector	49.97	10.53	49.91	27.73	49.98	12.78	49.96	12.14	49.97	4.16	48.08	1.04	49.96	32.61	45.13	0.16	54.39	0.11
CATCH	62.57	15.36	59.99	38.50	43.13	11.47	22.97	8.23	76.17	16.38	91.22	19.80	48.52	42.46	80.34	10.20	66.85	0.13
SEA-FGT	42.80	10.27	57.43	27.98	55.79	29.15	44.64	12.23	53.00	18.34	68.75	18.95	54.70	39.71	12.01	9.76	63.42	1.15
CoAD	57.10	15.80	70.20	48.91	50.76	19.50	85.31	63.23	71.30	19.20	82.87	22.14	62.75	47.76	78.20	11.38	50.27	0.09
LEFT	57.48	15.25	73.21	57.65	47.89	17.59	81.53	72.25	77.41	20.54	83.93	44.78	68.63	39.48	84.41	15.41	64.89	0.14
CAROTS	55.97	15.41	74.58	54.40	62.66	48.82	82.18	72.82	63.79	13.12	69.52	26.16	66.36	60.35	89.10	**48.82**	56.74	0.75
GCAD	62.70	**18.98**	68.06	52.32	43.97	**73.68**	45.34	11.25	72.77	28.42	93.44	50.38	56.98	62.68	62.93	13.33	52.65	0.45
AERCA	54.35	13.87	63.34	41.27	52.74	24.06	80.65	70.05	62.28	11.07	57.43	16.19	60.81	45.44	60.83	3.46	69.72	0.15
Linear-LoCo	71.93	17.17	79.28	54.28	62.90	49.55	85.42	77.45	**77.88**	28.71	**96.84**	**50.40**	73.40	63.14	96.17	10.63	**73.62**	1.16
Kernel-LoCo	**82.53**	17.48	**82.22**	**62.93**	**66.63**	45.15	**87.63**	**80.30**	77.72	14.22	78.68	28.34	**81.95**	**69.97**	**96.18**	10.41	72.14	**1.17**

Comparison with Generative Baselines. Recently, many prediction and reconstruction models have been used for AD, where anomalies are detected through forecasting or reconstruction. We refer to them as Generative Baselines, including models such as VAE (Kingma & Welling, 2013). Table 2 reports AUROC and AUPRC on nine benchmark datasets. Overall, LoCo achieves the best or second-best performance on most datasets and metrics, consistently outperforming or matching the generative baselines. These results confirm that LoCo remains highly competitive against generative baselines across diverse datasets.

Table 2: Performance comparison on nine benchmark datasets. The **best** result in each column is highlighted in bold, and the second result is underlined. * denotes the baseline detect anomaly via predicting, while the rest is via reconstructing.

Method	MSL		PSM		SMAP		SWaT		SMD		GECCO		SWAN		Creditcard		CICIDS	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
VAE	53.01	13.89	64.23	41.68	41.18	10.53	83.67	74.50	64.12	11.41	79.41	28.18	71.76	64.66	95.86	11.98	69.65	0.14
iTransform	58.95	14.06	58.26	37.23	40.90	11.43	23.45	8.28	74.61	14.58	78.00	8.10	51.16	41.86	90.90	3.58	66.40	0.14
Dlinear	60.14	13.87	55.89	35.16	39.06	10.35	30.42	8.52	70.81	13.34	79.09	11.61	49.43	44.60	91.91	7.00	**74.81**	0.18
Nlinear	59.56	13.95	57.04	35.91	43.74	10.99	23.69	8.07	72.33	13.82	86.85	15.53	0.00	0.00	0.00	0.00	72.52	0.16
PatchTST	62.21	14.87	58.15	38.59	44.11	11.13	24.17	8.65	73.64	14.65	90.80	21.05	54.39	46.70	95.68	10.18	69.83	0.14
GPT4TS*	59.61	13.76	58.33	38.00	47.90	11.80	22.98	8.25	70.76	13.19	84.83	17.23	50.27	45.38	95.67	8.00	71.51	0.15
LLMmixer	60.98	13.69	55.94	36.13	48.08	11.82	24.02	8.59	69.27	11.98	93.02	41.82	50.41	45.05	95.73	9.05	69.14	0.14
CALF*	59.60	14.44	58.12	35.97	39.73	10.63	28.47	8.80	74.71	14.46	86.39	20.55	51.68	41.19	65.24	0.81	56.60	0.11
Moment	61.41	13.71	61.41	13.71	57.63	36.49	43.56	10.88	71.99	13.60	90.28	29.37	48.14	43.36	95.70	8.34	68.75	0.14
Timer	63.65	15.45	55.59	35.89	47.66	11.65	28.62	11.92	69.12	12.75	95.47	46.95	62.45	51.09	95.82	9.56	67.84	0.13
UniTS	62.33	14.34	58.17	38.02	43.80	10.94	24.14	8.62	72.14	13.64	95.56	49.08	52.12	45.92	95.75	9.05	68.09	0.13
Chronos*	59.40	14.52	56.48	37.04	46.68	11.35	80.35	41.10	68.64	11.84	86.86	34.58	51.83	45.08	95.75	10.24	61.28	0.11
Linear-LoCo	71.93	17.17	79.28	54.28	62.90	**49.55**	85.42	77.45	**77.88**	**28.71**	**96.84**	**50.40**	73.40	63.14	96.17	**10.63**	73.62	1.16
Kernel-LoCo	**82.53**	**17.48**	**82.22**	**62.93**	**66.63**	45.15	**87.63**	**80.30**	77.72	14.22	78.68	28.34	**81.95**	**69.97**	**96.18**	10.41	72.14	**1.17**

Root Cause Analysis. Following conventional evaluation, we compare LoCo against ϵ-Diagnosis and AERCA on SWaT and MSDS. Table 3 reports $AC@K$. $AC^*@K$ are provided in Appendix H.7. On SWaT, Linear-LoCo achieves the best across all levels, outperforming both ϵ-Diagnosis and AERCA. On MSDS, Kernel-LoCo achieves the best $AC@1$, $AC@3$, and $Avg@10$, while Linear-LoCo remains competitive with AERCA.

Causal Discovery on CausalRivers. Table 4 reports the causal discovery results on CausalRivers. Linear-LoCo achieves the best on many subsets, and remains competitive on others, outperforming or matching most baselines. Kernel-LoCo performs stably and get top3 on this benchmark. This is consistent with the results in Stein et al. (2025) that simple linear baselines often outperform complex methods on real-world, because real-world series violate strong assumptions. On CausalRivers, the underlying hydrological dependencies are largely linear and low-rank, so the nonlinear capacity of Kernel brings limited additional benefit, and the linear already captures the main structure. This is same for PCMCI and Dynotears, which are more complex yet do not outperform VAR.

Table 3: Performance comparison on SWaT and MSDS. The **best** result in each column (per dataset) is highlighted in bold, and the second result is underlined.

Dataset	Method	AC@1	AC@3	AC@5	AC@10	Avg@10
SWaT	ϵ-Diagnosis	0.2000	0.2750	0.2750	0.5524	0.3547
	AERCA	0.3305	0.3980	0.5255	0.7860	0.5878
	Linear-LoCo	**0.3500**	**0.5250**	**0.7250**	**0.9000**	**0.6800**
	Kernel-LoCo	0.3000	0.4000	0.4751	0.7943	0.5988
MSDS	ϵ-Diagnosis	0.1500	0.0149	0.0677	**1.0000**	0.4399
	AERCA	0.3250	0.3250	**0.9500**	**1.0000**	0.7700
	Linear-LoCo	0.3366	0.7591	0.7789	**1.0000**	0.7703
	Kernel-LoCo	**0.4833**	**0.7880**	0.9168	**1.0000**	**0.8483**

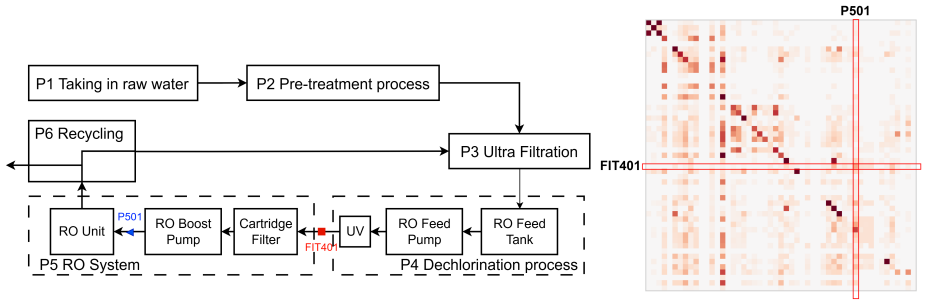
Table 4: Performance evaluation across different experimental settings in terms of AUROC and Individual AUROC (InROC). The **best** result is highlighted in bold, and the second result is underlined.

Method	Close		Root cause		Random+1		Confounder		Random	
	AUROC	InROC	AUROC	InROC	AUROC	InROC	AUROC	InROC	AUROC	InROC
VAR	0.7585	0.7641	**0.7099**	**0.7187**	0.7300	0.7195	**0.7099**	**0.7187**	0.7300	**0.7660**
Varlingam	0.6375	0.6507	0.6490	0.6653	0.7134	0.7167	0.5283	0.5315	0.6450	0.6572
PCMCI	0.5124	0.5235	0.5666	0.5803	0.6813	0.6902	0.5744	0.5425	0.5116	0.5180
Dynotears	0.4448	0.4648	0.4948	0.4939	0.4448	0.4648	0.4713	0.5032	0.4976	0.4968
Linear-LoCo	**0.8396**	**0.8327**	0.6779	0.7166	**0.7485**	**0.7508**	0.6395	0.6221	**0.7652**	0.7532
Kernel-LoCo	0.6638	0.6603	0.6496	0.6579	0.7036	0.7119	0.6468	0.6345	0.6558	0.6503

5.3 CASE STUDY

We present a case study on SWaT to illustrate the full detection-to-traceability pipeline. Fig. 3(a) shows a simplified SWaT process. For anomaly #2303, the system SPE reaches 6856.90, and the primary symptom variable is FIT401 with a standardized error of -29.713. The traceback identifies

P501 at lags 11–15 as the upstream root cause, with the highest attribution score 57.3828 at lag 11. Fig. 3(b) shows the compressed causal matrix, where the intersection of the FIT401 row and the P501 column is highlighted, indicating that a causal dependency between P501 and FIT401 exists in the learned graph. Fig. 4(a) shows the system SPE over $t = 14600$–15800, with a sharp increase around 15500. Fig. 4(b) marks the anomaly at 15518 for FIT401, while Figure 4(c) highlights the earlier interval 15503–15507 for P501, where a subtle deviation occurs before propagating downstream (even in 11503–11505 seems normal). This shows that the conspicuous FIT401 anomaly is caused by the earlier, less visible P501 behavior, with a delay of about 15 steps. A diagnostic report and a detailed analysis are provided in Appendix H.12 and H.13, respectively.



(a) Simplified process of the real case in SWaT (red is current anomaly, blue is the upstream root cause)

(b) The compressed causality discovered by Linear-LoCo

Figure 3: Case study on SWaT. (a) simplified process with the observed anomaly variable FIT401 and the traced upstream root cause P501; (b) compressed causal matrix, where the FIT401 row and P501 column intersect at a highlighted entry, indicating a learned causal dependency.

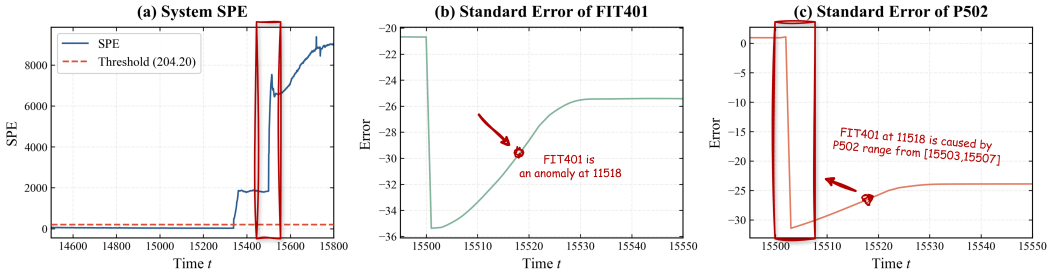


Figure 4: Case study on SWaT. (a) system SPE over $t = 14600$–15800, with a sharp increase around 15510; (b) standardized prediction error of FIT401, with the anomaly at 15518; (c) standardized prediction error of P501, highlighting the earlier subtle interval 15503–15507 that precedes the downstream FIT401 anomaly.

More Analysis. There are more experiments, including more metrics (Appendix H.1) and efficiency about LoCo (Appendix H.3), provided in Appendix H.

6 Conclusion

We proposed LoCo, a lightweight and interpretable framework for multivariate time series anomaly detection. Through Top-k gradient competition, LoCo learns a sparse time-lagged causal matrix that supports online detection, causal-driven root cause analysis, and time-lagged causal discovery with end-to-end interpretability, achieving competitive accuracy and efficiency across all three tasks. And theoretical analysis connects LoCo to gradient hard-thresholding and establishes population-level support screening under suitable assumptions. The main limitation is that existing benchmarks lack ground-truth labels for time-lagged root causes and causal edges. Future work will construct time-lagged causal graphs and corresponding benchmarks to enable fine-grained causal discovery and root cause analysis.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, assist with translation, propose or refine hypotheses, and clean and reformat datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, or interpret results, and generating synthetic data sets is not applicable to this work as we exclusively utilized real-world datasets. Additionally, we used generative AI tools to summarize or analyse existing literature and for sourcing/searching for information. We have reviewed all AI-assisted work. Specifically, all theoretical propositions and mathematical derivations were originally formulated by the authors, with AI utilized solely to verify and polish the proofs. Furthermore, all AI-assisted translations and literature summaries were manually cross-checked for accuracy and contextual relevance. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ACKNOWLEDGEMENTS

REFERENCES

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

Rafal Angryk, Petrus Martens, Berkay Aydin, Dustin Kempton, Sushant Mahajan, Sunitha Basodi, Azim Ahmadzadeh, Xumin Cai, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi, Michael Schuh, and Manolis Georgoulis. Multivariate time series dataset for space weather data analytics. *Scientific Data*, 7:227, 07 2020. doi: 10.1038/s41597-020-0548-x.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *J. Artif. Int. Res.*, 73, May 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13428. URL https://doi.org/10.1613/jair.1.13428.

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

Margret Bauer and Nina F. Thornhill. A practical method for identifying the propagation path of plant-wide disturbances. *Journal of Process Control*, 18(7):707–719, 2008. ISSN 0959-1524. doi: https://doi.org/10.1016/j.jprocont.2007.11.007. URL https://www.sciencedirect.com/science/article/pii/S0959152407001709.

Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2009.04.002. URL https://www.sciencedirect.com/science/article/pii/S1063520309000384.

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*, SIGMOD '00, pp. 93–104, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132174. doi: 10.1145/342009.335388. URL https://doi.org/10.1145/342009.335388.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems 37*, 2024. URL https://api.semanticscholar.org/CorpusID:268201715.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k38Th3x4d9.

Stephanie Holly, Robin Heel, Denis Katic, Leopold Schoeffl, Andreas Stiftinger, Peter Holzner, Thomas Kaufmann, Bernhard Haslhofer, Daniel Schall, Clemens Heitzinger, et al. Autoencoder based anomaly detection and explained fault localization in industrial cooling systems. *arXiv preprint arXiv:2210.08011*, 2022.

Wei Huang, Zhihong Wang, Yanyong Huang, Jia Liu, and Xiaocao Ouyang. Sea-fgt: Frequency-guided transformer with semantic expert augment for time series anomaly detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 1910–1921, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817823. URL https://doi.org/10.1145/3770855.3817823.

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

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

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *J. Mach. Learn. Res.*, 11:1709–1731, August 2010. ISSN 1532-4435.

Xingguo Jiang, Yue Wang, Chunpeng Wu, Xiaohui Wang, Zhenying Tai, Tianyu Chen, Mingliu Liu, and Jia Wu. Unsupervised anomaly detection for heterogeneous multivariate time series based on llm. In *2025 Thirteenth International Conference on Advanced Cloud and Big Data (CBD)*, pp. 125–130, 2025. doi: 10.1109/CBD69312.2025.00031.

Lutz Kilian. New introduction to multiple time series analysis, by helmut lütkepohl, springer, 2005. *Econometric theory*, 22(5):961–967, 2006.

HyunGi Kim, Jisoo Mok, Dongjun Lee, Jaihyun Lew, Sungjae Kim, and Sungroh Yoon. Causality-aware contrastive learning for robust multivariate time-series anomaly detection. *arXiv preprint arXiv:2506.03964*, 2025a.

Hyungi Kim, Jisoo Mok, Dongjun Lee, Jaihyun Lew, Sungjae Kim, and Sungroh Yoon. Causality-aware contrastive learning for robust multivariate time-series anomaly detection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30591–30608. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/kim25aa.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Md. Kowsher, Md. Shohanur, Islam Sobuj, Nusrat Jahan Prottasha, Emmanuel Alanis, Ozlem O. Garibay, Niloofar Yousefi, Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Xuan Zhang, Xun Liang, Aakas Zhiyuli, Shusen Zhang, Rui Xu, Bo Wu. 2019, Huaqin Zhao, Zheng Liu, Zihao Wu, Yiwei Li, Tian-Ze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wan Zhang. Llm-mixer: Multiscale mixing in llms for time series forecasting. *ArXiv*, abs/2410.11674, 2024. URL https://api.semanticscholar.org/CorpusID:273350827.

Geunho Lee, Jieun Lee, Tae-yong Kim, and Jongpil Jeong. Llm-ap: Llm-based anomaly detection with synthetic time-series data augmentation. *Applied Sciences*, 16(14), 2026. ISSN 2076-3417. doi: 10.3390/app16146851. URL https://www.mdpi.com/2076-3417/16/14/6851.

Long Li, Wencheng Zhang, Shi Yuan, Hongle Guo, and Wanghu Chen. Physdiff: A physically-guided diffusion model for multivariate time series anomaly detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ElTbpJp7b9.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 413–422, USA, 2008. IEEE Computer Society. ISBN 9780769535029. doi: 10.1109/ICDM.2008.17. URL https://doi.org/10.1109/ICDM.2008.17.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18915–18923, 2025a.

Xiangdong Liu, Jie Liu, Xiao-Hua Yang, Zhiqiang Wu, Ying Wei, Zhuoran Xu, and Juan Wen. Fault root cause analysis based on liang–kleeman information flow and graphical lasso. *Entropy*, 27, 2025b. URL https://api.semanticscholar.org/CorpusID:276499764.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lin-Tao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *ArXiv*, abs/2310.06625, 2023. URL https://api.semanticscholar.org/CorpusID:263830644.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:267412273.

Zehao Liu, Mengzhou Gao, and Pengfei Jiao. GCAD: Anomaly Detection in Multivariate Time Series from the Perspective of Granger Causality. In *AAAI Conference on Artificial Intelligence*, pp. 19041–19049, 2025c. doi: 10.1609/AAAI.V39I18.34096. URL https://mlanthology.org/aaai/2025/liu2025aaai-gcad/.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102301. URL https://www.sciencedirect.com/science/article/pii/S1566253524000794.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

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

Steffen Moritz, Frederik Rehbach, Sowmya Chandrasekaran, Margarita Rebolledo, and Thomas Bartz-Beielstein. Gecco industrial challenge 2018 dataset: A water quality dataset for the 'internet of things: Online anomaly detection for drinking water quality' competition at the genetic and evolutionary computation conference 2018, kyoto, japan., February 2018. URL https://doi.org/10.5281/zenodo.3884398.

Sasho Nedelkoski, Jasmin Bogatinovski, Ajay Kumar Mandapati, Soeren Becker, Jorge Cardoso, and Odej Kao. Multi-source distributed system data for ai-powered analytics. In Antonio Brogi, Wolf Zimmermann, and Kyriakos Kritikos (eds.), *Service-Oriented and Cloud Computing*, pp. 161–176, Cham, 2020. Springer International Publishing. ISBN 978-3-030-44769-4.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pp. 159–166, 2015. doi: 10.1109/SSCI.2015.33.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Jakob Runge, Peer Johannes Nowack, Marlene Kretschmer, Seth Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5, 2017. URL https://api.semanticscholar.org/CorpusID:88515817.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen (eds.), *Information Processing in Medical Imaging*, pp. 146–157, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59050-9.

Huasong Shan, Yuan Chen, Haifeng Liu, Yunpeng Zhang, Xiao Xiao, Xiaofeng He, Min Li, and Wei Ding. ?-diagnosis: Unsupervised and real-time diagnosis of small- window long-tail latency in large-scale microservice platforms. In *The World Wide Web Conference*, WWW '19, pp. 3215–3222, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313653. URL https://doi.org/10.1145/3308558.3313653.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. pp. 108–116, 01 2018. doi: 10.5220/0006639801080116.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers-scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations*, volume 2025, pp. 64516–64531, 2025.

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

Luis Enrique Sucar. *Causal Discovery from Temporal Data*, pp. 119–139. Springer Nature Switzerland, Cham, 2026. ISBN 978-3-031-98345-0. doi: 10.1007/978-3-031-98345-0_6. URL https://doi.org/10.1007/978-3-031-98345-0_6.

Qideng Tang, Chaofan Dai, Wubin Ma, Yahui Wu, Haohao Zhou, Tao Zhang, Huan Li, and Dalin Zhang. Bridging classification and reconstruction: Cooperative time series anomaly detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 4604–4615, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3818108. URL https://doi.org/10.1145/3770855.3818108.

Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *Proceedings of VLDB*, 15(6):1201–1214, 2022.

Dezheng Wang, Tong Chen, Guansong Pang, Congyan Chen, Shihua Li, and Hongzhi Yin. Left: Learnable fusion of tri-view tokens for unsupervised time series anomaly detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4800–4811, 2026a.

Fengling Wang, Yiyue Jiang, Rongjie Zhang, Aimin Wei, Jingming Xie, and Xiongwen Pang. A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. *Sensors*, 25(1), 2025. ISSN 1424-8220. doi: 10.3390/s25010190. URL https://www.mdpi.com/1424-8220/25/1/190.

Xin Wang, Yunshi Wen, Yanan He, Haotian Xu, Youlan Zhao, Michel Ferreira Cardia Haddad, and Tengfei Ma. Caad: Causality-aware multivariate time series anomaly detection via multi-scale alignment and structural causal consistency. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 5091–5101, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817978. URL https://doi.org/10.1145/3770855.3817978.

Renjie Wu and Eamonn 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, 2023. doi: 10.1109/TKDE.2021.3112126.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m08aK3xxdJ.

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

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

Ailing Zeng, Mu-Hwa Chen, L. Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:249097444.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All: Power general time series analysis by pretrained lm. In *NeurIPS*, 2023.

A Design Principles

The Finite Temporal Horizon principle states that causal influence is effective only within a finite window, and effects beyond L are negligible and hard-truncated. This justifies the sliding-window input $\mathbf{z}_t \in \mathbb{R}^{NL}$, which truncates the receptive field at $t - L$ and mitigates overfitting to spurious long-range correlations. Theoretically, it is formalized as Assumption C.3, which restricts the underlying SVAR to order L (Kilian, 2006).

The Causal Horizon principle states that, conditioned on direct causal parents, indirect and non-causal variables are conditionally independent of the target and offer no incremental predictive information. This motivates the Top-k gradient competition, since under MSE loss the gradient magnitude reflects incremental predictive contribution, and retaining only the top-k magnitudes forces the model to ignore conditionally independent variables and isolate direct causal parents. Theoretically, it is formalized via Assumption C.2 and Lemma 4.2, which guarantees that the expected gradient of non-causal variables is zero at the population level.

B Supplementary Details on LoCo

B.1 Linear LoCo Training Algorithm

We provide the full training procedure about Linear LoCo in Algorithm 1.

Algorithm 1 Linear LoCo Training with Top-k Gradient Competition

Require: $\{\mathbf{x}_t\}_{t=1}^{T}$, max lag L, competitive window k, learning rate η.
Ensure: Time-lagged causal graph $\mathcal{G}$ extracted from $\mathbf{W}$.

```
Initialize W ∈ R^{N×NL}.
for each mini-batch B do
    Construct z_t = [X_{1,t−1}, ..., X_{N,t−L}]^⊤ for t ∈ B.
    Forward: ŷ_t = W z_t.
    Compute L = 1/(2|B|) Σ_{t∈B} ||ŷ_t − y_t||_2^2 and ∇_W L.
    for each output row i do
        S_i = TopK_k({|[∇_W L]_ij| : j = 1, ..., NL})
        [∇̃_W L]_ij = [∇_W L]_ij if j ∈ S_i, else 0.
    end for
    W ← W − η ∇̃_W L.
end for
Extract time-lagged causal edges with |W_{i,(j,τ)}| > ε.
```

Remarks on Early Stopping. Although Lemma 4.2 guarantees zero *expected* gradient for non-causal variables, finite-sample gradients fluctuate. Early stopping on validation MSE acts as a temporal regularizer, preventing accumulation of spurious edges after true parents converge.

B.2 Detailed Derivation of Kernel LoCo

Variable-Wise Random Fourier Features. Standard kernel methods often project the entire multivariate vector $\mathbf{x}_t \in \mathbb{R}^N$ jointly into $\mathbb{R}^D$. However, joint projection causes cross-variable feature entanglement, making it impossible to disentangle individual causal contributions during backpropagation. By applying RFF independently per variable and per lag, therefore, Kernel LoCo isolates channel-specific non-linear transformations. For each lagged variable $X_{j,t-\tau}$,

$$\phi_{j,\tau}(x) = \sqrt{\tfrac{2}{D}} \cos(\boldsymbol{\omega}_{j,\tau} x + \mathbf{b}_{j,\tau}) \in \mathbb{R}^D, \tag{14}$$

with $\boldsymbol{\omega}_{j,\tau} \sim \mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I}_D)$, $\mathbf{b}_{j,\tau} \sim \mathrm{Uniform}(0, 2\pi)^D$, both frozen. Concatenating all mapped representations yields the global feature vector

$$\boldsymbol{\Phi}(\mathbf{z}_t) = [\phi_{1,1}(X_{1,t-1})^\top, \ldots, \phi_{N,L}(X_{N,t-L})^\top]^\top \in \mathbb{R}^{NLD}. \tag{15}$$

Variable-wise mapping avoids cross-variable feature entanglement, preserving per-channel attribution.

Group-Wise Weight Structure. The high-dimensional weight matrix $\mathbf{W} \in \mathbb{R}^{N \times NLD}$ can be partitioned column-wise into NL functional groups, each corresponding to a specific lagged variable $X_{j,t-\tau}$

$$\mathbf{W} = \left[\mathbf{W}_{(1,1)}, \mathbf{W}_{(1,2)}, \ldots, \mathbf{W}_{(N,L)}\right], \quad \text{where } \mathbf{W}_{(j,\tau)} \in \mathbb{R}^{N \times D}. \tag{16}$$

For a specific target output i, the weight sub-vector corresponding to input group (j, τ) is denoted as $\mathbf{w}_{i,(j,\tau)} \in \mathbb{R}^D$. The forward forecasting model is expressed as

$$\hat{\mathbf{y}}_t = \mathbf{W}\boldsymbol{\Phi}(\mathbf{z}_t). \tag{17}$$

Group-Level Top-k Competition. During backpropagation, the full gradient matrix $\mathbf{G} = \nabla_{\mathbf{W}}\mathcal{L} \in \mathbb{R}^{N \times NLD}$ mirrors the block structure of $\mathbf{W}$. For output i and feature group (j, τ), let $\mathbf{g}_{i,(j,\tau)} \in \mathbb{R}^D$ denote the gradient vector across the D kernel weights. Since the D features jointly represent one variable-lag entity, competition operates at the group level

$$v_{i,(j,\tau)} = \|\mathbf{g}_{i,(j,\tau)}\|_2, \quad \tilde{\mathbf{g}}_{i,(j,\tau)} = \begin{cases} \mathbf{g}_{i,(j,\tau)}, & \text{if } v_{i,(j,\tau)} \in \text{Top-}\, k, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{18}$$

Thus, each variable-lag entity is either holistically retained or fully pruned.

Causal Strength Extraction. Upon convergence, the non-linear causal strength from variable $X_{j,t-\tau}$ to target $X_{i,t}$ is measured by the L_2 norm of the learned weight vector

$$s_{i,(j,\tau)} = \left\| \mathbf{w}^*_{i,(j,\tau)} \right\|_2. \tag{19}$$

A time-lagged directed causal edge $X_{j,t-\tau} \longrightarrow X_{i,t}$ exists if $s_{i,(j,\tau)} > \epsilon$, where $\epsilon > 0$ is a small filtering threshold.

Comparison with Linear LoCo. While Linear LoCo enforces competition across individual scalar weights (NL candidates per row), Kernel LoCo enforces competition across feature groups (NL blocks of dimension D). This structural design ensures that an entire variable-lag entity (j, τ) is either holistically selected or fully pruned, preventing partial or corrupted feature representations.

B.3 Supplementary Details on Root Cause Analysis

This section provides the complete mathematical formulations, detailed derivations, algorithmic implementation, and report design for the RCA pipeline introduced in Section 3.4.

B.3.1 Complete Formulas for Adaptive Labeling.

To reliably distinguish genuine system anomalies from background operational noise without relying on manual threshold tuning, LoCo constructs variable-specific adaptive thresholds based on validation forecasting residual distributions under normal operating conditions.

Let $r_{i,t} = X_{i,t} - \hat{X}_{i,t}$ denote the raw prediction error, and $e_i(t) = r_{i,t}/\sigma_i$ denote the standardized error. The empirical quantile threshold θ_i for variable i is defined as

$$\theta_i = \max\left(\left|Q_{e_i}\left(\frac{\beta}{2}\right)\right|, \left|Q_{e_i}\left(1-\frac{\beta}{2}\right)\right|\right), \tag{20}$$

where $Q_{e_i}(\cdot)$ represents the empirical quantile function of the error e_i evaluated over normal historical data, and $\beta \in (0,1)$ is a hyperparameter regulating the significance level (false positive rate). In practice, setting $\beta = 0.01$ establishes a robust 99% confidence interval for nominal operation.

At runtime, the set of anomalous variables $\mathcal{A}_t$ at timestamp t is identified as

$$\mathcal{A}_t = \{i \mid |e_i(t)| > \theta_i\}. \tag{21}$$

When an anomaly is detected, the traceback algorithm initiates at the most salient symptom variable i^*, defined as the node exhibiting the largest normalized deviation relative to its individual tolerance threshold

$$i^* = \arg\max_{i\in\{1,\dots,N\}} \frac{|e_i(t)|}{\theta_i}. \tag{22}$$

B.3.2 Detailed Derivation of Counterfactual Perturbation

Once symptom variable i^* is pinpointed, LoCo isolates the root cause by evaluating the incremental contribution of its direct causal parents.

Linear Counterfactual Decomposition. For Linear LoCo, the prediction is

$$\hat{X}_{i,t} = \sum_{p=(j,\tau)} W_{i,p} X_p, \quad X_p \equiv X_{j,t-\tau}. \tag{23}$$

Let $\hat{X}_p^{\text{ref}}$ denote the normal reference value of parent p (in practice, the model's prediction $\hat{X}_{j,t-\tau}$). The deviation of parent p from its normal reference is

$$\Delta_p = X_p - \hat{X}_p^{\text{ref}}. \tag{24}$$

Because the predictor is linear, the contribution of parent p to the prediction residual is exactly

$$c_{i\leftarrow p} = W_{i,p} \cdot \frac{\Delta_p}{\sigma_i} = W_{i,p} \cdot \frac{X_p - \hat{X}_p^{\text{ref}}}{\sigma_i}, \tag{25}$$

where σ_i is the residual standard deviation of variable i on normal training data. This is an exact, closed-form attribution, and no approximation or perturbation is needed.

Kernel Counterfactual Perturbation. For Kernel LoCo, the non-linear RFF mapping $\mathbf{\Phi}(\mathbf{z}_t)$ prevents closed-form linear additive decomposition across individual scalar input channels. To resolve this, we employ explicit counterfactual intervention.

Let $\hat{X}_{i,t} = f_i(\mathbf{z}_t)$ denote the actual model forecast. We construct a counterfactual historical vector $\mathbf{z}_t^{(p)}$ wherein parent feature $p = (j,\tau)$ is replaced by its normal reference value $\hat{X}_{j,t-\tau}^{\text{ref}}$, while holding all other input features intact. The counterfactual forecast is $\hat{X}_{i,t}^{(p)} = f_i(\mathbf{z}_t^{(p)})$. The non-linear contribution score $c_{i\leftarrow p}^{\text{K}}$ is quantified by the normalized prediction shift under intervention

$$c_{i\leftarrow p}^{\text{K}} = \frac{\hat{X}_{i,t} - \hat{X}_{i,t}^{(p)}}{\sigma_i}. \tag{26}$$

Eq. 26 isolates the true counterfactual impact of parent channel p on target output i, bypassing non-linear feature entanglement.

B.3.3 Contribution-Guided Causal Traceback

Algorithm 2 formalizes the recursive procedure on the time-lagged graph $\mathcal{G}$.

Algorithm 2 Contribution-Guided Causal Traceback

Require: Anomaly timestamp t, graph $\mathcal{G}$, thresholds $\{\theta_i\}$, $\{\sigma_i\}$, cutoff ζ.
Ensure: Ranked root causes $\mathcal{R}$, propagation paths $\mathcal{P}$.

Compute $e_i(t) = |X_{i,t} - \hat{X}_{i,t}|/\sigma_i$; form $A_t = \{i : |e_i(t)| > \theta_i\}$.
$i^* \leftarrow \arg\max_i |e_i(t)|/\theta_i$.
$\mathcal{Q} \leftarrow \{(i^*, t, [(i^*, t)], 1.0)\}$; $\mathcal{R} \leftarrow \emptyset$; $\mathcal{P} \leftarrow \emptyset$.
while $\mathcal{Q} \neq \emptyset$ **do**
    Pop $(i, t_{\text{curr}}, \text{path}, \text{score})$ from $\mathcal{Q}$.
    for each parent $p = (j, \tau)$ of i in $\mathcal{G}$ with $t_{\text{prev}} = t_{\text{curr}} - \tau$ **do**
        Compute $c_{i \leftarrow p}$ via linear eq. 25 or kernel formula ep. 26.
        if $|c_{i \leftarrow p}| > \zeta$ **then**
            newpath $\leftarrow$ path $\oplus$ $[(j, t_{\text{prev}})]$; newscore $\leftarrow$ score $\cdot$ $|c_{i \leftarrow p}|$.
            if $j \in A_{t_{\text{prev}}}$ **then**
                Push $(j, t_{\text{prev}}, \text{newpath}, \text{newscore})$ into $\mathcal{Q}$.
            else
                $\mathcal{R} \leftarrow \mathcal{R} \cup \{(j, t_{\text{prev}}, \text{newscore})\}$; $\mathcal{P} \leftarrow \mathcal{P} \cup \{\text{newpath}\}$.
            end if
        end if
    end for
end while
Aggregate $\text{Attr}(j, t - \tau)$ for candidates in $\mathcal{R}$; sort descending.
return $\mathcal{R}, \mathcal{P}$.

Table 5: Industrial Root Cause Analysis Diagnostic Report Template

Diagnostic Metric	**System Output / Value**
Anomaly Timestamp	$t =$ [Timestamp]
Overall Error Index	$\text{SPE}(t) = \sum_{i=1}^{N} e_i(t)^2$
Primary Symptom Variable	X_{i^*} (Deviation ratio: $\frac{\lvert e_{i^*}(t)\rvert}{\theta_{i^*}}$)
Ranked Root Causes	1. $(X_k,\ t - \tau_1)$ – Attr $= 0.842$ 2. $(X_m,\ t - \tau_2)$ – Attr $= 0.517$
Primary Causal Propagation Path	$X_{k,t-2} \xrightarrow{c_{j \leftarrow k}=0.78} X_{j,t-1} \xrightarrow{c_{i^* \leftarrow j}=0.92} X_{i^*,t}$

B.3.4 Complete Output Report Template

For automated deployment in industrial monitoring platforms, LoCo synthesizes the traceback results into a structured Diagnostic Report. Below is the standard report template:

B.4 Supplementary Details on Causal Compression

This section elaborates on the temporal compression pipeline introduced in Section 3.5, which transforms fine-grained time-lagged matrices into a global variable-level causal graph $\mathbf{C} \in \mathbb{R}^{N \times N}$.

B.4.1 Aggregation Strategies

Given the reshaped 3rd-order tensor $\mathbf{W}' \in \mathbb{R}^{N \times N \times L}$, where $W'_{i,j,\tau}$ represents the causal impact of variable j at lag τ on target i, three temporal aggregation strategies are evaluated

(1) Average (AVG): Assumes equal temporal weighting across the light cone window

$$C_{i,j} = \frac{1}{L} \sum_{\tau=1}^{L} W'_{i,j,\tau}. \tag{27}$$

(2) Moving Average (MA): Imposes a decay parameter $\alpha > 0$ to prioritize recent temporal dependencies

$$w_\tau \propto \tau^{-\lambda}, \quad C_{i,j} = \sum_{\tau=1}^{L} w_\tau W'_{i,j,\tau}, \quad \text{s.t.} \sum_{\tau=1}^{L} w_\tau = 1. \tag{28}$$

(3) Maximum (MAX): Extracts the peak causal influence over the historical window

$$C_{i,j} = \max_{\tau \in \{1,\dots,L\}} W'_{i,j,\tau}. \tag{29}$$

The aggregation strategies are selected according to the reality and we set $\alpha = 0.5$ in experiments.

B.4.2 Signed Magnitude Aggregation for Kernel LoCo

For Kernel LoCo, each variable-lag entry corresponds to a weight vector $\mathbf{w}_{i,k} \in \mathbb{R}^D$, where index k maps to tuple (j, τ). Before applying temporal aggregation, we collapse the feature dimension D into a signed scalar magnitude.

For each weight vector $\mathbf{w}_{i,k}$, we compute its directional orientation and magnitude as follows

$$\text{dir}_{i,k} = \text{sgn}\left(\frac{1}{D}\sum_{d=1}^{D} W_{i,k,d}\right), \quad \text{mag}_{i,k} = \|\mathbf{w}_{i,k}\|_2 . \tag{30}$$

The signed magnitude is defined as $\text{dir}_{i,k} \cdot \text{mag}_{i,k}$. This mapping effectively collapses the kernel weight tensor back to $\mathbb{R}^{N\times NL}$, allowing Kernel LoCo to seamlessly utilize the exact same average aggregation, antisymmetrization, and thresholding pipeline as Linear LoCo.

B.4.3 Antisymmetrization and Thresholding

To eliminate spurious bidirectional edges arising from mutual correlation or unobserved confounders, we apply off-diagonal antisymmetrization:

$$\mathbf{C} \leftarrow \max\left(\mathbf{C} - \mathbf{C}^\top, 0\right), \quad \text{for } i \neq j. \tag{31}$$

This operation retains only the dominant causal flow between any pair of variables.

Preservation of Self-Autoregressive Effects. Crucially, the diagonal elements $C_{i,i}$, representing variable self-history dependencies $X_{i,t-\tau} \to X_{i,t}$ are excluded from antisymmetrization and fully preserved, as self-autoregression is essential for accurate time series forecasting.

Finally, a quantile threshold q (e.g., $q = 0.8$) is applied to filter residual weak edges, yielding a high-confidence sparse global causal graph $\mathbf{C} \in \mathbb{R}^{N\times N}$.

C Theoretical Assumptions and Proofs

C.1 Global Assumptions and Proposition of Finite-Sample Support Selection

We first state the global assumptions used in the theoretical analysis.

Assumption C.1 (Stationary Local SVAR (Kilian, 2006))**.** *The true physical system is governed by a stationary local structural vector autoregressive process:*

$$X_{i,t} = \sum_{j\in \text{Pa}_i} \sum_{\tau=1}^{L} W^*_{i,j,\tau} X_{j,t-\tau} + \epsilon_{i,t}, \tag{32}$$

where Pa_i *is the true causal parent set of* X_i*,* W^* *are the true causal weights, and* $\epsilon_{i,t}$ *are zero-mean, i.i.d. noises. Under normal conditions, the system is approximately stationary and variables have sparse, finite-lag linear causal dependencies.*

Assumption C.2 (Exogeneity)**.** *The noises* $\epsilon_{i,t}$ *are zero-mean and i.i.d., and strictly time-causal: for any* $\tau > 0$*,*

$$\epsilon_{i,t} \perp X_{j,t-\tau}. \tag{33}$$

Let $\boldsymbol{\epsilon}_t = [\epsilon_{1,t}, \dots, \epsilon_{N,t}]^\top \in \mathbb{R}^N$ *denote the innovation vector at time* t*.*

Assumption C.3 (Finite Lag (Kilian, 2006))**.** *The true lag* τ *is bounded by a maximum lag* L*. This justifies the hard truncation of the input vector* $\mathbf{z}_t$*.*

Assumption C.4 (Restricted Eigenvalue, RE). *There exist constants $\kappa_{\min}, \kappa_{\max} > 0$ such that for any $2k$-sparse vector v,*

$$\kappa_{\min}\|v\|_2^2 \le v^\top \Sigma_z v \le \kappa_{\max}\|v\|_2^2, \quad \Sigma_z = \mathbb{E}[\mathbf{z}_t \mathbf{z}_t^\top]. \tag{34}$$

This condition is analogous to the restricted isometry property (RIP) in compressed sensing, but is stated directly on the population covariance Σ_z (Negahban & Wainwright, 2012).

Assumption C.5 (Population Gradient Gap). *There exists $\Delta_{\text{pop}} > 0$ such that for the true support S_i^* and any non-support coordinate $q' \notin S_i^*$,*

$$|g_{i,q}(w_i^*)| \ge \Delta_{\text{pop}} + |g_{i,q'}(w_i^*)|, \quad \forall q \in S_i^*, \tag{35}$$

where $g_i(w) = \nabla \mathbb{E}[\frac{1}{2}(X_{i,t} - w^\top \mathbf{z}_t)^2]$ is the population gradient. This is a gradient-level beta-min condition.

Assumption C.6 (Sub-Gaussian Design). *The input vector $\mathbf{z}_t$ is sub-Gaussian, enabling finite-sample concentration of empirical gradients.*

This appendix then provides complete proofs for the theoretical results in Section 4, including the gradient–residual equivalence, the vanishing-gradient lemma, the population-level support screening theorem, and the finite-sample support selection proposition. We also clarify the relation to classical IHT, the population gradient gap, and the group-level extension for Kernel LoCo.

Proposition C.7 (Finite-Sample Support Selection). *Under Assumptions C.1–C.6, if the sample size T satisfies*

$$c\sqrt{\frac{\log(NL/\delta)}{T}} < \frac{\Delta_{\text{pop}}}{2}, \tag{36}$$

then with probability at least $1-\delta$, the empirical top-k gradient coordinates contain the true support S_i^*.

C.2 Proof of Theorem 4.1

Define the MSE loss for target variable i at time t

$$\mathcal{L} = \frac{1}{2}\mathbb{E}\big[(X_{i,t} - \hat{X}_{i,t})^2\big]. \tag{37}$$

Let $r_{i,t} = X_{i,t} - \hat{X}_{i,t}$. For a single-layer predictor,

$$\hat{X}_{i,t} = \sum_q W_{i,q} Z_{t,q}, \tag{38}$$

where $Z_{t,q}$ denotes the q-th component of the historical input $\mathbf{z}_t$, with q indexing the pair (j, τ). Taking the partial derivative with respect to $W_{i,q}$,

$$\nabla_{W_{i,q}} \mathcal{L} = \mathbb{E}\Big[(X_{i,t} - \hat{X}_{i,t}) \cdot \frac{\partial (X_{i,t} - \hat{X}_{i,t})}{\partial W_{i,q}}\Big]. \tag{39}$$

Since $\partial \hat{X}_{i,t} / \partial W_{i,q} = Z_{t,q}$, we have

$$\nabla_{W_{i,q}} \mathcal{L} = -\mathbb{E}\big[r_{i,t} Z_{t,q}\big]. \tag{40}$$

Because the data are centered and standardized, $\mathbb{E}[r_{i,t}] = 0$ and $\mathbb{E}[Z_{t,q}] = 0$. Therefore,

$$\mathbb{E}[\nabla_{W_{i,q}} \mathcal{L}] = -\text{Cov}(r_{i,t}, Z_{t,q}). \tag{41}$$

This proves the claimed proportionality. □

C.3 Proof of Lemma 4.2

We evaluate the population gradient at the true parameter w_i^*. Under Assumption C.1, the true data generating process is

$$X_{i,t} = \sum_{q \in S_i^*} W_{i,q}^* Z_{t,q} + \epsilon_{i,t}, \tag{42}$$

where $S_i^* = \mathrm{supp}(w_i^*)$ and $\mathbb{E}[\epsilon_{i,t} Z_{t,q}] = 0$ for all q by Assumption C.2. At w_i^*, the residual is exactly the innovation noise

$$r_{i,t} = X_{i,t} - \hat{X}_{i,t} = \epsilon_{i,t}. \tag{43}$$

By Theorem 4.1, for any $q \notin S_i^*$,

$$\mathbb{E}[\nabla_{W_{i,q}} \mathcal{L}] = -\mathrm{Cov}(\epsilon_{i,t}, Z_{t,q}) = 0, \tag{44}$$

where the last equality follows from $\epsilon_{i,t} \perp Z_{t,q}$ (Assumption C.2). Note that this is a statement about the *expected* gradient at the true parameter, not about stochastic gradients at arbitrary iterates. □

C.4 Proof of Theorem 4.3

By Theorem 4.1 and Lemma 4.2, at the true parameter w_i^* For any $q \in S_i^*$, Assumption C.5 gives

$$|g_{i,q}(w_i^*)| \geq \Delta_{\mathrm{pop}} > 0.$$

For any $q' \notin S_i^*$,

$$|g_{i,q'}(w_i^*)| = 0.$$

Therefore, the magnitude of the population gradient on the true support strictly exceeds that on any non-support coordinate. When selecting the top-k coordinates of $|g_i(w_i^*)|$ with $k \geq |S_i^*|$, all true support indices must be included

$$S_i^* \subseteq \mathrm{Top}_k\big(|g_i(w_i^*)|\big). \tag{45}$$

This proves the population-level screening property. □

C.5 Proof of Proposition C.7

We first establish a concentration inequality for the empirical gradient. For a fixed w, define the empirical gradient

$$\widehat{g}_i(w) = -\frac{1}{T} \sum_{t=1}^{T} \big(X_{i,t} - w^\top \mathbf{z}_t\big)\mathbf{z}_t. \tag{46}$$

Under Assumption C.6, both $\mathbf{z}_t$ and $\epsilon_{i,t}$ are sub-Gaussian. Standard sub-Gaussian concentration (see, e.g., Vershynin, 2018) yields that for any $t > 0$,

$$\Pr\Big(\|\widehat{g}_i(w) - g_i(w)\|_\infty \geq t\Big) \leq 2\exp\big(-cTt^2\big), \tag{47}$$

for some constant $c > 0$. Setting the right-hand side to δ and solving for t,

$$\|\widehat{g}_i(w) - g_i(w)\|_\infty \leq c\sqrt{\frac{\log(NL/\delta)}{T}} \tag{48}$$

with probability at least $1 - \delta$.

Now, by Assumption C.5, the population gradient gap satisfies

$$|g_{i,q}(w_i^*)| \geq \Delta_{\mathrm{pop}}, \quad \forall q \in S_i^*. \tag{49}$$

If the estimation error is smaller than $\Delta_{\mathrm{pop}}/2$, i.e.,

$$c\sqrt{\frac{\log(NL/\delta)}{T}} < \frac{\Delta_{\mathrm{pop}}}{2}, \tag{50}$$

then for any $q \in S_i^*$,

$$|\widehat{g}_{i,q}(w_i^*)| \geq |g_{i,q}(w_i^*)| - \|\widehat{g}_i - g_i\|_\infty \geq \frac{\Delta_{\mathrm{pop}}}{2} > 0. \tag{51}$$

Meanwhile, for any $q' \notin S_i^*$, the population gradient is zero, so $|\widehat{g}_{i,q'}(w_i^*)| \leq \|\widehat{g}_i - g_i\|_\infty < \Delta_{\mathrm{pop}}/2$. Hence, the empirical top-k coordinates cannot exclude any true support coordinate. Therefore, with probability at least $1 - \delta$,

$$S_i^* \subseteq \mathrm{Top}_k\big(|\widehat{g}_i(w_i^*)|\big). \tag{52}$$

□

C.6 Additional Theoretical Analysis

Relation to IHT. Classical Iterative Hard Thresholding (IHT) updates parameters by

$$W^{(m+1)} = \mathcal{H}_k\big(W^{(m)} - \eta \nabla_{W^{(m)}} \mathcal{L}\big), \tag{53}$$

which enforces k-sparsity on the *parameter* at every iteration. In contrast, LoCo uses gradient hard-thresholding (GHT)

$$W^{(m+1)} = W^{(m)} - \eta \mathcal{H}_k\big(\nabla_{W^{(m)}} \mathcal{L}\big), \tag{54}$$

which enforces sparsity on the *gradient direction* but does not guarantee that $W^{(m)}$ is k-sparse. Consequently, the standard IHT + RIP analysis, which relies on parameter projection, cannot be directly applied to LoCo. Our theory therefore establishes a population-level screening property and a finite-sample support selection guarantee for GHT, rather than parameter-level linear convergence.

RE vs. RIP. The restricted eigenvalue (RE) condition in Assumption C.4 is stated on the population covariance Σ_z. It is analogous to the restricted isometry property (RIP) in compressed sensing, but RIP is typically defined on a measurement matrix A via $\|Av\|_2^2$. For covariance-based analysis in high-dimensional statistics, the RE condition is the standard tool. It ensures that the Gram matrix is well-conditioned on sparse subspaces, which is sufficient for support selection.

Interpretation of the Population Gradient Gap. Assumption C.5 is a gradient-level beta-min condition. It requires that true causal parents produce a population gradient magnitude at least Δ_{pop} larger than any non-causal coordinate. In the ideal case where non-causal gradients are exactly zero, this reduces to a lower bound on the gradient magnitude of true parents. This condition is standard in support recovery literature and is necessary to distinguish true signals from noise.

Group-Level GHT for Kernel. For Kernel LoCo, each variable–lag entity corresponds to a group of D features. The group-level GHT update retains the top-k groups by gradient L_2 norm, zeroing out the rest. The same high-level argument applies at the group level: under a group-wise RE condition and a group-level gradient gap, the group-wise GHT retains all true groups at the population level, and a finite-sample version follows from concentration of group gradient norms. A rigorous treatment involves additional technicalities (feature dimension D, bandwidth σ, group-wise RE constants) and is left to future work.

D Discussion: Relation and Difference Between LoCo and Granger Causality

Granger causality is a classical framework for temporal causal discovery with mature statistical interpretation. LoCo shares its philosophical basis, that temporal precedence plus incremental predictive information, but differs in implementation, judgment timing, and source of information.

Definition D.1 (Granger Causality)**.** *X_j Granger-causes X_i if, conditional on X_i's own history, adding X_j's history significantly improves prediction of $X_{i,t}$ or $X_{i,t+1}$. In linear VAR, this reduces to comparing residual variance or likelihood across nested models, with judgment typically occurring after fitting via F-tests or likelihood ratio tests.*

Definition D.2 (LoCo Causal Support Recovery)**.** *LoCo is formalized as a gradient hard-thresholding (GHT) update:*

$$W^{(m+1)} = W^{(m)} - \eta\, \mathcal{H}_k\big(\nabla_{W^{(m)}} \mathcal{L}\big),$$

where $\mathcal{H}_k$ retains the top-k gradient coordinates and zeros out the rest. Rather than post-hoc significance tests, LoCo selects a stable support set through gradient competition during training. A variable is retained if it continuously receives sufficient gradient support; the goal is sparse support recovery, not only prediction improvement.

Both frameworks ask whether a historical variable provides irreplaceable predictive gain. Granger judges by comparing information sets, and LoCo recovers a sparse support via gradient competition. Thus Granger is a predictive-gain discrimination framework, while LoCo is a structure-recovery framework. The timing of structural judgment also differs. Granger follows a separated fit–test pipeline. Fit first, then test coefficient significance. LoCo internalizes selection into training. At

each iteration, only gradient-dominant coordinates are updated. Judgment is dynamic rather than one-shot, making LoCo suitable for online and streaming settings without multiple model fits.

Finally, the source of information differs. Granger emphasizes endpoint interpretability (final coefficient significance), and LoCo emphasizes causal survivability along the optimization. The two are complementary. Granger excels in statistical clarity, while LoCo couples structure selection with parameter learning.

E DATASETS

This section provides detailed descriptions of all datasets used in our experiments.

E.1 ANOMALY DETECTION DATASETS

We adopt five widely used multivariate time-series anomaly detection benchmarks and four additional datasets. Recent studies have pointed out that the five widely used benchmarks are flawed and may create an illusion of progress (Wu & Keogh, 2023). We therefore include four more challenging datasets to provide a more comprehensive evaluation. Table 6 summarizes their statistics.

Table 6: Statistics of the anomaly detection datasets. AR denotes the anomaly ratio.

Dataset	Source	Dim	Train	Test	AR (%)
MSL	NASA spacecraft telemetry	55	58,317	73,729	10.5
SMAP	NASA satellite telemetry	25	135,183	427,617	12.8
PSM	eBay server metrics	25	132,481	87,841	27.8
SMD	Internet server machines	38	708,405	708,420	4.20
SWaT	Industrial water-treatment system	51	495,000	449,919	12.1
NIPS-TS-SWAN	Space weather	38	6,000	6,000	32.6
NIPS-TS-Creditcard	Financial transactions	29	284,807	284,807	0.17
NIPS-TS-GECCO	Industrial water-treatment	9	69,260	69,261	1.05
CIC-IDS2017 subset	Network intrusion	72	85,115	85,116	0.09

- **MSL.** The Mars Science Laboratory dataset contains multivariate telemetry from NASA's Mars rover, including sensor and actuator readings. Anomaly labels come from incident surprise anomaly records, and it is commonly used for unsupervised multivariate time-series anomaly detection. (Hundman et al., 2018)
- **SMAP.** The Soil Moisture Active Passive dataset contains telemetry from NASA's SMAP satellite, recording sensor and system variables under normal and anomalous operation. It is typically used alongside MSL. (Hundman et al., 2018)
- **PSM.** Pooled Server Metrics contains monitoring metrics from multiple application server nodes at eBay. It is server-operations data rather than network-intrusion data. (Abdulaal et al., 2021)
- **SMD.** The Server Machine Dataset contains machine operational metrics from a large internet company's server cluster, covering 28 server-machine entities with a 1-minute sampling period. Statistics in Table 6 are aggregated over all 28 entities. (Su et al., 2019)
- **SWaT.** The Secure Water Treatment dataset is a scaled but operational water-treatment industrial control system testbed developed at iTrust, Singapore University of Technology and Design. It contains 51 sensor and actuator variables sampled at 1 Hz over 11 days, with 41 injected attack events in the last 4 days. (Mathur & Tippenhauer, 2016)
- **SWAN.** A space-weather dataset included in the NIPS-TS benchmark. (Angryk et al., 2020)
- **Creditcard.** A credit-card transaction dataset with extremely rare fraud labels. The reported train/test sizes follow the NIPS-TS preprocessing; the original dataset contains 284,807 transactions in total. (Pozzolo et al., 2015)
- **GECCO.** An industrial water-treatment dataset derived from the GECCO 2018 challenge on monitoring drinking-water quality. (Moritz et al., 2018)

- **CICIDS.** A subset of the CIC-IDS2017 network-intrusion dataset, containing normal traffic and multiple attack categories such as brute force, DoS, DDoS, Heartbleed, web attacks, infiltration, botnet, and port scan. (Sharafaldin et al., 2018)

E.2 Root Cause Analysis Datasets

For RCA, most public anomaly detection datasets do not provide fine-grained labels identifying which variable is responsible for an anomaly at a specific time step. We therefore use the following two datasets.

- **SWaT segment.** We use an annotated segment of SWaT containing component-level attack labels for root-cause localization at anomalous time steps. We emphasize that not all versions of SWaT provide such variable-level root-cause labels, our evaluation uses the annotated segment specifically.(Mathur & Tippenhauer, 2016) [1]
- **MSDS.** The Multi-Source Distributed System Data dataset is designed for AIOps and contains distributed metrics, application logs, and distributed traces from an OpenStack/microservice testbed. Controlled fault injection produces fault records suitable for root-cause analysis. (Nedelkoski et al., 2020)

E.3 Causal Discovery Datasets

RiversEastGermany. We use RiversEastGermany from the CausalRivers dataset, a real-world hydrological time series with a 15-minute sampling interval. It contains 666 nodes and 651 edges, with ground-truth causal directions and propagation lags derived from the river network and hydrological relationships. (Stein et al., 2025)

Subgraph construction. Following CausalRivers, we sample subgraphs with different structures and difficulty levels from the real causal graph. The final benchmark graphs are restricted to three nodes; we therefore omit node count and describe each graph set below.

- **Random**: samples all possible connected subgraphs. It covers the entire dataset and represents the broadest diversity and conditions.
- **Close**: samples all possible connected subgraphs where all edges have a geographic distance ≤ 0.3 (Euclidean). It excludes long-range effects to make causal effects more evident; all subgraphs are included in Random.
- **Random+1**: first samples 2-node connected subgraphs, then adds one fully disconnected node selected from coastal and border regions where stations are completely disconnected. It tests robustness to irrelevant variables.
- **Root cause**: samples all possible connected subgraphs where each node has at most one parent, forming a single-chain structure. It is used for root-cause analysis; all subgraphs are also included in Random.
- **Confounder**: first selects 4-node subgraphs, chooses one node with multiple children (natural or artificial river bifurcation), and then removes this multi-child node to simulate a permanently hidden confounder. It tests the ability to handle hidden confounders, making it the most challenging setting.

Each subgraph inherits the true causal edges and time lags from the underlying RiversEastGermany graph. The benchmark is used to evaluate causal edge identification, direction identification, and robustness under different structural conditions.

F Baselines

For **AD**, we compare against two groups of baselines. The anomaly detection-specific baselines include machine learning detectors, LOF (Breunig et al., 2000) and Isolation Forest (IForest)(Liu

[1] https://github.com/hanxiao0607/AERCA

et al., 2008), and recent deep detectors, including Anomaly Transformer (AnomalyTran), DCdetector (Yang et al., 2023), CATCH (Wu et al., 2025), SEA-FGT (Huang et al., 2026), CoAD (Tang et al., 2026), and LEFT (Wang et al., 2026a). We further include causality-aware detectors CAROTS (Kim et al., 2025b), GCAD (Liu et al., 2025c), and AERCA (Han et al., 2025).

The generative baselines include both forecasting-based and reconstruction-based detectors. Some deep models are adapted for reconstruction-based detection, including VAE (Kingma & Welling, 2013), iTransformer (Liu et al., 2023), DLinear (Zeng et al., 2022), NLinear (Zeng et al., 2022), and PatchTST (Nie et al., 2023). Others perform forecasting-based anomaly detection, such as the LLM-based detectors CALF (Liu et al., 2025a) and GPT4TS (Zhou et al., 2023), and the pretrained models Timer (Liu et al., 2024) and Chronos (Ansari et al., 2024). The remaining baselines are reconstruction-based detectors, including LLMMixer (Kowsher et al., 2024), Moment (Goswami et al., 2024), and UniTS (Gao et al., 2024). For **RCA**, we compare against ϵ-Diagnosis (Shan et al., 2019) and AERCA (Han et al., 2025). For **CD**, we compare against PCMCI with a linear conditional independence test (Runge et al., 2017), VarLiNGAM (Hyvärinen et al., 2010), Dynotears (Pamfil et al., 2020), a simple linear Granger causal approach (VAR) (Assaad et al., 2022).

G Details about metrics

G.1 Anomaly Detection Metrics

Threshold-Independent Metrics. For a score-based detector, let $s(x)$ denote the anomaly score of a sample x. Given a threshold θ, the true positive rate (TPR) and false positive rate (FPR) are

$$\mathrm{TPR}(\theta) = \frac{TP(\theta)}{TP(\theta) + FN(\theta)}, \quad \mathrm{FPR}(\theta) = \frac{FP(\theta)}{FP(\theta) + TN(\theta)}. \tag{55}$$

The ROC curve is

$$\mathrm{ROC} = \{(\mathrm{FPR}(\theta), \mathrm{TPR}(\theta)) \mid \theta \in \mathbb{R}\}, \tag{56}$$

and AUROC is

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}\, d(\mathrm{FPR}). \tag{57}$$

Equivalently,

$$\mathrm{AUROC} = P\big(s(x^+) > s(x^-)\big) + \frac{1}{2} P\big(s(x^+) = s(x^-)\big), \tag{58}$$

where x^+ is a true anomaly and x^- is a true normal sample.

The PRC curve is

$$\mathrm{PRC} = \{(\mathrm{Recall}(\theta), \mathrm{Precision}(\theta)) \mid \theta \in \mathbb{R}\}, \tag{59}$$

with

$$\mathrm{Precision}(\theta) = \frac{TP(\theta)}{TP(\theta) + FP(\theta)}, \quad \mathrm{Recall}(\theta) = \mathrm{TPR}(\theta). \tag{60}$$

AUPRC is

$$\mathrm{AUPRC} = \int_0^1 \mathrm{Precision}\, d(\mathrm{Recall}). \tag{61}$$

Affiliation Metrics. Classical point-wise Precision/Recall/F1 ignore temporal proximity and event duration (Huet et al., 2022). Affiliation metrics address this by assigning each prediction to the nearest ground-truth event and comparing, within a local affiliation zone, the observed temporal distance against that of a random prediction. Let $gt_1, \ldots, gt_n$ be ground-truth anomaly events, and let $pred_1, \ldots, pred_m$ be predicted anomaly events. For each ground-truth event gt_j, let I_j be its affiliation zone. Define the distance from a point x to a set Y as

$$\mathrm{dist}(x, Y) = \min_{y \in Y} |x - y|, \tag{62}$$

and the directed average distance from a set X to a set Y as

$$\mathrm{dist}(X, Y) = \frac{1}{|X|} \int_{x \in X} \mathrm{dist}(x, Y)\, dx. \tag{63}$$

For each ground-truth event gt_j, the individual Affiliation Precision is

$$P_{\mathrm{precision}_j} = \frac{1}{|\mathrm{pred} \cap I_j|} \int_{x \in \mathrm{pred} \cap I_j} \bar{F}_{\mathrm{precision}_j}\big(\mathrm{dist}(x, gt_j)\big)\, dx, \tag{64}$$

where $\bar{F}_{\mathrm{precision}_j}$ is the precision survival function. Similarly, the individual Affiliation Recall is

$$P_{\mathrm{recall}_j} = \frac{1}{|gt_j|} \int_{y \in gt_j} \bar{F}_{y,\mathrm{recall}_j}\big(\mathrm{dist}(y, \mathrm{pred} \cap I_j)\big)\, dy. \tag{65}$$

Let $S = \{j \mid \mathrm{pred} \cap I_j \neq \emptyset\}$. Global Affiliation Precision and Recall are

$$P_{\mathrm{aff}} = \frac{1}{|S|} \sum_{j \in S} P_{\mathrm{precision}_j}, \quad R_{\mathrm{aff}} = \frac{1}{n} \sum_{j=1}^{n} P_{\mathrm{recall}_j}. \tag{66}$$

Affiliation F1 follows the standard harmonic mean

$$F1_{\mathrm{aff}} = \frac{2\, P_{\mathrm{aff}}\, R_{\mathrm{aff}}}{P_{\mathrm{aff}} + R_{\mathrm{aff}}}, \tag{67}$$

with $F1_{\mathrm{aff}} = 0$ if the denominator is zero.

G.2 Root Cause Analysis Metrics

Following the evaluation of traditional RCA, we report two complementary metrics (Han et al., 2025). Let $\mathcal{X}$ be the set of evaluated time series.

Sequence-Level Root Cause Identification. For a time series $\mathbf{X} \in \mathcal{X}$, let $V_{\mathbf{X}}^{(RC)}$ be the set of true root-cause variables. Let $R_{\mathbf{X}}[k]$ denote the variable with the k-th highest root-cause score. Then

$$AC@K = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{X} \in \mathcal{X}} \frac{\left| V_{\mathbf{X}}^{(RC)} \cap \{R_{\mathbf{X}}[k] \mid k = 1, \ldots, K\} \right|}{\min\left(K, \left| V_{\mathbf{X}}^{(RC)} \right|\right)}. \tag{68}$$

This metric evaluates whether the top-K ranked variables cover the true root-cause variables over the entire sequence, without distinguishing individual time steps.

Time-Step-Level Root Cause Identification Let $V_{\mathbf{X},t}^{(RC)}$ be the set of true root-cause variables at time step t, and let $R_{\mathbf{X}}^{*}[k]$ denote the variable with the k-th highest root-cause score across all time steps. Then

$$AC^{*}@K = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{X} \in \mathcal{X}} \frac{\left| \bigcup_{t \in T} V_{\mathbf{X},t}^{(RC)} \cap \{R_{\mathbf{X}}^{*}[k] \mid k = 1, \ldots, K\} \right|}{\min\left(K, \left| \bigcup_{t \in T} V_{\mathbf{X},t}^{(RC)} \right|\right)}. \tag{69}$$

Unlike $AC@K$, this metric requires the top-K variables to cover root cause variables that occur at specific time steps.

The Averaged. The averaged variants aggregate performance across different truncation levels

$$Avg@K = \frac{1}{K} \sum_{k=1}^{K} AC@k, \quad Avg^{*}@K = \frac{1}{K} \sum_{k=1}^{K} AC^{*}@k. \tag{70}$$

We use uppercase K to denote the top-K truncation length (e.g., $K \in \{1, 3, 5, 10\}$), and lowercase k as the summation index. This is distinct from the lowercase k used for LoCo's gradient competition.

Existing public benchmarks provide root cause labels only at the current timestamp, i.e., which variables are responsible for the anomaly at the anomalous time step. We therefore evaluate RCA under this conventional definition. For baselines that cannot localize root causes at specific time steps, we follow AERCA and use the anomalous time series predicted at the last time step of the sliding window as the root cause, so that $AC^{*}@K$ can still be computed (Han et al., 2025).

G.3 Causal Discovery Metrics

Following the evaluation protocol of CausalRivers, we report AUROC at the level of directed edges. For each sampled subgraph g, let V_g be its variable set and

$$E_g = \{(i, j) \mid i, j \in V_g,\ i \neq j\} \tag{71}$$

the set of candidate directed edges. Self-loops $i \to i$ are ignored because they are always present and would otherwise skew the results. For each edge $e = (i, j)$, the ground-truth label is

$$y_e = \begin{cases} 1, & \text{if } i \to j \text{ exists in the true graph,} \\ 0, & \text{otherwise,} \end{cases} \tag{72}$$

and the method assigns a score s_e, where a higher score indicates stronger evidence for a causal edge.

Reported AUROC. Let G be the total number of sampled subgraphs. The reported AUROC is the mean of the Individual AUROC scores:

$$\mathrm{AUROC} = \frac{1}{G} \sum_{g=1}^{G} \mathrm{AUROC}_g. \tag{73}$$

This is a macro average, each subgraph contributes equally, regardless of its number of candidate edges. In contrast, a pooled (micro) AUROC would concatenate all edges from all subgraphs and compute a single AUROC, giving larger subgraphs more weight. We follow CausalRivers and report the macro average.

Individual AUROC. For subgraph g, let P_g be the set of true non-self-loop causal edges and N_g the set of true non-edges. The Individual AUROC is

$$\mathrm{AUROC}_g = \frac{1}{|P_g||N_g|} \sum_{e \in P_g} \sum_{e' \in N_g} \left[\mathbf{1}(s_e > s_{e'}) + \frac{1}{2} \mathbf{1}(s_e = s_{e'}) \right]. \tag{74}$$

Equivalently, AUROC_g is the area under the ROC curve of subgraph g. A value of 0.5 corresponds to random ranking.

LoCo produces a time-lagged graph $\mathcal{G}$. Since mainstream CD benchmarks provide ground-truth variable-level $N \times N$ graphs, we compress $\mathcal{G}$ into a global matrix $\mathbf{C}$ following Appendix B.4, and compare $\mathbf{C}$ against the ground-truth graph. The time-lagged edges themselves cannot be quantitatively evaluated due to the lack of ground-truth lag annotations, and are instead demonstrated through case studies.

H More analysis

H.1 Affiliation Results for AD-Specific Baselines

Table 7 reports Affiliation Precision, Recall, and F1 on the five benchmarks. Overall, LoCo achieves the best or second-best Affiliation F1 on most datasets, showing that it also performs well under the event-level Affiliation metric. Table 8 reports the Affiliation results on the four additional datasets. LoCo remains competitive across these more challenging datasets, further supporting its effectiveness under the Affiliation metric. We note that on some datasets, point-wise metrics may appear less favorable for LoCo, because many baselines select thresholds based on test-set results to get perfect results, whereas LoCo strictly determines its threshold from training prediction errors, making it more realistic and suitable for deployment.

H.2 Affiliation Results for Generative Baselines

Table 9 reports Affiliation Precision, Recall, and F1 on the five benchmarks. Although LoCo does not dominate on every dataset, it remains competitive overall, achieving the best or second-best Affiliation F1 on most datasets. Notably, these baselines typically adopt more complex architectures and larger parameter budgets to improve temporal representation, at the cost of interpretability. Table 10 reports the Affiliation results on the four additional datasets. LoCo remains competitive across these more challenging datasets, further supporting its effectiveness under the Affiliation metric.

Table 7: Performance comparison on benchmark datasets in terms of Affiliation Precision (Aff Prec.), Affiliation Recall (Aff Rec.), and Affiliation F1 (Aff F1) score. The **best** result in each column is highlighted in bold, and the second result is underlined.

Method	MSL			PSM			SMAP			SWaT			SMD		
	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1
Iforest	50.25	65.60	56.91	50.28	91.01	64.77	60.99	64.64	62.76	51.96	71.77	60.28	59.53	77.33	67.27
LOF	54.87	97.19	70.14	53.16	89.00	66.56	49.81	90.30	64.21	53.01	93.54	67.67	51.77	89.97	65.72
AnomalyTran	51.07	95.41	66.52	53.75	81.83	64.88	68.97	46.86	55.80	57.44	94.19	71.36	58.24	41.26	48.30
TranAD	56.32	78.11	65.45	**98.00**	30.77	46.84	**72.36**	76.54	74.39	68.09	78.81	73.06	**85.02**	51.57	64.20
DCdetector	50.80	93.33	65.79	51.64	94.51	66.79	51.04	92.33	65.74	52.17	96.14	67.64	50.31	74.74	60.14
CATCH	64.31	62.63	63.46	77.32	86.25	81.54	52.14	69.82	59.70	69.53	58.24	63.39	83.48	67.82	74.84
SEA-FGT	**64.32**	98.88	**77.94**	69.90	98.26	81.69	65.15	**99.17**	78.64	65.09	**98.57**	**78.40**	65.21	**97.84**	78.26
CoAD	52.12	98.22	68.10	64.88	91.15	75.80	55.83	98.31	71.22	79.09	51.37	62.29	63.30	28.43	39.24
LEFT	51.82	98.81	67.99	59.03	99.73	74.16	56.46	98.96	71.90	65.68	69.77	67.66	65.09	65.18	65.13
CAROTS	51.64	**99.73**	68.04	59.54	99.00	74.36	66.19	97.15	78.73	**79.30**	28.24	41.64	62.60	89.84	73.78
GCAD	55.27	99.50	71.06	57.42	**99.92**	72.92	62.69	90.31	74.01	53.20	97.87	68.93	67.28	76.14	71.44
AERCA	55.70	97.25	70.83	55.63	99.91	71.46	67.41	80.11	73.21	68.64	49.25	57.35	63.63	63.50	63.56
Linear-LoCo	58.54	95.60	72.62	80.48	86.33	**83.30**	71.87	88.72	**79.41**	66.21	95.52	77.17	67.54	79.35	72.97
Kernel-LoCo	57.71	99.02	72.92	78.28	88.42	83.04	70.66	89.19	78.85	67.83	94.50	78.29	69.27	92.40	**79.18**

Table 8: Performance comparison on benchmark datasets (GECCO, SWAN, Creditcard, and CICIDS) in terms of Affiliation Precision (Aff Prec.), Affiliation Recall (Aff Rec.), and Affiliation F1 (Aff F1) score. The **best** result is highlighted in bold, and the second result is underlined.

Method	GECCO			SWAN			Creditcard			CICIDS		
	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1
Iforest	64.69	31.51	42.38	60.64	36.09	45.25	65.76	61.23	63.42	54.02	7.72	13.50
LOF	50.19	89.82	64.40	58.56	91.12	71.30	50.16	83.21	62.59	52.12	74.03	61.17
AnomalyTran	55.59	90.88	68.98	55.43	38.82	45.66	49.65	81.73	61.77	51.27	25.16	33.75
TranAD	64.12	29.43	40.34	76.85	25.66	38.47	63.41	27.41	38.28	60.21	82.06	69.46
DCdetector	48.42	88.87	62.68	52.45	11.95	19.47	42.38	28.39	34.00	47.77	51.27	49.46
CATCH	75.06	99.33	85.51	66.92	26.65	38.12	54.53	94.64	69.19	46.92	51.09	48.91
SEA-FGT	59.51	98.90	74.30	58.11	95.76	72.33	53.34	**97.99**	69.08	51.10	92.97	65.95
CoAD	78.21	14.66	24.69	54.01	71.57	61.56	58.81	68.69	63.37	51.29	93.44	66.23
LEFT	85.98	98.46	91.80	68.67	56.54	62.02	65.50	31.12	42.19	58.07	97.05	72.66
CAROTS	86.78	95.49	90.93	79.86	97.44	**87.78**	**70.44**	91.12	**79.46**	60.05	90.08	72.06
GCAD	85.51	98.85	91.70	80.39	95.13	87.14	56.77	97.31	71.71	**60.61**	90.65	72.65
AERCA	66.92	97.94	79.51	55.23	**99.79**	71.11	59.56	43.43	50.23	52.76	47.18	49.81
Linear-LoCo	78.95	**99.96**	88.22	71.26	66.65	68.88	63.35	86.70	73.21	54.70	**98.67**	70.38
Kernel-LoCo	**86.92**	99.88	**92.95**	**85.18**	68.47	75.92	63.27	88.28	73.71	57.93	97.84	**72.77**

H.3 Efficiency Advantages

Fig. 5 compares LoCo against both AD-specific and generative baselines on PSM and SWaT in terms of training efficiency, model size, and performance. The x-axis is training time per iteration,

Table 9: Performance comparison on benchmark datasets (MSL, PSM, SMAP, SWaT, and SMD) in terms of Affiliation Precision (Aff Prec.), Affiliation Recall (Aff Rec.), and Affiliation F1 (Aff F1) score. The **best** result in each column is highlighted in bold, and the second-best result is underlined.

Method	MSL			PSM			SMAP			SWaT			SMD		
	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1
VAE	53.25	80.67	64.15	29.54	85.15	43.86	41.09	59.81	48.71	63.37	79.80	70.64	**87.77**	30.29	45.04
iTransform	58.93	74.59	65.84	77.96	80.30	79.11	49.92	65.03	56.48	57.22	87.94	69.33	68.15	79.44	73.36
Dlinear	57.01	88.51	69.35	76.90	87.02	81.65	42.60	68.10	52.41	56.66	89.37	69.36	74.52	81.01	77.63
Nlinear	56.60	87.66	68.78	76.08	87.85	81.54	53.20	69.14	60.13	56.05	88.98	68.78	72.61	82.59	77.28
PatchTST	56.53	80.80	66.52	78.29	73.20	75.66	56.16	73.76	63.77	60.32	70.41	64.98	80.32	72.04	75.96
GPT4TS	56.88	95.23	71.22	77.02	85.09	80.85	46.55	97.06	62.93	59.94	92.32	72.68	75.22	82.61	78.74
LLMmixer	55.95	97.63	71.13	64.96	96.20	77.55	49.39	**98.08**	65.70	56.27	**98.64**	71.66	76.46	83.57	79.86
CALF	**67.90**	40.41	50.67	78.87	61.23	68.94	51.16	63.98	56.85	63.58	81.68	71.50	71.50	63.58	67.31
Moment	56.20	97.59	71.33	65.93	95.35	77.96	51.60	70.53	59.60	45.34	92.79	60.91	67.81	**94.60**	78.99
Timer	56.61	97.61	71.66	66.76	**96.75**	79.00	46.16	94.53	62.03	57.13	97.69	72.10	75.53	92.61	**83.20**
UniTS	57.57	95.30	71.78	66.82	96.02	78.80	46.02	93.81	61.75	57.78	94.37	71.68	76.43	84.74	80.37
Chronos	58.10	94.67	72.01	76.19	79.13	77.63	49.21	97.57	65.43	58.22	95.57	72.36	72.86	94.25	82.18
Linear-LoCo	58.54	95.60	72.62	**80.48**	86.33	**83.30**	**71.87**	88.72	**79.41**	66.21	95.52	77.17	67.54	79.35	72.97
Kernel-LoCo	57.71	**99.02**	**72.92**	78.28	88.42	83.04	70.66	89.19	78.85	**67.83**	94.50	**78.29**	69.27	92.40	79.18

Table 10: Performance comparison on benchmark datasets (GECCO, SWAN, Creditcard, and CICIDS) in terms of Affiliation Precision (Aff Prec.), Affiliation Recall (Aff Rec.), and Affiliation F1 (Aff F1) score. The **best** result is highlighted in bold, and the second-best result is underlined.

Method	GECCO			SWAN			Creditcard			CICIDS		
	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1	Aff Prec.	Aff Rec.	Aff F1
VAE	29.48	78.01	42.79	**94.30**	13.93	24.27	62.32	80.82	70.37	45.72	3.36	6.27
iTransform	68.18	98.70	80.65	58.47	54.21	56.26	53.99	82.28	65.20	56.46	74.42	64.21
Dlinear	73.95	98.60	84.52	67.70	54.93	60.65	58.01	91.56	71.02	54.77	74.45	63.11
Nlinear	75.03	99.26	85.46	63.01	61.64	62.32	61.25	93.56	74.03	56.13	74.52	64.03
PatchTST	72.55	96.33	82.77	64.04	57.35	60.51	58.16	97.74	72.93	54.40	70.26	61.32
GPT4TS	76.83	97.68	86.01	66.95	63.82	65.35	56.40	99.37	71.96	55.18	70.95	62.08
LLMmixer	86.14	99.69	92.42	69.52	66.53	67.99	56.98	99.39	72.44	54.49	68.59	60.73
CALF	68.73	99.46	81.29	56.30	53.49	54.86	51.18	98.10	67.27	50.55	78.41	61.47
Moment	78.03	99.75	87.57	64.61	62.26	63.42	59.11	98.73	73.94	55.07	73.89	63.11
Timer	82.88	99.57	90.46	68.01	60.15	63.84	59.30	98.84	74.13	55.12	73.46	62.98
UniTS	82.26	99.82	90.20	67.13	59.80	63.25	57.08	**99.41**	72.52	54.73	71.81	62.12
Chronos	82.62	99.47	90.27	67.78	63.81	65.74	60.99	98.47	**75.32**	54.08	68.27	60.35
Linear-LoCo	78.95	**99.96**	88.22	71.26	66.65	68.88	**63.35**	86.70	73.21	54.70	**98.67**	70.38
Kernel-LoCo	**86.92**	99.88	**92.95**	85.18	**68.47**	**75.92**	63.27	88.28	73.71	**57.93**	97.84	**72.77**

the y-axis shows AUROC, and bubble size indicates parameter count, where the outer shell denotes total parameters and the inner dot denotes trainable parameters. Overall, LoCo achieves the highest AUROC on both datasets while requiring the least training time and the smallest parameter budget. Kernel-LoCo trades more time and parameters for higher AUROC, whereas large generative baselines offer no consistent accuracy advantage despite their much heavier cost. These results confirm that LoCo is suited for resource-constrained online deployment.

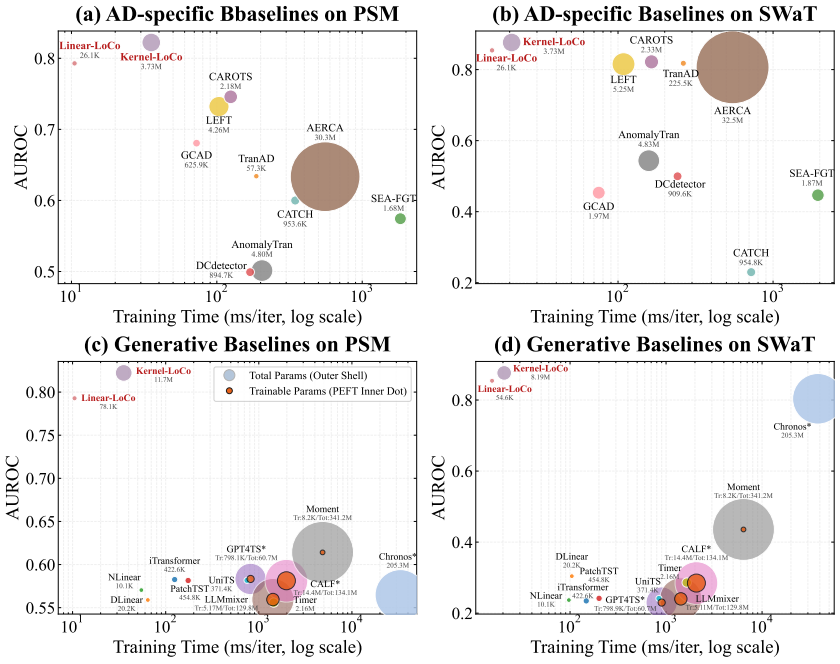


Figure 5: Efficiency, model size, and accuracy trade-off. The x-axis is training time per iteration (ms/iter, log scale), the y-axis is AUROC, and the bubble size indicates the number of parameters. The outer shell denotes total parameters, and the inner dot denotes trainable parameters. Methods marked with $*$ are forecasting-based. The other baselines are reconstruction-based.

H.4 Hyperparameter Sensitivity for AD

We analyze the sensitivity of LoCo to key hyperparameters on PSM and Creditcard. For stride, the x-axis denotes the sliding step as a percentage of the input length. For top-k, it denotes the percentage of retained gradient connections relative to the total number of trainable parameters ($L \times N$). D and σ apply only to Kernel LoCo.

Linear LoCo. Fig. 6 shows the sensitivity of Linear LoCo to L, stride, and top-k. Performance is stable across a wide range of strides, while L around 30–40 and top-k around 10%–30% yield the best AUROC. Both overly small and overly large values lead to a clear drop, indicating that a moderate temporal horizon and a moderately sparse gradient update are optimal.

Kernel LoCo. Fig. 7 reports the sensitivity of Kernel LoCo as Linear LoCo, plus kernel D and σ. Fig 7 reports the sensitivity of Kernel LoCo to L, stride, top-k, D, and σ. The trends are consistent with Linear LoCo, L around 40, stride around 30%–40%, and top-k around 10%–30% are preferable. For the kernel parameters, D around 100–150 and σ around 0.5–1.5 achieve the best AUROC. Overall, Kernel LoCo remains robust within a broad range of reasonable values.

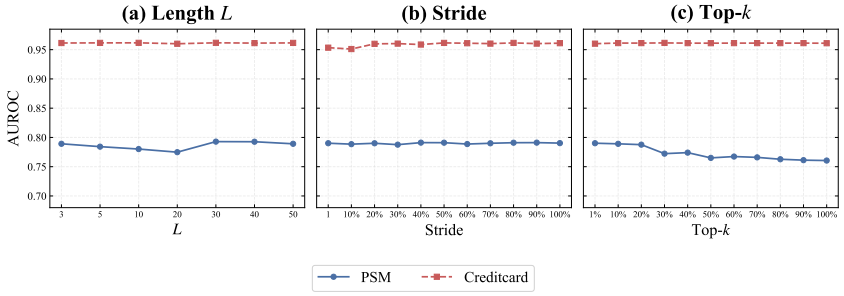


Figure 6: Hyperparameter sensitivity of Linear LoCo on PSM and Creditcard. The x-axis is equally spaced over parameter values. Stride is the sliding step as a percentage of the input length. Top-k is the percentage of retained gradient relative to the total number of trainable parameters ($L \times N$).

H.5 Weight Evolution under Different Top-k

Fig. 8 visualizes the Linear-LoCo weight matrix under five top-k settings (top-k $= 1$, 10%, 20%, 40%, 50%) and four training epochs $(0, 8, 16, 20)$. Colors encode weight values, with red and blue indicating positive and negative values and white near zero.

With top-k $= 1$, the weight matrix remains largely near zero even at epoch 20. Updating only one gradient per iteration is too restrictive to accumulate sufficient information, and most of the parameter space stays inactive. With top-k $= 10\%$, by contrast, clear horizontal banding emerges already at epoch 8. A small number of rows are repeatedly activated and their weight magnitudes grow quickly, while the remaining rows stay dormant, indicating that the gradient mask performs strong implicit feature selection. As top-k increases to 20%, 40%, and 50%, the sharp banding pattern becomes more fragmented, and red and blue pixels are distributed more evenly across the matrix. After the core causal connections are captured, the model gradually spreads updates to some non-causal directions to further reduce the residual. Across all settings, the spatial structure of the weight matrix is largely established between epoch 0 and epoch 8, and subsequent epochs mainly deepen the colors, i.e., increase the magnitude of already-activated weights, rather than changing which coordinates are active.

These observations are consistent with the analysis in Phase 4.2. The gradient hard-thresholding update performs sparse support selection in the early stage of training. Once the core support is selected, later iterations mainly refine the magnitudes on the active set. Increasing top-k does not necessarily increase the number of activated coordinates indefinitely, because non-causal directions have zero expected gradient. However, the model may still spread some updates to non-causal directions to fit the residual, which matches the behavior of competitive learning.

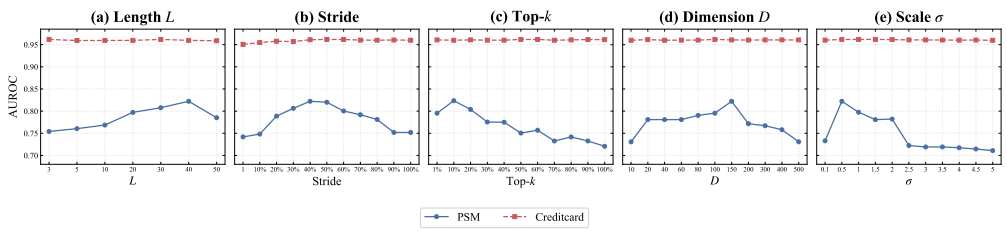


Figure 7: Hyperparameter sensitivity of Kernel LoCo on PSM and Creditcard. The x-axis is equally spaced over parameter values. Stride is the sliding step as a percentage of the input length. Top-k is the percentage of retained gradient connections relative to the total number of trainable parameters ($L \times N$). D and σ apply only to Kernel LoCo.

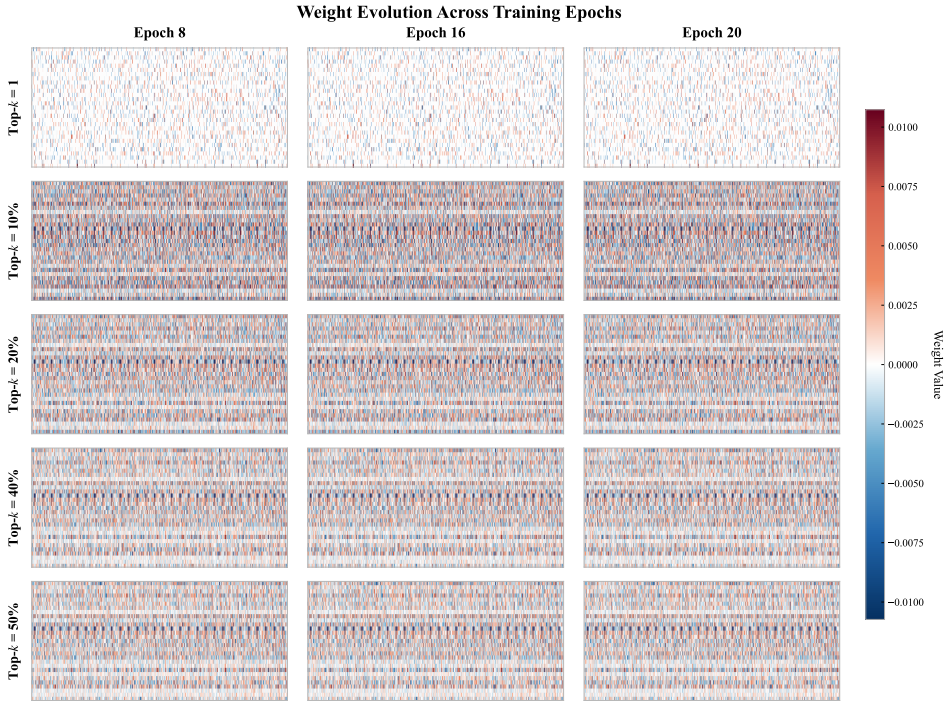


Figure 8: Weight evolution of Linear-LoCo across epochs under different top-k settings. Rows correspond to top-k values and columns to training epochs.

H.6 Convergence Analysis

Fig. 9 shows the training and test loss of Linear-LoCo and Kernel-LoCo under Top-k $= 10\%, 20\%, 40\%$. All configurations converge stably within around 10 epochs. For Linear-LoCo, larger Top-k yields lower final training and test loss. For Kernel-LoCo, a brief increase is followed by rapid convergence and larger Top-k yields lower test loss, while the train–test gap is more pronounced at Top-k $= 10\%$. These results indicate that LoCo remains optimizable under aggressive gradient sparsification, consistent with the GHT support selection in Appendix C and the weight evolution in Appendix H.5.

H.7 Time-Step-Level RCA Results

Table 11 reports $AC^*@K$ and $Avg^*@500$ on SWaT and MSDS. On SWaT, Kernel-LoCo achieves the best $AC^*@1$, and both LoCo variants reach 1.0 at $AC^*@100$ and $AC^*@500$, substantially improving $Avg^*@500$ over the baselines. On MSDS, all methods approach or reach 1.0 at $AC^*@10$ and above, with LoCo matching or slightly exceeding AERCA.

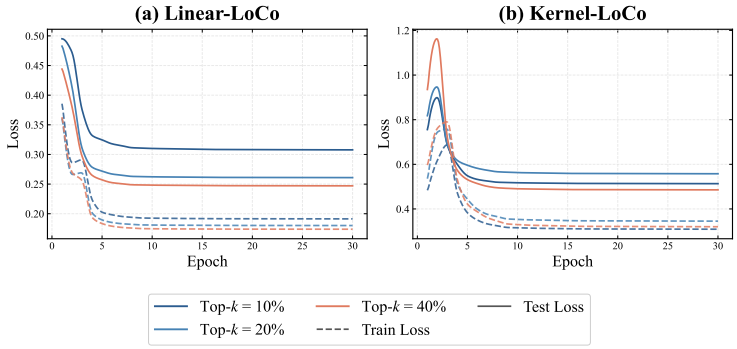


Figure 9: Convergence analysis of Linear-LoCo and Kernel-LoCo under different Top-k ratios. Dashed lines are train loss and solid lines are test loss.

Table 11: Performance comparison on SWaT and MSDS datasets in terms of AC*@k and Avg*@500. The **best** result is highlighted in bold, and the second result is underlined.

Dataset	Method	AC*@1	AC*@10	AC*@100	AC*@500	Avg*@500
SWaT	ϵ-Diagnosis	0.0000	0.0050	0.1825	0.5688	0.3344
	AERCA	0.3775	**0.4560**	0.5370	**1.0000**	0.8003
	Linear-LoCo	0.3500	0.4000	**1.0000**	**1.0000**	0.9561
	Kernel-LoCo	**0.4500**	0.4500	**1.0000**	**1.0000**	**0.9585**
MSDS	ϵ-Diagnosis	0.0465	0.2087	**1.0000**	**1.0000**	0.9243
	AERCA	0.1463	**1.0000**	**1.0000**	**1.0000**	0.9960
	Linear-LoCo	0.2179	**1.0000**	**1.0000**	**1.0000**	0.9968
	Kernel-LoCo	**0.2277**	**1.0000**	**1.0000**	**1.0000**	**0.9968**

H.8 Hyperparameter Sensitivity for RCA

We analyze the sensitivity of LoCo on MSDS for the RCA task, with the stride fixed to 1. Figures 10 and 11 report $AVG@10$ and $AVG^*@500$ under different hyperparameters for Linear LoCo and Kernel LoCo, respectively. For Linear LoCo, performance is best within a moderate range of temporal horizon and top-k, while extreme values lead to a clear drop. For Kernel LoCo, similar trends hold for the shared parameters, and the model remains relatively stable across reasonable values of the kernel-specific parameters. Overall, both variants are robust within a broad range of hyperparameters.

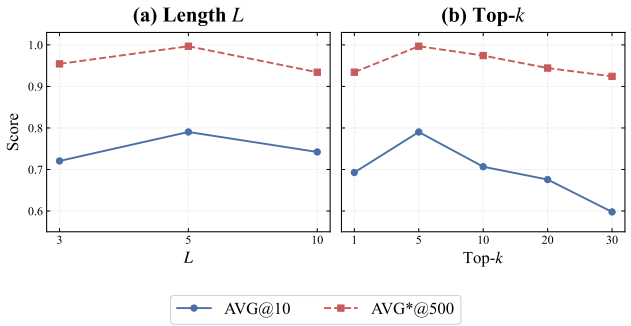


Figure 10: Hyperparameter sensitivity of Linear LoCo for RCA.

H.9 Hyperparameter Sensitivity on Causal Discovery

Fig. 12 reports the sensitivity of LoCo to key hyperparameters on the Confounder and Close subgraph sets, in terms of AUROC and Individual AUROC (InROC). Since the CausalRivers benchmark

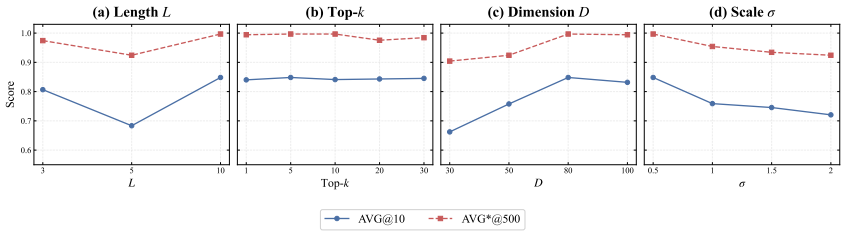


Figure 11: Hyperparameter sensitivity of Kernel LoCo for RCA.

uses three-node subgraphs, the number of candidate connections per output is limited and the effective Top-k range is bounded. When a larger update budget is needed, we increase L to expand the candidate connection space instead. For Linear-LoCo, smaller L and moderate Top-k yield the best results on both datasets. For Kernel-LoCo, the best configuration also appears at small L and moderate Top-k, but performance is more sensitive to D and σ. AUROC and InROC show consistent trends, and Linear-LoCo remains more robust than Kernel-LoCo on this benchmark.

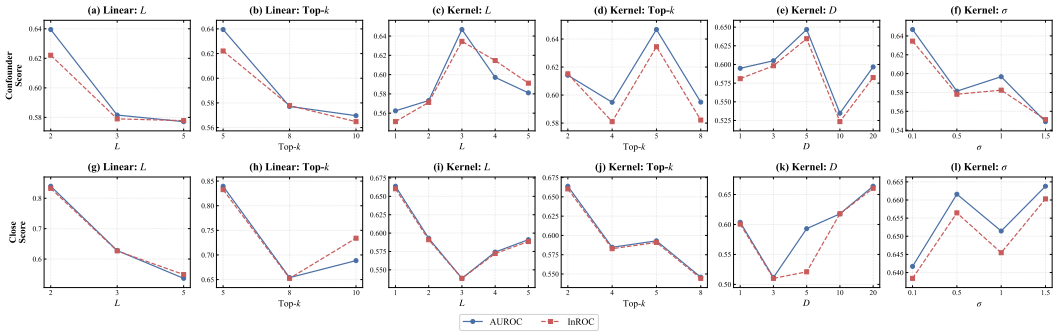


Figure 12: Hyperparameter sensitivity on CausalRivers for Confounder (top) and Close (bottom). Columns show Linear-LoCo (L, Top-k) and Kernel-LoCo (L, Top-k, D, σ).

H.10 Weight Evolution under Different Top-k

Fig. 8 shows the Linear-LoCo weight matrix under five top-k settings (top-k $=$ 1, 10%, 20%, 40%, 50%) and four training epochs $(0, 8, 16, 20)$. Colors encode weight values, with red and blue indicating positive and negative values and white near zero.

With top-k $=$ 1, the weight matrix remains largely near zero even at epoch 20. Updating only one gradient per iteration is too restrictive to accumulate sufficient information, and most of the parameter space stays inactive. With top-k $=$ 10%, by contrast, clear horizontal banding emerges already at epoch 8. A small number of rows are repeatedly activated and their weight magnitudes grow quickly, while the remaining rows stay dormant, indicating that the gradient mask performs strong implicit feature selection. As top-k increases to 20%, 40%, and 50%, the sharp banding pattern becomes more fragmented, and red and blue pixels are distributed more evenly across the matrix. After the core causal connections are captured, the model gradually spreads updates to some non-causal directions to further reduce the residual. Across all settings, the spatial structure of the weight matrix is largely established between epoch 0 and epoch 8 and subsequent epochs mainly deepen the colors, i.e., increase the magnitude of already-activated weights, rather than changing which coordinates are active.

These observations are consistent with the analysis in Phase 4.2. The gradient hard-thresholding update performs sparse support selection in the early stage of training. Once the core support is selected, later iterations mainly refine the magnitudes on the active set. Increasing top-k does not necessarily increase the number of activated coordinates indefinitely, because non-causal directions

have zero expected gradient. However, the model may still spread some updates to non-causal directions to fit the residual, which matches the behavior of LoCo's competitive learning.

H.11 Compressed Weight Matrices under Different Aggregation Strategies

Fig. 13 visualizes the compressed weight matrix of Linear-LoCo on PSM after training to Epoch 20, using three aggregation strategies (MAX, MA, AVG) under different Top-k settings. MAX retains larger magnitudes but is more sensitive to extreme peaks, resulting in noisier patterns with an unclear diagonal. In contrast, MA and AVG produce highly consistent results, revealing a clear main diagonal and a central high-response block, which indicates that both effectively suppress high-frequency noise and highlight the core structure. This supports the default choice of AVG in the method. Across Top-k, the matrix remains nearly empty at Top-k $=$ 1, while the main structure is already established from 10% onward and stays stable up to 50%, suggesting that core gradient information is concentrated in the top 10%–20% updates. This is consistent with the weight evolution in Appendix H.5.



Figure 13: Compressed weight matrices of Linear-LoCo on PSM at Epoch 20 under different aggregation strategies (MAX, MA, AVG) and Top-k settings. Colors encode compressed weight values.

H.12 Complete Output Report for Anomaly #2303

For automated deployment in industrial monitoring platforms, LoCo synthesizes the traceback results into a structured Diagnostic Report. Below is the diagnostic report for Anomaly #2303

Table 12: Industrial Root Cause Analysis Diagnostic Report for Anomaly #2303

Diagnostic Metric	**System Output / Value**
Anomaly Timestamp	$t =$ Anomaly #2303 (sample index: 15518)
Overall Error Index	$\mathrm{SPE}(t) = 6856.9019$
Primary Symptom Variable	FIT401 (Deviation ratio: 29.713; standardized error: -29.713)
Ranked Root Causes	1. (P501, $t-11$) – Attr $= 57.3828$; Cum. $= -0.1157$ 2. (P501, $t-12$) – Attr $= 36.9326$; Cum. $= -0.1172$ 3. (P501, $t-13$) – Attr $= 36.3696$; Cum. $= -0.1178$ 4. (P501, $t-15$) – Attr $= 35.7275$; Cum. $= -0.1201$ 5. (P501, $t-14$) – Attr $= 34.0908$; Cum. $= -0.1190$
Primary Causal Propagation Path	$\mathrm{FIT401}_t \xrightarrow{\text{Attr}=57.3828,\ \text{Cum.}=-0.1157} \mathrm{P501}_{t-11}$

H.13 Case Study Details in SWaT

For anomaly #2303 (sample index 15518), the SPE reaches 6856.90, and the primary symptom is FIT401 with a standardized error of -29.713. The traceback identifies P501 at lags 11–15 as the upstream root cause, with the highest attribution score 57.3828 at lag 11. Fig. 3(b) highlights the FIT401 row and P501 column, indicating a learned causal dependency between them.

Although FIT401 (P4) lies upstream of P501 (P5) in the physical water flow, the control causal direction in Swat differs from the flow direction. As described by Mathur & Tippenhauer (2016), sensors and actuators communicate with dedicated PLCs, implying feedback control. When P501 pumps dechlorinated water to RO, LIT401 decreases and the P4–P5 pressure–flow balance changes. The P4 PLC then adjusts P401/P402, valves, and UV401 via LIT401 and interlock logic, changing FIT401. Thus, P501 indirectly affects FIT401 through level feedback and pressure–flow coupling, not backward flow.

The 11–15-step delay (about 11–15 s at 1 Hz) matches tank buffering, PLC scan, actuator action, flow propagation, and sensor inertia. Fig. 4(c) shows P501 deviating slightly at 15503–15507, while Fig. 4(b) shows FIT401 clearly anomalous at 15518. The later FIT401 symptom thus stems from the earlier, subtler P501 behavior. This case shows that the most deviated variable at alarm time is not necessarily the root cause. Tracing backward along the learned time-lagged causal graph links the later symptom to the earlier upstream perturbation. We note that the traceback identifies P501 as the most likely upstream root cause of the FIT401 anomaly.