DYNAMIC CAUSAL CONTRASTIVE LEARNING FOR MULTIVARIATE TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

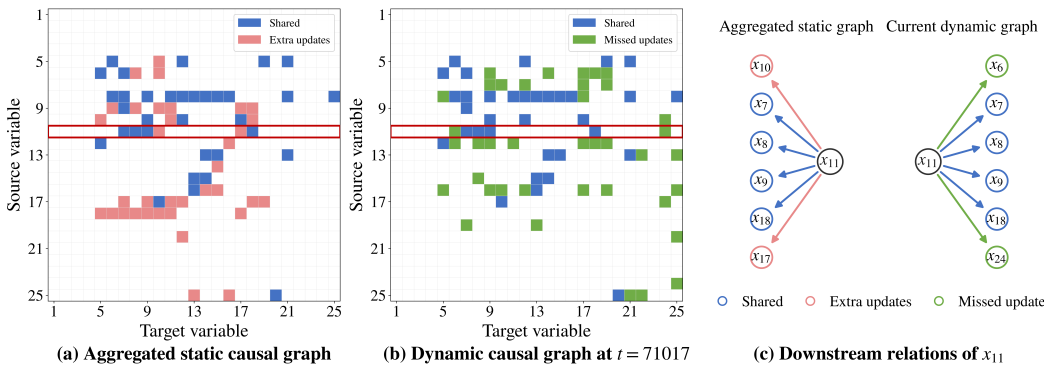
Multivariate time-series anomaly detection requires modeling temporal patterns and inter-variable relationships varying across operating states. Using static causal graphs to guide contrastive augmentation can misalign downstream variable updates in positive samples and overlook changes in outgoing connectivity when selecting negative perturbation targets. We propose DCCAD, a dynamic causal contrastive learning framework that constructs positive and negative samples using the causal relationships estimated for each window. DCCAD learns a shared causal predictor from blocks of normal data and freezes it to estimate dynamic graphs under a consistent construction procedure for training and unseen sequences. Positive augmentation selects sources with nonzero current out-degrees when available and coordinates their perturbations with updates to graph-specified downstream variables. Negative augmentation prioritizes variables with larger current out-degrees and samples a perturbation ratio to vary coverage. These augmentations encourage representations that distinguish variations consistent with the estimated causal relationships from synthetic disruptions of those relationships. During inference, DCCAD combines representation distance and causal prediction error with a graph change score quantifying changes in estimated causal influence and topology relative to a historical reference. Experiments on five real-world datasets and one synthetic dynamical system demonstrate strong detection performance, with ablation studies validating the contribution of each component.

1 INTRODUCTION

Multivariate time-series anomaly detection (MTSAD) plays a critical role in identifying failures and attacks in industrial monitoring and cyber-physical systems (Zhang et al., 2019; Li et al., 2019; Zhou et al., 2019; Su et al., 2019). These systems exhibit complex temporal dependencies and cross-variable relationships that may emerge, disappear, or vary in strength across operating conditions (Deng & Hooi, 2021; Zhao et al., 2020; Huang et al., 2015; 2019; 2020). Because these relationships can vary even during normal operation, dependency changes alone do not necessarily indicate anomalies. Effective MTSAD approaches should therefore capture normal temporal dynamics while assessing deviations in the context of inter-variable relationships in the current operating state.

Many MTSAD methods learn normal behavior through reconstruction or prediction and use resulting errors as key anomaly detection signals (Audibert et al., 2020; Xu et al., 2022; Wu et al., 2023; Li et al., 2025). Although these methods capture temporal patterns and cross-variable dependencies, reconstruction and prediction objectives alone do not explicitly encourage representations to distinguish normal variations from potential anomalies. Contrastive learning provides another route by contrasting positive and negative samples to learn discriminative representations (Chen et al., 2020; Sehwag et al., 2021; Tack et al., 2020; Darban et al., 2025). For anomaly detection, however, its effectiveness depends on sample construction: positive augmentations should preserve the relationships characterizing normal behavior, whereas negative augmentations should introduce meaningful departures from them. CAROTS incorporates causal discovery into this process: it perturbs source variables and uses a learned causal predictor to update their affected downstream variables, constructing causality-preserving positives, while perturbing causally related variable subsets to generate causality-disrupted negatives (Kim et al., 2025). Here, causal relationships refer to Granger-causal predictive dependencies rather than interventional or physical causality (Granger, 1969). By linking sample construction to estimated causal relationships, this design makes the correspondence

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.



(a) Aggregated static causal graph **(b) Dynamic causal graph at $t = 71017$** **(c) Downstream relations of x_{11}**

Figure 1: Comparison of estimated causal structures and downstream sets on PSM. (a) Temporally aggregated static binary causal graph. (b) Dynamic binary causal graph at the endpoint of the selected window. (c) Direct downstream sets of source variable x_{11} in the two graphs.

between the causal graph and the dependencies in each window important for constructing meaningful contrastive samples.

However, CAROTS uses a single causal graph learned from normal training data, so its graph-guided augmentation cannot account for changes in inter-variable relationships across windows. A single graph may retain relationships associated with other operating stages while weakening those specific to the current stage. Figure 1 illustrates this potential mismatch by comparing a temporally aggregated static graph with the estimated dynamic graph for a selected PSM window. For the same source variable, the temporally aggregated static graph includes downstream nodes absent from the current dynamic graph and omits nodes present in it. Relative to the current estimate, these differences imply *extra* or *missed* downstream updates during positive augmentation, potentially reducing the consistency of the augmented sample with the dependencies in the current window. Since such samples are treated as positives during contrastive learning, this mismatch may weaken the intended supervision that positive augmentation should represent variations consistent with normal dependencies. Additionally, CAROTS selects source variables for positive augmentation from the full variable set, potentially including zero-out-degree nodes. Perturbing these nodes contributes no downstream update targets under the estimated graph, limiting the opportunity to construct coordinated source-downstream variations. Meanwhile, CAROTS constructs negatives by perturbing causally related subsets selected through depth-first search. Although these subsets can vary in size, their selection relies on a static graph and does not reflect changes in outgoing connectivity across windows. *These observations motivate using the causal relationships estimated for each window as the structural reference for contrastive augmentation. Positive augmentation should select source variables with downstream targets and coordinate source perturbations with downstream updates according to this reference. Negative augmentation should perturb variables selected according to the current structure with varying coverage to introduce violations of the estimated causal relationships.*

To address these challenges, we propose **DCCAD**, a **D**ynamic **C**ausal **C**ontrastive learning framework for multivariate time-series **A**nomaly **D**etection. DCCAD conditions sample construction on the causal relationships estimated for each window, aligning source selection, downstream updates, and negative perturbation targets with the current estimated causal structure. We develop a graph estimation procedure that builds on perturbation-based causal estimation (Bi et al., 2025), using a predictor trained on normal data and fixed estimation criteria to obtain window-specific graphs for training and inference. During training, Causal-Dynamics-Aware Positive Augmentation (**CDA-PA**) selects source variables with nonzero out-degrees in the current estimated causal graph, perturbs them, and uses the causal predictor to update their graph-specified direct downstream variables. Causal-Difficulty-Aware Negative Augmentation (**CDA-NA**) prioritizes variables with larger current out-degrees and samples a perturbation ratio to vary the perturbation coverage. These augmentations encourage representations that distinguish variations consistent with the estimated causal relationships from synthetic disruptions of those relationships. During inference, DCCAD measures changes in estimated causal influence and graph topology relative to a historical reference. The graph change score provides complementary relational evidence alongside representation distance and causal prediction error. The **contributions** of this paper are summarized as follows.

- We propose DCCAD, a dynamic causal contrastive learning framework for MTSAD that constructs positive and negative samples according to the causal relationships estimated for each window. CDA-PA selects sources with nonzero current out-degrees when available and updates their graph-specified downstream variables, while CDA-NA combines current out-degree-based variable selection with varying perturbation coverage.
- We introduce a graph change score that jointly quantifies changes in estimated causal influence and graph topology relative to a historical reference. This score provides relational evidence that complements representation distance and causal prediction error in anomaly detection.
- Extensive experiments on five real-world datasets and one synthetic dynamical system demonstrate competitive detection performance. Ablation studies and further analyses support the effectiveness of dynamic graph guidance and the proposed augmentation and scoring components.

2 Related Work

Deep Learning-Based MTSAD. Many deep MTSAD methods learn normal temporal patterns and inter-variable dependencies, scoring anomalies through reconstruction, prediction, or association discrepancies. Representative approaches use adversarially trained autoencoders (Audibert et al., 2020), combine reconstruction errors with association discrepancy (Xu et al., 2022), or model multi-period and cross-scale patterns (Wu et al., 2023; Li et al., 2025). GCAD measures deviations of dynamic Granger-causal graphs from normal causal patterns (Liu et al., 2025), whereas DCCAD uses dynamic causal graphs for both anomaly scoring and constructing contrastive training samples.

Contrastive Learning-Based MTSAD. Sample construction is central to contrastive representation learning. SimCLR constructs positive pairs from augmented views (Chen et al., 2020), while SSD and CSI use self-supervised representations for anomaly detection (Sehwag et al., 2021; Tack et al., 2020). CARLA uses synthetic anomalies as negatives for time-series contrastive learning (Darban et al., 2025). These approaches do not explicitly use inter-variable causal relationships to guide augmentation. CAROTS constructs causality-preserving positives and causality-disrupting negatives (Kim et al., 2025), but uses a fixed graph to determine downstream update targets and may select zero-out-degree sources. DCCAD selects sources with nonzero current out-degrees when available, updates their graph-specified downstream variables, and constructs negatives using current out-degree rankings and varying perturbation coverage.

Causal Discovery. Granger causality provides a basis for discovering directed predictive dependencies in time series (Granger, 1969; Assaad et al., 2022). Existing approaches infer lagged relationships through conditional independence tests, neural prediction, or differentiable structure learning (Runge et al., 2019; Tank et al., 2022; Pamfil et al., 2020). A single graph summarizes relationships across a sequence but cannot resolve their temporal variation. Non-stationary causal discovery instead examines time- or regime-dependent relationships (Huang et al., 2019; Saggioro et al., 2020). UnCLe estimates time-resolved causal influences from perturbation-induced changes in pointwise prediction errors (Bi et al., 2025). DCCAD modifies its training and graph construction procedures for inductive MTSAD, combining shared training across normal-data blocks with frozen-model graph estimation and a threshold calibrated on normal training data.

3 Methodology

3.1 Problem Definition

Let $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T \times N}$ denote an N-variable time series of length T, where $\boldsymbol{x}_t \in \mathbb{R}^N$ is the observation at time t. A length-w sliding window ending at time t is denoted by $X_t = [\boldsymbol{x}_{t-w+1}, \ldots, \boldsymbol{x}_t]^\top \in \mathbb{R}^{w \times N}$. Given the window X_t ending at time t, the task is to compute an anomaly score $S_t \in \mathbb{R}$ for the observation $\boldsymbol{x}_t$. A larger S_t indicates greater abnormality at time t.

3.2 Framework Overview

As illustrated in Figure 2, DCCAD integrates dynamic causal discovery, causality-aware contrastive learning, and anomaly scoring. Perturbation-based analysis of the causal predictor yields weighted graphs, which are thresholded into binary graphs. Each window X_t uses the graphs at its endpoint

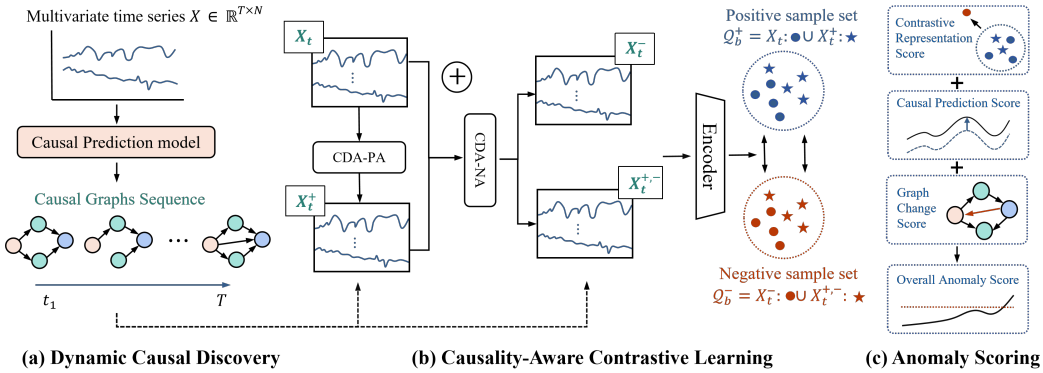


Figure 2: The framework of DCCAD. (a) Dynamic causal graphs are estimated through perturbation-based analysis of a causal prediction model. (b) These graphs guide CDA-PA and CDA-NA to construct positive and negative samples for contrastive representation learning. (c) Representation distance, causal prediction error, and graph change are combined to compute the anomaly score at each time step.

to guide augmentation. CDA-PA selects sources with nonzero out-degrees and coordinates their perturbations with updates to their direct downstream variables. CDA-NA prioritizes variables with larger current out-degrees and samples a perturbation ratio to vary coverage. The resulting samples support contrastive representation learning. During inference, DCCAD measures changes in causal influence and topology relative to a historical reference, then combines the graph change score with representation distance and causal prediction error to obtain the anomaly score S_t at time t.

3.3 Dynamic Causal Discovery

DCCAD estimates time-varying causal graphs for window-specific augmentation and relational anomaly scoring through perturbation-based analysis (Bi et al., 2025). The predictor and binarization threshold remain fixed during inference, providing consistent estimation criteria while allowing graphs to vary with observations.

Causal Prediction Model. Under Granger causality, x_j Granger-causes x_i if its history improves the prediction of x_i beyond the histories of the remaining variables (Granger, 1969; Shojaie & Fox, 2022). The Uncoupler extracts variable-specific temporal representations, learnable dependency matrices model inter-variable interactions, and the Recoupler maps predicted representations to observations. For long multivariate time series, we partition normal training sequences into non-overlapping blocks of length s and jointly train a shared predictor f_{θ_c} across blocks. After selection on normal validation data, the predictor is frozen for causal graph estimation and anomaly detection.

Dynamic Causal Graph Construction. For each source variable x_j, we randomly permute its observations preceding t within the current causal-inference block, leaving other variables unchanged. Denoting the original and perturbed histories by $X_{<t}$ and $X_{<t}^{\setminus j}$, the frozen predictor produces

$$\widehat{\boldsymbol{x}}_t = f_{\theta_c}(X_{<t}), \qquad \widehat{\boldsymbol{x}}_t^{\setminus j} = f_{\theta_c}(X_{<t}^{\setminus j}). \tag{1}$$

For target variable x_i, the original and perturbed prediction errors are $\epsilon_{i,t} = (\widehat{x}_{i,t} - x_{i,t})^2$ and $\epsilon_{i,t}^{\setminus j} = (\widehat{x}_{i,t}^{\setminus j} - x_{i,t})^2$, respectively. We quantify the estimated predictive influence from x_j to x_i at time t as

$$(A_t^{\text{raw}})_{j,i} = \max\left(0, \epsilon_{i,t}^{\setminus j} - \epsilon_{i,t}\right), \tag{2}$$

where rows and columns index source and target variables, respectively. Repeating this procedure across variables and time steps yields pointwise weighted graphs. We reduce short-term fluctuations using a backward moving average over current and past estimates, obtaining $\mathcal{A} = \{A_t\}_{t=1}^T$. Each graph is then binarized using a global quantile threshold δ, calibrated solely on normal training data and fixed during inference:

$$(B_t)_{j,i} = \mathbb{I}[(A_t)_{j,i} > \delta], \qquad \mathcal{B} = \{B_t\}_{t=1}^T. \tag{3}$$

For X_t, the endpoint graphs A_t and B_t represent its current estimated causal structure (evaluated in Appendix C.3); B_t determines augmentation targets, while both support graph-change scoring.

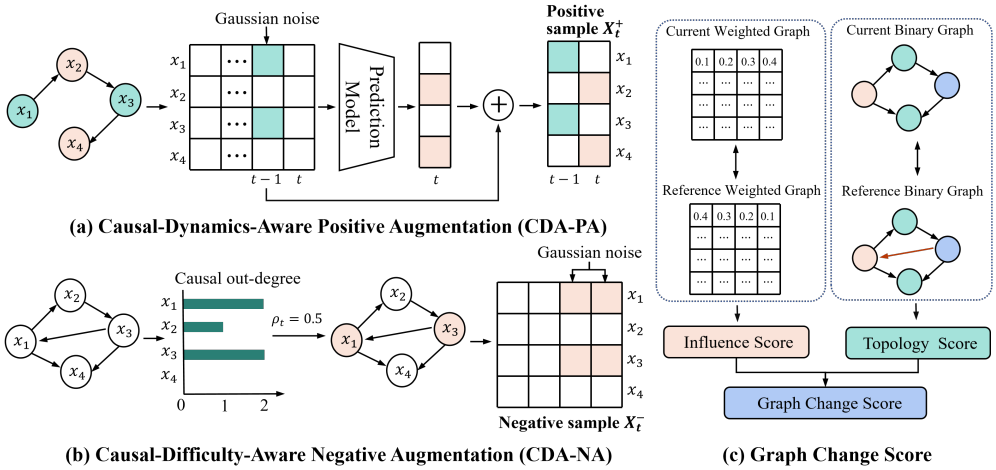


Figure 3: (a) CDA-PA perturbs sources with nonzero out-degrees and updates their current direct downstream variables using the causal predictor. (b) CDA-NA selects variables by current out-degrees and varies perturbation coverage to construct negatives. (c) The graph change score compares current and reference causal graphs in terms of estimated causal influence and topology.

3.4 Dynamic Causal Graph-Guided Data Augmentation

Following the paradigm of causality-preserving positive and causality-disrupting negative augmentation (Kim et al., 2025), DCCAD uses current causal graphs to select augmentation targets and varies negative perturbation coverage to encourage diversity in sample difficulty.

Causal-Dynamics-Aware Positive Augmentation (CDA-PA).

As illustrated in Figure 3(a), for a window X_t, let $d_{t,j}^{\text{out}} = \sum_{i=1}^{N} (B_t)_{j,i}$ for $i \neq j$, and define the candidate source set as $\mathcal{V}_t^{\text{out}} = \{j \mid d_{t,j}^{\text{out}} > 0\}$. When $\mathcal{V}_t^{\text{out}} \neq \emptyset$, CDA-PA selects $\mathcal{S}_t \subseteq \mathcal{V}_t^{\text{out}}$; otherwise, $\mathcal{S}_t$ contains one variable sampled uniformly from all variables (specified in Appendix B.3). Its direct downstream set is $\mathcal{C}_t = \{i \mid \exists j \in \mathcal{S}_t : j \neq i, (B_t)_{j,i} = 1\}$, ensuring downstream targets whenever available. Gaussian noise $\epsilon^+ \sim \mathcal{N}(0, \sigma^2)$ is then added to the variables in $\mathcal{S}_t$ at time $t-1$, yielding the perturbed history $\widetilde{X}_{<t}$. The frozen causal predictor produces $\hat{x}_t^+ = f_{\theta_c}(\widetilde{X}_{<t})$. For $r \in \{t-w+1, \ldots, t\}$, the entries of the positive window X_t^+ are

$$x_{j,r}^+ = \begin{cases} x_{j,r} + \epsilon^+, & r = t-1,\ j \in \mathcal{S}_t, \\ \hat{x}_{j,t}^+, & r = t,\ j \in \mathcal{C}_t, \\ x_{j,r}, & \text{otherwise.} \end{cases} \tag{4}$$

This retains source perturbations and restricts predicted downstream updates to B_t, avoiding extra or missed targets relative to the estimated graph.

Causal-Difficulty-Aware Negative Augmentation (CDA-NA).

As illustrated in Figure 3(b), CDA-NA aims to diversify negative sample difficulty by varying perturbation coverage. Let $M_t = \sum_{j=1}^{N} \mathbb{I}[d_{t,j}^{\text{out}} > 0]$ be the number of variables with nonzero out-degrees. For $M_t > 0$, a variable perturbation ratio ρ_t is sampled from $[\rho_{\min}, \rho_{\max}]$, and $K_t = \lceil \rho_t M_t \rceil$. Using the K_t-th largest nonzero out-degree as the threshold γ_t, CDA-NA selects $\mathcal{D}_t = \{j \mid d_{t,j}^{\text{out}} \geq \gamma_t\}$, including all variables tied at the threshold. This prioritizes variables with more direct downstream connections in the current graph. Varying ρ_t changes the selection cutoff and hence the perturbation coverage, although ties can make $|\mathcal{D}_t|$ exceed K_t.

CDA-NA randomly selects $\text{round}(pw)$ time indices from $\{t-w+1, \ldots, t\}$ to form $\mathcal{T}_t$, where p is the temporal coverage ratio. Gaussian noise scaled by ξ is injected at $\mathcal{D}_t \times \mathcal{T}_t$, giving

$$x_{j,r}^- = \begin{cases} x_{j,r} + \xi \epsilon_{j,r}^-, & (j,r) \in \mathcal{D}_t \times \mathcal{T}_t, \\ x_{j,r}, & \text{otherwise,} \end{cases} \tag{5}$$

where $\epsilon_{j,r}^{-} \sim \mathcal{N}(0,1)$. Applying the same negative augmentation to the original sample X_t and the positive sample X_t^{+} yields X_t^{-} and $X_t^{+,-}$, respectively, for contrastive representation learning.

3.5 Causality-Aware Contrastive Learning

For mini-batch b, original windows X_t and CDA-PA positives X_t^{+} form $\mathcal{Q}_b^{+}$, while CDA-NA negatives X_t^{-} and $X_t^{+,-}$ form $\mathcal{Q}_b^{-}$. The encoder h_θ and projection head g_ϕ produce representations that are ℓ_2-normalized to compute cosine similarity $\mathrm{sim}_{i,j}$. For each anchor $i \in \mathcal{Q}_b^{+}$, $\mathcal{P}(i) \subseteq \mathcal{Q}_b^{+} \setminus \{i\}$ includes its paired view of the same window and other candidates with similarity above α. Increasing α linearly during training tightens positive selection. All samples in $\mathcal{Q}_b^{-}$ serve as negatives. The contrastive objective is

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{|\mathcal{Q}_b^{+}|} \sum_{i \in \mathcal{Q}_b^{+}} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp\left(\mathrm{sim}_{i,j}/\tau\right)}{\sum_{a \in \mathcal{P}(i)} \exp\left(\mathrm{sim}_{i,a}/\tau\right) + \sum_{k \in \mathcal{Q}_b^{-}} \exp\left(\mathrm{sim}_{i,k}/\tau\right)}, \tag{6}$$

where τ is the temperature parameter.

3.6 Anomaly Scoring

DCCAD evaluates anomalies using three complementary signals: representation distance, causal prediction error, and graph change. They capture deviations from normal representations, prediction patterns, and inter-variable relationships, respectively. The three scores are computed separately, standardized using training-set statistics, and fused into the final anomaly score.

Contrastive Representation Score. The normal representation center $\boldsymbol{c}$ is the mean of the unnormalized projected representations of normal training windows and their CDA-PA positive samples. For a test window X_t, let $\boldsymbol{z}_t = g_\phi(h_\theta(X_t))$ denote its unnormalized projected representation. The representation score assigned to time t is

$$S_t^{\mathrm{con}} = \left\|\boldsymbol{z}_t - \boldsymbol{c}\right\|_2. \tag{7}$$

Causal Prediction Score. The frozen causal predictor estimates the current observation from its history as $\widehat{\boldsymbol{x}}_t = f_{\theta_c}(X_{<t})$. The prediction score is the mean squared error across all variables:

$$S_t^{\mathrm{pred}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{x}_{i,t} - x_{i,t}\right)^2. \tag{8}$$

Graph Change Score. In Figure 3(c), the graph change score measures changes in estimated causal influence and topology relative to a historical reference. At t, the reference weighted graph is $A_t^{\mathrm{ref}} = H^{-1} \sum_{h=1}^{H} A_{t-h}$. The influence change is measured by the relative Frobenius distance:

$$S_t^{\mathrm{inf}} = \frac{\left\|A_t - A_t^{\mathrm{ref}}\right\|_F}{\left\|A_t^{\mathrm{ref}}\right\|_F + \varepsilon}, \tag{9}$$

where $\varepsilon > 0$ ensures numerical stability. The reference graph is binarized using the threshold δ from dynamic graph construction: $(B_t^{\mathrm{ref}})_{j,i} = \mathbb{I}[(A_t^{\mathrm{ref}})_{j,i} > \delta]$. Let $\mathcal{E}_t$ and $\mathcal{E}_t^{\mathrm{ref}}$ denote the edge sets of B_t and B_t^{ref}, respectively, excluding self-loops. Topology change is measured by the Jaccard distance:

$$S_t^{\mathrm{top}} = 1 - \frac{\left|\mathcal{E}_t \cap \mathcal{E}_t^{\mathrm{ref}}\right|}{\left|\mathcal{E}_t \cup \mathcal{E}_t^{\mathrm{ref}}\right|}. \tag{10}$$

When both edge sets are empty, $S_t^{\mathrm{top}} = 0$. We combine the two measures into a pointwise score and assign the maximum over the recent w time steps to time t:

$$S_t^{\mathrm{gc,pt}} = \beta S_t^{\mathrm{inf}} + (1-\beta) S_t^{\mathrm{top}}, \qquad S_t^{\mathrm{gc}} = \max_{r=t-w+1,\ldots,t} S_r^{\mathrm{gc,pt}}, \tag{11}$$

where $\beta \in [0,1]$ balances causal influence and topology changes.

Table 1: Mean performance across random seeds on real-world datasets (standard deviations in Appendix C.4). Best and second-best results are bold and underlined, respectively.

Dataset	Metric	AT	TN	USAD	CrossAD	GCAD	SimCLR	SSD	CSI	CARLA	CAROTS	DCCAD
MSL_P-14	AUROC	0.4880	0.7324	0.7350	0.7541	**0.8185**	0.7512	0.6923	0.5377	0.5286	0.7642	0.7831
	AUPRC	0.0326	0.3085	0.3086	0.2573	0.4427	0.3118	0.1443	0.0453	0.2720	0.3721	**0.4459**
	F1	0.0605	0.5018	0.5018	0.4906	0.4697	0.4603	0.1594	0.1326	0.3887	**0.5448**	0.5324
MSL_P-15	AUROC	0.4939	0.5748	0.5928	0.6718	0.7264	0.5590	0.4930	0.5963	0.5363	0.6911	**0.7735**
	AUPRC	0.0053	0.0169	0.0203	0.0202	0.1949	0.0174	0.0105	0.0211	0.1225	0.0213	**0.3738**
	F1	0.0209	0.0739	0.0908	0.0688	0.4787	0.0787	0.0279	0.0810	0.0431	0.0863	**0.5502**
SMD_3-6	AUROC	0.5361	0.9691	0.9256	0.9490	0.9199	0.9549	0.5445	0.9492	0.6312	0.9201	**0.9696**
	AUPRC	0.3105	**0.7274**	0.4978	0.5982	0.5368	0.5283	0.1535	0.4797	0.4949	0.4657	0.6180
	F1	0.1315	**0.6679**	0.5432	0.5847	0.5317	0.5097	0.2526	0.4796	0.3828	0.4345	0.6317
SMD_3-11	AUROC	0.5179	**0.9123**	0.8815	0.8859	0.8869	0.7950	0.4751	0.7878	0.6323	0.8724	0.9000
	AUPRC	0.0255	0.4121	**0.4382**	0.3542	0.2745	0.0452	0.0069	0.0286	0.2873	0.2378	0.2897
	F1	0.0426	0.4457	**0.5534**	0.4191	0.3293	0.0943	0.0203	0.0639	0.1596	0.4473	0.4441
PSM	AUROC	0.5001	0.7425	0.5836	0.6434	0.7019	0.7278	0.6222	0.6908	0.4820	0.7278	**0.7506**
	AUPRC	**0.6423**	0.5649	0.4341	0.4406	0.5424	0.4800	0.3821	0.4779	0.5156	0.5301	0.5900
	F1	0.4431	0.5576	0.4729	0.4604	0.5118	0.5741	0.4773	0.5396	0.4448	0.5304	**0.5893**
SWaT	AUROC	0.4588	0.8124	0.7908	0.8382	0.5946	0.3680	0.7695	0.6413	0.5000	0.8512	**0.8694**
	AUPRC	0.2711	0.7193	0.6885	0.4182	0.1637	0.0899	0.3857	0.2227	0.3747	**0.7599**	0.7496
	F1	0.2176	0.7707	0.7338	0.6118	0.2782	0.2467	0.4654	0.3470	0.4518	0.7866	**0.7903**
WADI	AUROC	0.4585	0.4956	0.4865	0.4465	0.5134	0.4957	0.4078	0.4895	0.4572	0.4879	**0.6785**
	AUPRC	0.0466	0.0547	0.0655	0.0478	0.1097	0.0846	0.0753	0.0529	**0.2905**	0.0545	0.2373
	F1	0.1156	0.1376	0.1296	0.1116	0.1176	0.1829	0.1147	0.1277	0.1156	0.1339	**0.3236**
Average AUROC		0.4933	0.7484	0.7137	0.7413	0.7374	0.6645	0.5721	0.6704	0.5382	0.7592	**0.8178**

Overall Anomaly Score. We standardize each component using the mean and standard deviation of its normal training scores, then combine them as

$$S_t = \widetilde{S}_t^{\mathrm{con}} + \widetilde{S}_t^{\mathrm{pred}} + \lambda_{\mathrm{gc}} \widetilde{S}_t^{\mathrm{gc}}. \tag{12}$$

Here, tildes denote standardized scores and λ_{gc} weights graph change. A larger S_t indicates greater abnormality at time t.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate DCCAD on five widely used real-world MTSAD datasets: MSL (Hundman et al., 2018), SMD (Su et al., 2019), PSM (Abdulaal et al., 2021), SWaT (Goh et al., 2017), and WADI (Ahmed et al., 2017). For MSL and SMD, the main comparisons use selected subsets with differing anomaly ratios and stationarity characteristics. We also use the Lorenz-96 system (Lorenz, 1996) to generate synthetic data at two anomaly strengths, yielding two detection difficulty settings.

Baselines. Non-contrastive baselines include Anomaly Transformer (AT) (Xu et al., 2022), TimesNet (TN) (Wu et al., 2023), USAD (Audibert et al., 2020), CrossAD (Li et al., 2025), and GCAD (Liu et al., 2025). Contrastive baselines include SimCLR (Chen et al., 2020), SSD (Sehwag et al., 2021), CSI (Tack et al., 2020), CARLA (Darban et al., 2025), and CAROTS (Kim et al., 2025). Performance is evaluated using AUROC, AUPRC, and Best-F1, denoted as F1 throughout the paper.

Implementation. Each normal training sequence is divided into 80% for optimization and 20% for validation. We use an LSTM encoder with a window length of $w = 10$. Unless otherwise specified, the historical reference length, contrastive temperature, and negative perturbation intensity are set to $H = 10$, $\tau = 0.1$, and $\xi = 1.0$, respectively, and the graph change score weight is set to $\lambda_{\mathrm{gc}} = 0.1$. Additional implementation and hyperparameter details are provided in Appendix B.

4.2 Main Results

Results on Real-World Datasets. Table 1 reports performance on the real-world datasets. DCCAD achieves the highest average AUROC of 0.8178 and ranks first or second in 19 of the 21 dataset-

Table 2: Detection performance (AUROC) on Lorenz-96 synthetic datasets. Lorenz#1 and Lorenz#2 denote low and high difficulty settings, respectively. PG, PC, CT, CG, and CS represent point global, point contextual, collective trend, collective global, and causal shift anomalies, respectively; Avg. denotes their average.

	Lorenz#1						Lorenz#2					
Model	PG	PC	CT	CG	CS	Avg.	PG	PC	CT	CG	CS	Avg.
AT	0.6790	0.5588	0.7174	0.4797	0.4381	0.5746	0.5659	0.5168	0.6269	0.4737	0.4484	0.5263
TN	0.9939	0.9574	0.9156	0.5652	0.6084	0.8081	0.8500	0.7632	0.8199	0.5544	0.5677	0.7110
USAD	0.8869	0.7891	0.9137	0.5119	0.3748	0.6953	0.5023	0.4948	0.7942	0.4978	0.4362	0.5451
CrossAD	0.8580	0.7232	**0.9706**	0.5152	0.3851	0.6904	0.5006	0.4637	0.7955	0.5017	0.4399	0.5403
GCAD	0.9936	0.9522	0.6291	0.5459	0.5629	0.7367	0.8855	0.7712	0.5846	0.5479	0.5450	0.6668
SimCLR	0.4624	0.5708	0.7926	0.5515	0.5034	0.5761	0.3989	0.5049	0.6017	0.5469	0.5271	0.5159
SSD	0.4749	0.5603	0.6296	0.5118	0.5035	0.5360	0.4562	0.5115	0.4912	0.5127	0.5076	0.4958
CSI	0.5250	0.5307	0.7589	0.5155	0.4854	0.5631	0.4711	0.4864	0.6110	0.5113	0.4994	0.5158
CARLA	0.4990	0.4860	0.6878	0.4529	0.4374	0.5126	0.4958	0.4889	0.6031	0.4523	0.4461	0.4972
CAROTS	0.9991	0.9856	0.9176	0.5435	0.6863	0.8264	0.9679	0.8673	0.8316	0.5253	0.5818	0.7548
DCCAD	**0.9995**	**0.9960**	0.9221	**0.5841**	**0.7270**	**0.8457**	**0.9789**	**0.9141**	**0.8598**	**0.5850**	**0.6679**	**0.8011**

Table 3: Ablation of graph and augmentation strategies on PSM. All/Nonzero indicate source selection from all variables/nonzero-out-degree variables, while DFS/Degree indicate selecting variables through depth-first search/according to causal out-degree, respectively.

Graph	CDA-PA	CDA-NA	AUROC	AUPRC	F1
Static	All	DFS	0.7278	0.5301	0.5304
Dynamic	All	DFS	0.7293	0.5523	0.5332
Dynamic	Nonzero	DFS	0.7325	0.5524	0.5359
Dynamic	Nonzero	Degree	0.7453	0.5822	0.5789

Table 4: Ablation of anomaly score components on PSM. PS and GCS denote causal prediction score and graph change score, respectively. The last row corresponds to the complete DCCAD.

PS	GCS	AUROC	AUPRC	F1
×	×	0.7378	0.5672	0.5640
×	✓	0.7394	0.5663	0.5649
✓	×	0.7453	0.5822	0.5789
✓	✓	0.7506	0.5900	0.5893

metric combinations. It leads all three metrics on MSL_P-15 and achieves the best AUROC and F1 on PSM, SWaT, and WADI, while remaining competitive on the SMD subsets. The improvement on MSL_P-15 is notable because DCCAD achieves an AUPRC of 0.3738 despite an anomaly ratio of only 0.74%, indicating improved ranking of sparse anomalies under severe class imbalance. These results support DCCAD's overall effectiveness. Appendix C.7 compares DCCAD with CAROTS across all MSL and SMD entities.

Results on Synthetic Datasets. Table 2 reports AUROC results under two Lorenz-96 difficulty settings. DCCAD achieves the highest average AUROC of 0.8457 on Lorenz#1 and 0.8011 on Lorenz#2. Under both settings, DCCAD performs best on PG, PC, and CS and performs best on CT in the more challenging Lorenz#2. DCCAD attains the highest CG AUROC, but the modest absolute values indicate that detecting this anomaly remains difficult. For CS anomalies involving changes in the underlying causal mechanism, DCCAD achieves AUROC values of 0.7270 and 0.6679, outperforming the strongest corresponding baselines by 0.0407 and 0.0861, respectively. These consistent improvements suggest DCCAD is more sensitive to causal shifts than the evaluated baselines.

4.3 Model Analysis

Ablation of Graph and Augmentation Strategies. Table 3 progressively replaces CAROTS's graph and augmentation strategies under a fixed scoring configuration. Replacing static causal graphs with dynamic ones increases AUPRC from 0.5301 to 0.5523, showing that window-specific causal relationships provide better augmentation guidance. Restricting CDA-PA sources to variables with nonzero out-degrees avoids selecting variables that cannot trigger downstream updates. For CDA-NA, replacing DFS-based with degree-based selection increases F1 from 0.5359 to 0.5789, indicating that prioritizing variables with broader causal influence generates more informative negatives. The first row represents CAROTS, while the last represents DCCAD without GCS.

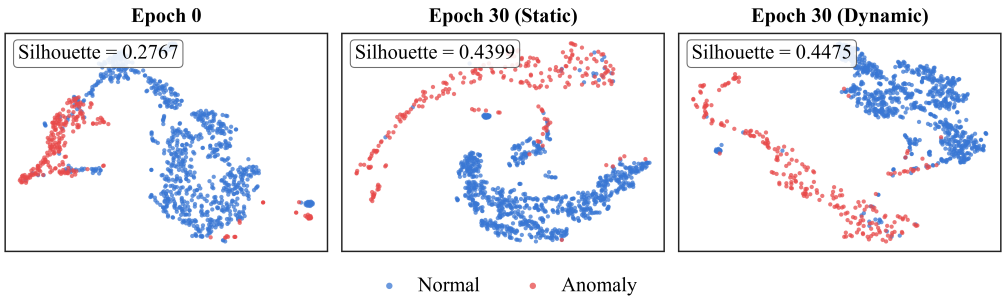


Figure 4: t-SNE visualizations of learned representations and the corresponding silhouette scores.

Ablation of Anomaly Score Components. Table 4 evaluates PS and GCS while retaining the contrastive representation score. Adding GCS to the representation score yields only marginal AUROC and F1 gains, with a slight decrease in AUPRC. Adding PS improves AUPRC from 0.5672 to 0.5822 and F1 from 0.5640 to 0.5789, showing that prediction errors capture deviations from normal causal dynamics. Combining PS with GCS further raises F1 to 0.5893, demonstrating that causal influence and topology changes provide complementary anomaly evidence. Appendices C.1 and C.2 provide cross-dataset and finer-grained score ablations.

Parameter Sensitivity. Table 5 evaluates the effects of historical reference length H, contrastive temperature τ, and negative perturbation intensity ξ. Performance varies only modestly with H, with $H = 10$ yielding the strongest overall results. A temperature of $\tau = 0.1$ performs best on PSM and remains competitive on SWaT, whereas $\tau = 0.2$ notably degrades the SWaT results. For ξ, moderate perturbation at 1.0 provides the best balance across both datasets; weaker perturbations reduce AUPRC and F1 on SWaT, while increasing ξ to 2.0 lowers AUROC on both datasets and F1 on PSM. Overall, the default configuration $(H, \tau, \xi) = (10, 0.1, 1.0)$ is consistently effective across the two datasets.

Table 5: Hyperparameter sensitivity on PSM and SWaT. H, τ, and ξ denote reference length, contrastive temperature, and negative perturbation intensity; ROC/PRC denote AUROC/AUPRC. Best results per group are bold.

		PSM			SWaT		
		ROC	PRC	F1	ROC	PRC	F1
H	1	0.73	0.57	0.56	0.87	0.72	0.79
	5	0.74	0.57	0.56	0.87	0.72	0.79
	10	**0.75**	**0.59**	**0.59**	**0.87**	**0.75**	**0.79**
	20	0.74	0.58	0.56	0.86	0.73	0.79
τ	0.05	0.74	0.57	0.55	0.87	**0.76**	0.78
	0.10	**0.75**	**0.59**	**0.59**	**0.87**	0.75	**0.79**
	0.20	0.74	0.58	0.56	0.84	0.61	0.75
	0.50	0.72	0.56	0.55	0.86	0.76	0.79
ξ	0.25	0.74	0.57	0.55	0.86	0.68	0.76
	0.50	0.73	0.56	0.55	0.87	0.71	0.76
	1.00	**0.75**	**0.59**	**0.59**	**0.87**	**0.75**	**0.79**
	2.00	0.70	0.59	0.53	0.83	0.74	0.79

Representation Visualization. Figure 4 qualitatively visualizes the learned representations using t-SNE (van der Maaten & Hinton, 2008), while the silhouette score (Rousseeuw, 1987) quantifies within-class compactness and between-class separation. Before training, normal and anomalous samples overlap substantially, with a score of 0.2767. After 30 epochs, the score increases to 0.4399 with a static causal graph and 0.4475 with dynamic causal graphs, which also yield more compact normal representations and clearer class separation. These results support the effectiveness of causality-aware contrastive learning and dynamic causal modeling.

5 Conclusion

We presented DCCAD, a dynamic causal contrastive learning framework for unsupervised multivariate time-series anomaly detection. DCCAD models evolving dependencies using estimated dynamic causal graphs and employs them to guide causality-preserving and causality-disrupting augmentations, thereby learning discriminative normal representations. It further integrates representation distance, causal prediction error, and graph change scores for complementary anomaly assessment. Experiments across diverse datasets demonstrate its effectiveness. Nevertheless, variable-wise perturbation introduces computational overhead for high-dimensional data, while the current offline training setting may limit adaptation to rapidly changing systems. Future work will investigate more efficient causal graph estimation and extend DCCAD to online anomaly detection.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code generation. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodologies or experiments, or interpret experimental results. The remaining tasks requiring disclosure are not applicable to this work.

Additionally, we used generative AI tools to improve the grammar, clarity, and readability of the manuscript. All AI-assisted outputs were carefully reviewed and verified by the authors. AI-assisted code was examined, modified where necessary, and tested within our experimental pipeline to ensure correctness and consistency. Language-editing suggestions were reviewed to ensure that they did not alter the original scientific meaning or introduce unsupported claims.

The authors take full responsibility for the final content of this work, including all text, scientific claims, experimental results, and research artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The model formulation and anomaly scoring procedure are described in Section 3, while the datasets, evaluation protocol, and experimental settings are presented in Section 4. The real-world datasets used in this work are publicly available. Further dataset and implementation details are provided in Appendices A and B, with additional experimental results in Appendix C. The anonymized supplementary repository provides the core code and supports evaluation-level reproducibility under the included configuration: `https://anonymous.4open.science/r/DCCAD-4626`.

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 and Data Mining*, pp. 2485–2494. ACM, 2021. doi: 10.1145/3447548.3467174.

Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P. Mathur. WADI: A water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, pp. 25–28. ACM, 2017. doi: 10.1145/3055366.3055375.

Charles K. Assaad, Emilie Devijver, and Éric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022. doi: 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 and Data Mining*, pp. 3395–3404, 2020. doi: 10.1145/3394486.3403392. URL `https://doi.org/10.1145/3394486.3403392`.

Tingzhu Bi, Yicheng Pan, Xinrui Jiang, Huize Sun, Meng Ma, and Ping Wang. UnCLe: Towards scalable dynamic causal discovery in non-linear temporal systems. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106563–106592, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL `https://proceedings.mlr.press/v119/chen20j.html`.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. CARLA: Self-supervised contrastive representation learning for time series anomaly detection. *Pattern Recognition*, 157:110874, 2025. doi: 10.1016/j.patcog.2024.110874.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021. doi: 10.1609/aaai.v35i5.16523.

Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. In *Critical Information Infrastructures Security*, volume 10242 of *Lecture Notes in Computer Science*, pp. 88–99. Springer, 2017. doi: 10.1007/978-3-319-71368-7_8.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Biwei Huang, Kun Zhang, and Bernhard Schölkopf. Identification of time-dependent causal model: A Gaussian process treatment. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3561–3568. AAAI Press, 2015.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2901–2910. PMLR, 2019.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

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 and Data Mining*, pp. 387–395. ACM, 2018. doi: 10.1145/3219819.3219845.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30591–30608, 2025.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. CrossAD: Time series anomaly detection with cross-scale associations and cross-window modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, volume 11730 of *Lecture Notes in Computer Science*, pp. 703–716. Springer, 2019. doi: 10.1007/978-3-030-30490-4_56.

Zehao Liu, Mengzhou Gao, and Pengfei Jiao. GCAD: Anomaly detection in multivariate time series from the perspective of granger causality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19041–19049, 2025. doi: 10.1609/aaai.v39i18.34096.

Edward N. Lorenz. Predictability: A problem partly solved. In *Proceedings of the Seminar on Predictability*, volume 1, pp. 1–18. ECMWF, 1996.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure learning from time-series data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605, 2020.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Sejdinovic Dinoj. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11):113115, 2020. doi: 10.1063/5.0020538.

M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, and Marios Koulakis. Position: Quo vadis, unsupervised time series anomaly detection? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.02678.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=v5gjXpmR8J.

Ali Shojaie and Emily B. Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319, 2022. doi: 10.1146/annurev-statistics-040120-010930.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2828–2837, 2019. doi: 10.1145/3292500.3330672.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11839–11852, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8965f76632d7672e7d3cf29c87ecaa0c-Abstract.html.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022. doi: 10.1109/TPAMI.2021.3065601.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Trans. Knowl. Data Eng.*, 35(3):2421–2429, 2023.

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_.

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V. Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1409–1416, 2019. doi: 10.1609/aaai.v33i01.33011409.

Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE International Conference on Data Mining*, pp. 841–850. IEEE, 2020.

Bin Zhou, Shenghua Liu, Bryan Hooi, Xueqi Cheng, and Jing Ye. BeatGAN: Anomalous rhythm detection using adversarially generated time series. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4433–4439, 2019. doi: 10.24963/ijcai.2019/616.

A Dataset Description

A.1 Real-World Datasets

MSL. The Mars Science Laboratory (MSL) dataset contains spacecraft telemetry collected from the Curiosity rover by NASA. Each telemetry channel records the temporal behavior of a spacecraft component or subsystem, together with correlated measurements that provide contextual information. Its training sequences mainly represent normal operating conditions, whereas the test sequences contain anomalies labeled by domain experts. These anomalies arise from unexpected subsystem behaviors and operational events, making MSL a widely used benchmark for evaluating anomaly detection in complex cyber-physical systems (Hundman et al., 2018).

SMD. The Server Machine Dataset (SMD) consists of monitoring data collected from multiple server machines in a large Internet company over five weeks. Each machine is represented by a multivariate sequence containing system indicators such as CPU load, memory usage, network activity, and other operational measurements. The training portion records normal server operation, while anomalies in the test portion are labeled by system operators. Since different machines exhibit distinct operating patterns and anomaly distributions, SMD provides a heterogeneous benchmark for evaluating model robustness across server environments (Su et al., 2019).

PSM. The Pooled Server Metrics (PSM) dataset contains monitoring metrics collected from multiple application server nodes at eBay. The measurements describe the operating states of interacting server components and exhibit complex temporal and cross-variable dependencies. The training data represent normal system behavior, while the test data include labeled anomalous intervals caused by abnormal server operations. The relatively high anomaly prevalence and strong interactions among monitoring variables make PSM suitable for evaluating whether a method can distinguish anomalous system states from substantial normal variations (Abdulaal et al., 2021).

SWaT. Secure Water Treatment (SWaT) is a cyber-physical testbed that reproduces a six-stage industrial water treatment process. The dataset records sensor measurements and actuator states under normal operation and a series of cyberattacks. Its variables describe tightly coupled physical processes, including water flow, pressure, tank levels, and control actions. Attacks may affect individual devices or propagate across multiple processing stages, producing both direct measurement deviations and changes in inter-variable dependencies. These properties make SWaT particularly suitable for evaluating anomaly detection methods in industrial control systems (Goh et al., 2017).

WADI. The Water Distribution (WADI) dataset was collected from a large-scale water distribution testbed operating under normal conditions and simulated cyberattacks. It contains measurements from numerous sensors and actuators distributed across different components of the system. Compared with SWaT, WADI involves a larger system scale and more variables, resulting in more complex temporal dependencies and attack propagation patterns. Its attack scenarios manipulate sensing or control components and can influence multiple downstream processes, making WADI a challenging benchmark for multivariate time-series anomaly detection (Ahmed et al., 2017).

Recent studies have shown that commonly used time-series anomaly detection benchmarks may contain trivial, inconsistent, or unreliable cases, highlighting the need for transparent and data-driven subset selection (Wu & Keogh, 2023; Sarfraz et al., 2024). Following CAROTS (Kim et al., 2025), we characterize individual series according to anomaly prevalence and stationarity. Accordingly, the main text reports results on MSL_P-14 and MSL_P-15, which represent non-stationary spacecraft telemetry with different anomaly sparsity, and SMD_3-6 and SMD_3-11, which exhibit contrasting stationarity characteristics and anomaly ratios. These subsets are selected based on dataset characteristics rather than DCCAD's performance, and all methods are evaluated on the same subsets. For completeness, Appendix C.7 additionally compares CAROTS and DCCAD across all MSL and SMD entities.

Table 6 summarizes the training and test lengths, number of channels, anomaly points and ratios, and stationarity characteristics of each dataset. A more negative ADF statistic with a small p-value indicates stronger evidence of stationarity, whereas a larger ADF statistic with a high p-value suggests non-stationarity. Based on these characteristics, MSL_P-14 and MSL_P-15 represent non-stationary

telemetry sequences with different levels of anomaly sparsity, while SMD_3-6 and SMD_3-11 provide a contrast between stationary and non-stationary server behaviors with distinct anomaly ratios.

Table 6: Statistics of the real-world datasets used in the experiments.

Dataset	Train Length	Test Length	Channels	Anomaly Points	Anomaly Ratio (%)	ADF Statistic	p-value
MSL_P-14	2, 880	6, 100	55	181	2.97	-0.352	0.918
MSL_P-15	3, 682	2, 856	55	21	0.74	-0.621	0.866
SMD_3-6	28, 726	28, 726	38	1, 194	4.16	-9.799	0.000
SMD_3-11	28, 695	28, 696	38	198	0.69	-0.424	0.906
PSM	132, 481	87, 841	25	24, 381	27.76	-3.493	0.010
SWaT	495, 000	449, 919	51	54, 621	12.14	-0.780	0.830
WADI	784, 537	172, 801	123	9, 977	5.77	-0.347	0.920

A.2 Synthetic Datasets

We use Lorenz-96, a nonlinear dynamical system originally developed to model atmospheric processes, as the synthetic dataset (Lorenz, 1996). Its cyclic interactions and chaotic dynamics provide a controlled environment for evaluating anomaly detection under complex temporal and inter-variable dependencies. Following CAROTS (Kim et al., 2025) and its adopted anomaly injection schemes (Lai et al., 2021), we generate point global (PG), point contextual (PC), collective trend (CT), and collective global (CG) anomalies. We additionally introduce causal shift (CS) to simulate abrupt changes in the causal mechanism of a multivariate time series by directly rewiring the causal parents and dynamics of Lorenz-96.

Within a CS interval, the dynamics are defined as

$$\dot{x}_i^{\mathrm{CS}}(t) = (1-\omega_\lambda)(x_{i+1}-x_{i-2})x_{i-1} + \omega_\lambda(x_{i+2}-x_{i-3})x_{i-1} - x_i + F, \tag{13}$$

where $\omega_\lambda = \min(\lambda/3, 1)$, F is the constant forcing term, and variable indices are interpreted cyclically over the N variables. This formulation interpolates from the normal Lorenz-96 coupling to a rewired mechanism as the anomaly strength increases. Specifically, the causal parent set of x_i changes from $\{x_{i+1}, x_{i-1}, x_{i-2}\}$ to $\{x_{i+2}, x_{i-1}, x_{i-3}\}$: the parent x_{i-1} is retained, whereas x_{i+1} and x_{i-2} are replaced by x_{i+2} and x_{i-3}. Thus, CS alters the underlying causal mechanism rather than directly perturbing observations. Numerically integrating $\dot{x}_i^{\mathrm{CS}}(t)$ produces the anomalous sequence within the CS interval.

The anomaly strength factor λ controls the deviation from normal dynamics for all anomaly types. In the main experiments, we set $\lambda = 2.0$ and $\lambda = 1.0$ to construct the lower-difficulty Lorenz#1 and higher-difficulty Lorenz#2 settings, respectively. Appendix C further reports AUROC, AUPRC, and F1 under $\lambda \in \{0.5, 1.0, 2.0, 3.0\}$ to evaluate performance across a broader range of anomaly strengths.

B Implementation Details

B.1 Data Preprocessing and Sliding Windows

For each multivariate time series, the first 80% of the normal training sequence is used for model optimization, while the remaining 20% is retained for validation and checkpoint selection. A standard scaler is fitted exclusively on the optimization portion and subsequently applied to the validation and test sequences to avoid data leakage. Each variable is standardized independently as

$$\widetilde{x}_{i,t} = \frac{x_{i,t} - \mu_i^{\mathrm{tr}}}{\sigma_i^{\mathrm{tr}} + \varepsilon}, \tag{14}$$

where μ_i^{tr} and σ_i^{tr} denote the mean and standard deviation of variable i calculated from the optimization portion of the normal training sequence, and ε is a small constant for numerical stability. For notational simplicity, the standardized observations continue to be denoted by $\boldsymbol{x}_t$ and X_t throughout the paper.

The standardized sequences are divided into overlapping windows of length $w = 10$ with a stride of one during both training and testing. The window ending at time step t is represented as

$$X_t = [\boldsymbol{x}_{t-w+1}, \ldots, \boldsymbol{x}_t]^\top \in \mathbb{R}^{w \times N}, \tag{15}$$

where N denotes the number of variables. The window X_t provides the temporal context required to compute an anomaly score rather than serving as a window-level detection unit. With a stride of one, successive windows produce a continuous anomaly-score sequence, and each score is aligned with its corresponding ground-truth label for evaluation. Since the training sequences contain only normal observations, all training scores correspond to normal data.

B.2 Dynamic Causal Graph Learning

The dynamic causal graphs are estimated using an inductive UnCLe model. For each dataset, the first 80% of the normal training sequence is used for model optimization, while the remaining 20% is retained for validation. The standard scaler is also fitted exclusively on the optimization portion. The sequences are then divided into non-overlapping causal inference blocks, whose lengths are reported in Table 7. Blocks of the same length are organized into mini-batches with a batch size of four. All mini-batches jointly update a single shared model rather than training a separate model for each block. For datasets containing multiple entities, each entity is segmented independently, and no block crosses entity boundaries.

Table 7: Causal inference block lengths used for different datasets.

Dataset	Block Length s
MSL_P-14	10, 000
MSL_P-15	10, 000
SMD_3-6	5, 000
SMD_3-11	5, 000
PSM	5, 000
SWaT	5, 000
WADI	2, 500

UnCLe employs a shared TCN encoder to extract the temporal features of individual variables and learnable inter-variable coupling parameters to model cross-variable dependencies. Training consists of two stages. The model is first pretrained using a self-reconstruction objective and then jointly optimized with one-step prediction loss, self-reconstruction loss, and an L_1 regularization term on the inter-variable coupling parameters. During joint training, the prediction and reconstruction losses are evaluated on the temporally held-out validation portion. The checkpoint with the lowest validation loss is retained, resulting in one selected model for each dataset.

The selected model is subsequently frozen and used with the same parameters to process both the complete training and test sequences; test observations are never used for parameter optimization or checkpoint selection. After the raw causal influence estimates are obtained, negative values are clipped to zero, followed by a backward moving average of length five along the temporal dimension. $\mathcal{A} = \{A_t\}_{t=1}^T$ denotes the resulting smoothed weighted causal graph sequence. The global binarization threshold δ is determined exclusively from the distribution of smoothed causal influence estimates in the training portion and is applied unchanged to both training and test graphs. This procedure yields the time-resolved binary graph sequence $\mathcal{B} = \{B_t\}_{t=1}^T$. Since causal relationships are assumed to remain approximately stable within a short anomaly-detection window, each window X_t uses the dynamic causal graph B_t corresponding to its ending time step.

Specifically, let $\mathcal{T}_{\text{opt}}$ denote the first 80% of the normal training sequence used for model optimization. Since A_t denotes the nonnegative causal influence after backward moving-average smoothing, the threshold and binary graph are computed as

$$\delta = Q_q(\{(A_t)_{j,i} : t \in \mathcal{T}_{\text{opt}},\ 1 \le i, j \le N\}), \qquad (B_t)_{j,i} = \mathbb{I}[(A_t)_{j,i} > \delta]. \tag{16}$$

The quantile calculation includes diagonal entries; off-diagonal masking is used only when counting edges. To account for differences in variable dimensionality, causal-strength distributions, and system coupling complexity, we use dataset-family-specific quantile thresholds rather than enforcing

a common edge-retention ratio across all datasets. Specifically, q is set to 0.97, 0.93, 0.75, 0.96, 0.96, and 0.85 for MSL, SMD, PSM, SWaT, WADI, and Lorenz-96, respectively. Higher quantiles are used for datasets with more variables or numerous weak relationships to control graph expansion caused by the quadratic growth of candidate edges and to suppress numerical noise. Relatively lower quantiles are used for datasets with fewer variables or where moderate-strength dependencies should be retained, avoiding overly sparse or degenerate graphs. The resulting graphs remain sparse while preserving sufficient cross-variable connections.

For each dataset family, the numerical threshold δ is estimated from the normal training graphs after applying a length-five backward moving average to the causal influences. It is then held fixed across all entities, random seeds, and test samples. No test labels are used in threshold estimation.

B.3 CDA-PA Implementation

For each sample, CDA-PA constructs $\mathcal{S}_t$ by uniformly selecting one node from those with nonzero out-degree in the current binary causal graph; thus, $|\mathcal{S}_t| = 1$. Gaussian noise with an intensity of 0.1 is added to the selected source variable, and all its direct downstream variables are updated by the causal prediction model. If the current graph contains no node with nonzero out-degree, one source variable is sampled uniformly from all variables.

B.4 CDA-NA Implementation

CDA-NA samples the variable perturbation ratio ρ_t from $[0.2, 0.8]$ and sets the temporal coverage ratio and negative perturbation intensity to $p = 0.5$ and $\xi = 1.0$, respectively. It uses the K_t-th largest nonzero out-degree as the selection threshold γ_t and perturbs all variables whose out-degrees are at least γ_t. Consequently, when multiple variables are tied at the threshold, the actual number of selected variables can exceed K_t; no additional tie-breaking or truncation is applied.

B.5 Representation Encoder and Projection Head

For each multivariate time-series window $X_t \in \mathbb{R}^{w \times N}$, we employ an LSTM-based representation encoder to capture its temporal dynamics and inter-variable dependencies. The LSTM contains one layer with a hidden dimension of 512 and a dropout rate of 0. Its final hidden state is used as the window-level representation:

$$\boldsymbol{h}_t = f_{\text{enc}}(X_t) \in \mathbb{R}^{512}. \tag{17}$$

The original windows and their causality-aware positive and negative augmentations are processed by the same encoder with shared parameters, ensuring that all representations lie in a common feature space.

The encoder output is subsequently passed through a two-layer multilayer perceptron projection head. The first linear layer maps the 512-dimensional representation to a 1024-dimensional hidden space, followed by batch normalization and a GELU activation. The second linear layer maps the hidden representation to a 512-dimensional contrastive embedding:

$$\boldsymbol{z}_t = W_2\,\mathrm{GELU}\left(\mathrm{BN}\left(W_1\boldsymbol{h}_t + \boldsymbol{b}_1\right)\right) + \boldsymbol{b}_2, \tag{18}$$

where $W_1 \in \mathbb{R}^{1024\times512}$ and $W_2 \in \mathbb{R}^{512\times1024}$. The resulting projected representation is $\boldsymbol{z}_t \in \mathbb{R}^{512}$.

Before computing the contrastive objective, the projected representation is normalized using the ℓ_2 norm:

$$\overline{\boldsymbol{z}}_t = \frac{\boldsymbol{z}_t}{\|\boldsymbol{z}_t\|_2}. \tag{19}$$

The similarity between two windows is then measured by the inner product of their normalized representations, which is equivalent to cosine similarity.

B.6 Training Details

DCCAD is implemented in PyTorch and evaluated on a single NVIDIA GeForce RTX 4090 GPU. The representation model is trained for 30 epochs with a mini-batch size of 256. We use the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 3×10^{-4} and a weight decay of

10^{-3}. The learning rate follows a cosine decay schedule with a five-epoch warm-up period, and the gradient norm is clipped to 0.5. The temperature is set to 0.1, while the similarity threshold increases linearly from 0.5 to 0.9. The model is evaluated every five epochs, and the checkpoint with the lowest validation loss is retained. The pretrained causal prediction model remains frozen during representation learning.

All comparison methods are rerun on the same computational platform and evaluated using the same protocol. We use the same data splits whenever compatible with the method implementations. Because different methods impose distinct requirements on input sequence length, sampling procedures, and model architectures, method-specific preprocessing and training hyperparameters follow their original implementations or recommended configurations. All methods use the same test labels and metric computation procedures, including AUROC, AUPRC, and Best-F1. We do not apply point adjustment because it uses ground-truth anomaly intervals to modify point-wise predictions and may substantially overestimate detection performance (Kim et al., 2022). For a controlled comparison with CAROTS (Kim et al., 2025), we further use identical data preprocessing, encoder architecture, window size, training epochs, and threshold selection strategy. Training data are shuffled at each epoch, whereas validation and test data retain their temporal order.

B.7 Anomaly Scoring Computation

In the implementation, the trained representation encoder and projection head are first used to extract the representations of the normal training windows. Their mean is computed as the normal representation center $\boldsymbol{c}$. Causality-aware positive samples are also treated as normal samples and included in the estimation of $\boldsymbol{c}$. For a test window X_t with representation $\boldsymbol{z}_t$, the contrastive representation score is computed as the Euclidean distance between $\boldsymbol{z}_t$ and the normal representation center:

$$S_t^{\mathrm{con}} = \left\| \boldsymbol{z}_t - \boldsymbol{c} \right\|_2 . \tag{20}$$

The unnormalized output of the projection head is used for this distance calculation. A larger distance indicates a greater deviation from the normal representation region.

The causal prediction score is computed using the frozen UnCLe causal prediction model. Given a test window X_t, the model predicts the current observation from its preceding historical observations:

$$\widehat{\boldsymbol{x}}_t = f_{\theta_c}(X_{<t}), \tag{21}$$

where θ_c denotes the frozen causal prediction model parameters selected according to the validation loss. The causal prediction score is defined as the mean squared error between the predicted observation $\widehat{\boldsymbol{x}}_t$ and the actual observation $\boldsymbol{x}_t$ over all N variables:

$$S_t^{\mathrm{pred}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{x}_{i,t} - x_{i,t} \right)^2 . \tag{22}$$

A larger prediction error indicates that the current observation is less consistent with the normal causal dynamics learned from the training data.

The graph change score S_t^{gc} is computed as described in the main text, using a historical reference length of $H = 10$ and a graph fusion coefficient of $\beta = 0.5$. It is combined with the representation and causal prediction scores after standardization, with its contribution weighted by $\lambda_{\mathrm{gc}} = 0.1$.

Before score fusion, the three score components are independently standardized using the mean and standard deviation of their corresponding training scores:

$$\widetilde{S}_t^m = \frac{S_t^m - \mu_{\mathrm{tr}}^m}{\sigma_{\mathrm{tr}}^m}, \qquad m \in \{\mathrm{con}, \mathrm{pred}, \mathrm{gc}\} , \tag{23}$$

where μ_{tr}^m and σ_{tr}^m denote the mean and standard deviation of score component m on the training set.

B.8 Baselines

For non-contrastive baselines, Anomaly Transformer (AT) detects anomalies by combining reconstruction error with association discrepancy (Xu et al., 2022). TimesNet (TN) models multi-period

Table 8: Full ablation of graph and augmentation strategies on all real-world datasets. All configurations use the contrastive representation and causal prediction scores, with GCS excluded. The first and last configurations correspond to CAROTS and DCCAD w/o GCS, respectively.

Graph	CDA-PA	CDA-NA	Metric	MSL_P-14	MSL_P-15	SMD_3-6	SMD_3-11	PSM	SWaT	WADI
Static	All	DFS	AUROC	0.7642	0.6911	0.9201	0.8724	0.7278	0.8512	0.4879
			AUPRC	0.3721	0.0213	0.4657	0.2378	0.5301	0.7599	0.0545
			F1	0.5448	0.0863	0.4345	0.4473	0.5304	0.7866	0.1339
Dynamic	All	DFS	AUROC	0.7667	0.7497	0.9451	0.8946	0.7293	0.8397	0.5959
			AUPRC	0.3778	0.1118	0.5281	0.2496	0.5523	0.7562	0.1875
			F1	0.5324	0.2745	0.4902	0.3867	0.5332	0.8045	0.2584
Dynamic	Nonzero	DFS	AUROC	0.7668	0.7735	0.9614	0.9173	0.7325	0.8498	0.5319
			AUPRC	0.3777	0.1936	0.5806	0.2671	0.5524	0.6926	0.1616
			F1	0.5324	0.4211	0.5260	0.4112	0.5359	0.7699	0.2578
Dynamic	Nonzero	Degree	AUROC	0.7679	0.7733	0.9674	0.8877	0.7453	0.8596	0.6416
			AUPRC	0.3977	0.1969	0.5964	0.3014	0.5822	0.7487	0.2005
			F1	0.5324	0.4167	0.6194	0.4754	0.5789	0.7901	0.2740

temporal variations through two-dimensional representations (Wu et al., 2023), while USAD employs adversarially trained autoencoders (Audibert et al., 2020). CrossAD reconstructs time-series data by exploiting cross-scale and cross-window dependencies (Li et al., 2025). GCAD dynamically infers Granger-causal graphs from prediction gradients and detects deviations from normal causal patterns (Liu et al., 2025).

For contrastive baselines, SimCLR learns representations by maximizing agreement between augmented views (Chen et al., 2020). SSD combines self-supervised representations with distance-based anomaly detection (Sehwag et al., 2021), whereas CSI introduces distributionally shifted samples as negatives (Tack et al., 2020). CARLA employs synthetic anomaly generation and neighborhood-based contrastive learning for time-series anomaly detection (Darban et al., 2025). CAROTS incorporates a static causal graph into contrastive augmentation (Kim et al., 2025). SimCLR, SSD, and CSI are adapted to multivariate time-series windows for comparison.

The source-code repositories are listed below:

- AT: `https://github.com/thuml/Anomaly-Transformer`
- TN: `https://github.com/thuml/TimesNet`
- USAD: `https://github.com/manigalati/usad`
- CrossAD: `https://github.com/decisionintelligence/CrossAD`
- GCAD: `https://github.com/Tc99m/GCAD`
- SimCLR: `https://github.com/google-research/simclr`
- SSD: `https://github.com/inspire-group/SSD`
- CSI: `https://github.com/alinlab/CSI`
- CARLA: `https://github.com/zamanzadeh/CARLA`
- CAROTS: `https://github.com/kimanki/CAROTS`

C Additional Experiments

C.1 Ablation Results across All Datasets

We extend the two ablation studies in Section 4.3 to all real-world datasets. Table 8 compares the graph and augmentation strategies under the same scoring condition, while Table 9 evaluates PS and GCS with the contrastive representation score retained.

Table 9: Full-dataset ablation of anomaly score components. PS and GCS denote the causal prediction and graph change scores, respectively; the contrastive representation score is retained in all configurations. Blank entries indicate results that are not reported.

PS	GCS	Metric	MSL_P-14	MSL_P-15	SMD_3-6	SMD_3-11	PSM	SWaT	WADI
		AUROC	0.7497	0.6307	0.9664	0.8263	0.7378	0.8639	0.5439
×	×	AUPRC	0.2844	0.0218	0.5677	0.1429	0.5672	0.6257	0.1351
		F1	0.4853	0.0966	0.6289	0.1937	0.5640	0.7461	0.1282
		AUROC	0.7388	0.7103	0.9662	0.8282	0.7394	0.8621	0.4601
×	✓	AUPRC	0.1030	0.3721	0.5683	0.0780	0.5663	0.6216	0.0490
		F1	0.1767	0.5000	0.6287	0.1113	0.5649	0.7435	0.1227
		AUROC	0.7679	0.7733	0.9674	0.8877	0.7453	0.8596	0.6416
✓	×	AUPRC	0.3977	0.1969	0.5964	0.3014	0.5822	0.7487	0.2005
		F1	0.5324	0.4167	0.6194	0.4754	0.5789	0.7901	0.2740
		AUROC	0.7831	0.7735	0.9696	0.9000	0.7506	0.8694	0.6785
✓	✓	AUPRC	0.4459	0.3738	0.6180	0.2897	0.5900	0.7496	0.2373
		F1	0.5324	0.5502	0.6317	0.4441	0.5893	0.7903	0.3236

Graph and Augmentation Strategies. As shown in Table 8, the successive introduction of dynamic graphs, source selection restricted to nodes with nonzero out-degrees, and degree-based negative augmentation generally improves detection performance. Compared with the CAROTS configuration, DCCAD w/o GCS achieves higher AUROC on all seven datasets and higher AUPRC and F1 on six datasets each. Although individual modifications exhibit some dataset-dependent variation, the overall results support the combined effectiveness of DCCAD's graph and augmentation designs.

Anomaly Score Components. Table 9 shows that PS provides a consistent complementary signal, improving AUPRC on all seven datasets when added to the representation score. Incorporating GCS on top of PS further improves AUROC on every dataset and AUPRC on six. Consequently, the complete DCCAD score attains the highest AUROC on all seven datasets, the highest AUPRC on six, and the highest F1 on five, with an additional F1 tie on MSL_P-14. These results confirm the primary contribution of PS and the complementary value of GCS across datasets.

C.2 Fine-Grained Score Ablation

To further examine the anomaly scoring mechanism, we evaluate different combinations of its components on PSM. As shown in Table 10, the contrastive representation score provides the primary anomaly-discrimination capability, achieving the strongest standalone performance. Among the causal indicators, the causal prediction score performs best when used alone. Adding it to the representation score improves AUROC, AUPRC, and F1 from 0.7378, 0.5672, and 0.5640 to 0.7453, 0.5822, and 0.5789, respectively.

The causal influence and topology change scores show limited standalone performance because causal relationships may also vary during normal operating transitions. Nevertheless, they provide complementary evidence when combined with the representation and prediction scores. Specifically, adding influence change to these two scores improves AUPRC and F1, whereas topology change improves all three metrics and contributes more strongly to AUROC. Integrating all four components achieves the best AUROC, AUPRC, and F1 of 0.7506, 0.5900, and 0.5893, respectively. These results indicate that representation deviation provides the main detection signal, while prediction inconsistency and causal changes refine anomaly ranking from complementary perspectives.

C.3 Selection of the Window-Level Causal Graph

Each anomaly-detection window contains a sequence of binary causal graphs. We use the endpoint graph as the window-level graph because it is aligned with the observation being evaluated. As shown in Table 11, PSM achieves edge-wise consistency of 95.80% and 96.29%, average pairwise

Table 10: Extended ablation of anomaly score components on PSM. CON and PS denote the contrastive representation and causal prediction scores, respectively, while INF and TOP denote the causal influence and topology change components of GCS.

CON	PS	INF	TOP	AUROC	AUPRC	F1
✓	×	×	×	0.7378	0.5672	0.5640
×	✓	×	×	0.6600	0.4905	0.4694
×	×	✓	×	0.4877	0.2828	0.4437
×	×	×	✓	0.5349	0.3150	0.4434
×	×	✓	✓	0.5216	0.3043	0.4431
✓	✓	×	×	0.7453	0.5822	0.5789
✓	×	✓	✓	0.7394	0.5663	0.5649
✓	✓	✓	×	0.7430	0.5857	0.5877
✓	✓	×	✓	0.7484	0.5833	0.5849
✓	✓	✓	✓	**0.7506**	**0.5900**	**0.5893**

Table 11: Comparison of window-level causal graph strategies on PSM.

(a) Within-window graph consistency			
Split	Edge-wise Consistency	Jaccard Similarity	Endpoint-Graph Agreement
Train	95.80%	0.7604	94.95%
Test	96.29%	0.7699	95.54%
(b) Anomaly detection performance			
Graph Strategy	AUROC	AUPRC	F1
Window Aggregation	0.7263	0.5613	0.5400
Endpoint Graph	**0.7506**	**0.5900**	**0.5893**

Jaccard similarities of 0.7604 and 0.7699, and endpoint-graph agreement of 94.95% and 95.54% on the training and test sets, respectively. These results indicate smooth within-window graph evolution and support the representativeness of the endpoint graph.

Compared with aggregating all ten graphs, the endpoint-graph strategy improves AUROC from 0.7263 to 0.7506, AUPRC from 0.5613 to 0.5900, and F1 from 0.5400 to 0.5893. Window aggregation may retain earlier relationships and dilute transient changes associated with the current observation, whereas the endpoint graph preserves the most recent causal state.

C.4 Stability across Random Seeds

Table 12 reports the standard deviations of AUROC, AUPRC, and F1 over multiple runs with different random seeds. Lower values indicate lower sensitivity to random initialization and data shuffling. DCCAD exhibits particularly low variability on MSL_P-15, SMD_3-6, and SWaT, with AUROC standard deviations of 0.0005, 0.0021, and 0.0013, respectively. Its AUPRC and F1 deviations on SWaT are also small, at 0.0011 and 0.0036. Although the degree of variation differs across datasets, the results demonstrate that DCCAD provides stable and reproducible performance on multiple benchmarks, while also highlighting the importance of multi-seed evaluation under different data characteristics.

C.5 Computational Efficiency

Table 13 reports the total time required by each method for model training and inference on PSM. For DCCAD and CAROTS, the causal graphs are constructed in a separate offline causal discovery stage and can be reused during subsequent training and inference without being recomputed for every run. The cost of this stage is therefore excluded from the table; the reported results reflect the computational cost given preconstructed causal graphs, rather than the complete end-to-end cost of an initial run that includes graph construction.

Table 12: Standard deviations of detection performance on real-world datasets over multiple runs with different random seeds.

Dataset	Metric	AT	TN	USAD	CrossAD	GCAD	SimCLR	SSD	CSI	CARLA	CAROTS	DCCAD
MSL_P-14	AUROC	0.0067	0.0005	0.0001	0.0002	0.0036	0.0228	0.0489	0.2949	0.2420	0.0002	0.0123
	AUPRC	0.0144	0.0000	0.0000	0.0000	0.0133	0.1746	0.0552	0.0239	0.1767	0.0000	0.1179
	F1	0.0000	0.0000	0.0000	0.0000	0.0291	0.1025	0.0619	0.0626	0.2515	0.0000	0.0000
MSL_P-15	AUROC	0.0060	0.0010	0.0126	0.0063	0.0070	0.0299	0.0319	0.0133	0.1313	0.0101	0.0005
	AUPRC	0.0000	0.0002	0.0002	0.0052	0.0213	0.0055	0.0013	0.0035	0.1907	0.0014	0.0021
	F1	0.0000	0.0008	0.0004	0.0150	0.0242	0.0286	0.0036	0.0148	0.0237	0.0063	0.0170
SMD_3-6	AUROC	0.0131	0.0008	0.0091	0.0005	0.0054	0.0017	0.0840	0.0036	0.1858	0.0114	0.0021
	AUPRC	0.0627	0.0035	0.0107	0.0002	0.0651	0.0125	0.0452	0.0360	0.1303	0.0340	0.0194
	F1	0.0424	0.0008	0.0186	0.0005	0.0372	0.0138	0.0369	0.0125	0.2721	0.0367	0.0051
SMD_3-11	AUROC	0.0118	0.0000	0.0000	0.0001	0.0068	0.0111	0.0399	0.0039	0.2193	0.0036	0.0101
	AUPRC	0.0251	0.0001	0.0000	0.0001	0.0081	0.0456	0.0007	0.0200	0.2337	0.0098	0.0130
	F1	0.0321	0.0000	0.0000	0.0015	0.0029	0.0736	0.0002	0.0214	0.2242	0.0202	0.0362
PSM	AUROC	0.0001	0.0035	0.0028	0.0137	0.0025	0.0182	0.0134	0.0138	0.0311	0.0179	0.0391
	AUPRC	0.0001	0.0028	0.0101	0.0107	0.0024	0.0356	0.0179	0.0187	0.2194	0.0142	0.0374
	F1	0.0000	0.0043	0.0011	0.0130	0.0021	0.0141	0.0029	0.0262	0.0029	0.0102	0.0447
SWaT	AUROC	0.0792	0.0005	0.0002	0.0005	0.0081	0.0118	0.0320	0.0915	0.1475	0.0009	0.0013
	AUPRC	0.2327	0.0017	0.0025	0.0046	0.0021	0.0011	0.1676	0.0340	0.1790	0.0024	0.0011
	F1	0.0000	0.0000	0.0071	0.0040	0.0088	0.0049	0.0886	0.0721	0.2309	0.0029	0.0036
WADI	AUROC	0.0208	0.0013	0.0032	0.0038	0.0058	0.0274	0.0122	0.0110	0.0607	0.0048	0.0866
	AUPRC	0.0466	0.0001	0.0003	0.0003	0.0040	0.0278	0.0322	0.0018	0.1397	0.0004	0.0706
	F1	0.0000	0.0001	0.0014	0.0023	0.0042	0.0462	0.0053	0.0016	0.0000	0.0005	0.0989

Table 13: Combined training and inference runtime on PSM. For DCCAD and CAROTS, the offline causal discovery stage is excluded.

Method	Runtime (s)
AT	54.75
TN	751.00
USAD	924.19
CrossAD	2638.50
GCAD	158.08
SimCLR	191.31
SSD	1156.40
CSI	306.15
CARLA	> 4500
CAROTS	696.77
DCCAD	1620.78

DCCAD requires 1620.78 seconds, approximately $2.33\times$ the runtime of CAROTS. Since offline causal discovery is excluded, this difference mainly reflects the additional operations required to utilize window-specific causal graphs, perform dynamic causal graph-guided augmentation, and compute multi-component anomaly scores. Nevertheless, DCCAD remains substantially faster than CrossAD and CARLA: its runtime is approximately 61.4% of that of CrossAD and less than 36.1% of the lower bound reported for CARLA. CARLA had not completed after 4500 seconds, so its actual runtime exceeds the value reported in the table. Overall, when causal graphs are constructed in advance and reused, DCCAD maintains practical computational efficiency while supporting dynamic causal graph-guided augmentation and relation-aware anomaly scoring.

C.6 Dynamic Causal Graph Visualization

Figure 5 illustrates the temporal evolution of the dynamic causal graph around an anomalous interval on SMD_3-6. As shown in Figure 5(a), several variables exhibit noticeable state changes after the anomaly begins. Correspondingly, the causal influence change and fused graph change score in Figure 5(b) show prominent peaks at the anomaly onset and during subsequent state transitions. Figure 5(c) further shows that the ratio of changed causal edges increases during the anomalous in-

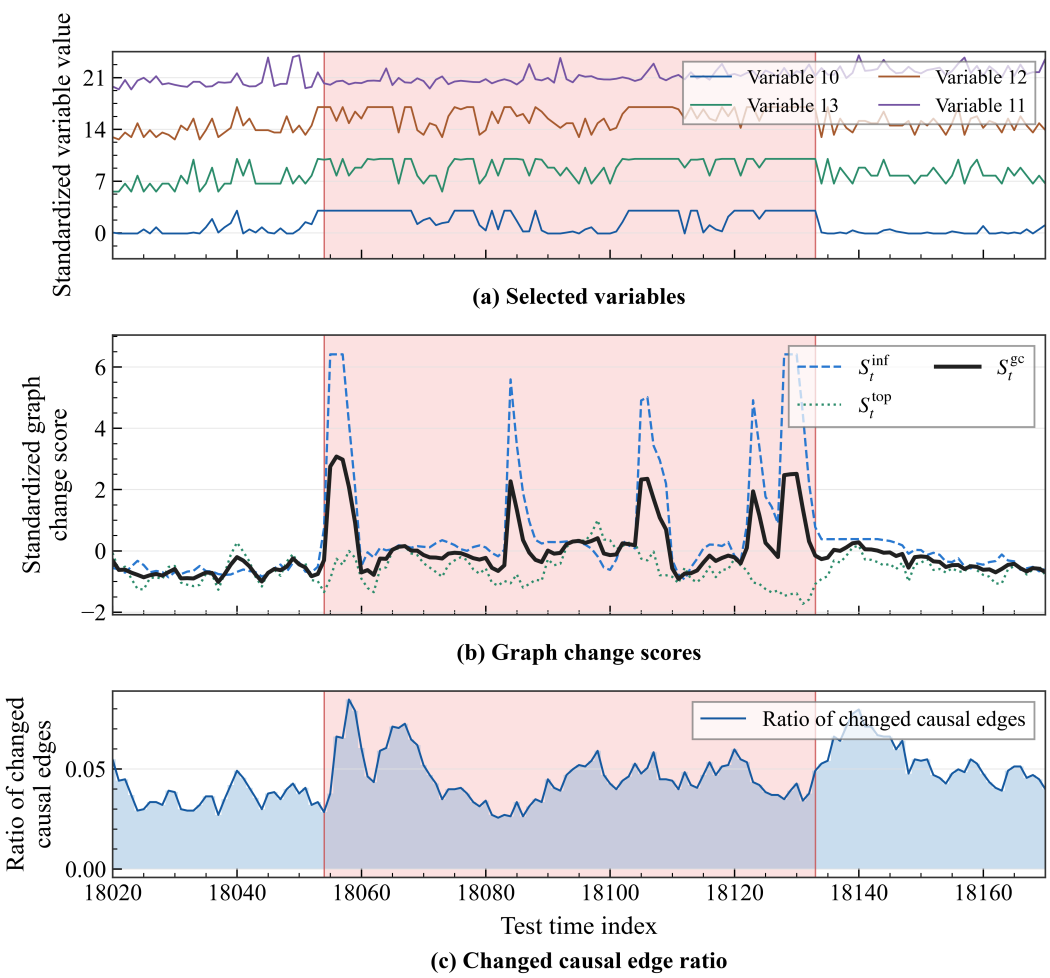


Figure 5: Dynamic causal graph changes around an anomalous interval on SMD_3-6. The shaded region denotes the ground-truth anomaly.

terval, indicating that the anomaly affects not only individual observations but also the dependencies among variables. The edge-change ratio remains temporarily elevated after the anomaly, reflecting the reorganization of causal dependencies as the system returns to normal operation. These results demonstrate that dynamic causal graphs can capture transient dependency changes that would be averaged out by a static graph, thereby providing useful complementary information for anomaly detection.

Figure 6 further compares the mean weighted causal graphs before, during, and after the same anomalous interval. All three matrices use the same logarithmic color scale, allowing their causal influence strengths to be compared directly. During the anomaly, causal influences become substantially stronger and more widespread than those observed before the anomaly, indicating a pronounced disruption of the normal dependency structure. After the anomaly, the overall influence strength decreases and the graph pattern returns close to its pre-anomaly state. This transition from a normal graph to a disrupted graph and then back toward the original structure provides direct evidence that inter-variable causal relationships vary over time. It also supports the use of dynamic causal graphs and graph change information for detecting anomalies that alter the underlying system dynamics.

C.7 Entity-Level Results on MSL and SMD

Table 14 compares CAROTS and DCCAD on all MSL and SMD entities. On MSL, DCCAD improves the average AUROC, AUPRC, and F1 from 0.5880, 0.2211, and 0.3413 to 0.6619, 0.2591,

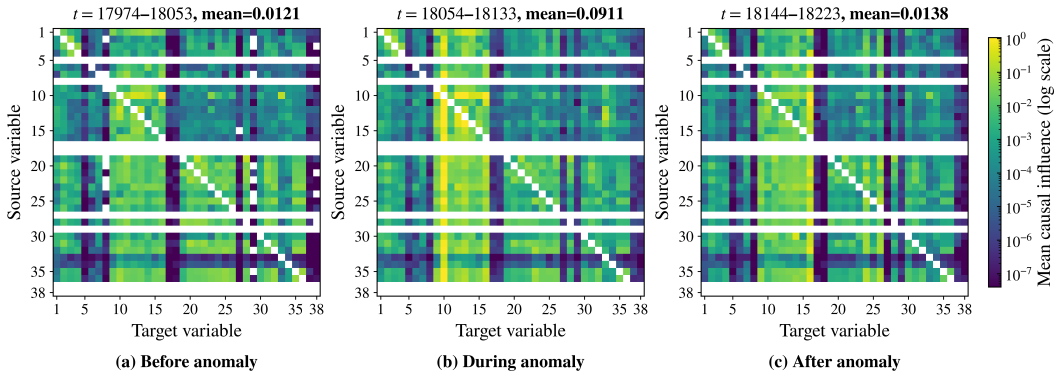


Figure 6: Mean weighted causal graphs before, during, and after an anomalous interval on SMD_3-6. All matrices share the same logarithmic color scale.

Table 14: Entity-level comparison between CAROTS and DCCAD on MSL and SMD. Each entry reports CAROTS/DCCAD, with the higher value in bold.

MSL				SMD			
Entity	**AUROC**	**AUPRC**	**F1**	**Entity**	**AUROC**	**AUPRC**	**F1**
M-1	0.2728 / **0.5866**	0.3776 / **0.5606**	0.6729 / **0.7062**	1-1	**0.8886** / 0.8798	0.5195 / **0.5268**	**0.5206** / 0.5030
M-2	0.7444 / **0.8272**	0.6894 / **0.7874**	0.7743 / **0.7767**	1-2	0.8840 / **0.8992**	**0.2398** / 0.1548	**0.3672** / 0.2524
M-3	**0.5553** / 0.5083	**0.1845** / 0.1232	**0.3386** / 0.2385	1-3	0.6569 / **0.7240**	0.1641 / **0.2773**	0.1983 / **0.3482**
M-4	**0.6011** / 0.5597	**0.2688** / 0.1428	**0.4691** / 0.2604	1-4	**0.7392** / 0.6896	**0.1795** / 0.1631	**0.2944** / 0.2624
M-5	**0.5852** / 0.5694	0.1670 / **0.1789**	**0.2823** / 0.2822	1-5	**0.9235** / 0.9180	0.5174 / **0.5908**	0.5962 / **0.6377**
M-6	0.9686 / **0.9785**	0.8980 / **0.9327**	**0.9415** / **0.9415**	1-6	0.8866 / **0.9067**	0.8066 / **0.8583**	0.7624 / **0.8572**
M-7	0.7274 / **0.7556**	0.5104 / **0.5160**	**0.5976** / 0.5949	1-7	0.8660 / **0.8800**	0.6200 / **0.6680**	**0.7126** / 0.7080
C-1	0.4865 / **0.5530**	0.1525 / **0.1784**	0.2568 / **0.2886**	1-8	**0.6146** / 0.5941	**0.2423** / 0.2304	**0.2787** / 0.2627
C-2	0.5301 / **0.6174**	0.0964 / **0.1239**	0.2243 / **0.2257**	2-1	**0.6992** / 0.6153	0.1689 / **0.1774**	**0.2479** / 0.2338
D-14	**0.8086** / 0.5046	**0.2104** / 0.1246	**0.4857** / 0.1724	2-2	**0.7457** / 0.6543	**0.2364** / 0.1853	**0.3633** / 0.2888
D-15	0.4751 / **0.5957**	0.3100 / **0.3636**	0.4663 / **0.5799**	2-3	**0.9099** / 0.8579	**0.5256** / 0.3286	**0.5427** / 0.3984
D-16	0.6225 / **0.6576**	0.3788 / **0.4197**	0.4973 / **0.5231**	2-4	**0.7298** / 0.6966	0.3438 / **0.3643**	0.3512 / **0.3944**
F-4	**0.7840** / 0.7462	**0.0887** / 0.0512	**0.1957** / 0.1241	2-5	0.8591 / **0.9350**	0.4803 / **0.5421**	0.4797 / **0.5851**
F-5	0.3940 / **0.7225**	0.0375 / **0.0851**	0.0897 / **0.2473**	2-6	**0.8853** / 0.8805	**0.4381** / 0.3525	**0.4698** / 0.3974
F-7	0.3443 / **0.6292**	0.0629 / **0.1139**	0.1700 / **0.2711**	2-7	0.8176 / **0.9497**	0.5385 / **0.7060**	0.5998 / **0.6680**
F-8	0.4541 / **0.4962**	0.2079 / **0.2231**	0.3578 / **0.3581**	2-8	0.9896 / **0.9903**	**0.9648** / 0.9592	0.9451 / **0.9538**
P-10	0.8075 / **0.8726**	0.1571 / **0.3045**	0.3581 / **0.3610**	2-9	0.9645 / **0.9706**	0.8095 / **0.8413**	0.7428 / **0.7725**
P-11	0.4247 / **0.7363**	0.0675 / **0.1325**	0.1542 / **0.2811**	3-1	0.7091 / **0.8222**	0.3792 / **0.5895**	0.4665 / **0.6441**
P-14	0.7642 / **0.7831**	0.3721 / **0.4459**	**0.5448** / 0.5324	3-2	**0.6360** / 0.5059	**0.1003** / 0.0923	**0.1373** / 0.1353
P-15	0.6911 / **0.7735**	0.0213 / **0.3738**	0.0863 / **0.5502**	3-3	0.6498 / **0.8034**	0.1197 / **0.1930**	0.2217 / **0.3114**
S-2	0.5556 / **0.5964**	0.0145 / **0.0176**	0.0533 / **0.0822**	3-4	**0.9142** / 0.9086	**0.8474** / 0.8197	**0.8780** / 0.8504
T-4	**0.5607** / 0.4504	**0.0415** / 0.0299	**0.0944** / 0.0736	3-5	0.8242 / **0.8361**	0.2696 / **0.2876**	0.3483 / **0.4151**
T-5	0.4503 / **0.6483**	0.0165 / **0.0261**	0.0475 / **0.0909**	3-6	0.9201 / **0.9696**	0.4657 / **0.6180**	0.4345 / **0.6317**
T-8	0.4897 / **0.5538**	0.0834 / **0.1152**	0.1520 / **0.1638**	3-7	0.7347 / **0.7617**	**0.4220** / 0.2219	**0.5399** / 0.3135
T-9	**0.7467** / 0.6816	**0.3516** / 0.2865	**0.5132** / 0.4221	3-8	0.8528 / **0.9142**	0.2858 / **0.3194**	0.3367 / **0.4613**
T-12	0.6589 / **0.7486**	0.0888 / **0.1202**	0.1896 / **0.2193**	3-9	0.6861 / **0.8185**	0.3207 / **0.5499**	0.4311 / **0.6266**
T-13	0.3712 / **0.7204**	0.1161 / **0.2182**	0.2031 / **0.3693**	3-10	**0.8085** / 0.7764	0.5310 / **0.5796**	0.6353 / **0.6570**
				3-11	0.8724 / **0.9000**	0.2378 / **0.2897**	**0.4473** / 0.4441
Average	0.5880 / **0.6619**	0.2211 / **0.2591**	0.3413 / **0.3606**	**Average**	0.8096 / **0.8235**	0.4205 / **0.4460**	0.4768 / **0.5005**

and 0.3606, respectively. It obtains higher AUROC on 20 of 27 entities, higher AUPRC on 21, and higher F1 on 17, with one F1 tie. The gains are therefore broad but not uniform: CAROTS remains stronger on several entities, including D-14, F-4, T-4, and T-9, whereas DCCAD performs particularly well on M-6.

On SMD, DCCAD also yields higher average AUROC, AUPRC, and F1, increasing them from 0.8096, 0.4205, and 0.4768 to 0.8235, 0.4460, and 0.5005, respectively. The margins are smaller than on MSL, reflecting more mixed entity-level results: DCCAD leads on 16 of 28 entities in AUROC, 18 in AUPRC, and 15 in F1. For example, on 2-8, DCCAD achieves higher AUROC and F1, while CAROTS is slightly better in AUPRC. Overall, DCCAD provides better average performance across both datasets, while CAROTS remains competitive on particular entities and metrics. These results complement the representative subsets in the main text with a more complete comparison across heterogeneous entities.

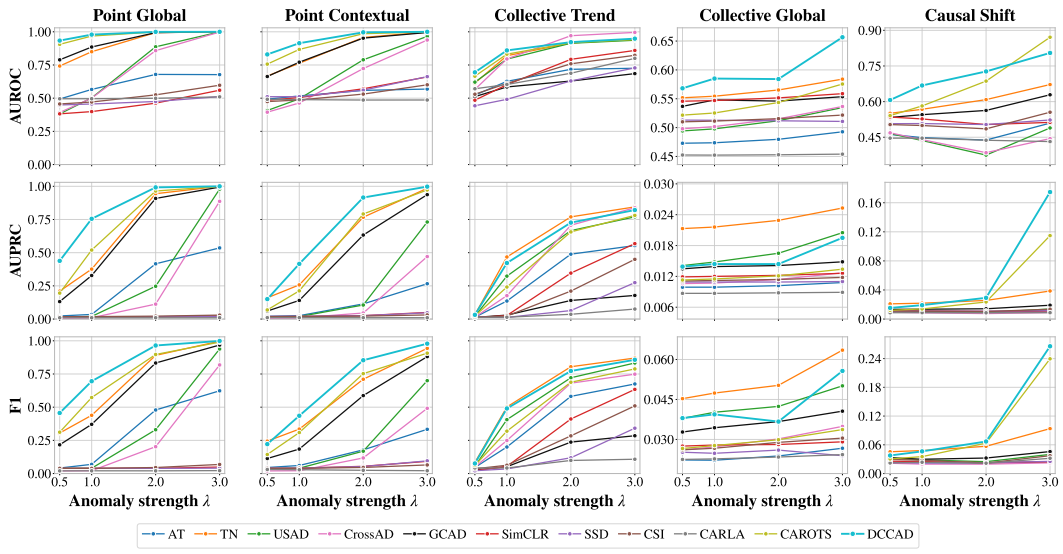


Figure 7: Performance comparison on the Lorenz-96 synthetic datasets across different anomaly types and evaluation metrics.

C.8 Lorenz-96 Results across Anomaly Strengths

Figure 7 compares all methods on the Lorenz-96 synthetic dataset across five anomaly types and four anomaly strengths. The rows report AUROC, AUPRC, and F1, while the columns correspond to point global (PG), point contextual (PC), collective trend (CT), collective global (CG), and causal shift (CS) anomalies. Performance generally improves with anomaly strength, but the trend varies substantially across anomaly types, metrics, and methods.

DCCAD shows the most consistent overall advantage on PG and PC, particularly under weak perturbations, indicating greater sensitivity to subtle deviations while preserving normal inter-variable dependencies. It also achieves leading or competitive performance on CT across different strengths, showing its ability to identify persistent temporal changes. On CG, DCCAD achieves the strongest AUROC across all strengths and improves markedly at $\lambda = 3.0$. However, AUPRC and F1 remain relatively low across all methods, indicating that precise detection of CG anomalies remains challenging despite DCCAD's improved anomaly ranking.

For CS, all methods achieve relatively low AUPRC and F1 at weak strengths, confirming the difficulty of precisely detecting subtle causal-mechanism changes. As the shift strengthens, DCCAD's AUPRC and F1 increase sharply and exceed those of the baselines at $\lambda = 3.0$. Its AUROC is also the highest or competitive across the evaluated strengths. These results show that DCCAD is sensitive to sufficiently strong changes in the underlying dependency mechanism, while weak causal shifts remain challenging.