JoCE: Joint Counterfactual Explanation for Interpretable Time Series Anomaly Detection

Woojin Jeong[1], **Geonwoo Shin**[1], **Jaewook Lee**[1]

[1]Department of Industrial Engineering, Seoul National University, Republic of Korea
jwj7955@snu.ac.kr, shin0621@snu.ac.kr, jaewook@snu.ac.kr

Abstract

Interpreting time series anomaly detection (TSAD) models remains challenging due to complex temporal dependencies and multivariate feature interactions. While counterfactual explanations (CE) provide intuitive and actionable interpretability by suggesting minimal input changes, existing CE methods are primarily designed for classification and do not fully capture the structural characteristics of time series. A few recent studies have explored CE for TSAD, but they often ignore key temporal properties or remain tied to specific detection models. To address these limitations, we propose **JoCE** (**Jo**int **C**ounterfactual **E**xplanation), a model-agnostic framework for generating realistic counterfactuals in TSAD. JoCE identifies minimal, causally relevant regions across temporal and feature dimensions, producing counterfactuals that selectively repair anomalies while preserving temporal consistency. Experiments on real-world TSAD benchmarks show that JoCE generates coherent and plausible explanations, offering deeper insights into model behavior.

Introduction

Anomaly detection in multivariate time series is crucial in applications such as industrial monitoring, healthcare, and finance. Transformer-based models have shown strong performance by capturing complex temporal dependencies and feature interactions (Xu et al. 2021; Yang et al. 2023), yet they typically act as black boxes, making it difficult for domain experts to interpret or trust their predictions. As deep learning-based anomaly detectors are increasingly deployed in automated monitoring systems, the need for interpretable and actionable explanations grows.

For instance, in industrial manufacturing, when a machine sensor reading suddenly spikes, a maintenance engineer needs to know exactly which combination of variables was minimally responsible for the alarm and, more importantly, an actionable explanation like 'by how much should the temperature be lowered for the next 10 minutes' to bring the system back to normal. Similarly, in healthcare, a model flagging a patient's vital signs as anomalous requires explanations that guide immediate clinical intervention. These scenarios underscore that simply detecting an anomaly is insufficient; interpretable models must provide concrete, minimal, and actionable guidance for repair or diagnosis, making the model trustworthy for high-stakes decisions.

Post-hoc interpretability methods help explain black-box models without modifying their architecture (Mochaourab et al. 2022). Among them, counterfactual explanations (CE) provide an intuitive approach by identifying minimal, plausible changes to the input that would alter model predictions (Wachter, Mittelstadt, and Russell 2017; Sulem et al. 2022). Unlike saliency maps (Simonyan, Vedaldi, and Zisserman 2013) or Grad-CAM (Selvaraju et al. 2017), which highlight important input regions, CE answers "what-if'' questions with concrete alternative instances. Similarly, feature attribution methods like SHAP (Lundberg and Lee 2017) and LIME (Ribeiro, Singh, and Guestrin 2016) quantify feature importance but do not generate actual counterfactuals, limiting actionable guidance.

Although CE has been studied in tabular and image domains (Wachter, Mittelstadt, and Russell 2017; Goyal et al. 2019), its application to time series anomaly detection remains underexplored. Time series anomalies are often point- or segment-level within the same sequence, unlike classification tasks where labels exist per instance. Existing classification-oriented CE methods cannot directly repair localized anomalous regions while preserving normal patterns.

Time series present additional challenges: temporal continuity, autocorrelation, spectral characteristics, and interdependent features must be preserved to generate realistic counterfactuals (Rojat et al. 2021; Giannoulis, Harris, and Barra 2023; Younis, Hakmeh, and Ahmadi 2024). Naïve perturbations ignoring these properties often yield implausible results. Existing CE methods either fail to preserve temporal properties or are tied to specific detection models (Ji et al. 2024; Lee, Malacarne, and Aune 2024), limiting general applicability.

To address these gaps, we introduce JoCE, a novel counterfactual explanation framework designed for time series anomaly detection. First, it ensures *time-series plausibility* by enforcing temporal consistency and spectral fidelity. In this way, JoCE preserves basic statistical properties such as stationarity and smoothness, while producing sparse, segment-level perturbations that yield coherent and interpretable repairs. Second, it emphasizes *model-agnostic practicality*, as it operates solely through the anomaly score function. This design makes JoCE broadly compatible with diverse anomaly detection models without the need for retraining or access to internal parameters.

The main contributions of this work are as follows:

- We propose JoCE, a model-agnostic framework for generating counterfactual explanations in time series anomaly detection that preserves both temporal continuity and minimality.
- We design a loss function that balances multiple objectives, ensuring explanations are valid, sparse, and plausible.
- We demonstrate the effectiveness of JoCE on real-world time series anomaly detection datasets, showing its ability to generate coherent and plausible counterfactual explanations across diverse scenarios.

Related Work

Interpretability in Time Series Anomaly Detection

Transformer-based models have shown strong performance in time series anomaly detection (Wen et al. 2022; Tuli, Casale, and Jennings 2022; Zamanzadeh Darban et al. 2024). For instance, Anomaly Transformer (Xu et al. 2021) leverages self-attention to detect deviations in temporal associations, while DCdetector (Yang et al. 2023) uses dual attention contrastive learning to distinguish normal and abnormal patterns. Despite their effectiveness, these models often remain black boxes, making it difficult for non-experts to trust predictions and for domain experts to understand model reasoning, especially in high-stakes areas like finance (Sabharwal et al. 2024) and medicine (Abououf et al. 2023).

Post-hoc interpretability methods, such as feature attribution (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), highlight influential variables but offer limited actionable guidance. They identify where anomalies occur but not why or how to prevent them (Ates et al. 2021). Counterfactual explanations address this limitation by answering causal "what-if'' questions, showing minimal temporal or feature-level changes needed to alter anomaly outcomes, which can provide actionable insights in domains like healthcare and manufacturing.

Counterfactual Explanations

Counterfactual explanations generate minimal input modifications that change model predictions, forming a key approach in explainable AI (XAI). Foundational methods like WachterCF (Wachter, Mittelstadt, and Russell 2017) and DiCE (Mothilal, Sharma, and Tan 2020) frame counterfactual generation as an optimization problem, while FACE (Poyiadzi et al. 2020) and EACE (Zhou et al. 2025) improve realism and diversity using neighborhood graphs or density-aware objectives.

Extending counterfactual reasoning to multivariate time series is challenging due to temporal continuity and inter-feature dependencies (Tripathy et al. 2022). Methods such as Sparse DPE (Sulem et al. 2022), AR-Pro (Ji et al. 2024), and TimeVQVAE-AD (Lee, Malacarne, and Aune 2024) adapt counterfactuals for time series anomaly detection, providing interpretable explanations without sacrificing detection accuracy. Our work proposes a model-agnostic framework applicable to various TSAD models, aiming for broader usability across domains.

Problem Formulation

Anomaly detection in multivariate time series aims to identify abnormal time points or segments that deviate significantly from expected system behavior. Let $\mathcal{X} \in \mathbb{R}^{D\times T}$ denote a multivariate time series with D features and length T. Given an anomaly scoring function

$$f : \mathbb{R}^{D\times T} \to \mathbb{R}^{T},$$

which assigns an anomaly score $s_t = f(\mathcal{X})_t$ to each time step t, an observation at time t is flagged as anomalous if s_t exceeds a threshold τ. To explain why a specific time series $\mathcal{X}$ is detected as anomalous, we focus on generating a counterfactual time series $\tilde{\mathcal{X}}$ that modifies $\mathcal{X}$ minimally while lowering the anomaly scores. Following the conceptual definitions in previous works (Mothilal, Sharma, and Tan 2020; Huang et al. 2024; Zhou et al. 2025), we define the counterfactual explanation task as follows:

Definition 1 (Counterfactual Explanation for Time Series Anomaly Detection)**.** *Given an input time series $\mathcal{X} \in \mathbb{R}^{D\times T}$ with anomaly scores $f(\mathcal{X}) \in \mathbb{R}^{T}$, the counterfactual explanation $\tilde{\mathcal{X}}^*$ is defined as the solution to the following constrained optimization problem:*

$$\tilde{\mathcal{X}}^* = \arg\min_{\tilde{\mathcal{X}}\in\mathbb{R}^{D\times T}} \quad \|\tilde{\mathcal{X}} - \mathcal{X}\|_F^2 + \lambda\,\mathcal{R}(\tilde{\mathcal{X}}) \tag{1}$$

$$\textit{subject to} \quad f(\tilde{\mathcal{X}})_t \le \tau, \quad \forall t \in \mathcal{T}_{\text{anom}} \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm ensuring minimal deviation from the original input, and $\mathcal{R}(\cdot)$ is a regularization term promoting temporal smoothness and plausibility. The parameter λ balances the trade-off between fidelity to the original input and regularity. The set $\mathcal{T}_{\text{anom}} = \{t \mid f(\mathcal{X})_t > \tau\}$ denotes the indices of detected anomalies, and the constraint Eq. (2) *ensures that the anomaly score at each detected anomaly point is reduced in the counterfactual.*

To ensure interpretability, we restrict modifications to features with sufficient temporal variability, keeping stable features unchanged. This approach balances anomaly reduction with interpretability and temporal consistency, enabling the generation of meaningful counterfactual explanations for time series anomalies.

Proposed Method

Generating counterfactual explanations for anomaly detection in time series requires methods that produce realistic, interpretable, and temporally coherent modifications. Existing approaches (Zhou et al. 2025; Poyiadzi et al. 2020) often depend on heuristic or gradient-free optimization strategies that are computationally expensive and tend to generate temporally inconsistent counterfactuals. Moreover, these methods typically treat feature-wise and temporal dimensions independently, neglecting the joint structure of anomalies that manifest across both axes in multivariate time series.

We propose **JoCE** (**Jo**int **C**ounterfactual **E**xplanation) for multivariate time series anomaly detection, a novel framework that addresses these limitations by jointly modeling temporal and feature dimensions to produce faithful and interpretable counterfactuals. JoCE formulates the generation

process as a constrained optimization problem (see Definition 1) aimed at simultaneously suppressing anomaly scores and preserving proximity to the original input. Unlike prior methods, JoCE captures the joint dependencies of time and features, enabling localized, sparse, and temporally smooth modifications. The overall architecture is illustrated in Figure 1.

JoCE is designed to generate counterfactuals that (i) maintain the original temporal dynamics of the input sequence, and (ii) make minimal changes only to the most relevant time-feature areas with high anomaly scores. The design goals and optimization framework of JoCE are explained in Section , and the detailed design of its modules is described in Section .

Design Objectives of JoCE

To explicitly generate a counterfactual time series $\tilde{\mathcal{X}}$, we introduce two learnable components: a perturbation tensor $\boldsymbol{\delta} \in \mathbb{R}^{D\times T}$ and a mask tensor $\mathbf{M} \in [0,1]^{D\times T}$. The mask tensor regulates which elements of the original input $\mathcal{X}$ are subject to modification. The counterfactual is constructed via a mask-controlled additive perturbation as follows:

$$\tilde{\mathcal{X}} = (1-\mathbf{M}) \odot \mathcal{X} + \mathbf{M} \odot (\mathcal{X}_{\text{base}} + \boldsymbol{\delta}), \tag{3}$$

where $\mathbf{M}$ softly selects the editable regions and modulates the perturbation's effect during optimization, while $\boldsymbol{\delta}$ determines the magnitude and direction of modifications. The baseline sequence $\mathcal{X}_{\text{base}}$, typically defined as the mean or a smoothed version of $\mathcal{X}$, ensures that the generated counterfactual remains semantically coherent and temporally consistent, thereby avoiding unrealistic alterations.

The objective is to produce counterfactual time series that preserve the intrinsic temporal structure of the data while fulfilling task-specific desiderata. To this end, we formulate a composite optimization objective that balances anomaly suppression, sparsity, and temporal consistency:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\underbrace{\mathbb{E}_t\left[f(\tilde{\mathcal{X}})_t(1-w_t)\right]}_{\text{Anomaly Suppression}} + \underbrace{\lambda_1\|\mathbf{M}\|_1}_{\text{Sparsity}} \\ &+ \underbrace{\lambda_2\left(\|\mathcal{F}(\tilde{\mathcal{X}}) - \mathcal{F}(\mathcal{X})\|_2^2 + \log\left(1+\text{Var}(\nabla_t\tilde{\mathcal{X}})\right)\right)}_{\text{Temporal Consistency}},\end{aligned} \tag{4}$$

where $\tilde{\mathcal{X}}$ denotes the counterfactual sequence, $\mathcal{F}(\cdot)$ represents the discrete Fourier transform, and f is a fixed, pretrained anomaly scoring function. The weight w_t is computed by applying a smooth, differentiable transformation to the change in anomaly score at time t before and after applying perturbations, thereby enabling adaptive reweighting based on detection sensitivity.

This objective function Eq. (4) instantiates a time series–specific counterfactual generation loss, incorporating regularization terms tailored to capture fundamental properties of time series data such as temporal continuity, stationarity, and frequency characteristics. The perturbation $\boldsymbol{\delta}$ and mask $\mathbf{M}$ are jointly optimized subject to the following principles:

- **Anomaly suppression**: The counterfactual $\tilde{\mathcal{X}}$ should reduce the anomaly scores assigned by the pre-trained anomaly detection function $f(\cdot)$.
- **Sparsity**: The modifications should be sparse and localized to enhance interpretability.
- **Temporal Consistency**: Temporal smoothness and consistency must be preserved to maintain realistic dynamical behavior.

Each component of the objective aligns with these design principles. The anomaly suppression term encourages $\tilde{\mathcal{X}}$ to reside closer to the normal data manifold as defined by $f(\cdot)$. The sparsity term imposes an ℓ_1 penalty on the mask $\mathbf{M}$, incentivizing minimal and focused interventions. The temporal consistency term combines frequency-domain alignment, enforced via the Fourier transform, to retain global temporal patterns such as periodicity, with a gradient variance penalty that discourages abrupt local fluctuations, thereby fostering smoothness and stationarity. This formulation emphasizes the central role of the mask $\mathbf{M}$ and perturbation $\boldsymbol{\delta}$ in facilitating controlled, minimal, and interpretable modifications of the original time series. In the subsequent subsection, we elaborate on the learning mechanisms through which JoCE optimizes these components to generate effective counterfactual time series.

Masking and Perturbation Mechanism

Building on the design objectives, JoCE identifies and modifies critical regions in time series to explain anomalies effectively. It generates semantically meaningful counterfactuals by perturbing only the minimal set of input regions necessary to flip anomaly detection outcomes, preserving overall input structure and interpretability. JoCE is model-agnostic and operates post hoc with black-box access to a scoring function $f : \mathbb{R}^{D\times T} \to \mathbb{R}^T$, requiring no knowledge of the detector's internals or ground truth labels for training. The architecture consists of a *dual-axis masking module* to identify candidate regions and a *perturbation module* to generate plausible counterfactuals.

Dual-axis Masking Module To identify salient time-feature regions relevant to detected anomalies, we propose a mask generator that operates over both temporal and feature dimensions, promoting sparsity and interpretability. The temporal axis is partitioned into non-overlapping segments of fixed length, and for each segment-feature pair, the module learns a binary segment-level mask $\mathbf{M}_{\text{seg}} \in 0, 1^{B\times K\times D}$, where B is the batch size, K is the number of temporal segments, and D is the feature dimension. This coarse-grained mask is upsampled to the original temporal resolution by uniformly repeating each segment's mask across its corresponding time steps, yielding a full-resolution mask $\mathbf{M} \in 0, 1^{B\times T\times D}$.

To enable the stable training, we apply the Gumbel-Softmax trick (Herrmann, Bowen, and Zabih 2020) with a sigmoid activation and straight-through gradient estimation. A learnable logit tensor $\boldsymbol{\theta} \in \mathbb{R}^{1\times K\times D}$ is shared across

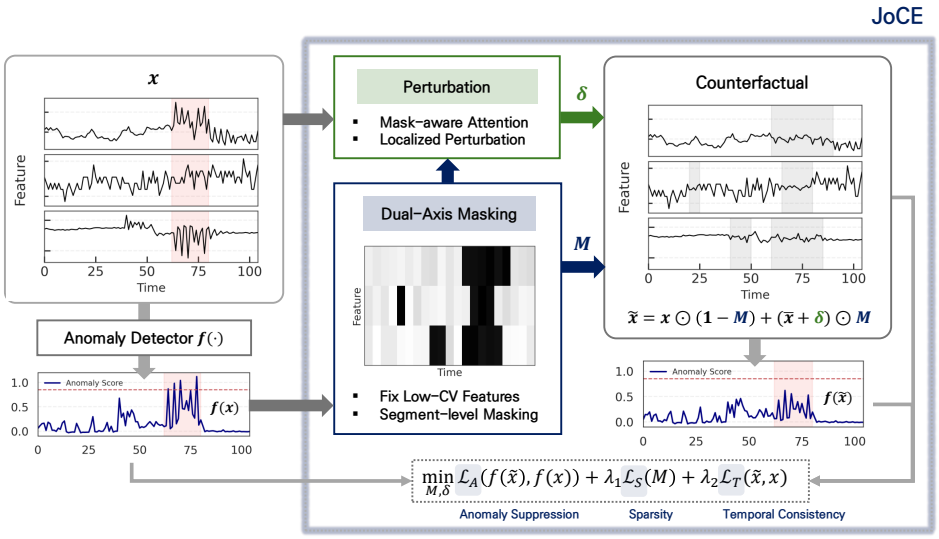


Figure 1: **Overview of JoCE**. JoCE provides counterfactual explanations for the behavior of an anomaly detection model f through two key modules: (1) a dual-axis masking module that identifies the most relevant time-feature regions influencing the model's output, and (2) a perturbation module that generates realistic and interpretable counterfactuals by selectively modifying those regions. These counterfactuals lower the anomaly score below a threshold, flipping anomalous segments to normal.

batches and expanded during forward passes. Random Gumbel noise is added to these logits, and the result is passed through a sigmoid function to produce soft, probabilistic masks. During training, hard masks can be obtained by thresholding at 0.5 and applying the straight-through estimator for gradient propagation. Additionally, a static feature-wise binary mask can be applied to prevent low-variance or uninformative features from being selected, enforcing the model's focus on meaningful dimensions. This design enables the model to identify compact, informative regions in both time and feature space that are most responsible for the anomaly.

Perturbation Module Given the input $\mathbf{X} \in \mathbb{R}^{B \times T \times D}$ and the binary mask $\mathbf{M} \in \{0,1\}^{B \times T \times D}$, the perturbation module predicts an additive perturbation $\delta \in \mathbb{R}^{B \times T \times D}$ to generate the counterfactual output $\mathbf{X}'$ as defined in Equation (3). To construct the base input $\bar{\mathbf{X}}$, we apply a low-pass filter to $\mathbf{X}$, which smooths out high-frequency components and enhances the realism of the generated counterfactuals. This formulation ensures that only the regions indicated by the mask $\mathbf{M}$ are modified, while the rest of the input remains unchanged, thereby preserving the overall structure and interpretability of the original time series.

To produce the perturbation δ, we employ a Transformer-based encoder that takes as input the concatenated tensor

$$[\mathbf{X}, \mathbf{M}] \in \mathbb{R}^{B \times T \times 2D},$$

where $\mathbf{M}$ explicitly encodes the regions designated for modification. This concatenated tensor is first linearly projected into a hidden representation, which is then processed by multiple Transformer blocks. Each block incorporates a self-attention mechanism augmented with an attention bias that emphasizes interactions among masked time steps. The bias is computed by averaging $\mathbf{M}$ over the feature dimension to obtain a per-time-step score, and then constructing a bias matrix via an outer product, scaled by a learnable weight. This mechanism guides the model to concentrate its representational capacity on the masked regions, enabling targeted and meaningful perturbations. Finally, the hidden representation is projected back to the input dimension and passed through a scaled $\tanh$ activation, which constrains the magnitude of the perturbation and ensures stable training while maintaining sufficient flexibility for effective counterfactual generation.

These modules enable JoCE to generate sparse, precise, and temporally coherent counterfactuals using only anomaly scores, making it widely applicable and interpretable.

Experiments

Datasets

We evaluate our method on four widely used real-world datasets that serve as standard benchmarks in time series anomaly detection research (Xu et al. 2021; Yang et al. 2023; Dai et al. 2024; Liu et al. 2024). These datasets are: (1) SMD[1], which contains five weeks of data from multiple server machines at an internet company, (2) SWaT[2], derived from a secure water treatment system, (3) SMAP[3], comprising soil sample and telemetry data from NASA's Mars rover, and (4) PSM[4], provided by eBay and consisting of 25-dimensional server machine metrics. Each dataset includes

[1]https://github.com/NetManAIOps/OmniAnomaly
[2]https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/
[3]https://nsidc.org/data/smap/data
[4]https://github.com/eBay/RANSynCoders/tree/main/data

anomaly labels that indicate whether each time point is normal or anomalous. Table 1 summarizes key characteristics of the datasets used in this work.

Table 1: Summary of benchmark datasets.

Dataset	# Features	# Samples	Anomaly Ratio
SMD	38	6,656	4.20%
SWaT	51	4,224	11.94%
SMAP	25	3,840	12.70%
PSM	25	768	24.56%

Experimental Setup

Baselines We compare JoCE against several baselines. Among existing counterfactual generation methods for TSAD, **AR-Pro** (Ji et al. 2024) is the only approach with publicly available code and is included in our evaluation; it generates counterfactuals via a diffusion-based repair process that reconstructs inputs under non-anomalous conditions. Although Sparse DPE (Sulem et al. 2022) is conceptually relevant, its implementation was not publicly released and is thus excluded. To supplement our evaluation, we implement three naive baselines inspired by prior work (Sulem et al. 2022): (1) **Flat**, replacing the sequence with its mean, (2) **KS**, applying kernel smoothing to preserve local trends, and (3) **LPF**, using low-pass filtering to remove high-frequency components. Compared to KS, LPF more effectively suppresses sharp, peak-shaped anomalies, making it especially suitable for point anomaly reduction.

Evaluation Metrics We assess counterfactual time series generation using metrics that capture effectiveness, realism, and feasibility, reflecting the properties outlined in Section . Specifically, we evaluate whether counterfactuals (1) flip anomalous points to normal (validity), (2) introduce minimal, localized changes (sparsity, FCR), and (3) preserve the original data distribution (TS-FID, MMD). These metrics are commonly applied in counterfactual studies for vision (Jeanneret, Simon, and Jurie 2023; Dash, Balasubramanian, and Sharma 2022; Boreiko et al. 2022; Jeanneret, Simon, and Jurie 2024) and tabular data (Wachter, Mittelstadt, and Russell 2017; Zhou et al. 2025).

Validity The proportion of time points in the generated counterfactuals that are flipped from anomalous to normal, as determined by an anomaly score and a predefined threshold. Higher validity indicates that the counterfactuals effectively suppress anomalies while adhering to domain-specific constraints, thereby making them more actionable and meaningful.

Sparsity The proportion of elements in the counterfactual time series that differ from the original input. Lower sparsity indicates that the counterfactual introduces fewer changes, preserving most of the original input.

TS-FID The difference between real and generated time series, measured by applying the Fréchet Inception Distance (FID) (Salimans et al. 2016) to flattened time series data. Lower TS-FID values indicate that the generated data better matches the distribution of real time series.

MMD The Maximum Mean Discrepancy (MMD) quantifies the distributional difference between real and generated samples via a kernel-based distance between their mean embeddings (Gretton et al. 2012). Lower MMD values suggest higher distributional similarity and fidelity.

FCR Feature Change Ratio (FCR) is the proportion of features that are changed in a multivariate counterfactual. In realistic settings, not all variables can be freely altered. Therefore, we report the ratio of changed features to reflect how feasible a counterfactual is in practice. Lower FCR indicates more realistic and constrained counterfactuals.

Implementation Details For generating counterfactuals, we employ pretrained anomaly scoring functions derived from established detection models. In particular, we utilize a pretrained GPT-2 model following the approach in (Ji et al. 2024), along with a pretrained DCDetector (Yang et al. 2023). Both anomaly detection models are pretrained using the hyperparameters reported in their original works to maintain consistency.

For JoCE, we create dataloaders by selecting only sequences that contain at least one anomalous point from each dataset, then split the data into training and testing sets with an 80:20 ratio. A fixed window length of 105 is applied across all datasets, and counterfactual generation is performed only on these subsequences. The masking segment length is set to 5, and a feature masking constraint excludes features with a coefficient of variation below 0.2. Training is conducted for 20 epochs using the Adam optimizer with a learning rate of 10^{-3}. The batch size is set to 256 (and 16 for GPT-2), following the batch size settings used during the training of each detection model. The parameters λ_1 and λ_2 in Equation (4) are set to 10^{-1} and 10^{-3}, respectively. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU using PyTorch.

Results and Analysis

Counterfactual Quality Comparison

We evaluate the performance of the proposed method against several baseline counterfactual explanation techniques on multiple real-world time series anomaly detection datasets. The evaluations use two different anomaly scoring models: a GPT-2-based model (Radford et al. 2019) and DCdetector (Yang et al. 2023). The GPT-2-based model is included to ensure a fair comparison with prior work (Ji et al. 2024). To assess the robustness and generalizability of JoCE, we also include DCdetector. Both detectors generate pointwise anomaly scores over time.

Tables 2 and 3 present the quantitative evaluation results under each experimental setting. Among the reported metrics, *Validity* is evaluated on a higher-is-better basis, whereas the remaining metrics are interpreted as lower-is-better. To provide a comprehensive and robust comparison, we additionally report the *mean rank* of each method across all metrics, where a lower mean rank indicates superior overall performance.

Table 2: **Performance of JoCE with GPT-2.** Evaluation metrics (Validity, Sparsity, TS-FID, MMD, FCR) across datasets using GPT-2 based AD model. Mean-rank-based rankings shown; best results in bold. † indicates statistically significant improvement over all baselines (Wilcoxon signed-rank test with Holm correction, $p_{\text{Holm}} < 0.05$).

Dataset	Method	Validity	Sparsity†	TS-FID†	MMD†	FCR†	Rank
SMD	Flat	**0.9898**	0.7799	20.7584	0.0199	0.7799	2.8
	KS	0.8930	0.7404	8.3114	0.0199	0.9482	2.8
	LPF	0.8687	0.7794	13.3602	0.0200	0.7799	3.4
	AR-Pro	0.8203	0.6663	40.4954	0.1142	0.6667	4.4
	JoCE	0.9106	**0.0099**	**5.1418**	**0.0148**	**0.1429**	**1.2**
SWaT	Flat	**0.9888**	0.4543	12.0374	0.0514	0.4515	2.8
	KS	0.9799	0.3125	2.2423	0.0514	0.7898	2.6
	LPF	0.9789	0.4536	3.2629	0.0511	0.4515	2.6
	AR-Pro	0.9775	0.8942	78.0373	0.0934	0.8947	4.8
	JoCE	0.9701	**0.0171**	**0.8119**	**0.0299**	**0.2237**	**1.8**
SMAP	Flat	**0.9903**	0.1280	42.7180	0.1144	0.1280	3.0
	KS	0.9888	0.0955	**18.6799**	0.0328	0.9606	2.8
	LPF	0.9833	0.1277	38.2659	0.0287	0.1280	2.8
	AR-Pro	0.9811	0.7106	64.8541	0.2443	0.6743	4.8
	JoCE	0.9848	**0.0121**	18.0707	**0.0125**	**0.0986**	**1.4**
PSM	Flat	**0.9910**	0.9941	8.0640	**0.0869**	0.9948	2.2
	KS	0.9662	0.9827	2.1564	0.0870	1.0000	2.6
	LPF	0.9346	0.9945	2.8703	0.0870	0.9948	3.2
	AR-Pro	0.9909	1.0000	29.7267	0.3334	1.0000	4.2
	JoCE	0.9388	**0.3731**	**0.9303**	0.0870	**0.9930**	**1.8**

Table 3: **Performance of JoCE with DCdetector.** Evaluation metrics (Validity, Sparsity, TS-FID, MMD, FCR) across datasets using DCdetector (Yang et al. 2023). Mean-rank-based rankings shown; best results in bold. † indicates statistically significant improvement over all baselines (Wilcoxon signed-rank test with Holm correction, $p_{\text{Holm}} < 0.05$).

Dataset	Method	Validity	Sparsity†	TS-FID†	MMD†	FCR†	Rank
SMD	Flat	0.9893	0.7799	20.7584	**0.0199**	0.7799	3.4
	KS	0.9898	0.7404	8.3114	**0.0199**	0.9482	2.6
	LPF	**0.9908**	0.7794	13.3602	0.0200	0.7799	3.0
	AR-Pro	0.9870	0.6663	40.0645	0.1142	**0.6667**	3.6
	JoCE	0.9898	**0.3760**	**8.0070**	**0.0199**	0.7775	**1.4**
SWaT	Flat	0.9901	0.4543	12.0374	0.0514	0.4515	3.4
	KS	0.9904	0.3125	2.2423	0.0514	0.7898	2.6
	LPF	0.9904	0.4536	3.2629	0.0511	0.4515	2.6
	AR-Pro	**0.9905**	0.8941	78.2465	0.0934	0.8947	4.8
	JoCE	0.9900	**0.2379**	**1.8791**	**0.0506**	**0.4480**	**1.8**
SMAP	Flat	0.9899	0.1280	42.7180	0.1144	**0.1280**	3.0
	KS	0.9899	0.0955	**18.6799**	0.0328	0.9606	2.6
	LPF	0.9899	0.1277	38.2659	0.0287	**0.1280**	2.2
	AR-Pro	**0.9905**	0.7092	65.3774	0.2448	0.6114	4.0
	JoCE	0.9893	**0.0709**	27.1979	**0.0242**	**0.1280**	**2.0**
PSM	Flat	0.9910	0.9941	8.0640	**0.0869**	0.9948	**2.0**
	KS	0.9887	0.9827	2.1564	0.0870	1.0000	3.2
	LPF	0.9880	0.9945	2.8703	0.0870	0.9948	3.0
	AR-Pro	**0.9955**	1.000	29.4511	0.3334	1.0000	3.8
	JoCE	0.9868	**0.6062**	**1.8726**	0.0870	**0.9930**	**2.0**

The results in Tables 2 and 3 show that JoCE consistently achieves the best overall performance, with the lowest average ranks across five metrics: TS-FID, MMD, validity, sparsity, and feature change ratio (FCR). This highlights JoCE's ability to generate realistic, minimally altered, and effective counterfactual explanations across diverse TSAD datasets and two anomaly scoring models.

JoCE attains the lowest or near-lowest TS-FID and MMD scores, reflecting high fidelity and distributional realism, and excels in sparsity metrics, producing counterfactuals with fewer modifications and affecting fewer feature dimensions. While not always the highest in validity, JoCE remains competitive, offering a favorable trade-off between validity, sparsity, and realism. Competing methods, such as Flat and LPF, often achieve high validity at the cost of excessive perturbations, which reduces interpretability.

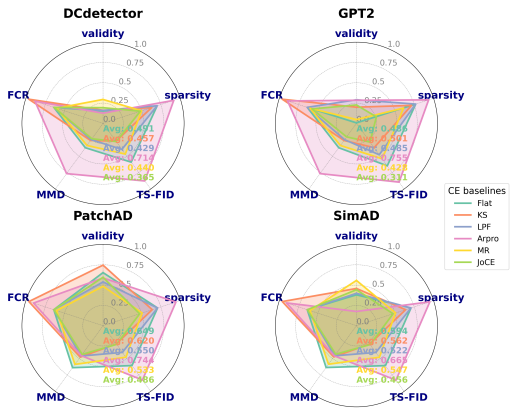


Figure 2: **Radar plots of normalized metrics.** Each polygon shows the average performance of a counterfactual generator across datasets for two anomaly detectors. Smaller areas indicate better trade-offs, with values displayed near the edges in matching colors.

Under GPT-2, JoCE generalizes well on challenging datasets such as SMAP and SWaT, reducing TS-FID and MMD while maintaining low FCR. Under DCdetector, it balances performance, improving validity without compromising sparsity or fidelity—for instance, on SWaT, achieving the highest validity (0.950) and lowest TS-FID (0.792). Figure 2 presents radar plots of normalized metrics for each generator under both detectors. Validity was transformed so that lower values indicate better performance, and the average polygonal area quantifies overall balance: smaller areas reflect generators with high validity, low sparsity, and improved fidelity.

Qualitative Evaluation of JoCE

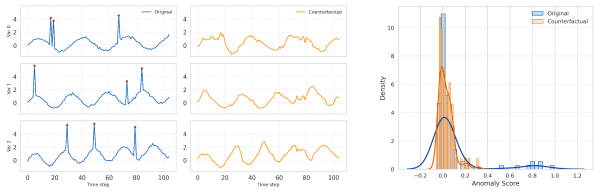
To qualitatively assess JoCE, we construct a synthetic multivariate time series with $D = 3$ variables. Each variable is a sinusoid with a distinct frequency and a small positive linear trend, combined with Gaussian noise to emulate realistic fluctuations. Sparse point anomalies are injected as abrupt, large-magnitude deviations at random time points.

Specifically, the d-th variable is

$$x_t^{(d)} = \sin(\omega_d t) + \alpha_d t + \epsilon_t^{(d)} + A_t^{(d)},$$

where $\omega_d = 0.2 + 0.05d$, $\alpha_d = 0.005 + 0.005d$, $\epsilon_t^{(d)} \sim \mathcal{N}(0, 0.1^2)$, and $A_t^{(d)}$ is nonzero only at sparse anomaly points with magnitude sampled from a uniform distribution.

Figure 3 illustrates JoCE's effectiveness on a synthetic dataset. Left panel shows the original signals with anomalies highlighted in red and the corresponding counterfactual time series generated by JoCE, which removes anomalous peaks



(a) Original and Counterfactual Time Series (b) Distribution of Anomaly scores

Figure 3: **Visualization of JoCE on a synthetic time series. (a) Left:** original signals with injected point anomalies (red circles) and the corresponding counterfactual time series generated by JoCE. **(b) Right:** anomaly score distributions, showing how JoCE suppresses anomalies while retaining normal characteristics.

while preserving temporal dynamics. The counterfactual trajectories remain smooth and consistent, indicating that JoCE disentangles anomalies from normal patterns. Right panel compares anomaly score distributions before and after counterfactual generation, showing a general shift toward lower scores and demonstrating meaningful suppression of anomalous features, despite some overlap between the distributions.

Comparison with CE Baselines We evaluated JoCE on real-world benchmark datasets to compare its counterfactual reconstruction behavior with existing CE baselines. For each dataset–detector pair, counterfactual time series were generated using five baseline models and JoCE.

We focus on two representative anomaly types: *peak* and *transient fluctuation*. Peak anomalies refer to short, abrupt deviations, while transient fluctuations correspond to temporary oscillatory behavior during a level shift. Peak anomalies test a model's ability to suppress localized outliers without distorting surrounding context, whereas transient fluctuations test preservation of baseline and temporal continuity (Blázquez-García et al. 2021; Tsay 1988).

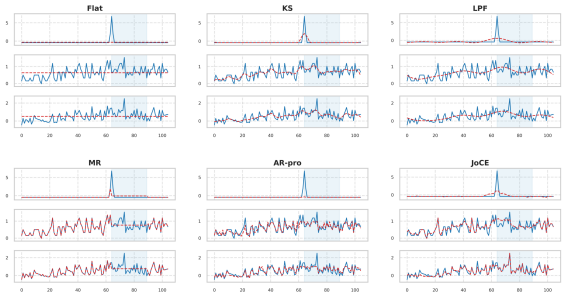


Figure 4: **Peak anomalies.** Blue solid lines show the original series, and red dashed lines indicate counterfactuals generated by each method. Blue-shaded regions highlight the anomalous intervals.

Figure 4 shows that all baselines reduce spike magnitude. Although anomaly labels mark entire variable sets as anomalous, AR-pro and JoCE better maintain local consistency in unaffected variables, preserving temporal characteristics while mitigating peaks.

For transient fluctuations, *Flat* and *AR-pro* produce nearly constant reconstructions, whereas *KS*, *LPF*, and JoCE generate smoother transitions. Figure 5 compares AR-pro and JoCE using heatmaps. Dark points in variables 2 and 19 indicate transient fluctuations. AR-pro largely flattens the signal, while JoCE reduces the anomalous level but preserves the original series outside anomalous segments, better maintaining temporal continuity.

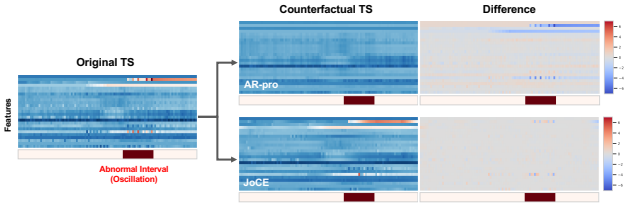


Figure 5: **Transient fluctuation during level shift (AR-Pro vs. JoCE).** Heatmaps of original and counterfactual multivariate time series. Horizontal and vertical axes represent time and features. Each row shows the counterfactual series and its difference from the original. Top row: AR-Pro; bottom row: JoCE.

Analysis of M Selection in JoCE A key component in generating realistic counterfactuals in JoCE is the masking module M, which determines the time–variable regions to be modified. We assessed its effectiveness using anomaly labels as reference masks. In multivariate time series, an anomaly in one variable may spuriously mask normal variables, causing false positives. By employing macro metrics, such as macro F1 and macro precision, we evaluate each variable independently, providing a robust measure of M's ability to identify relevant regions while preserving normal temporal patterns, as summarized in Table 4.

Table 4: **Macro F1 scores for the masking module M in JoCE** (with corresponding macro precision in parentheses). Higher values indicate better alignment between modified regions and true anomalies.

Dataset	GPT-2	DCdetector	PatchAD	SimAD
SMD	0.846 (0.915)	0.259 (0.289)	0.264 (0.294)	0.268 (0.298)
SWAT	0.831 (0.874)	0.592 (0.628)	0.590 (0.625)	0.588 (0.623)
SMAP	0.826 (0.908)	0.762 (0.841)	0.772 (0.852)	0.760 (0.838)
PSM	0.076 (0.129)	0.055 (0.093)	0.064 (0.104)	0.056 (0.095)

Across most datasets, JoCE achieves macro F1 scores above 0.25, showing that it modifies relevant regions while preserving normal temporal patterns elsewhere. For SMD, SWaT, and SMAP, macro F1 under GPT-2 ranges from 0.826 to 0.846, reflecting balanced precision and recall. For PSM, the macro F1 is low (0.076), indicating few false positives but many missed anomalies, likely due to the sparse anomaly distribution despite a higher apparent anomaly ratio (Table 1). Overall, M identifies relevant regions well, though sparse-anomaly datasets may need improved sensitivity.

Conclusion

In this work, we present JoCE, a novel framework for generating counterfactual explanations in multivariate time series anomaly detection. By jointly modeling feature-wise and temporal dependencies, JoCE generates counterfactuals that balance anomaly score reduction, sparsity, and temporal consistency. Experiments on real-world datasets demonstrate that JoCE produces realistic and interpretable explanations, providing domain experts with actionable insights into black-box anomaly detectors. Future work will explore domain-specific constraints, interactive generation, and manifold-based analyses to investigate the geometric relationship between original and counterfactual trajectories. These directions aim to deepen understanding of how counterfactual and adversarial perturbations differ in their impact on anomaly detection, paving the way for more robust and theoretically grounded explanation methods.

References

Abououf, M.; Singh, S.; Mizouni, R.; and Otrok, H. 2023. Explainable AI for event and anomaly detection and classification in healthcare monitoring systems. *IEEE Internet of Things Journal*, 11(2): 3446–3457.

Ates, E.; Aksar, B.; Leung, V. J.; and Coskun, A. K. 2021. Counterfactual explanations for multivariate time series. In *2021 international conference on applied artificial intelligence (ICAPAI)*, 1–8. IEEE.

Blázquez-García, A.; Conde, A.; Mori, U.; and Lozano, J. A. 2021. A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, 54(3): 1–33.

Boreiko, V.; Augustin, M.; Croce, F.; Berens, P.; and Hein, M. 2022. Sparse visual counterfactual explanations in image space. In *DAGM German Conference on Pattern Recognition*, 133–148. Springer.

Dai, Z.; He, L.; Yang, S.; and Leeke, M. 2024. SARAD: Spatial association-aware anomaly detection and diagnosis for multivariate time series. *Advances in Neural Information Processing Systems*, 37: 48371–48410.

Dash, S.; Balasubramanian, V. N.; and Sharma, A. 2022. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 915–924.

Giannoulis, M.; Harris, A.; and Barra, V. 2023. DITAN: A deep-learning domain agnostic framework for detection and interpretation of temporally-based multivariate ANomalies. *Pattern Recognition*, 143: 109814.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, 2376–2384. PMLR.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.

Herrmann, C.; Bowen, R. S.; and Zabih, R. 2020. Channel selection using gumbel softmax. In *European conference on computer vision*, 241–257. Springer.

Huang, Q.; Kitharidis, S.; Bäck, T.; and van Stein, N. 2024. TX-Gen: Multi-Objective Optimization for Sparse Counterfactual Explanations for Time-Series Classification. *arXiv preprint arXiv:2409.09461*.

Jeanneret, G.; Simon, L.; and Jurie, F. 2023. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16425–16435.

Jeanneret, G.; Simon, L.; and Jurie, F. 2024. Text-to-image models for counterfactual explanations: a black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4757–4767.

Ji, X.; Xue, A.; Wong, E.; Sokolsky, O.; and Lee, I. 2024. AR-Pro: Counterfactual Explanations for Anomaly Repair with Formal Properties. *Advances in Neural Information Processing Systems*, 37: 16133–16159.

Lee, D.; Malacarne, S.; and Aune, E. 2024. Explainable time series anomaly detection using masked latent generative modeling. *Pattern Recognition*, 156: 110826.

Liu, J.; Zhang, C.; Qian, J.; Ma, M.; Qin, S.; Bansal, C.; Lin, Q.; Rajmohan, S.; and Zhang, D. 2024. Large language models can deliver accurate and interpretable time series anomaly detection. *arXiv preprint arXiv:2405.15370*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mochaourab, R.; Venkitaraman, A.; Samsten, I.; Papapetrou, P.; and Rojas, C. R. 2022. Post hoc explainability for time series classification: Toward a signal processing perspective. *IEEE signal processing magazine*, 39(4): 119–129.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617.

Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; and Flach, P. 2020. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rojat, T.; Puget, R.; Filliat, D.; Del Ser, J.; Gelin, R.; and Díaz-Rodríguez, N. 2021. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*.

Sabharwal, R.; Miah, S. J.; Wamba, S. F.; and Cook, P. 2024. Extending application of explainable artificial intelligence for managers in financial organizations. *Annals of Operations Research*, 1–31.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sulem, D.; Donini, M.; Zafar, M. B.; Aubet, F.-X.; Gasthaus, J.; Januschowski, T.; Das, S.; Kenthapadi, K.; and Archambeau, C. 2022. Diverse counterfactual explanations for anomaly detection in time series. *arXiv preprint arXiv:2203.11103*.

Tripathy, S. M.; Chouhan, A.; Dix, M.; Kotriwala, A.; Klöpper, B.; and Prabhune, A. 2022. Explaining Anomalies in Industrial Multivariate Time-series Data with the help of eXplainable AI. In *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 226–233. IEEE.

Tsay, R. S. 1988. Outliers, level shifts, and variance changes in time series. *Journal of forecasting*, 7(1): 1–20.

Tuli, S.; Casale, G.; and Jennings, N. R. 2022. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*.

Yang, Y.; Zhang, C.; Zhou, T.; Wen, Q.; and Sun, L. 2023. 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*, 3033–3045.

Younis, R.; Hakmeh, A.; and Ahmadi, Z. 2024. MTS2Graph: Interpretable multivariate time series classification with temporal evolving graphs. *Pattern Recognition*, 152: 110486.

Zamanzadeh Darban, Z.; Webb, G. I.; Pan, S.; Aggarwal, C.; and Salehi, M. 2024. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1): 1–42.

Zhou, P.; Tong, Q.; Chen, S.; Zhang, Y.; and Wu, X. 2025. EACE: Explain Anomaly via Counterfactual Explanations. *Pattern Recognition*, 164: 111532.