# Causality-Aware Contrastive Learning for Robust Multivariate Time-Series Anomaly Detection

HyunGi Kim [1] Jisoo Mok [1] Dongjun Lee [2] Jaihyun Lew [2] Sungjae Kim [3] Sungroh Yoon [1 2 4]

## Abstract

Utilizing the complex inter-variable causal relationships within multivariate time-series provides a promising avenue toward more robust and reliable multivariate time-series anomaly detection (MTSAD) but remains an underexplored area of research. This paper proposes **C**ausality-**A**ware contrastive learning for **RO**bust multivariate **T**ime-**S**eries (CAROTS), a novel MTSAD pipeline that incorporates the notion of causality into contrastive learning. CAROTS employs two data augmentors to obtain causality-preserving and -disturbing samples that serve as a wide range of normal variations and synthetic anomalies, respectively. With causality-preserving and -disturbing samples as positives and negatives, CAROTS performs contrastive learning to train an encoder whose latent space separates normal and abnormal samples based on causality. Moreover, CAROTS introduces a similarity-filtered one-class contrastive loss that encourages the contrastive learning process to gradually incorporate more semantically diverse samples with common causal relationships. Extensive experiments on five real-world and two synthetic datasets validate that the integration of causal relationships endows CAROTS with improved MTSAD capabilities. The code is available at https://github.com/kimanki/CAROTS.

## 1. Introduction

Multivariate time-series anomaly detection (MTSAD) aims to identify anomalies in time-series collected across multiple interdependent variables (Li & Jung, 2023; Choi et al., 2021). Applications of MTSAD are far-reaching, spanning from detecting network intrusion in cyber-physical systems (Nandanwar & Katarya, 2024) to monitoring patient vitals in healthcare systems (Clifford et al., 2017; Ben-Moshe et al., 2024). Since undetected anomalies can cause serious harm in mission-critical systems, MTSAD plays a crucial role in ensuring their operational efficiency and security. Furthermore, as such systems continue to grow in size and complexity, with an increasing number of variables, understanding the intricate inter-variable relationships rises as a critical challenge in MTSAD (Kang & Kang, 2024).

Causality offers a structured approach to capture the underlying relationships within multivariate time-series (Runge et al., 2019). In multivariate time-series, the observed variables are governed by causal relationships that dictate how changes in one variable affect others (Granger, 1969; Cheng et al., 2024b). For instance, the increase in temperature directly causes the duration for which the air conditioner stays on to be longer. Once an anomaly arises, provoking an unexpected behavior from one or more variables, these causal relationships break down (*e.g.*, a broken air conditioner will remain turned off regardless of the temperature). Except in the case of such an abnormal event, the inter-variable causal relationships generally remain consistent, and thus, a violation of causal relationships can be used as a key indicator of anomalies (Liu et al., 2025; Febrinanto et al., 2023).

Understanding the complex inter-variable dynamics from the perspective of causality enables a more reliable and robust MTSAD by allowing discernment of causality-preserving variations of normal instances (*e.g.,* the air-conditioner staying on for longer than usual due to heat-waves) from true anomalies with broken causality. However, current unsupervised MTSAD methods, such as reconstruction- or contrastive learning-based approaches, focus instead on superficial differences in data values or distributions and overlook inter-variable causal relationships (Dai et al., 2024; Xu et al., 2022; Kim et al., 2023). Failing to account for inter-variable causal relationships can lead to

[1] Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea [2] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea [3] Hyundai Motor Company, Gyeonggi-do, Republic of Korea [4] AIIS, ASRI, and INMC, Seoul National University, Seoul, Republic of Korea. Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

*Proceedings of the $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

misidentifying normal variations as anomalies, resulting in false alarms and reduced detection accuracy.

In this work, we propose a novel causality-aware MTSAD framework dubbed CAROTS, which is a shorthand for **C**ausality-**A**ware contrastive learning for **RO**bust M**TS**AD. CAROTS first extracts the causal relationships among normal multivariate time-series by leveraging a forecasting-based causal discovery model. Afterward, it augments the training data with both causality-preserving and -disturbing samples. The former represents a diversified set of normal, causality-preserving operations, while the latter simulates anomalies that violate causal relationships. To explicitly separate normal operations from anomalies based on the causal structure, CAROTS utilizes contrastive learning, where causality-preserving and -disturbing samples serve as positive and negative examples, respectively. In the embedding space of this contrastively-trained encoder, the causality-preserving and -disturbing samples form their own clusters that are clearly distinguishable from each other.

In addition, we introduce a similarity-filtered one-class contrastive loss, which guides the contrastive learning process to gradually incorporate more semantically diverse samples that share common normal causal relationships. With each sample in a mini-batch as an anchor, the similarity filter excludes low-similarity positive samples, which fall below a pre-set threshold, from the loss computation. Thus, the similarity filter only treats samples with highly similar patterns as positives in early iterations. As training progresses, positive samples with more diverse patterns are incorporated, allowing the trained encoder to include a broader range of normal variations. Finally, CAROTS performs anomaly detection by ensembling two causality-aware anomaly scores: the distance between a test time-series and the centroid of normal, causality-preserving embeddings obtained with the contrastively-trained encoder and the forecasting error of the forecasting-based causal discovery model.

We validate the effectiveness of CAROTS across five real-world MTSAD datasets, where it consistently outperforms existing MTSAD methods. On two synthetic datasets with explicit linear and non-linear causal relationships, CAROTS again significantly outperforms methods that overlook causal relationships. In particular, CAROTS successfully detects even the most difficult synthetic anomalies, which other baselines all fail to identify; the remarkable improvement in AUROC up to 50% verifies that the integration of causal knowledge endows CAROTS with the ability to reliably detect anomalies. We demonstrate that CAROTS is robust to changes in hyperparameters and encoder architecture to showcase its practical applicability. Lastly, the ablation study on each technical component of CAROTS verifies its individual contributions.

We summarize our contributions as follows:

- We propose CAROTS, a novel MTSAD pipeline that leverages causality-preserving and -disturbing augmentors to induce causal-awareness in contrastive learning. The trained encoder successfully separates normal and abnormal samples based on causal relationships.
- We propose a similarity-filtered one-class contrastive loss, which enforces the latent space to respect diverse semantic patterns by progressively incorporating more dissimilar samples throughout training.
- An extensive empirical verification of CAROTS on five real-world and two synthetic datasets highlights its outstanding ability to detect anomalies, especially those that previous baselines thoroughly struggle with.

## 2. Related Works

### 2.1. Multivariate Time-series Anomaly Detection

**Reconstruction-based methods.** Reconstruction-based MTSAD methods train a reconstruction model exclusively on normal data (Wang et al., 2025; Dai et al., 2024). Anomalies are detected based on the reconstruction error of this model, under the assumption that they will yield higher reconstruction errors. This line of research mostly focuses on utilizing advanced architectures. AnomalyTransformer (Xu et al., 2022) introduces anomaly attention which models prior and series associations to distinguish normal and anomalous points. TimesNet (Wu et al., 2023) exploits TimesBlock that learns multi-scale temporal dependencies based on Fourier transform. Aside from studying architectural variants, USAD (Audibert et al., 2020) uses adversarial training to build a more reliable reconstruction model. However, the sole reliance on normal training samples limits their ability to fully consider the variability of normal operations or integrate insights related to potential anomalies.

**Contrastive learning-based methods.** Contrastive learning is a self-supervised learning approach that aims to learn latent representations of data by contrasting positive and negative sample pairs (Chen et al., 2020; He et al., 2020; Lee et al., 2024). Its usefulness in anomaly detection has been shown across various domains (Tack et al., 2020; Sehwag et al., 2021). In anomaly detection, contrastive learning is used to train an encoder that can distinguish normal and abnormal patterns in a self-supervised manner. The negative samples in contrastive MTSAD allow the integration of simulated anomalies during the training phase, unlike in reconstruction-based methods (Darban et al., 2025; Kim et al., 2023). The efforts to advance contrastive MTSAD can be categorized into: data augmentation techniques (Ky et al., 2024; Choi et al., 2024), integration of one-class learning (Kim et al., 2023; Xu et al., 2024), and self-supervised classification (Yang et al., 2023; Ngu & Lee, 2023).

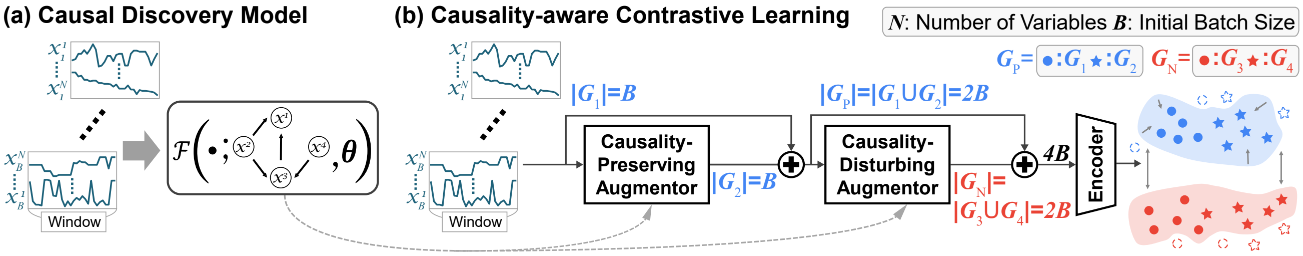
(a) Causal Discovery Model
(b) Causality-aware Contrastive Learning
N: Number of Variables B: Initial Batch Size
$G_P$= ●:$G_1$ ★:$G_2$
$G_N$= ●:$G_3$ ★:$G_4$
$x_1^1$
$x_1^N$
$x_B^N$
$x_B^1$
Window
$\mathcal{F}$
θ
$|G_1|=B$
$|G_2|=B$
$|G_P|=|G_1 \cup G_2|=2B$
$|G_N|=|G_3 \cup G_4|=2B$
Causality-Preserving Augmentor
Causality-Disturbing Augmentor
4B
Encoder

*Figure 1.* An overall pipeline of CAROTS. **(a)** First, a causal discoverer is trained to learn the normal causal relationships in the training data. **(b)** Based on the causal discoverer, the causality-preserving and -disturbing augmentors construct a mini-batch of four groups of samples: original ($G_1$), causality-preserving augmentations ($G_2$), causality-disturbing augmentations of $G_1$ ($G_3$), and causality-disturbing augmentations of $G_2$ ($G_4$). Contrastive learning with the similarity-filtered one-class contrastive loss clusters positive groups ($G_P \coloneqq G_1 \cup G_2$) while separating them from ($G_N \coloneqq G_3 \cup G_4$) in the embedding space. $N$, $B$, and $\oplus$ denotes the number of variables, initial batch size, and concatenation operation, respectively.

Our work sets itself apart from previous contrastive MTSAD by integrating the concept of causality into the data augmentation process. In contrastive MTSAD, data augmentation functions that reflect the traits of time-series are preferable for creating positive and negative examples that simulate normals and abnormal samples. In this vein, CL-TAD (Ngu & Lee, 2023) leverages encoder-decoder-based reconstruction with randomly masked inputs, while CTAD (Kim et al., 2023) offers guidelines for using general time-series augmentations for contrastive MTSAD. CARLA (Darban et al., 2025) introduces various types of synthetic anomalies, such as point, contextual, and collective anomalies, to improve anomaly detection. In contrast, our approach employs causality-preserving and causality-disturbing augmentors, allowing the model to explicitly exploit causal relationships, an essential characteristic of multivariate time-series.

### 2.2. Causal Discovery in Multivariate Time-series

Causal discovery uncovers causal relationships among interdependent variables in multivariate time-series (Assaad et al., 2022; Cheng et al., 2024b). The Granger causality test (Granger, 1969) has been widely adopted to infer causal relationships by testing whether past values of one variable improve the prediction of another. Recently, deep learning approaches have been used for more robust causal discovery (Tank et al., 2021; Cheng et al., 2023; 2024a). NGC (Tank et al., 2021) infers non-linear causal relationships in multivariate time-series by combining structured neural networks with sparsity-inducing regularization. CUTS (Cheng et al., 2023) jointly imputes missing data and constructs causal graphs, and CUTS+ (Cheng et al., 2024a) extends this by improving scalability for high-dimensional data with a Coarse-to-fine Discovery technique and a Message-Passing-based Graph Neural Network. Causalformer (Kong et al., 2024) integrates causal discovery into Transformer, learning sparse adjacency matrices to capture temporal and causal relationships. Contrary to the exhaustive efforts toward causal discovery in time-series, leveraging causal relationships to distinguish anomalies from normal operations in MTSAD remains underexplored (Qiu et al., 2012; Xing et al., 2023; Liu et al., 2025).

## 3. Preliminaries

### 3.1. Causal Discovery Model

A multivariate time-series is defined as a collection of time-series obtained from $N$ variables $\{x^i\}_{i=1}^N$, where $N > 1$. According to the Granger causality (Granger, 1969), $x^i$ causes $x^j$ if the past values of $x^i$ affect the future values of $x^j$. In this case, $x^i$ is the cause of $x^j$ while $x^j$ is the effect of $x^i$. We denote a set of causes and effects of $x^i$ as $\text{Pa}(x^i)$ and $\text{Ch}(x^i)$, respectively. The objective of a causal discovery model is to search for $\text{Pa}(x^i)$ and $\text{Ch}(x^i)$ for every $x^i$ and to learn the function $f^i$ that maps the causal relationships between $x^i$ and $\text{Pa}(x^i)$. Given $\text{Pa}(x^i)$ and $f^i$, the value of $x^i$ at time $t$, denoted as $x_t^i$, can be represented as $x_t^i = f^i(\boldsymbol{x}_{<t}^i, \text{Pa}(x^i)_{<t})$ where $\boldsymbol{x}_{<t}^i$ denotes the values of $x^i$ preceding $t$.

The forecasting-based causal discovery model $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ represents the causal relationships with the causality matrix $\boldsymbol{A} \in \{0,1\}^{N\times N}$ and causal relationship functions $\{f^i(\cdot)\}_{i=1}^N$. The causality matrix $\boldsymbol{A}$ represents $\text{Pa}(x^i)$ and $\text{Ch}(x^i)$, where $\boldsymbol{A}_{i,j} = 1$ holds if and only if $x^i \in \text{Pa}(x^j)$. To train $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$, a single, contiguous multivariate time-series $\{\boldsymbol{x}_t\}_{t=1}^T$ is segmented with a sliding window with window size $w$ such that $\boldsymbol{X}_t = \{\boldsymbol{x}_{t-w+1}, \dots, \boldsymbol{x}_t\} \in \mathbb{R}^{w\times N}$, where $\boldsymbol{x}_t \in \mathbb{R}^N$ denotes the values of $N$ variables at time $t$. $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ is trained to minimize the forecasting Mean Squared Error (MSE) as follows:

$$\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{A}} = \underset{\boldsymbol{\theta},\boldsymbol{A}}{\text{argmin}} \underset{\boldsymbol{X}_t}{\mathbb{E}} \left[\text{MSE}\left(\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}\left(\boldsymbol{X}_{<t}\right), \boldsymbol{x}_t\right) + \lambda ||\boldsymbol{A}||_1\right], \tag{1}$$

where $\boldsymbol{X}_{<t}$ denotes values in $\boldsymbol{X}_t$ preceding $t$. $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ approximates $\{f^i(\cdot)\}_{i=1}^N$ that predicts $\boldsymbol{x}_t$ based on the preceding values $\boldsymbol{X}_{<t}$ and $\boldsymbol{A}$. Our work utilizes CUTS+ (Cheng

et al., 2024a), a representative forecasting-based causal discovery model, to extract meaningful causal structures.

### 3.2. MTSAD through the Lens of Causality

Following the general assumption in time-series analysis (Absar & Zhang, 2021; Absar et al., 2023), we assume the causal relationships in normal multivariate time-series remain consistent over time (*i.e.,* $\boldsymbol{A}$ and $\{f^i(\cdot)\}_{i=1}^N$ are time-invariant). From the causal perspective, multivariate time-series anomalies can then be interpreted as values that deviate from the normal causal relationships:

$$\exists x_t^i \,:\, |x_t^i - f^i(x_{<t}^i, \mathrm{Pa}(x^i)_{<t})| > \delta. \tag{2}$$

In this study, we harness this causal interpretation of anomalies by integrating the notion of causality into contrastive MTSAD. By exploiting synthetic samples that either preserve normal, consistent causal relationships or violate them as in Eq. 2, we enable an explicitly causal-aware discrimination of anomalies in MTSAD.

## 4. Methodology

We now introduce CAROTS, **C**ausality-**A**ware contrastive learning for **RO**bust **MTS**AD. Section 4.1 discusses how the causal relationships, extracted by a causal discoverer, are employed for causality-aware data augmentation. Section 4.2 elaborates how the causality-informed discriminative embedding space is obtained through contrastive learning with the proposed similarity-filtered one-class contrastive loss. Lastly, Section 4.3 describes how the anomalies are scored based on the overall framework. Figure 1 illustrates the overall pipeline of CAROTS.

### 4.1. Causality-aware Data Augmentation

The process in Section 3 yields a causality matrix $\boldsymbol{A}$, which encodes the causal interactions among variables, and a causality-integrated forecaster $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ that predicts future values given these causal relationships. Because unsupervised MTSAD assumes that training data are void of anomalies, the learned $\boldsymbol{A}$ and $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ represent normal causal relationships. With $\boldsymbol{A}$ and $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$, CAROTS generates causality-preserving and -disturbing samples with respective augmentors, visualized in Figure 2.

**Causality-Preserving Augmentor (CPA)** generates positive samples with intact causal relationships. Causality-preserving samples from CPA increase the diversity of normal data that maintain causal relationships and simulate wide-ranging normal data not included in the training set. For a given input $\boldsymbol{X}_t$, CPA randomly selects a set of $M < N$ causing variables $\boldsymbol{C}$ such that $\boldsymbol{C} \subset \{x^i\}_{i=1}^N$, $|\boldsymbol{C}| = M$. Then, a set of variables directly affected by $\boldsymbol{C}$, defined as $\boldsymbol{E} := \{\mathrm{Ch}(x^j)\,|\,x^j \in \boldsymbol{C}\}$, is identified based on $\boldsymbol{A}$. CPA adds small Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$ to trigger variations in the causing variables. The modified causing variables are fed into $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$, which predicts new values of $\boldsymbol{E}$ according to the learned causal relationship. By replacing the original values of $\boldsymbol{E}$ with the predictions of $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$, CPA obtains new causality-preserving samples that maintain the original causal relationships as the following:

$$\begin{aligned}
(\boldsymbol{X'_{<t}})^i &= \begin{cases} (\boldsymbol{X}_{<t})^i + z & \text{if } \boldsymbol{x}^i \in \boldsymbol{C} \\ (\boldsymbol{X}_{<t})^i & \text{otherwise,} \end{cases} \\
(\boldsymbol{x'_t})^j &= \begin{cases} (\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\boldsymbol{X'_{<t}}))^j & \text{if } \boldsymbol{x}^j \in \boldsymbol{E} \\ \boldsymbol{x}_t^j & \text{otherwise,} \end{cases} \\
\mathrm{CPA}(\boldsymbol{X}_t) &= \mathrm{Concat}\left(\{\boldsymbol{X'_{<t}}, \boldsymbol{x'_t}\}\right).
\end{aligned} \tag{3}$$

**Causality-Disturbing Augmentor (CDA)** generates negative samples with broken causal relationships. Causality-disturbing samples from CDA act as proxies for anomalies caused by disrupted causal relationships. Multivariate time-series can be represented in a directed graph, whose nodes are all variables and edges are formed according to $\boldsymbol{A}$. Like in CPA, CDA first selects a random set of causing variables $\boldsymbol{C}$. Beginning from $\boldsymbol{C}$, CDA performs a depth-first search (DFS) to generate multiple subgraphs as shown in Figure 2. Each iteration of DFS terminates with a cut-off probability of $p$, allowing subgraphs of varying sizes to be extracted. CDA perturbs the variables in the extracted subgraph by injecting random biases so that the original causal relationships no longer hold.

### 4.2. Causality-aware Contrastive Learning with Similarity-filtered One-class Contrastive Loss

With samples from CDA and CPA, CAROTS performs contrastive learning to train an encoder $\mathcal{E}_{\boldsymbol{\phi}}(\cdot)$ that explicitly separates normal, causality-preserving samples from abnormal, causality-disturbing samples. In the latent space of the resulting encoder, trained with causality-aware contrastive learning, the embeddings of causality-preserving and -disturbing samples appear far apart from each other.

Given a mini-batch of $B$ samples ($G_1$), the CPA first generates $B$ additional causality-preserving samples ($G_2$). Afterward, the CDA is applied simultaneously to $G_1$ and $G_2$ to obtain causality-disturbing samples from both original and causality-preserving augmented samples. As a result, CDA yields a total of $2B$ causality-disturbing samples. The four types of data – $G_1$, $G_2$, $B$ causality-disturbing samples from original samples ($G_3$), and $B$ causality-disturbing samples from causality-preserving augmented samples ($G_4$) – form an augmented mini-batch $\mathcal{B}$:

$$\mathcal{B} = \mathrm{Concat}\left(\{G_1, G_2, G_3, G_4\}\right), \ |\mathcal{B}| = 4B. \tag{4}$$

These four groups of data are used for positive and negative

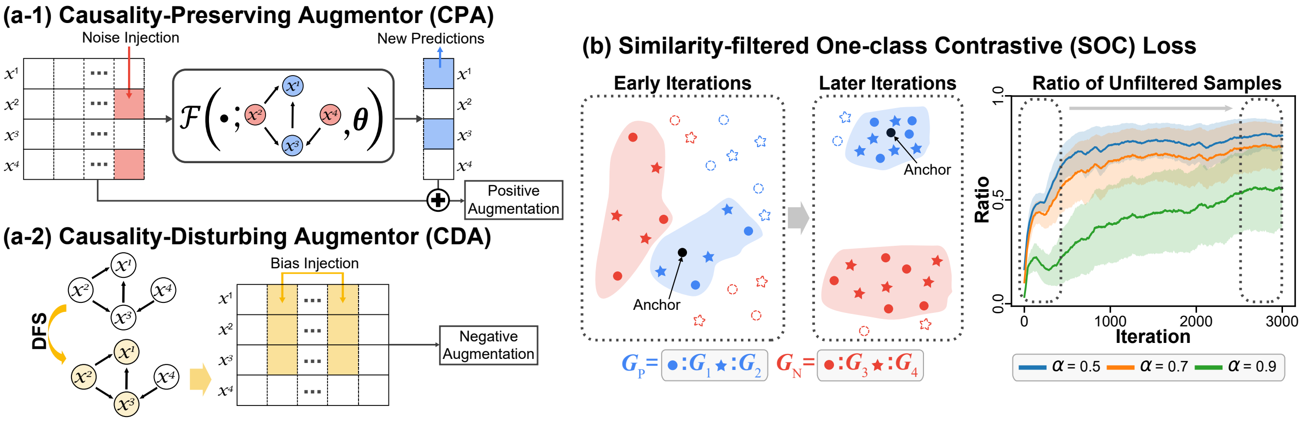

*Figure 2.* **(a-1)** CPA adds Gaussian noise to randomly selected variables (Red) and uses the causality-integrated forecaster to predict and replace the values of affected variables (Blue). **(a-2)** CDA selects a random variable (Yellow) and extracts a directed subgraph of a causality matrix through DFS. Random perturbations that are unrelated to causal relationships are injected into the selected variables.**(b)** [Left] In early iterations, SOC loss filters out low-similarity positive samples (Dashed) based on a threshold $\alpha$, ensuring that the loss is applied only to high-similarity positive samples and their causality-disturbed counterparts (Solid-colored). As training progresses, more diverse samples engage in the SOC loss computation. [Right] shows that a ratio of unfiltered samples steadily increases during training.

samples in contrastive learning as:

$$\begin{aligned} G_P &\coloneqq G_1 \cup G_2 \\ G_N &\coloneqq G_3 \cup G_4, \end{aligned} \tag{5}$$

where $G_P$ and $G_N$ denote a set of positive and negative samples, respectively. $G_P$ allows the encoder to effectively embed a wide variety of causality-preserving samples, including those not immediately available as a part of the original training data. $G_N$ enables the encoder to distinguish anomalies caused by disturbances in causal relationships. In particular, we note that $G_2$ and its causality-disturbing counterparts $G_4$ equip the encoder to generalize to normal variations and anomalies that arise from them.

In $G_P$, semantically diverse samples that share common causal relationships exist (*e.g.* the causal relationship between the temperature and the air conditioner usage is preserved, but diverse patterns are generated depending on the air conditioner setting, such as the wind strength). Therefore, $G_P$ can be grouped into multiple clusters based on the semantic similarity of individual samples. However, the naïve one-class contrastive loss forces all samples in the $G_P$ to collapse into a single cluster, which in turn prevents the encoder's embedding space from encoding the semantic diversity of samples belonging to $G_P$.

To address this pitfall, we introduce **Similarity-filtered One-class Contrastive Loss** (SOC), whose overall scheme is illustrated in Figure 2. At each update step, SOC filters out positive samples in $G_P$ whose similarity with an $i$-th anchor sample is lower than a pre-defined threshold $\alpha$. Then, SOC applies the one-class contrastive loss only to the remaining positive and corresponding negative samples, whose indices are denoted as $P_i$ and $N_i$, respectively. The proposed SOC is formulated as follows:

$$\begin{aligned} \mathcal{L} = \frac{1}{2B} \sum_{i=1}^{2B} \frac{1}{|P_i|} \sum_{j \in P_i} - \log \frac{\exp(\mathcal{S}_{i,j})}{\exp(\mathcal{S}_{i,j}) + \sum_{k \in N_i} \exp(\mathcal{S}_{i,k})}, \\ P_i = \{ j \in \{1, 2, \ldots, 2B\} \mid \mathcal{S}_{i,j} \geq \alpha / \tau \}, \\ N_i = \{ j + 2B \mid j \in P_i \}, \end{aligned} \tag{6}$$

where $\mathcal{S}_{i,j}$ denotes cosine similarity between embeddings of $i$-th and $j$-th sample in $\mathcal{B}$ divided by temperature $\tau$. SOC encourages each anchor sample to be aggregated with semantically similar samples first, allowing the encoder to respect different semantic clusters within $G_p$. As positive embeddings are aggregated during training, the proportion of filtered positive samples decreases, gradually incorporating more diverse samples sharing causal relationships.

### 4.3. Anomaly Score

CAROTS conducts anomaly detection by combining two causality-aware anomaly scores: the distance between a test time-series and the centroid of normal, causality-preserving data embeddings obtained with the contrastively-trained encoder ($\mathcal{A}_{\text{CL}}$) and the forecasting error of the forecasting-based causal discovery model ($\mathcal{A}_{\text{CD}}$).

To compute $\mathcal{A}_{\text{CL}}$, CAROTS relies on the latent space of the contrastively-trained encoder $\mathcal{E}_\phi(\cdot)$ from Section 4.2. Given this latent space, CAROTS obtains the centroid $\boldsymbol{\mu}_P$ of embeddings of all positive samples, including original data. Since the encoder is trained to cluster the embeddings of $G_P$ closely while pushing them away from those of causality-disturbing $G_N$, the distance between a test time-series and $\boldsymbol{\mu}_P$ is indicative of how much the test time-series deviates from causality-preserving positive samples. Consequently, larger $\mathcal{A}_{\text{CL}}$ signifies that the corresponding test time-series is more likely to be anomalies.

CAROTS additionally incorporates $\mathcal{A}_{\text{CD}}$, the forecasting error of forecasting-based causal discovery model $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ in Section 3.1, as an auxiliary anomaly score. $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$, which is trained to predict future values based on learned normal causal relationships, struggles to forecast future values of abnormal samples that deviate from normal causal relationships. Therefore, the forecasting error of $\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}(\cdot)$ provides another crucial signal for causality-driven detection of anomalies following Eq. 2. Like $\mathcal{A}_{\text{CL}}$, test-time time-series with large $\mathcal{A}_{\text{CD}}$ are considered to be anomalies.

To ensemble the two scores, CAROTS normalizes each score to match the scales. Specifically, we calculate the mean and standard deviation of each score over the training data and apply z-normalization. The final anomaly score is obtained by summing the two z-scores as the following:

$$\begin{aligned}
\mathcal{A}(\boldsymbol{X}_t) &= \mathcal{A}_{\text{CL}}^{\text{norm}}(\boldsymbol{X}_t) + \mathcal{A}_{\text{CD}}^{\text{norm}}(\boldsymbol{X}_t), \\
\mathcal{A}_{\text{CL}}(\boldsymbol{X}_t) &= \mathcal{D}\left(\mathcal{E}_{\boldsymbol{\phi}}(\boldsymbol{X}_t), \boldsymbol{\mu}_P\right), \\
\mathcal{A}_{\text{CD}}(\boldsymbol{X}_t) &= \text{MSE}\left(\mathcal{F}_{\boldsymbol{\theta},\boldsymbol{A}}\left(\boldsymbol{X}_{<t}\right), \boldsymbol{x}_t\right),
\end{aligned} \tag{7}$$

where $\mathcal{A}_{\text{CL}}^{\text{norm}}$ and $\mathcal{A}_{\text{CD}}^{\text{norm}}$ denote the z-normalized $\mathcal{A}_{\text{CL}}$ and $\mathcal{A}_{\text{CD}}$, respectively, and $\mathcal{D}$ denotes distance function.

## 5. Experiments

### 5.1. Experimental Setup

**Datasets.** We demonstrate the effectiveness of CAROTS on five widely used real-world MTSAD datasets - SWaT (Goh et al., 2016), WADI (Ahmed et al., 2017), PSM (Abdulaal et al., 2021), SMD (Su et al., 2019), and MSL (Hundman et al., 2018) - and two synthetic datasets - VAR and Lorenz96 (Karimi & Paul, 2010). The real-world datasets are characterized by the intricate causal relationships among variables, which are critical for understanding and detecting anomalies. For example, in SWaT and WADI, the behavior of actuators, such as pumps and valves, is directly influenced by sensor readings such as flow rates and water levels. The two synthetic datasets are generated based on explicit linear and non-linear causal relationships. Following the approach of Lai et al. (2021), test sets for these datasets are constructed by injecting four different types of synthetic anomalies, enabling evaluation of the model performance under diverse anomaly scenarios and difficulties. Detailed descriptions of each dataset are provided in the Appendix.

**Baselines and evaluation metrics.** We select three representative reconstruction-based models - AnomalyTransformer (Xu et al., 2022), TimesNet (Wu et al., 2023), and USAD (Audibert et al., 2020) - and five contrastive learning-based models - SimCLR (Chen et al., 2020), SSD (Sehwag et al., 2021), CSI (Tack et al., 2020), CTAD (Kim et al., 2023), and CARLA (Darban et al., 2025) - as strong baselines for comparison with CAROTS. The MTSAD performance is evaluated using AUROC, AUPRC, and F1 scores.

**Implementation details.** We use 20% of the training data as validation data and apply standard normalization to the entire dataset using the mean and standard deviation of training data. We construct training, validation, and test sets using a sliding window; windows containing at least one anomaly point are labeled as anomalies. Unless specified otherwise, we train models with a window size of 10 and a batch size of 256 for 30 epochs. We report the average value for each metric obtained over three random seeds. For contrastive learning-based models including CAROTS, we primarily employ an LSTM encoder (Hochreiter & Schmidhuber, 1997) following CTAD (Kim et al., 2023) and set the temperature parameter to 0.1. The threshold for similarity filtering is initialized to 0.5 and linearly increased to 0.9.

### 5.2. Results on Real-world Datasets

The results in Table 1 showcase the effectiveness of CAROTS and CAROTS$^\dagger$ (without $\mathcal{A}_{\text{CD}}$) across various real-world MTSAD datasets. CAROTS consistently achieves the highest detection scores in nearly all metrics and datasets, while CAROTS$^\dagger$ comes in a close second. Both CAROTS and CAROTS$^\dagger$ surpassing other baselines underscore that $\mathcal{A}_{\text{CL}}$ from causality-aware contrastive learning is a robust MTSAD score. For instance, CAROTS$^\dagger$ already outperforms other baselines on the particularly challenging WADI and MSL_P-15 datasets, highlighting that it is capable of identifying anomalies in complex environments. Yet, the slight superiority of CAROTS to CAROTS$^\dagger$ implies that $\mathcal{A}_{\text{CD}}$ makes CAROTS even more robust. Notably, in SWaT, CAROTS, with an AUROC of 0.852, AUPRC of 0.764, and F1 score of 0.791, outperforms all competitors. We report full results including standard deviations in the Appendix.

### 5.3. Results on Synthetic Datasets

In Table 2, We compare CAROTS against various models on Lorenz96 and VAR, synthetically designed to exhibit clear causal relationships. These datasets provide a controlled environment to evaluate how effectively models leverage causal relationships for MTSAD. On the Lorenz96 dataset, CAROTS excels in detecting Point Global (PG) and Point Contextual (PC) anomalies, achieving AUROC scores of 0.998 and 0.975, respectively, significantly outperforming all other models. It also achieves the highest AUROC for Collective Global (CG) anomalies, highlighting its ability to capture both local and global causal relationships. On the VAR dataset, CAROTS is particularly effective at detecting Point Contextual (PC) and Collective Global (CG) anomalies with AUROC scores of 0.648 and 0.997, respectively.

*Table 1.* Evaluation on widely-used real-world MTSAD datasets. AT and TN denote AnomalyTransformer and TimsNet, respectively. CAROTS$^\dagger$ refers to the CAROTS without ensemble scoring, which only employs contrastive learning-based anomaly score.

| Dataset | Metric | AT | TN | USAD | SimCLR | SSD | CSI | CTAD | CARLA | CAROTS | CAROTS$^\dagger$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SWaT | AUROC | 0.501 | 0.808 | 0.812 | 0.528 | 0.486 | 0.540 | 0.820 | 0.807 | 0.852 | **0.861** |
| | AUPRC | 0.477 | 0.713 | 0.720 | 0.136 | 0.264 | 0.146 | 0.702 | 0.691 | **0.764** | 0.760 |
| | F1 | 0.222 | 0.762 | 0.755 | 0.289 | 0.394 | 0.297 | 0.755 | 0.742 | **0.791** | 0.789 |
| WADI | AUROC | 0.430 | 0.493 | 0.493 | 0.519 | 0.471 | 0.512 | 0.599 | 0.533 | 0.502 | **0.622** |
| | AUPRC | 0.045 | 0.055 | 0.054 | 0.106 | 0.246 | 0.297 | **0.327** | 0.103 | 0.056 | 0.260 |
| | F1 | 0.116 | 0.140 | 0.133 | 0.183 | 0.130 | 0.141 | 0.374 | 0.175 | 0.143 | **0.391** |
| PSM | AUROC | 0.500 | 0.742 | 0.694 | 0.766 | 0.689 | 0.695 | 0.708 | 0.445 | **0.783** | 0.729 |
| | AUPRC | **0.642** | 0.547 | 0.538 | 0.593 | 0.514 | 0.511 | 0.527 | 0.257 | 0.595 | 0.535 |
| | F1 | 0.443 | 0.589 | 0.506 | 0.554 | 0.498 | 0.498 | 0.515 | 0.444 | **0.603** | 0.534 |
| SMD_2-1 | AUROC | 0.503 | 0.702 | 0.707 | 0.655 | 0.602 | 0.662 | **0.744** | 0.546 | 0.703 | 0.726 |
| | AUPRC | **0.530** | 0.321 | 0.349 | 0.227 | 0.184 | 0.231 | 0.229 | 0.156 | 0.184 | 0.193 |
| | F1 | 0.103 | 0.353 | **0.390** | 0.240 | 0.207 | 0.242 | 0.254 | 0.202 | 0.260 | 0.299 |
| SMD_3-7 | AUROC | 0.584 | 0.767 | 0.727 | 0.642 | 0.640 | 0.664 | 0.659 | 0.483 | 0.769 | **0.779** |
| | AUPRC | 0.225 | 0.340 | 0.300 | 0.101 | 0.044 | 0.251 | 0.120 | 0.171 | **0.430** | 0.392 |
| | F1 | 0.190 | 0.436 | 0.380 | 0.197 | 0.093 | 0.329 | 0.193 | 0.254 | **0.564** | 0.542 |
| MSL_P-14 | AUROC | 0.484 | 0.734 | 0.734 | 0.398 | 0.469 | 0.496 | 0.761 | 0.712 | 0.764 | **0.782** |
| | AUPRC | 0.042 | 0.309 | 0.309 | 0.072 | 0.076 | 0.064 | 0.303 | **0.521** | 0.372 | 0.449 |
| | F1 | 0.064 | 0.502 | 0.502 | 0.182 | 0.115 | 0.104 | 0.406 | **0.639** | 0.545 | 0.599 |
| MSL_P-15 | AUROC | 0.500 | 0.572 | 0.580 | 0.504 | 0.527 | 0.537 | 0.641 | 0.572 | **0.701** | 0.683 |
| | AUPRC | **0.339** | 0.017 | 0.016 | 0.014 | 0.014 | 0.015 | 0.018 | 0.150 | 0.022 | 0.019 |
| | F1 | 0.021 | 0.072 | 0.070 | 0.053 | 0.053 | 0.072 | 0.067 | **0.272** | 0.087 | 0.079 |

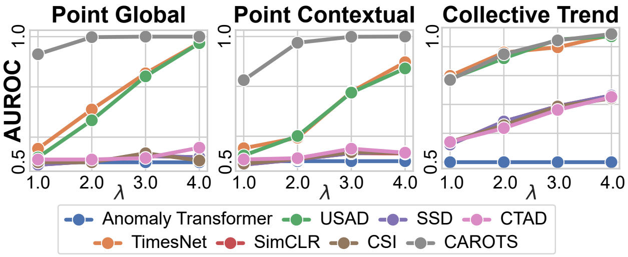

*Figure 3.* Anomaly detection performance measured by AUROC across different anomaly types in the Lorenz96 dataset, evaluated based on anomaly difficulty levels controlled by the factor $\lambda$.

Figure 3 shows the MTSAD performance for three anomaly types in the Lorenz96 dataset across four difficulty levels. $\lambda$ determines the difficulty of synthetic anomalies, where a smaller $\lambda$ corresponds to injecting more challenging anomalies. We highlight that CAROTS demonstrates successful dtection capability even for the most difficult anomalies ($\lambda$=1.0), where other methods struggle.

### 5.4. Analysis of CAROTS

**Robustness to hyperparameters.** Table 3 analyzes the robustness of CAROTS to variations in the temperature hyperparameter ($\tau$) and batch size ($B$), both critical in contrastive learning. CAROTS shows stable performance across a wide range of $\tau$. This suggests that CAROTS is resilient to changes in how similarity scores are scaled. Similarly, CAROTS shows consistent performance across varying batch sizes, with optimal results observed in the range of $B = 256$ to $B = 512$. The stability across these hyperparameters implies that CAROTS can be practically and reliably applied in real-world scenarios.

**Ablation studies.** Table 4 studies the effect of key technical components in CAROTS. Without CPA, AUPRC noticeably drops from 0.764 to 0.721 and the F1 score declines from 0.792 to 0.775, which implies that observing augmented normal variations during training was crucial in CAROTS. Similarly, the absence of SOC results in a significant decrease in AUROC from 0.852 to 0.819 and F1 score from 0.792 to 0.769. The contrastive learning-based anomaly score ($\mathcal{A}_{\text{CL}}$) is shown to play an essential role as the primary anomaly score, as its exclusion leads to one of the largest performance drops, reducing AUROC from 0.852 to 0.814. When CDA is removed, all variables are selected for perturbation without performing DFS, and the results indicate that CDA, along with ensemble scoring, contributes to additional robustness. The contrastive learning-based anomaly score ($\mathcal{A}_{\text{CL}}$) is also essential, as excluding it leads to one of the largest performance drops, reducing AUROC from 0.852 to 0.814.

**Choice of encoder architecture.** Table 6 presents the results of CAROTS with various encoder architectures, ranging from representative recurrent architectures (Hochre-

*Table 2.* Evaluation on synthetic datasets with explicit causal relationships: Lorenz96 and VAR. We report separate and averaged AUROC scores across four anomaly types: Point Global (PG), Point Contextual (PC), Collective Trend (CT), and Collective Global (CG).

| | Lorenz96 | | | | | VAR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PG | PC | CT | CG | AVG. | PG | PC | CT | CG | AVG. |
| AT | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 | 0.615 | 0.526 | 0.535 |
| TimesNet | 0.710 | 0.594 | **0.880** | 0.617 | 0.700 | 0.635 | 0.610 | 0.970 | **0.997** | 0.803 |
| USAD | 0.667 | 0.601 | 0.860 | 0.506 | 0.659 | **0.663** | 0.610 | **0.968** | **0.997** | 0.810 |
| SimCLR | 0.505 | 0.506 | 0.648 | 0.531 | 0.548 | 0.525 | 0.515 | 0.838 | 0.854 | 0.683 |
| SSD | 0.504 | 0.507 | 0.641 | 0.526 | 0.545 | 0.525 | 0.516 | 0.837 | 0.853 | 0.683 |
| CSI | 0.500 | 0.509 | 0.628 | 0.527 | 0.541 | 0.520 | 0.516 | 0.837 | 0.850 | 0.681 |
| CTAD | 0.511 | 0.512 | 0.618 | 0.486 | 0.532 | 0.527 | 0.522 | 0.897 | 0.842 | 0.697 |
| CAROTS | **0.998** | **0.975** | 0.874 | **0.788** | 0.909 | 0.612 | **0.648** | 0.963 | **0.997** | 0.805 |

*Table 3.* Performance evaluation on the SWaT and PSM datasets under varying temperature ($\tau$) and batch size ($B$). ROC and PRC represent AUROC and AUPRC, respectively.

| | | SWaT | | | PSM | | |
|---|---|---|---|---|---|---|---|
| | | ROC | PRC | F1 | ROC | PRC | F1 |
| $\tau$ | 0.1 | 0.85 | 0.76 | 0.79 | 0.73 | 0.54 | 0.53 |
| | 0.2 | 0.85 | 0.76 | 0.79 | 0.76 | 0.56 | 0.57 |
| | 0.4 | 0.84 | 0.76 | 0.80 | 0.69 | 0.47 | 0.53 |
| | 0.8 | 0.84 | 0.75 | 0.78 | 0.62 | 0.43 | 0.50 |
| | 1.6 | 0.85 | 0.77 | 0.81 | 0.61 | 0.44 | 0.46 |
| $B$ | 128 | 0.85 | 0.75 | 0.77 | 0.73 | 0.54 | 0.54 |
| | 256 | 0.85 | 0.76 | 0.79 | 0.76 | 0.56 | 0.57 |
| | 512 | 0.86 | 0.78 | 0.80 | 0.74 | 0.55 | 0.54 |
| | 1024 | 0.84 | 0.76 | 0.79 | 0.73 | 0.54 | 0.52 |

*Table 4.* Ablation studies on the SWaT dataset, evaluating the impact of key components in the CAROTS framework, including CPA, CDA, similarity filtering, and score ensemble.

| | AUROC | AUPRC | F1 |
|---|---|---|---|
| w/o CPA | 0.850 | 0.721 | 0.775 |
| w/o CDA | 0.842 | 0.740 | 0.786 |
| w/o similarity filtering | 0.819 | 0.733 | 0.769 |
| w/o $\mathcal{A}_{\text{CL}}$ | 0.814 | 0.722 | 0.769 |
| w/o $\mathcal{A}_{\text{CD}}$ | 0.861 | 0.760 | 0.789 |
| CAROTS | 0.852 | 0.764 | 0.792 |

iter & Schmidhuber, 1997; Cho et al., 2014) to a recent Transformer-based architecture (Liu et al., 2024). While GRU achieves the highest overall performance, the difference in performance from LSTM, GRU, and iTransformer is small. Such a result suggests that the strong MTSAD capability of CAROTS is not contingent on a specific choice of architecture, and thus, CAROTS is compatible with different temporal modeling approaches.

**Choice of causal discovery model.** Table 5 presents results of CAROTS with different causal discovery models: NGC (Tank et al., 2021), CUTS (Cheng et al., 2023), and CUTS+ (Cheng et al., 2024a). We find that CAROTS consistently performs well across all causal discovery methods, with only modest performance variation. While each method performs best on different datasets (e.g., CUTS+ on WADI and SMD_3-7; CUTS on SWaT and SMD_2-1; NGC on MSL_P-15), the overall performance remains robust and competitive. This indicates that CAROTS does not overly depend on a particular discovery algorithm or exact causal graph structure. Instead of solely relying on the causal graph searched by a causal discoverer for anomaly detection, CAROTS uses the causal graph as a guide for generating semantically meaningful causality-preserving or disturbing augmentations for contrastive learning.

*Table 5.* Results on how CAROTS perform under different causal discovery methods. We report the averaged AUROC for runs with three different seeds.

| | SWaT | WADI | SMD_2-1 | SMD_3-7 | MSL_P-15 |
|---|---|---|---|---|---|
| NGC | 0.85 | 0.49 | 0.68 | 0.69 | 0.76 |
| CUTS | 0.86 | 0.49 | 0.73 | 0.69 | 0.66 |
| CUTS+ | 0.85 | 0.50 | 0.70 | 0.77 | 0.70 |

*Table 6.* Anomaly detection performance of CAROTS with different encoder architectures on the SWaT and PSM datasets.

| | SWaT | | | PSM | | |
|---|---|---|---|---|---|---|
| | ROC | PRC | F1 | ROC | PRC | F1 |
| LSTM | 0.85 | 0.76 | 0.79 | 0.78 | 0.60 | 0.60 |
| GRU | 0.86 | 0.78 | 0.81 | 0.82 | 0.63 | 0.65 |
| iTransformer | 0.85 | 0.75 | 0.78 | 0.81 | 0.60 | 0.63 |

### 5.5. Qualitative Analysis with t-SNE Visualization

Figure 4 presents the t-SNE plots (Van der Maaten & Hinton, 2008) of learned embeddings during CAROTS training. Blue, green, and red points represent original samples, positive samples from CPA, and negative samples from CDA, respectively. Prior to training, the data embeddings lie according to natural clusters based on the semantic similarity;

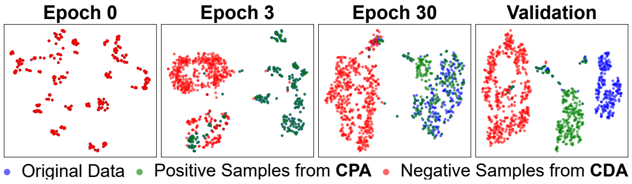

*Figure 4.* T-SNE visualizations of embeddings during training and validation after training.

thus, they show a significant overlap and appear inseparable from a causal viewpoint. As training progresses, positive and negative samples begin to separate, showing that the proposed causality-aware contrastive learning successfully distinguishes the two. Also, the embeddings of positive samples form distinct clusters, showing that the model effectively encodes information about different semantic patterns in the data. In final iterations, the positive samples become more integrated, focusing on shared causal relationships. CPA-generated samples lie outside the original data distribution, providing diverse extrapolated information that enriches the training process and enhances the model's understanding of causal structures. The t-SNE plot from validation data again shows a clear separation of positive and negative samples, confirming the encoder's ability to extend causality-aware discrimination to unseen data.

## 6. Conclusion

In this study, we proposed CAROTS, a novel framework that enables causality-aware multivariate time-series anomaly detection (MTSAD). With its unique causality-preserving and -disturbing augmentation schemes, CAROTS performs contrastive learning to train an encoder that distinguishes normal operations from anomalies based on the causal structure. CAROTS replaces the vanilla contrastive loss with a similarity-filtered one-class contrastive loss to preserve the diversity of semantic patterns in normal operations. The strong empirical results of CAROTS across an array of real-world and synthetic datasets validate that the consideration of causal relationships indeed equips CAROTS with enhanced anomaly detection capabilities. Potential directions of future research include extending CAROTS to handle more complex causal scenarios, such as confounding or exogenous variables, or non-stationary causal relationships.

## Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2022-II220959; No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720; 2022R1A5A708390811), the BK21 FOUR program of the Education and the Research Program for Future ICT Pioneers, Seoul National University in 2025, Hyundai Motor Company, and a grant from Yang Young Foundation.

## Impact Statement

Effective anomaly detection in multivariate time series requires an understanding of the causal relationships among variables, as anomalies often disrupt these stable inter-variable dependencies rather than simply deviating from statistical norms. CAROTS pioneers a causality-aware contrastive learning framework that leverages both causality-preserving and causality-disturbing augmentations to better differentiate normal variations from true anomalies. This approach highlights the growing importance of causality in time-series analysis and offers a promising direction for further research—such as extending to dynamic causal structures, integrating latent confounders, or adapting to non-stationary environments—paving the way toward more robust and reliable anomaly detection systems.

## References

Abdulaal, A., Liu, Z., and Lancewicki, T. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2485–2494, 2021.

Absar, S. and Zhang, L. Discovering time-invariant causal structure from temporal data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2807–2811, 2021.

Absar, S., Wu, Y., and Zhang, L. Neural time-invariant causal discovery from time series data. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2023.

Ahmed, C. M., Palleti, V. R., and Mathur, A. P. Wadi: A water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, CySWATER '17, pp. 25–28, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349758. doi: 10.1145/3055366.3055375. URL https://doi.org/10.1145/3055366.3055375.

Assaad, C. K., Devijver, E., and Gaussier, E. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., and Zuluaga, M. A. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3395–3404, 2020.

Ben-Moshe, N., Tsutsui, K., Biton, S., Zvuloni, E., Sörnmo, L., and Behar, J. A. Rawecgnet: Deep learning generalization for atrial fibrillation detection from the raw ecg. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Cheng, Y., Yang, R., Xiao, T., Li, Z., Suo, J., He, K., and Dai, Q. CUTS: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UG8bQcD3Emv.

Cheng, Y., Li, L., Xiao, T., Li, Z., Suo, J., He, K., and Dai, Q. Cuts+: High-dimensional causal discovery from irregular time-series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10): 11525–11533, Mar. 2024a. doi: 10.1609/aaai.v38i10.29034. URL https://ojs.aaai.org/index.php/AAAI/article/view/29034.

Cheng, Y., Wang, Z., Xiao, T., Zhong, Q., Suo, J., and He, K. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=iad1yyyGme.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W. (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Choi, K., Yi, J., Park, C., and Yoon, S. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE access*, 9:120043–120065, 2021.

Choi, K., Yi, J., Mok, J., and Yoon, S. Self-supervised time-series anomaly detection using learnable data augmentation. *arXiv preprint arXiv:2406.12260*, 2024.

Clifford, G. D., Liu, C., Moody, B., Li-wei, H. L., Silva, I., Li, Q., Johnson, A., and Mark, R. G. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pp. 1–4. IEEE, 2017.

Dai, Z., He, L., Yang, S., and Leeke, M. SARAD: Spatial association-aware anomaly detection and diagnosis for multivariate time series. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gmf5Aj01Hz.

Darban, Z. Z., Webb, G. I., Pan, S., Aggarwal, C. C., and Salehi, M. Carla: Self-supervised contrastive representation learning for time series anomaly detection. *Pattern Recognition*, 157:110874, 2025.

Dickey, D. A. and Fuller, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a): 427–431, 1979.

Febrinanto, F. G., Moore, K., Thapa, C., Liu, M., Saikrishna, V., Ma, J., and Xia, F. Entropy causal graphs for multivariate time series anomaly detection. *arXiv preprint arXiv:2312.09478*, 2023.

Goh, J., Adepu, S., Junejo, K. N., and Mathur, A. A dataset to support research in the design of secure water treatment systems. In *International conference on critical information infrastructures security*, pp. 88–99. Springer, 2016.

Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., and Soderstrom, T. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 387–395, 2018.

Kang, H. and Kang, P. Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism. *Knowledge-Based Systems*, 290:111507, 2024.

Karimi, A. and Paul, M. R. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4), 2010.

Kim, H., Kim, S., Min, S., and Lee, B. Contrastive time-series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Kim, S., Choi, K., Choi, H.-S., Lee, B., and Yoon, S. Towards a rigorous evaluation of time-series anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7194–7201, Jun. 2022. doi: 10.1609/aaai.v36i7.20680. URL https://ojs.aaai.org/index.php/AAAI/article/view/20680.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kong, L., Li, W., Yang, H., Zhang, Y., Guan, J., and Zhou, S. Causalformer: An interpretable transformer for temporal causal discovery. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Ky, J. R., Mathieu, B., Lahmadi, A., and Boutaba, R. Cats: Contrastive learning for anomaly detection in time series. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1352–1359. IEEE, 2024.

Lai, K.-H., Zha, D., Xu, J., Zhao, Y., Wang, G., and Hu, X. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=r8IvOsnHchr.

Lee, S., Park, T., and Lee, K. Soft contrastive learning for time series. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pAsQSWlDUf.

Li, G. and Jung, J. J. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. *Information Fusion*, 91:93–102, 2023.

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

Liu, Z., Gao, M., and Jiao, P. Gcad: Anomaly detection in multivariate time series from the perspective of granger causality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 19041–19049, Apr. 2025. doi: 10.1609/aaai.v39i18.34096. URL https://ojs.aaai.org/index.php/AAAI/article/view/34096.

Loshchilov, I. and Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Nandanwar, H. and Katarya, R. Deep learning enabled intrusion detection system for industrial iot environment. *Expert Systems with Applications*, 249:123808, 2024.

Ngu, H. C. V. and Lee, K. M. Cl-tad: A contrastive-learning-based method for time series anomaly detection. *Applied Sciences*, 13(21):11938, 2023.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/pascanu13.html.

Qiu, H., Liu, Y., Subrahmanya, N. A., and Li, W. Granger causality for time-series anomaly detection. In *2012 IEEE 12th international conference on data mining*, pp. 1074–1079. IEEE, 2012.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

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

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., and Pei, D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, 2019.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. B. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, F., Jiang, Y., Zhang, R., Wei, A., Xie, J., and Pang, X. A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. *Sensors (Basel, Switzerland)*, 25(1):190, 2025.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Xing, S., Niu, J., and Ren, T. Gcformer: Granger causality based attention mechanism for multivariate time series anomaly detection. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1433–1438. IEEE, 2023.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., et al. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 World Wide Web Conference*, pp. 187–196, 2018.

Xu, H., Wang, Y., Jian, S., Liao, Q., Wang, Y., and Pang, G. Calibrated one-class classification for unsupervised time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

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

Yang, Y., Zhang, C., Zhou, T., Wen, Q., and Sun, L. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023.

# A. Dataset Description

## A.1. Real-world Datasets

The datasets SWaT (Goh et al., 2016), WADI (Ahmed et al., 2017), PSM (Abdulaal et al., 2021), MSL (Hundman et al., 2018), and SMD (Su et al., 2019) are widely used in time series anomaly detection research, each tailored to specific domains such as industrial control systems, equipment maintenance, and environmental monitoring. These datasets provide labeled multivariate time series data and are designed to test the effectiveness of anomaly detection methods under real-world conditions. Their diverse domains and unique characteristics make them valuable resources for developing and benchmarking state-of-the-art algorithms.

The Secure Water Treatment (SWaT) dataset, developed by iTrust at the Singapore University of Technology and Design, simulates a six-stage water treatment process within an industrial control system (ICS). It consists of time series data collected from sensors and actuators, featuring both normal and attack data, where attacks simulate intentional anomalies introduced to disrupt the system. The multivariate nature of the dataset, combined with its complex sensor-actuator relationships, makes it a benchmark for ICS anomaly detection. Similarly, the Water Distribution (WADI) dataset, also created by iTrust, extends the scope to urban water distribution systems. It provides a larger and more complex simulation of long-term water network operations, capturing anomalies within a broader and more intricate system compared to SWaT.

The Proactive System Maintenance (PSM) dataset focuses on anomaly detection in industrial equipment maintenance. Derived from real-world industrial systems, this dataset contains multivariate time series data from various sensors monitoring equipment states and processes. PSM emphasizes fault detection and proactive maintenance, where anomalies represent potential equipment failures. The strong interdependencies among sensors in this dataset make it particularly useful for testing models that require context-aware anomaly detection.

In the domain of space exploration, the Mars Science Laboratory (MSL) dataset is derived from telemetry data collected by NASA's Curiosity Rover. It contains labeled normal and anomalous data based on real incidents recorded during the rover's operations. The dataset includes multivariate time series data from various sensors, reflecting the complexity and interconnectivity of space exploration systems.

Finally, the Server Machine Dataset (SMD) represents time series data from large-scale server systems. It includes measurements from server components such as CPU, memory, and disk, along with labeled normal and anomalous states. SMD is particularly relevant for IT infrastructure and server monitoring, as it simulates realistic operational anomalies in server clusters.

*Table 7.* Summary of dataset characteristics, including the number of variables, training and test sequence lengths, anomaly ratio, and stationarity assessment using Augmented Dickey-Fuller (ADF) test statistics.

| Dataset | Variables | Train Steps | Test Steps | Anomaly Ratio | ADF Statistics | $p$-value |
|---|---|---|---|---|---|---|
| SWaT | 51 | 495,000 | 449,919 | 0.121 | -0.780 | 0.83 |
| WADI | 123 | 784,537 | 172,801 | 0.058 | -0.347 | 0.92 |
| PSM | 25 | 132,481 | 87,841 | 0.278 | -3.493 | 0.01 |
| SMD_2-1 | 38 | 23,693 | 23,694 | 0.049 | -13.649 | 0.00 |
| SMD_3-7 | 38 | 28,705 | 28,705 | 0.015 | 0.128 | 0.97 |
| MSL_P-14 | 55 | 2,880 | 6,100 | 0.030 | -1.228 | 0.66 |
| MSL_P-15 | 55 | 3,682 | 2,856 | 0.007 | -0.621 | 0.87 |

Table 7 summarizes the key characteristics of the datasets used in our experiments. The datasets vary in the number of variables, the length of training and test sequences, and the proportion of anomalies. The anomaly ratio represents the fraction of anomalous data points in the test set, highlighting the imbalance in certain datasets. Additionally, we report the Augmented Dickey-Fuller (ADF) test (Dickey & Fuller, 1979) statistics and corresponding p-values to assess the stationarity of each dataset, where higher p-values indicate non-stationary time-series.

For SMD and MSL, we selected subsets that exhibit stronger non-stationarity, making them more challenging for anomaly detection. These subsets better reflect real-world scenarios where complex temporal dependencies and distribution shifts pose difficulties for standard detection methods.

## A.2. Synthetic Datasets

The Lorenz96 system is a chaotic dynamical system widely used for modeling time-series with complex temporal dependencies (Karimi & Paul, 2010). Originally designed to represent simplified atmospheric dynamics, it has become a benchmark for evaluating causality-aware models, including those based on Granger causality. The system consists of $N$ coupled variables $x^i$ evolving according to the differential equation:

$$\frac{dx_t^i}{dt} = (x_t^{i+1} - x_t^{i-2})x_t^{i-1} - x_t^i + F, \quad i = 1, 2, \ldots, N, \tag{8}$$

where $F$ is an external forcing parameter controlling the level of chaos. The indices follow cyclic boundary conditions such that $x_t^{-1} = x_t^{N-1}$, $x_t^0 = x_t^N$, $x_t^{N+1} = x_t^1$. From a Granger causality perspective, the Lorenz96 system exhibits strong directional dependencies between variables. Each variable $x^i$ is influenced by its two preceding variables ($x^{i-1}$ and $x^{i-2}$) and its one succeeding variable ($x^{i+1}$). We set $F = 10.0$ following the values in (Cheng et al., 2024a).

The VAR dataset is generated using a Vector Autoregressive model. The VAR model captures linear dependencies among multiple time-series variables by modeling each variable as a function of both its past values and the past values of other variables in the system. A VAR model is defined as:

$$\boldsymbol{x}_t = \sum_{i=1}^{p} \boldsymbol{A}\boldsymbol{x}_{t-i} + \boldsymbol{\epsilon}_t, \tag{9}$$

where $\boldsymbol{x}_t \in \mathbb{R}^N$ denotes the value of $N$ variables observed at time $t$, $\boldsymbol{A}$ is coefficient matrices that define the relationships between past and current values, and $p$ is the lag order, determining how many past time steps are considered. $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma^2 I)$ represents a multivariate Gaussian noise. We set $\boldsymbol{A}$, $p$, and $\sigma$ following the values in (Cheng et al., 2024a).

Following the approach of Lai et al. (2021), we constructed the synthetic datasets with anomalies by injecting four types of synthetic anomalies, each designed to evaluate different aspects of anomaly detection. The following introduces how we generated each type of synthetic anomaly: Point Global (PG), Point Contectual (PC), Collective Trend (CT), and Collective Global (CG).

For Point Global (PG) anomalies, anomalous time steps were randomly selected based on a predefined anomaly ratio, and affected variables were also chosen randomly. For each selected variable $x^i$, the global mean and global standard deviation were computed as:

$$\mu_{\text{global}} = \frac{1}{T}\sum_{t=1}^{T} x_t^i, \quad \sigma_{\text{global}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t^i - \mu_{\text{global}})^2}. \tag{10}$$

The value at the anomalous time step $t_a$ was then replaced as follows:

$$x_{t_a}^i = \mu_{\text{global}} + \lambda \cdot \sigma_{\text{global}}, \tag{11}$$

where the factor $\lambda$ controls the magnitude of deviation, allowing adjustment of anomaly difficulty.

For Point Contextual (PC) anomalies, the approach was similar, but instead of using global statistics, local mean and local standard deviation were computed over a surrounding window of radius $r$ = 5:

$$\mu_{\text{local}} = \frac{1}{2r+1}\sum_{t=t_a-r}^{t_a+r} x_t^i, \quad \sigma_{\text{local}} = \sqrt{\frac{1}{2r+1}\sum_{t=t_a-r}^{t_a+r}(x_t^i - \mu_{\text{local}})^2} \tag{12}$$

.

The anomaly was then injected as:

$$x_{t_a}^i = \mu_{\text{local}} + \lambda \cdot \sigma_{\text{local}}, \tag{13}$$

ensuring that the anomaly is context-dependent rather than an absolute deviation.

For Collective Trend (CT) anomalies, values within the selected radius $r$ = 5 were modified by adding a linear trend with a randomly assigned slope:

$$x_t^i = x_t^i + \text{sign} \cdot \lambda \cdot (t - (t_a - r)), \quad t \in [t_a - r, t_a + r], \tag{14}$$

*Table 8.* Standard deviations for evaluation on widely-used real-world MTSAD datasets. AT and TN denote AnomalyTransformer and TimsNet, respectively. CAROTS$^\dagger$ refers to the CAROTS without ensemble scoring, which only employs contrastive learning-based anomaly score.

| | | AT | TN | USAD | SimCLR | SSD | CSI | CTAD | CARLA | CAROTS | CAROTS$^\dagger$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SWaT | AUROC | 0.001 | 0.002 | 0.000 | 0.015 | 0.015 | 0.054 | 0.031 | 0.034 | 0.008 | 0.003 |
| | AUPRC | 0.013 | 0.005 | 0.000 | 0.008 | 0.008 | 0.002 | 0.160 | 0.015 | 0.003 | 0.020 |
| | F1 | 0.000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.014 | 0.075 | 0.022 | 0.008 | 0.009 |
| WADI | AUROC | 0.025 | 0.005 | 0.003 | 0.068 | 0.068 | 0.095 | 0.016 | 0.056 | 0.007 | 0.042 |
| | AUPRC | 0.002 | 0.000 | 0.000 | 0.022 | 0.022 | 0.019 | 0.033 | 0.047 | 0.001 | 0.021 |
| | F1 | 0.000 | 0.001 | 0.002 | 0.031 | 0.031 | 0.019 | 0.043 | 0.058 | 0.002 | 0.076 |
| PSM | AUROC | 0.000 | 0.021 | 0.015 | 0.060 | 0.060 | 0.015 | 0.089 | 0.041 | 0.008 | 0.018 |
| | AUPRC | 0.000 | 0.010 | 0.004 | 0.050 | 0.050 | 0.012 | 0.026 | 0.012 | 0.007 | 0.012 |
| | F1 | 0.000 | 0.040 | 0.013 | 0.043 | 0.043 | 0.014 | 0.027 | 0.001 | 0.011 | 0.022 |
| SMD_2-1 | AUROC | 0.000 | 0.008 | 0.001 | 0.006 | 0.006 | 0.016 | 0.021 | 0.157 | 0.021 | 0.023 |
| | AUPRC | 0.000 | 0.007 | 0.001 | 0.017 | 0.017 | 0.023 | 0.050 | 0.078 | 0.016 | 0.018 |
| | F1 | 0.000 | 0.005 | 0.002 | 0.026 | 0.026 | 0.027 | 0.055 | 0.087 | 0.025 | 0.026 |
| SMD_3-7 | AUROC | 0.036 | 0.006 | 0.001 | 0.006 | 0.006 | 0.044 | 0.004 | 0.075 | 0.011 | 0.045 |
| | AUPRC | 0.235 | 0.015 | 0.000 | 0.050 | 0.050 | 0.135 | 0.001 | 0.050 | 0.015 | 0.008 |
| | F1 | 0.099 | 0.016 | 0.000 | 0.061 | 0.061 | 0.157 | 0.001 | 0.069 | 0.011 | 0.012 |
| MSL_P-14 | AUROC | 0.031 | 0.000 | 0.000 | 0.204 | 0.204 | 0.294 | 0.195 | 0.165 | 0.000 | 0.028 |
| | AUPRC | 0.008 | 0.000 | 0.000 | 0.010 | 0.010 | 0.226 | 0.229 | 0.154 | 0.000 | 0.030 |
| | F1 | 0.005 | 0.000 | 0.000 | 0.030 | 0.030 | 0.295 | 0.274 | 0.113 | 0.000 | 0.051 |
| MSL_P-15 | AUROC | 0.000 | 0.001 | 0.000 | 0.077 | 0.077 | 0.055 | 0.048 | 0.117 | 0.008 | 0.012 |
| | AUPRC | 0.289 | 0.000 | 0.000 | 0.007 | 0.007 | 0.005 | 0.004 | 0.115 | 0.001 | 0.002 |
| | F1 | 0.000 | 0.001 | 0.000 | 0.038 | 0.038 | 0.037 | 0.024 | 0.136 | 0.004 | 0.013 |

where $\mathrm{sign}$ is randomly chosen as $+1$ or $-1$ ensuring that the trend can either increase or decrease, introducing deviations from the expected behavior.

For Collective Global (CG) anomalies, values within a predefined radius $r$ = 5 around the selected time step were replaced using a square sine wave, defined as:

$$x_t^i = x_t^i + \sum_{k=0}^{L-1} \frac{1}{2k+1} A \sin\left(2\pi f (2k+1) t\right), \quad t \in [t_a - r, t_a + r], \tag{15}$$

where $A$ and $f$ control the amplitude and frequency of the anomaly. We used $A = 1.5,\ f = 0.04, L = 5$ following (Lai et al., 2021). As with other anomaly types, only a subset of variables was affected.

In this experiment, synthetic datasets were generated with $N = 128$ and a total length of 40,000 time steps, which were split into 16,000, 4,000, and 20,000 steps for the train, validation, and test sets, respectively. Following the unsupervised time-series anomaly detection setting, synthetic anomalies were injected only into the test set, with an anomaly ratio of 0.01. Among the 128 variables, 10 were randomly selected, and anomalies were introduced only into these variables. The factor $\lambda$ was set to 2.0 by default for all anomaly types except Collective Global (CG), which does not require a factor.

## B. Additional Implementation Details

**Baselines.** We re-implement all baseline methods for a fair comparison. For the contrastive learning-based baselines, we follow the augmentation strategies of (Kim et al., 2023).

**Causality-preserving augmentation.** For causality-preserving augmentation, we randomly select a single causing variable $(M = 1)$. Gaussian noise with a standard deviation of 0.2 is applied at the last time step of $\boldsymbol{X}_{<t}$ to maintain stability.

**Causality-disturbing augmentation.** Perturbations were introduced by adding a random bias to increase anomaly diversity and difficulty. The bias was randomly chosen from (-0.5, -0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4, 0.5), creating anomalies ranging from easy to hard. To enhance augmentation diversity, perturbations were applied to 50% of randomly selected time

*Table 9.* Anomaly detection performance on the SWaT dataset for various values of the similarity filtering threshold $\alpha$, with and without $\alpha$ scheduling. In the scheduled setting, $\alpha$ is linearly increased up to 0.9 over epochs.

| $\alpha$ scheduling | $\alpha$ | AUROC | AUPRC | F1 |
|---|---|---|---|---|
| no | -1.0 | 0.820 | 0.733 | 0.770 |
| no | -0.5 | 0.820 | 0.734 | 0.766 |
| no | 0.0 | 0.828 | 0.744 | 0.786 |
| no | 0.5 | 0.857 | 0.761 | 0.778 |
| no | 0.75 | 0.855 | 0.778 | 0.801 |
| yes | -1.0 | 0.820 | 0.736 | 0.773 |
| yes | -0.5 | 0.825 | 0.743 | 0.784 |
| yes | 0.0 | 0.845 | 0.759 | 0.779 |
| yes | 0.5 | 0.852 | 0.764 | 0.791 |
| yes | 0.75 | 0.845 | 0.767 | 0.799 |

steps per variable. To enhance the diversity of the extracted subgraph, a cut-off probability $p$ of $0.1$.

**Model optimization and evaluation.** Each model was optimized using Adam optimizer (Kingma & Ba, 2015). We applied gradient clipping (Pascanu et al., 2013) with a maximum norm of 1.0 for training stability. The learning rate followed a cosine learning rate scheduling (Loshchilov & Hutter, 2017) with warm up, starting at 0.0001 and increasing linearly over 5 epochs as a warm-up phase. A hyperparameter search was conducted over the learning rate in $\{0.001, 0.0003, 0.0001\}$ and weight decay in $\{0.001, 0.0001, 0.0\}$. The model with the lowest validation loss was selected for evaluation. Training was performed on a single NVIDIA A40 GPU. Following standard evaluation protocols, we reported the best F1-score. We excluded the point-adjusted F1 (Xu et al., 2018), which considers a segment as detected if any anomaly point within it is identified. This is because the point-adjusted F1 overestimates anomaly detection performance (Kim et al., 2022), and a robust detector should be capable of identifying arbitrary anomaly point effectively. For computational efficiency, the SWaT and WADI datasets were downsampled by a factor of 5. In Lorenz96 and VAR, the window sizes were set to 2 and 4 following (Cheng et al., 2024a), respectively. L2 distance was used for anomaly scoring of $\mathcal{A}_{\mathrm{CL}}$, except for WADI and VAR, where cosine distance provided better performance.

## C. Additional Experimental Results

**Standard deviations.** Table 8 shows the standard deviations for experiments presented in Table 1, which is averaged for three random seeds.

**Effectiveness of similarity filtering.** Table 9 presents anomaly detection performance on the SWaT dataset across different values of the similarity filtering threshold $\alpha$ and examines the impact of $\alpha$ scheduling. In the non-scheduled setting, $\alpha$ is fixed throughout training, while in the scheduled setting, $\alpha$ is linearly increased up to 0.9 over epochs. The results indicate that increasing $\alpha$ generally improves AUROC and AUPRC, suggesting that a more relaxed similarity filtering threshold enhances anomaly detection.

Applying $\alpha$ scheduling consistently leads to better performance, particularly in AUPRC and F1 score, highlighting its role in improving model robustness. The performance gains are more pronounced at intermediate $\alpha$ values, where scheduling stabilizes results and enhances the model's ability to distinguish anomalies. Overall, a moderate similarity filtering threshold combined with progressive $\alpha$ scheduling yields the best results. This underscores the importance of adaptively tuning $\alpha$ over training to balance anomaly separation and representation diversity.

**Extended ablation studies.** Table 10 shows the extended ablation studies of Table 4 on 7 real-world datasets and 2 synthetic datasets, validating the effectiveness of each component of CAROTS. Table 11 presents the performance across different values of $\sigma$ in CPA. Performance remains stable across different $\sigma$'s, indicating that the generated samples do not degrade model quality, even with higher noise levels. These results suggest that CPA is robust to the choice of $\sigma$ within a reasonable range.

**Causal relationships over time.** CAROTS assumes that causal relationships among variables remain consistent over time, following existing literature on causal discovery (Tank et al., 2021; Cheng et al., 2023; 2024a). To empirically assess whether

*Table 10.* Ablation studies on 7 real-world datasets and 2 synthetic datasets. We report the averaged AUROC for runs with three different seeds. For Lorenz96 and VAR, the reported values represent the average performance across the four different synthetic anomaly types.

| | SWaT | WADI | PSM | SMD_2-1 | SMD_3-7 | MSL_P-14 | MSL_P-15 | Lorenz96 | VAR |
|---|---|---|---|---|---|---|---|---|---|
| w/o CPA | 0.850 | 0.486 | 0.786 | 0.700 | 0.756 | 0.764 | 0.740 | 0.918 | 0.767 |
| w/o CDA | 0.842 | 0.488 | 0.789 | 0.623 | 0.732 | 0.764 | 0.609 | 0.917 | 0.785 |
| w/o similarity filtering | 0.819 | 0.493 | 0.706 | 0.758 | 0.719 | 0.764 | 0.719 | 0.923 | 0.765 |
| w/o $\mathcal{A}_{\text{CL}}$ | 0.814 | 0.494 | 0.778 | 0.602 | 0.701 | 0.768 | 0.694 | 0.943 | 0.732 |
| CAROTS$^\dagger$ (w/o $\mathcal{A}_{\text{CD}}$) | 0.861 | 0.622 | 0.729 | 0.726 | 0.779 | 0.782 | 0.683 | 0.919 | 0.769 |
| CAROTS | 0.852 | 0.502 | 0.783 | 0.703 | 0.769 | 0.764 | 0.701 | 0.909 | 0.805 |

*Table 11.* Performance metrics across different $\sigma$ in CPA on SWaT dataset.

| $\sigma$ | AUROC | AUPRC | F1 |
|---|---|---|---|
| 0 | 0.850±0.001 | 0.761±0.002 | 0.798±0.001 |
| 0.05 | 0.853±0.003 | 0.762±0.002 | 0.797±0.004 |
| 0.1 | 0.852±0.008 | 0.764±0.003 | 0.791±0.008 |
| 0.2 | 0.849±0.007 | 0.759±0.009 | 0.795±0.000 |
| 0.4 | 0.848±0.002 | 0.762±0.007 | 0.792±0.001 |

this statement holds in our setting, we further analyze the evolution of causal structures in three benchmark datasets: SWaT, WADI, and PSM. For each dataset, we split the normal training data into four disjoint, time-ordered segments (quarters 1 to 4), train a causal discovery model on each, and compute pairwise cosine similarities between the resulting graphs. As shown in Table 12, the consistently high similarity indicates that the learned causal relationships remain stable across time segments, supporting the validity of our approach.

**Computational cost.** We compare the training times of the evaluated anomaly detectors in Table 13. The pre-training time of the causal discovery model is excluded to ensure a fair comparison of the anomaly detector training times. Even with causal modules, CAROTS is efficient due to a lightweight one-layer LSTM. The train time of CAROTS is comparable to baselines and cheaper than heavier models like TimesNet, indicating that leveraging causal structure can reduce reliance on deeper architectures.

**Learned causality matrix.** Figure 5 presents the learned causality matrix from the causal discovery model for each dataset.

## D. Limitations and Future Work

CAROTS demonstrates strong performance in MTSAD by effectively leveraging causal relationships. At the same time, there remain opportunities to further enhance its robustness and applicability in more complex scenarios. First, CAROTS assumes that causal relationships in normal operations remain largely time-invariant. While this assumption holds in many stable environments, real-world systems often experience gradual or abrupt causal shifts due to evolving conditions or external interventions. Adapting CAROTS to dynamically capture and adjust to non-stationary causal structures would further strengthen its applicability to real-world settings. Second, CAROTS focuses on direct causal relationships but does not explicitly model confounding variables, which can introduce hidden biases. Addressing this challenge by developing mechanisms to identify and mitigate confounding effects could improve the reliability of causality-aware anomaly detection. Third, exogenous variables, which influence the system externally without being directly modeled, may impact the learned causal relationships. Incorporating external contextual factors, such as domain knowledge or auxiliary data sources, could expand CAROTS' effectiveness in handling more complex causal interactions. Despite these challenges, CAROTS introduces a novel and pioneering approach to MTSAD by integrating causal awareness into contrastive learning. To the best of our knowledge, this is the first framework to explicitly incorporate causality into contrastive MTSAD, providing a new perspective on how anomalies can be characterized through deviations from stable causal structures.

*Table 12.* Comparison of cosine similarity of causality matrix across different quarters for SWaT, WADI, and PSM datasets.

| Quarters | SWaT | WADI | PSM |
|---|---|---|---|
| Q1vsQ2 | 0.911 | 0.965 | 0.955 |
| Q1vsQ3 | 0.923 | 0.966 | 0.953 |
| Q1vsQ4 | 0.928 | 0.959 | 0.918 |
| Q2vsQ3 | 0.978 | 0.973 | 0.952 |
| Q2vsQ4 | 0.978 | 0.964 | 0.898 |
| Q3vsQ4 | 0.981 | 0.963 | 0.915 |

*Table 13.* Comparison of training time for different anomaly detectors on SWaT dataset.

| Method | Time (min) |
|---|---|
| CAROTS | 25 |
| AnomalyTransformer | 12 |
| TimesNet | 56 |
| USAD / SimCLR / SSD | 6 |
| CSI / CTAD | 10 |

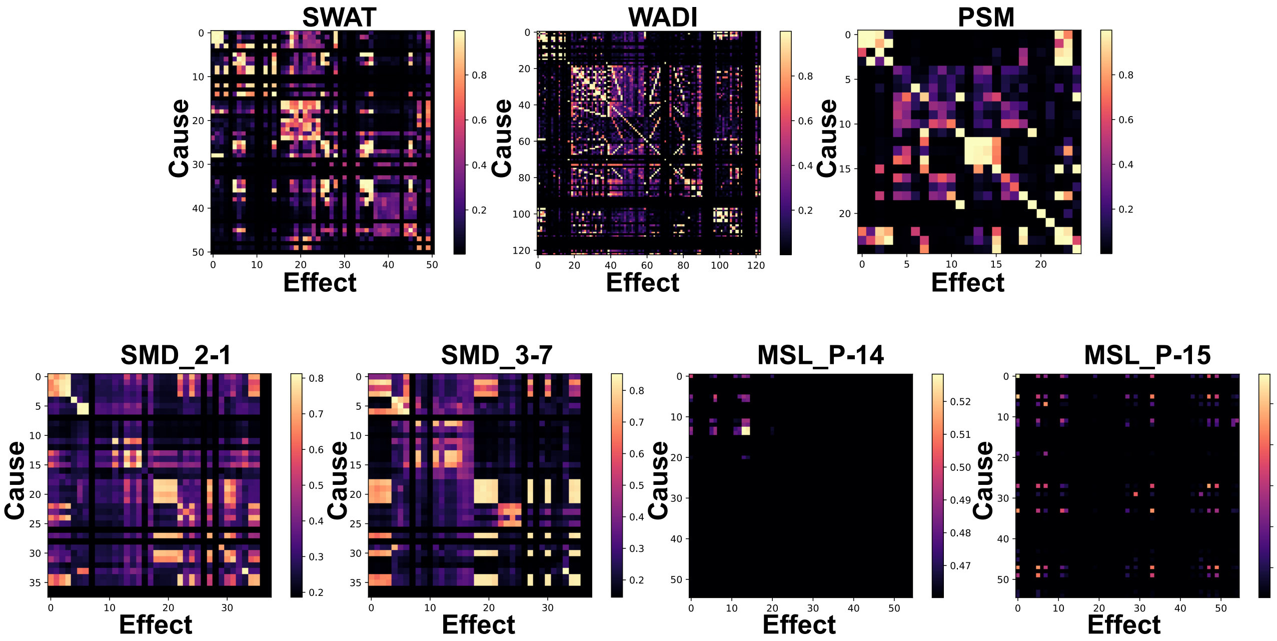

*Figure 5.* Learned causality matrix from the causal discovery model for each dataset.