

Towards A Unified Information Bottleneck Framework for Time Series Explanations

Xu Zheng*, Zichuan Liu*, Zhuomin Chen, Mayur Akewar, Janki Bhimani, Jason Liu, Mo Sha, Jingchao Ni, Wei Cheng, Dongsheng Luo

Abstract—Explaining deep learning models operating on time series data is crucial in various applications that require transparent and interpretable insights into model behavior. Existing explanation methods generally fall into two categories: attribution-based explanations, which identify the temporal regions most responsible for a prediction, and counterfactual explanations, which reveal how an input should be modified to alter the model’s decision. Despite valuable insights, these two fields are largely studied independently. This disconnect leaves attribution methods lacking causal validation, while counterfactual methods suffer from severe instability, producing adversarial-like noise instead of meaningful explanations. In this work, we revisit time-series explainability from an information-theoretic perspective and show that existing explainers are vulnerable to trivial solutions and distributional shifts. To address these limitations, we propose a unified objective function for explainable time series learning that bridges attribution and counterfactual reasoning within a single framework. Building upon the Information Bottleneck principle, our formulation explicitly prevents trivial explanations and out-of-distribution counterfactuals. Based on this objective function, we introduce TimeX++, a novel explanation framework that learns a parametric transformation network to construct explanation-embedded instances, where preserved information yields attribution explanations and controlled information removal produces stable counterfactual explanations. We evaluate TimeX++ on synthetic and real-world benchmarks against state-of-the-art baselines. Extensive quantitative and qualitative results show that TimeX++ consistently outperforms competing methods, yielding faithful attributions and stable counterfactual explanations.

Index Terms—Explainable AI, Time Series, Deep Learning

I. INTRODUCTION

Deep learning has become a cornerstone technology in analyzing time series data, prevalent in scenarios such as finance [4], healthcare [24], and environmental science [60]. Despite its predictive success, the black-box nature of deep neural networks remains a critical limitation. The lack of explainability hinders user trust and prevents domain experts from gaining actionable insights, which are essential in these sensitive applications [49, 15, 47].

X. Zheng, Z. Chen, M. Akewar, J. Bhimani, J. Liu, and M. Sha are with Florida International University. E-mail: {xzhen019, zchen051, makew001, jbhimani, liux, msha}@fiu.edu

Z. Liu is with Carnegie Mellon University. E-mail: zichuanl@andrew.cmu.edu

J. Ni is with the University of Houston. E-mail: jni7@uh.edu

W. Cheng is with NEC Labs America. E-mail: weicheng@nec-labs.com

D. Luo is with Singapore Management University. This work was partially conducted while he was at Florida International University. E-mail: luodongsheng01@gmail.com

* indicates equal contribution.

Current efforts to enhance the explainability of time series models primarily fall into two disconnected paradigms: attribution and counterfactual explanations. Attribution methods focus on a post-hoc pinpointing of the most salient temporal regions that dominate the model’s prediction [20, 9, 54]. Popular perturbation-based attribution methods, such as Dynamask [9] and Extrmask [12], evaluate feature importance by masking non-salient regions. Nevertheless, these methods often rely on ad-hoc objectives lacking a solid theoretical foundation and frequently suffer from out-of-distribution (OOD) issues when the masked sub-instances are evaluated by the classifier [50, 59, 43]. Conversely, counterfactual explanations answer “what-if” scenarios by finding the minimal perturbation required to change the model’s prediction [6]. Nevertheless, generating counterfactuals in the vast and continuous space without structural guidance often leads to trivial solutions, such as mode collapse in conditional generation [40, 39, 23]. Pursuing label changes alone can result in adversarial attacks that satisfy the classifier’s decision boundary while failing to capture meaningful and interpretable changes [28, 25, 42].

Crucially, existing research typically treats attribution and counterfactual explanations as completely isolated tasks, which introduces fundamental flaws [26]. When treated separately, attributions remain purely descriptive, highlighting salient features but not causally verifying that altering those specific features actually changes the model’s decision [22, 37]. Conversely, without the structural guidance provided by attributions, counterfactual generation lacks semantic grounding. If a counterfactual method optimizes for a label flip across the entire time series without knowing which features are semantically responsible for the original prediction, it inevitably produces adversarial noise rather than a valid, interpretable pattern shift.

To overcome these limitations, we establish a theoretical foundation that unifies attribution and counterfactual explanations through the lens of the Information Bottleneck (IB) principle [53]. Formally, given a time series instance X and its label Y , the IB principle formulates the extraction of an *explanation* sub-instance X' as an optimization problem that balances compactness and informativeness: X' minimizes $I(X; X') - \alpha I(X'; Y)$, where $I(\cdot; \cdot)$ denotes mutual information and $\alpha > 0$ controls the trade-off [36]. In this context, the **attribution explanation** equates exactly to solving this IB objective. By minimizing $I(X; X')$, it discards redundant temporal dynamics, and by maximizing $I(X'; Y)$, it retains the sufficient statistics necessary for predicting Y . Consequently, the attribution X' acts as the optimal information bottleneck, defining a semantically salient sub-manifold of the original

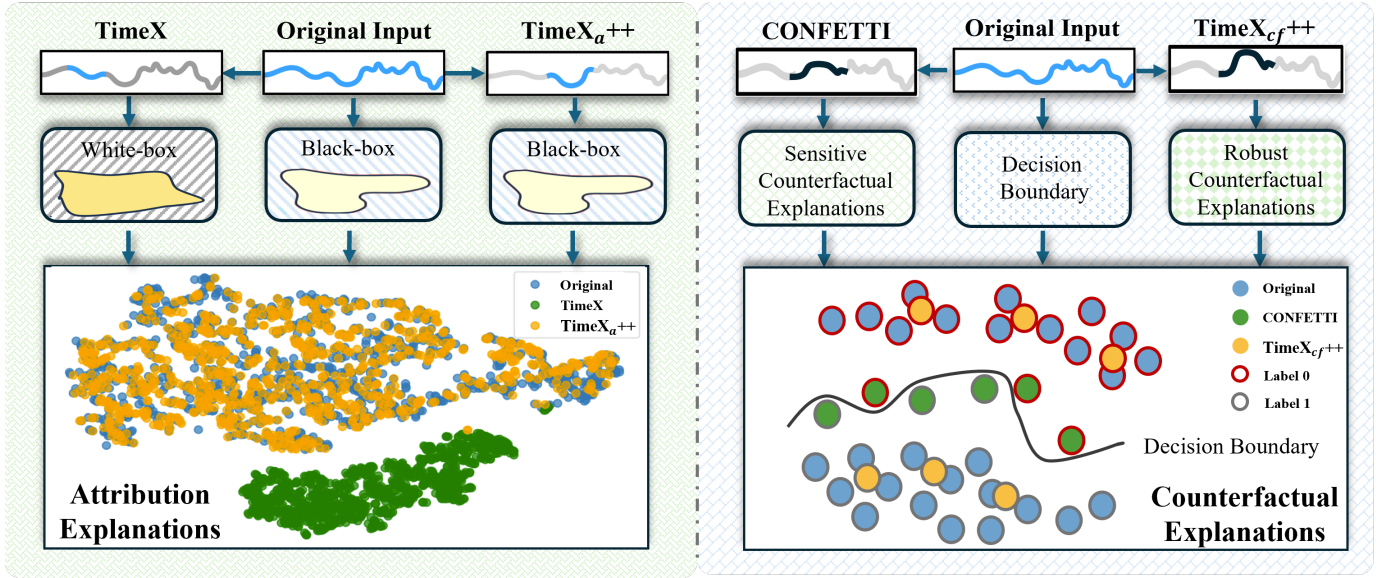


Fig. 1: A comparison between our model and previous work. For the attribution explanations, the latent embeddings of explanations are learned from the ECG dataset. Our explanations are within the original distribution, whereas the reference model’s explanations are not. For the counterfactual explanations, our framework represents more robust results.

high-dimensional temporal space.

Building upon this, we analyze the **counterfactual explanation** within the same information-theoretic framework. Generating a counterfactual requires finding a perturbed instance X'' that alters the classifier’s prediction to a target label Y'' . However, optimizing this perturbation directly in the unconstrained space of X inevitably explores the whole decision space, which could cause injecting adversarial noise into the redundant regions $X \setminus X'$ that the IB explicitly discarded. To ensure a semantically valid distribution shift, the counterfactual perturbation must be strictly bounded within the support of the information bottleneck X' . From an information-theoretic perspective, a legitimate counterfactual transition $Y \rightarrow Y''$ necessitates a corresponding manipulation within the sufficient statistics subspace defined by $I(X'; Y)$, rather than arbitrary deviations in the redundant space. Through this IB-driven formulation, attribution and counterfactual explanations are intrinsically unified as complementary mechanisms over the same information bottleneck. The attribution identifies the exact semantic support X' , while the counterfactual manipulation within this bounded support provides the causal verification that X' indeed dictates the predictive behavior.

However, directly applying the standard IB principle to time series presents critical challenges. Primarily, the exact computation of mutual information in continuous, high-dimensional temporal spaces is computationally intractable [35]. Furthermore, attempting to evaluate the extracted sub-instances X' directly with the original classifier often violates the underlying data manifold, inevitably leading to OOD predictions and unreliable evaluations [18, 61]. To address these challenges, we propose TimeX++, a novel and unified framework designed to simultaneously extract both attribution and counterfactual explanations. It optimizes a practical, unified objective function. We replace the compactness quantifier $I(X; X')$ with a tractable variational upper bound comprising minimality

and discrete constraints, and substitute the informativeness quantifier $I(X'; Y)$ with a measure of label consistency that strictly preserves the underlying data distribution. The key advantage of our unified architecture lies in its dual capability. First, for attribution explanations, it generates in-distribution, explanation-embedded instances, entirely circumventing the severe OOD issues that plague reference models (as illustrated in Figure 1). Second, for counterfactual explanations, by tightly coupling the counterfactual search with the in-distribution attribution bottleneck, TimeX++ constrains the counterfactual trajectory along meaningful feature dimensions. This guarantees the generation of stable, semantically valid counterfactuals that are highly resistant to noise vulnerabilities. We summarize our contributions as follows:

- We formally investigate the limitations of treating attribution and counterfactual explanations as isolated tasks in time series. By analyzing them through the Information Bottleneck principle, we establish their theoretical connection and propose a practical, unified objective function.
- We propose a novel explanation framework, TimeX++, which successfully addresses the out-of-distribution shifting issue in attribution explanations by generating in-distribution sub-instances and leveraging this attribution bottleneck to generate stable, semantically grounded counterfactuals.
- We achieve state-of-the-art performance for both attribution and counterfactual explanation tasks on synthetic and real-world datasets and demonstrate their effectiveness.

II. RELATED WORK

a) Attribution Explanations in Time Series: Existing explainable AI (XAI) methods for time series primarily focus on feature attribution, which aims to identify the most salient temporal regions dictating a well-trained model’s decision [12, 4]. While early approaches relied on attention weights [7] or gradient signals [51], perturbation-based methods

have become the dominant paradigm. These methods evaluate feature importance by masking or altering data segments using static baselines [52], generative models [54], or information diminution [9, 33]. However, a critical flaw in perturbation-based attribution is the OOD problem that when salient features are masked or substituted, the resulting sub-instances often violate the underlying data manifold [18]. Consequently, querying the original classifier with these OOD samples leads to unreliable evaluations [59, 50].

b) Counterfactual Explanations for Time Series: In contrast to descriptive attribution, counterfactual explanations seek minimal modifications to an input required to alter the model's prediction. For time series data, existing paradigms are broadly classified into three categories. First, replacement-based methods (e.g., CoMTE [2], Native Guide [11]) substitute contiguous segments using nearest unlike neighbors or temporal shapelets [3, 32]. Second, gradient-based approaches (e.g., LatentCF++ [57], Glacier [56]) directly optimize perturbations in the input or learned latent spaces. Third, genetic frameworks (e.g., TSEvo [17], Sub-SpaCE [45]) explore the vast space of temporal segment edits using evolutionary objectives. Despite these advances, generating robust time series counterfactuals remains challenging. Because continuous temporal space is unconstrained, counterfactual optimization often exploits local decision boundary vulnerabilities. This results in the injection of meaningless noise that acts as an adversarial attack rather than providing a semantically interpretable pattern shift [28, 25].

III. PROBLEM FORMULATION

This work focuses on explainability in time series classification. Let $X \in \mathbb{R}^{T \times D}$ be a multivariate time series instance of length T with D features, and $\mathcal{X}$ is input space. A multivariate time series is one for which $D > 1$, or univariate. The value of the feature indexed d at time t is denoted by $X[t, d]$. A training set $\mathcal{T} = \{(X_i, Y_i) | i \in [N]\}$ consists of N time series instances X_i along with their associated labels Y_i , where $Y_i \in \mathcal{C}$ and $\mathcal{C} = \{1, 2, \dots, |\mathcal{C}|\}$ is the label space. A pre-trained black-box classifier $f: \mathcal{X} \rightarrow \mathcal{C}$ maps X to a predicted label $Y \in \mathcal{C}$.

Our objective is to formally unify feature attribution and counterfactual generation under the IB principle. In this unified view, the attribution acts as the optimal information bottleneck, and the counterfactual acts as a semantic perturbation strictly bounded within that bottleneck. To develop a general explainability framework, we focus on *post-hoc*, *instance-level* methods that are task-agnostic and treat the model $f(\cdot)$ as a black box [58]. Furthermore, we study model explainability, where the extracted sub-instance is required to be sufficient for the model output rather than the ground-truth label [13, 33].

Problem 1 (Attribution Explanations). *Given a trained model f and input X , the objective in post-hoc instance-level time series explanation is to find a sub-instance X' that 'explains' the prediction of f on X . The sub-instance X' is obtained by applying a binary mask $M \in \mathcal{M} = \{0, 1\}^{T \times D}$ on X , i.e., $X' = X \odot M$, where $\odot$ is the element-wise multiplication.*

To transition this discrete extraction into a differentiable continuous optimization, we define an attribution extractor $g(\cdot)$

that maps the input X to a mask $M \in [0, 1]^{T \times D}$. Each element is independently sampled from a Bernoulli distribution $\pi_{t,d}$.

Problem 2 (Counterfactual Explanations). *Given a trained model f , an input time series X with predicted label $Y = f(X)$, and a target label $Y'' \in \mathcal{C}$ ($Y'' \neq Y$), the objective is to find a counterfactual instance X'' such that $f(X'') = Y''$, while minimizing the perturbation distance $d(X, X'')$, where the counterfactual search space is restricted to the semantic support set defined by the attribution mask M .*

IV. EXPLAINING TIME SERIES LEARNING VIA INFORMATION BOTTLENECK

The IB principle extracts an attribution explanation sub-instance $X' = X \odot M$ by maximizing $I(X'; Y) - \alpha I(X; X')$. However, directly estimating the mutual information $I(X'; Y)$ for informativeness in high-dimensional continuous spaces is computationally intractable. To overcome and extend this, following established practices in explainable learning (e.g., PGExplainer [34]), we substitute the informativeness term with a tractable Label Consistency (LC) measure (i.e., cross-entropy), yielding the modified IB formulation:

$$\min_{M \sim \text{Bern}(\pi)} -\text{LC}(Y^{\text{expl}}, f(\tilde{X})) + \alpha I(X; X'), \quad (1)$$

where Y^{expl} is the objective explanation label, and $\tilde{X}$ is the explanation (X' for attribution and X'' for counterfactual explanation). While Eq. (1) bypasses the informativeness estimation, directly applying it to time series explainability introduces three critical interconnected problems. **First**, regarding the intractability of compactness, estimating the mutual information $I(X; X')$ remains computationally prohibitive and suffers from severe sample complexity bounds in high-dimensional temporal spaces. **Second**, a severe OOD problem arises during attribution evaluation. When Y^{expl} equals Y , computing $f(X')$ requires passing a discretely masked instance directly to the classifier. This drastically violates the underlying data manifold, causing OOD predictions and destroying the reliability of the LC term [18]. **Third**, counterfactual generation suffers from adversarial instability. When Y^{expl} is set to a target label Y'' to seek a perturbed instance X'' , unconstrained optimization of LC over the continuous temporal space inevitably exploits the classifier's vulnerabilities. This generates meaningless adversarial noise rather than semantically valid shifts [25]. To resolve these interconnected problems, we reconstruct the objective into a unified framework that enforces distribution consistency and structural stability.

A. Tractable Compactness Bound

To bypass the intractable mutual information $I(X; X')$, we replace it with a tractable variational upper bound. Since the physical goal of minimizing $I(X; X')$ is to enforce sparsity and determinism in the explanation bottleneck, we directly penalize the expected size $|M|$ and the entropy $H(M)$ of the mask. By defining a prior Bernoulli distribution $\mathbb{Q}(M)$ with a parameter r , we analytically bound the compactness term as a

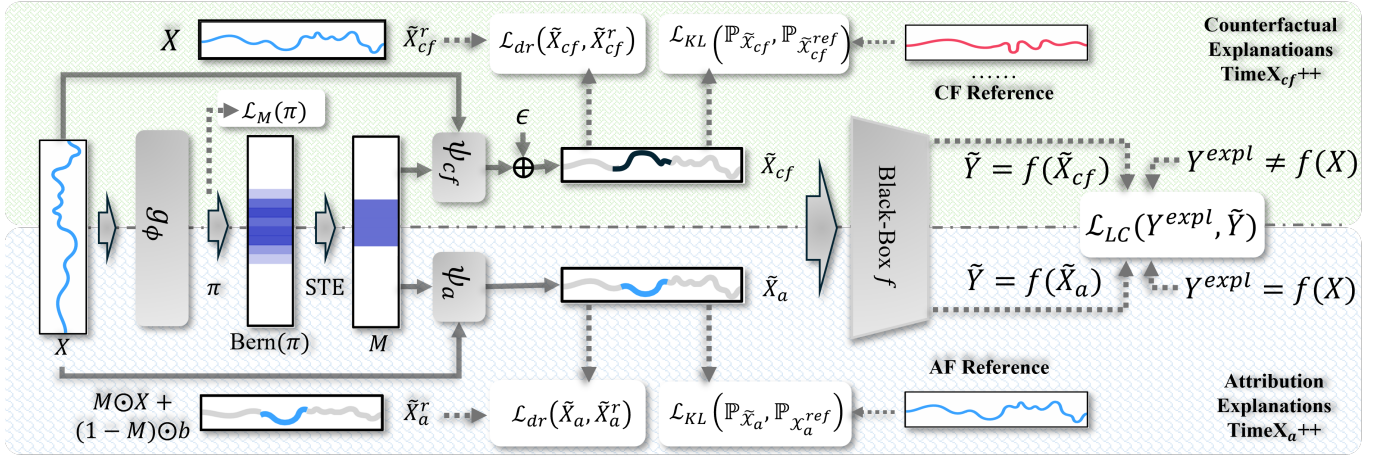


Fig. 2: The overall architecture of TimeX++, a unified framework for attribution and counterfactual explanations supported by Information Bottleneck. TimeX_a++ and TimeX_{cf}++ have similar components and loss functions but different implementations.

Kullback-Leibler (KL) divergence between the posterior mask distribution $\mathbb{P}(M|X)$ and the prior:

$$\mathcal{L}_{compact} = \mathbb{E}_X [D_{KL}(\mathbb{P}(M|X) \parallel \mathbb{Q}(M))] \quad (2)$$

Minimizing this term efficiently compresses the bottleneck without requiring complex mutual information estimators.

B. Distribution-Preserving and Bounded Generation

To simultaneously solve the attribution OOD problem and the counterfactual instability problem, we project the bottleneck into an *in-distribution* generated instance $\tilde{X} \in \mathcal{X}$, and evaluate $f(\tilde{X})$. The modified informativeness objective becomes:

$$\mathcal{L}_{info} = \mathcal{L}_{LC}(Y^{expl}, f(\tilde{X})) + \beta \mathcal{L}_{KL}(\mathbb{P}_{\tilde{X}} \parallel \mathbb{P}_{\mathcal{X}}) + \lambda \mathcal{L}_{bound}(\tilde{X}, X, M), \quad (3)$$

where $\mathcal{L}_{LC}(\cdot, \cdot)$ equals to $LC(\cdot, \cdot)$. This introduces two crucial regularization terms. The distribution preservation penalty, $\mathcal{L}_{KL}(\mathbb{P}_{\tilde{X}} \parallel \mathbb{P}_{\mathcal{X}})$, explicitly regularizes the distribution shift between the generated $\tilde{X}$ and the true data manifold $\mathcal{X}$. This mathematically eliminates the OOD problem, ensuring the classifier evaluates valid, realistic samples. Concurrently, the bottleneck constraint, $\mathcal{L}_{bound}(\tilde{X}, X, M)$, acts as the causal structural anchor. It forces the generated perturbation ($\tilde{X} - X$) to be strictly bounded by the spatial and temporal active regions of the attribution mask M . By penalizing changes in the redundant background, this constraint fundamentally prevents unconstrained adversarial search.

C. Unified Architecture Integration

Combining the tractable compactness bound and the regularized generation, the final unified objective of TimeX++ is:

$$\min_{M, \tilde{X}} \mathcal{L}_{info} + \alpha \mathcal{L}_{compact}.$$

This single cohesive architecture systematically integrates attribution and counterfactual explanations through the shared bottleneck M . For **attribution** extraction, we set $Y^{expl} = Y$. The optimization minimizes $\mathcal{L}_{compact}$ to find the tightest

possible mask M , while $\mathcal{L}_{KL}$ ensures the generated evaluation instance $\tilde{X}$ remains strictly in-distribution, resolving the OOD problem. Conversely, for **counterfactual** generation, we set $Y^{expl} = Y''$. Although the optimization seeks to flip the label, $\mathcal{L}_{bound}$ explicitly locks the counterfactual trajectory within the semantic support of M . This guarantees that the counterfactual only manipulates the causal features identified by the attribution, resolving the adversarial instability problem. Through this integration, TimeX++ provides causally verified attributions and semantically grounded counterfactuals.

V. THE TIMEX++ UNIFIED FRAMEWORK

Translating the tractable unified Information Bottleneck (IB) objective from Section IV into a practical neural architecture, we introduce the TimeX++ framework (depicted in Figure 2). The framework is constructed upon two primary neural modules working in tandem: the shared **Attribution Bottleneck Extractor** (g_ϕ) and the paradigm-specific **Conditioned Generators** (ψ_a and ψ_{cf}). By dynamically configuring the objective explanation label Y^{expl} and the specific generator, TimeX++ seamlessly instantiates into two operational modes: TimeX_a++ and TimeX_{cf}++.

A. Attribution Bottleneck Extractor

The extractor $g_\phi : \mathbb{R}^{T \times D} \mapsto [0, 1]^{T \times D}$ serves as the shared semantic foundation for both explanation paradigms. It is implemented via an encoder-decoder Transformer that maps the input time series X into a stochastic probability matrix π . To bridge the gap between continuous network optimization and the discrete mask M required by the IB formulation, we employ the Straight-Through Estimator (STE) [21]. During the forward pass, we sample a deterministic binary mask $M = \text{STE}(\text{Bern}(\pi)) \in \{0, 1\}^{T \times D}$. During the backward pass, gradients bypass the discrete sampling operator, allowing the parameters ϕ to be optimized smoothly.

To satisfy the tractable compactness bound $\mathcal{L}_{compact}$ (Eq. 2) without requiring complex mutual information estimators, we apply two structural regularizations to the predicted probability π . First, to enforce sparsity and determinism, we explicitly

minimize the KL divergence against a sparse prior $\mathbb{Q}(M)$ parameterized by r . For attribution explanations (TimeX_a++), we defaultly set $r = 0.5$. For counterfactual generation (TimeX_{cf}++), we enforce a strictly narrower bottleneck by default setting $r = 0.1$, isolating the most critical causal features to restrict the counterfactual search space. Second, to ensure temporal continuity, we penalize irregular mask fragmentation across the time axis. Because true semantic features in time series typically occur as contiguous temporal segments rather than isolated anomalous spikes, the continuity loss is formulated as:

$$\mathcal{L}_{\text{con}} = \frac{1}{T \times D} \sum_{d=1}^D \sum_{t=1}^{T-1} \sqrt{(\pi_{t,d} - \pi_{t+1,d})^2}. \quad (4)$$

The total bottleneck extraction loss is defined as $\mathcal{L}_M = \mathcal{L}_{\text{compact}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}$, formulated as:

$$\mathcal{L}_M = \sum_{t,d} \left[\pi_{t,d} \log \left(\frac{\pi_{t,d}}{r} \right) + (1 - \pi_{t,d}) \log \left(\frac{1 - \pi_{t,d}}{1 - r} \right) \right] + \lambda_{\text{con}} \mathcal{L}_{\text{con}}. \quad (5)$$

where the first summation [36] expands the KL divergence $\mathcal{L}_{\text{compact}}$, and λ_{con} controls the temporal continuity penalty.

B. Conditioned Generators

The conditioned generators translate the structural bottleneck M into the generated explanation-embedded instance $\tilde{X} \in \mathcal{X}$. They are explicitly designed to resolve the OOD problem for attribution explanations (using ψ_a) and the robustness problem for counterfactual explanations (using ψ_{cf}).

a) *TimeX_a++*: When evaluating feature attribution, directly passing the masked instance $X \odot M$ to the classifier destroys the temporal manifold, causing arbitrary OOD predictions. TimeX_a++ solves this by projecting the masked data back into the true distribution. We first define a baseline reference $\tilde{X}_a^r = X \odot M + b \odot (1 - M)$, where background noise b is sampled from the dataset distribution $\mathbb{B}_{\mathcal{X}}$. The attribution generator ψ_a then maps $[X, M]$ into a smoothed, generated instance $\tilde{X}_a = \psi_a(X, M)$. To guarantee it resolves the OOD problem, we enforce distribution preservation via $\mathcal{L}_{\text{KL}}(\mathbb{P}_{\tilde{X}_a} \parallel \mathbb{P}_{\mathcal{X}_a^r})$. To ensure the generator strictly focuses on smoothing the OOD artifacts rather than ungrounded features, we apply the bottleneck constraint $\mathcal{L}_{\text{bound}}^a$ as a Euclidean penalty against the reference:

$$\mathcal{L}_{\text{bound}}^a = \frac{1}{TD} \sum_{d=1}^D \sum_{t=1}^T \left\| \tilde{X}_a[t, d] - \tilde{X}_a^r[t, d] \right\|^2 \quad (6)$$

The attribution is then validated by matching the original prediction: $\mathcal{L}_{\text{LC}}(f(X), f(\tilde{X}_a))$.

b) *TimeX_{cf}++*: Unconstrained counterfactual generation in continuous temporal space inevitably exploits classifier null spaces, producing adversarial noise. TimeX_{cf}++ solves this robustness problem by strictly locking the perturbation within the semantic bottleneck M . Instead of generating an entirely

Algorithm 1: Training Pipeline of TimeX++

```

1: Input: Training set  $\mathcal{T}$ , Trained classifier  $f$ , mode
    $\in \{A, CF\}$ , Target  $Y''$  (if CF), Hyperparams
    $\{\alpha, \beta, r, \lambda_{\text{con}}\}$ 
2: Init: Extractor  $g_\phi$ , Generators  $\psi_a, \psi_{cf}$ , and Variation
   function  $\psi_n$ 
3: for  $e = 1$  to  $E$  do
4:   for  $i = 1$  to  $N$  do
5:      $\pi_i = g_\phi(X_i)$ ; Sample  $M_i = \text{STE}(\text{Bern}(\pi_i))$ 
6:     if mode == A then
7:       Set reference  $\tilde{X}_{a,i}^r = M_i \odot X_i + (1 - M_i) \odot b$ 
8:       Generate instance  $\tilde{X}_{a,i} = \psi_a(X_i, M_i)$ 
9:        $Y^{\text{expl}} = f(X_i)$ ;  $\tilde{X}_i = \tilde{X}_{a,i}$ ;  $\tilde{X}_i^r = \tilde{X}_{a,i}^r$ 
10:    else
11:      Set reference  $\tilde{X}_{cf,i}^r = X_i$ 
12:      Predict perturbation  $E_i = \psi_{cf}(X_i, M_i)$ 
13:      Sample  $X^{\text{ref}}$  with label  $Y''$ , compute
         $\epsilon = \psi_n(X_i, M_i, X^{\text{ref}})$ 
14:      Generate instance  $\tilde{X}_{cf,i} = X_i + M_i \odot E_i + \epsilon$ 
15:       $Y^{\text{expl}} = Y''$ ;  $\tilde{X}_i = \tilde{X}_{cf,i}$ ;  $\tilde{X}_i^r = \tilde{X}_{cf,i}^r$ 
16:    end if
17:  end for
18:  Compute compactness  $\mathcal{L}_M = \mathcal{L}_{\text{compact}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}$ 
19:  Compute bottleneck constraint  $\mathcal{L}_{\text{bound}}$  against  $\tilde{X}^r$ 
20:  Compute distribution preservation  $\mathcal{L}_{\text{KL}}$ 
21:  Compute Label Consistency  $\mathcal{L}_{\text{LC}}(Y^{\text{expl}}, f(\tilde{X}))$ 
22:  Total loss:  $\tilde{\mathcal{L}} = \mathcal{L}_{\text{LC}} + \alpha \mathcal{L}_M + \beta (\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{bound}})$ 
23:  Update  $\phi$  and active generators ( $\psi_a$ , or  $\psi_{cf}, \psi_n$ ) via
    gradient descent
24: end for

```

new instance, the counterfactual generator ψ_{cf} learns a target-driven perturbation matrix $E = \psi_{cf}(X, M)$. The counterfactual instance is structurally formulated as:

$$\tilde{X}_{cf} = X + M \odot E + \epsilon \quad (7)$$

where ϵ introduces target-class reference noise. Specifically, we define an auxiliary variation function $\epsilon = \psi_n(X, M, X^{\text{ref}})$, which incorporates fine-grained target-class dynamics from a randomly sampled training instance X^{ref} belonging to the target label Y'' . During inference, this simplifies deterministically to $\tilde{X}_{cf} = X + M \odot E$. Crucially, to mathematically alleviate adversarial vulnerabilities, the bottleneck constraint $\mathcal{L}_{\text{bound}}^{cf}$ heavily penalizes any deviation from the original reference ($\tilde{X}_{cf}^r = X$) in the non-bottleneck background regions $(1 - M)$:

$$\mathcal{L}_{\text{bound}}^{cf} = \frac{1}{TD} \sum_{d=1}^D \sum_{t=1}^T \left\| (1 - M[t, d]) \odot (\tilde{X}_{cf}[t, d] - \tilde{X}_{cf}^r[t, d]) \right\|^2 \quad (8)$$

This formulation acts as a hard causal anchor, which prevents the model from injecting imperceptible adversarial noise into redundant features, forcing the targeted prediction shift $\mathcal{L}_{\text{LC}}(\tilde{Y}, f(\tilde{X}_{cf}))$ to rely on robust semantic manipulations.

TABLE I: The description of synthetic and real-world datasets.

DATASET	# OF SAMPLES	LENGTH	DIMENSION	CLASSES
FREQSHAPES	6,100	50	1	4
SEQCOMB-UV	6,100	200	1	4
SEQCOMB-MV	6,100	200	4	4
LOWVAR	6,100	200	2	4
ECG	92,511	360	1	2
PAM	5,333	600	17	8
EPILEPSY	11,500	178	1	2
BOILER	160,719	36	20	2
WAFER	7,164	152	1	2
FREEZERREGULAR	3,000	301	1	2

C. End-to-End Optimization

The entire TimeX++ framework is optimized end-to-end. By sharing the underlying architecture and unifying the mathematical formulation, both TimeX_{a++} and TimeX_{cf++} minimize the same overarching objective derived in Section IV:

$$\tilde{\mathcal{L}} = \mathcal{L}_{\text{LC}}(Y^{\text{expl}}, f(\tilde{X})) + \alpha \mathcal{L}_M + \beta (\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{bound}}) \quad (9)$$

This allows our method to seamlessly alternate between extracting robust attributions and generating causally constrained counterfactuals simply by toggling the objective label Y^{expl} and the specific generator (ψ_a or ψ_{cf}). The unified train procedure is summarized in Algorithm 1.

VI. EXPERIMENTS

In this section, we empirically evaluate the proposed TimeX++ framework. By unifying attribution and counterfactual explanations through the lens of the Information Bottleneck, we aim to demonstrate three key aspects of their effectiveness. **First**, we assess TimeX++’s ability to generate faithful attribution explanations and effective counterfactual explanations. **Second**, we examine the extent to which the proposed distribution-matching constraint alleviates the out-of-distribution shift. **Third**, we evaluate the robustness of counterfactual explanations and investigate whether TimeX++ effectively guides counterfactual generation toward semantically meaningful outcomes, preventing it from degenerating into trivial adversarial perturbations.

A. Experimental Setup

To rigorously evaluate our framework, we establish a comprehensive experimental setup encompassing diverse datasets, established baselines, and strict evaluation metrics.

1) *Datasets*: We utilize ten datasets encompassing both synthetic and real-world time-series scenarios, adopting the standard configurations from established benchmarks [43]. The synthetic datasets, including FreqShapes, SeqComb-UV, SeqComb-MV, and LowVar, are meticulously crafted with inherent ground-truth explanations to encapsulate temporal dynamics, guarding against heuristic learning shortcuts [14]. The real-world datasets cover various domains, such as electrocardiogram classification (ECG) [38], action recognition (PAM) [46], electroencephalogram analysis (Epilepsy) [1], mechanical fault detection (Boiler)¹, and sensor classifications (Wafer, FreezerRegular) [10]. The Transformer [55] is employed as the underlying black-box classifier across all tasks.

2) *Baselines*: For comprehensive comparative analysis, we benchmark TimeX++ against eleven state-of-the-art explainers meticulously selected across both the attribution and counterfactual paradigms. For **attribution explanations**, we evaluate against six popular baselines: Integrated Gradients (IG) [51] as a fundamental gradient-based explainer; Dynamask [9] and WinIT [29] as recent occlusion-based methods specifically designed for time series; CoRTX [8] leveraging contrastive learning; SGT + GRAD [19] demonstrating an *in-hoc* approach; and TIMEX [43], which serves as the strongest information-theoretic baseline. For **counterfactual explanations**, we compare against five advanced methodologies: CoMTE [2], AB-CF [30], M-CELS [31], and CONFETTI [5]. These baselines encapsulate a diverse range of optimization-based, evolutionary, saliency-guided, and multi-objective paradigms, ensuring a robust evaluation of counterfactual generation capabilities.

3) *Evaluation Metrics*: To assess explanation quality across distinct paradigms, we use specific quantitative metrics.

Metrics for Attribution Explanations. Given that the precise salient features are known for the synthetic datasets, we utilize them as the ground truth for evaluation. At each time step, features causing prediction label changes are attributed an explanation of 1, whereas those that do not affect such changes are 0. Following the previous paper [9], we evaluate the quality of salient features using Area Under Precision (AUP) and Area Under Recall (AUR), framing it as a binary classification task. We also employ AUPRC for consistency with TimeX [43], which combines the results from both AUP and AUR. Higher values indicate better performance. Let $M \in [0, 1]^{T \times D}$ be the obtained explanation mask, and $Q \in \{0, 1\}^{T \times D}$ be the ground-truth matrix where $Q[t, d] = 1$ if feature $X[t, d]$ is salient. Given a detection threshold $\tau \in (0, 1)$, we convert M into a binary estimator $\hat{Q}[t, d](\tau)$:

$$\hat{Q}[t, d](\tau) = \begin{cases} 1 & \text{if } M[t, d] \geq \tau \\ 0 & \text{else.} \end{cases} \quad (10)$$

Considering the truly salient index set $A = \{(t, d) \mid Q[t, d] = 1\}$ and the selected index set $\hat{A}(\tau) = \{(t, d) \mid \hat{Q}[t, d](\tau) = 1\}$, the precision (P) and recall (R) curves are defined as:

$$P(\tau) = \frac{|A \cap \hat{A}(\tau)|}{|\hat{A}(\tau)|}, \quad R(\tau) = \frac{|A \cap \hat{A}(\tau)|}{|A|}. \quad (11)$$

AUP and AUR are derived via integration over all thresholds:

$$\text{AUP} = \int_0^1 P(\tau) d\tau, \quad \text{AUR} = \int_0^1 R(\tau) d\tau. \quad (12)$$

Metrics for Counterfactual Explanations. To comprehensively evaluate the quality of the generated counterfactual instances (X''), we assess them across five distinct dimensions: Validity, Confidence, Sparsity, Proximity-L1, and Proximity-L2. For Validity, Confidence, and Sparsity, higher values are better ($\uparrow$). For both Proximity metrics, lower values are better ($\downarrow$).

• **Validity ($\uparrow$)**: Measures the ratio of successful counterfactuals that successfully alter the black-box model’s prediction to the desired target label Y'' . For a dataset of N instances:

$$\text{Validity} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(f(X''_i) = Y'') \quad (13)$$

¹<https://dx.doi.org/10.21227/awav-bn36>

TABLE II: Attribution explanation performance on univariate and multivariate datasets.

METHOD	AUPRC	AUP	AUR	AUPRC	AUP	AUR	AUPRC	AUP	AUR
FREQSHAPES			SEQCOMB-UV			ECG			
IG	0.752±0.003	0.691±0.003	0.598±0.002	0.576±0.002	0.816±0.002	0.287±0.002	0.418±0.001	0.595±0.002	0.320±0.001
DYNAMASK	0.220±0.001	0.295±0.004	0.504±0.002	0.442±0.002	0.878±0.004	0.103±0.001	0.328±0.001	0.525±0.003	0.108±0.008
WINIT	0.507±0.002	0.555±0.003	0.456±0.002	0.457±0.002	0.787±0.003	0.225±0.002	0.305±0.001	0.443±0.003	0.347±0.001
CoRTX	0.698±0.016	0.494±0.000	0.326±0.001	0.564±0.002	0.824±0.003	0.175±0.001	0.374±0.001	0.497±0.002	0.303±0.001
SGT + GRAD	0.531±0.002	0.414±0.001	0.393±0.002	0.573±0.002	0.783±0.001	0.214±0.001	0.314±0.001	0.424±0.002	0.264±0.001
TIMEX	0.832±0.003	0.722±0.003	0.638±0.002	0.712±0.002	0.941±0.001	0.338±0.001	0.472±0.002	0.566±0.003	0.446±0.002
TIMEX _a ++	0.891±0.002	0.781±0.001	0.662±0.002	0.847±0.001	0.907±0.000	0.406±0.001	0.660±0.001	0.726±0.001	0.460±0.001
SEQCOMB-MV			LOWVAR						
IG	0.330±0.002	0.748±0.003	0.258±0.003	0.869±0.004	0.483±0.003	0.817±0.002			
DYNAMASK	0.314±0.002	0.548±0.005	0.195±0.003	0.139±0.001	0.164±0.003	0.211±0.002			
WINIT	0.281±0.002	0.759±0.002	0.208±0.002	0.167±0.002	0.114±0.002	0.384±0.002			
CoRTX	0.363±0.002	0.563±0.001	0.346±0.002	0.498±0.001	0.328±0.003	0.471±0.001			
SGT + GRAD	0.489±0.001	0.497±0.001	0.429±0.002	0.345±0.001	0.213±0.003	0.353±0.002			
TIMEX	0.688±0.002	0.833±0.001	0.387±0.002	0.867±0.003	0.545±0.003	0.900±0.002			
TIMEX _a ++	0.759±0.001	0.878±0.001	0.391±0.001	0.947±0.002	0.806±0.002	0.833±0.002			

TABLE III: Performance report on different real-world datasets by masking the top 10% of salient features. The masked portion is substituted with an average of this feature or with zeros.

METHOD	SUBSTITUTION	PAM	EPILEPSY	BOILER	WAFER	FREEZER	RANK
RANDOM	MEAN	0.979±0.002	0.939±0.005	0.814±0.056	0.993±0.002	0.772±0.093	7.2
	ZERO	0.979±0.002	0.940±0.004	0.895±0.012	0.992±0.002	0.775±0.093	7.8
DYNAMASK	MEAN	0.778±0.015	0.816±0.019	0.808±0.036	0.488±0.267	0.452±0.129	4.6
	ZERO	0.777±0.015	0.361±0.073	0.555±0.146	0.488±0.267	0.355±0.110	3.6
TIMEX	MEAN	0.749±0.049	0.852±0.018	0.563±0.041	0.463±0.145	0.469±0.061	4.4
	ZERO	0.754±0.049	0.454±0.077	0.416±0.048	0.465±0.057	0.465±0.057	3.6
TimeX_{a++}	MEAN	0.717±0.019	0.845±0.029	0.385±0.123	0.410±0.150	0.379±0.074	2.8
	ZERO	0.775±0.023	0.355±0.146	0.361±0.073	0.400±0.061	0.377±0.074	2.0

- **Confidence** (↑): Evaluates the average predicted probability assigned by the classifier to the target class, indicating the robustness of the class flip:

$$\text{Confidence} = \frac{1}{N} \sum_{i=1}^N P_f(Y'' | X_i'') \quad (14)$$

- **Sparsity** (↓): Quantifies the minimality of the intervention by measuring the proportion of features left unperturbed. It is derived from the L_1 norm of the continuous mask M :

$$\text{Sparsity} = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{T \times D} \sum_{t=1}^T \sum_{d=1}^D M_i[t, d] \right) \quad (15)$$

- **Proximity-L1** (↓): Measures the feature-wise Mean Absolute Error (MAE) between the counterfactual explanation and the original input to ensure realistic generation:

$$\text{Proximity}_{L1} = \frac{1}{N} \sum_{i=1}^N \frac{1}{T \times D} \|X_i'' - X_i\|_1 \quad (16)$$

- **Proximity-L2** (↓): Measures the feature-wise Mean Squared Error (MSE) to heavily penalize large, unrealistic deviations from the original input X :

$$\text{Proximity}_{L2} = \frac{1}{N} \sum_{i=1}^N \frac{1}{T \times D} \|X_i'' - X_i\|_2^2 \quad (17)$$

B. Efficacy of Unified Explanations

We first examine the fidelity and efficacy of TimeX++ in producing both attribution and counterfactual explanations via the single unified training process.

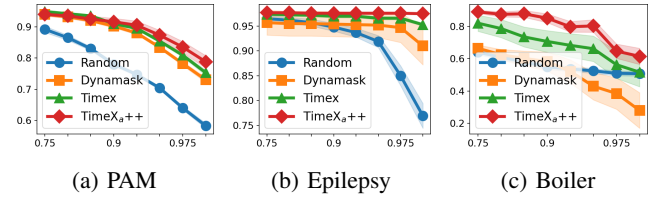


Fig. 3: Occlusion experiments on real-world datasets. Higher values indicate better performance. The x-axis is the Bottom Proportion Perturbed, and the y-axis is the Prediction AUROC.

For **attribution** explanations, Table II summarizes the performance across univariate and multivariate synthetic datasets. TimeX_{a++} outperforms competing explainers in 9 out of 12 evaluation cases. Compared to the strongest baseline, TIMEX, our framework achieves average improvements of 11.01% in AUPRC, 10.87% in AUP, and 1.25% in AUR. When analyzing the global metric AUPRC specifically, TimeX_{a++} significantly improves upon the identification of ground-truth explanations by 6.97% on FreqShapes, 18.86% on SeqComb-UV, 10.33% on SeqComb-MV, and 8.91% on LowVar over the strongest baselines. The statistical significance of these improvements is corroborated by the Friedman Test across all methods and cases, yielding a statistic of $F_F = 51.32$ with $p < 0.001$. It supports our theoretical claim that optimizing via the Information Bottleneck principle reliably extracts precise temporal explanation instances without compromising predictive accuracy.

Furthermore, to evaluate attribution fidelity on real-world datasets where ground-truth explanations are unavailable, we conduct occlusion experiments. Specifically, we occlude

TABLE IV: Targeted counterfactual explanation performance on univariate and multivariate datasets.

METHOD	VALIDITY	CONFIDENCE	SPARSITY	PROX.-L1	PROX.-L2	VALIDITY	CONFIDENCE	SPARSITY	PROX.-L1	PROX.-L2
FREQSHAPES						SEQCOMB-UV				
CoMTE	1.000 ± 0.000	0.965 ± 0.006	1.000 ± 0.000	0.573 ± 0.004	0.828 ± 0.006	1.000 ± 0.000	0.784 ± 0.022	0.999 ± 0.000	0.209 ± 0.002	0.401 ± 0.008
AB-CF	1.000 ± 0.000	0.871 ± 0.008	0.733 ± 0.025	0.441 ± 0.013	0.718 ± 0.011	1.000 ± 0.000	0.911 ± 0.013	0.650 ± 0.037	0.151 ± 0.007	0.348 ± 0.014
M-CELS	0.486 ± 0.053	0.466 ± 0.051	0.061 ± 0.008	0.064 ± 0.011	0.196 ± 0.031	0.361 ± 0.037	0.353 ± 0.036	0.025 ± 0.003	0.016 ± 0.001	0.090 ± 0.007
CONFETTI	0.990 ± 0.003	0.512 ± 0.002	0.349 ± 0.010	0.242 ± 0.011	0.538 ± 0.012	0.894 ± 0.034	0.477 ± 0.017	0.209 ± 0.019	0.059 ± 0.008	0.222 ± 0.021
TimeX _{cf} ++	0.974 ± 0.004	0.964 ± 0.005	0.214 ± 0.005	0.339 ± 0.021	0.757 ± 0.030	0.890 ± 0.053	0.887 ± 0.052	0.516 ± 0.074	0.288 ± 0.085	0.487 ± 0.082
SEQCOMB-MV						LOWVAR				
CoMTE	0.990 ± 0.004	0.820 ± 0.020	0.443 ± 0.018	0.081 ± 0.003	0.247 ± 0.006	1.000 ± 0.000	0.985 ± 0.003	0.733 ± 0.007	0.780 ± 0.007	1.121 ± 0.005
AB-CF	1.000 ± 0.000	0.954 ± 0.008	0.641 ± 0.029	0.099 ± 0.004	0.246 ± 0.007	1.000 ± 0.000	0.916 ± 0.009	0.627 ± 0.011	0.662 ± 0.012	1.003 ± 0.011
M-CELS	0.350 ± 0.043	0.348 ± 0.042	0.011 ± 0.001	0.008 ± 0.001	0.072 ± 0.007	0.365 ± 0.017	0.357 ± 0.015	0.029 ± 0.001	0.039 ± 0.001	0.232 ± 0.005
CONFETTI	0.910 ± 0.031	0.475 ± 0.013	0.246 ± 0.017	0.039 ± 0.003	0.153 ± 0.011	0.998 ± 0.001	0.514 ± 0.000	0.190 ± 0.004	0.193 ± 0.005	0.537 ± 0.007
TimeX _{cf} ++	0.983 ± 0.007	0.979 ± 0.008	0.305 ± 0.041	0.152 ± 0.016	0.353 ± 0.033	0.985 ± 0.003	0.980 ± 0.003	0.175 ± 0.024	0.212 ± 0.010	0.548 ± 0.013
ECG						EPILEPSY				
CoMTE	1.000 ± 0.000	0.936 ± 0.008	1.000 ± 0.000	0.275 ± 0.006	0.380 ± 0.010	1.000 ± 0.000	0.852 ± 0.027	0.996 ± 0.000	0.633 ± 0.130	0.832 ± 0.162
AB-CF	1.000 ± 0.000	0.841 ± 0.008	0.701 ± 0.059	0.189 ± 0.017	0.303 ± 0.018	1.000 ± 0.000	0.739 ± 0.036	0.788 ± 0.039	0.539 ± 0.132	0.762 ± 0.161
M-CELS	0.231 ± 0.075	0.242 ± 0.068	0.022 ± 0.005	0.013 ± 0.003	0.061 ± 0.011	0.097 ± 0.049	0.103 ± 0.046	0.007 ± 0.005	0.009 ± 0.006	0.027 ± 0.017
CONFETTI	0.977 ± 0.005	0.505 ± 0.003	0.287 ± 0.008	0.080 ± 0.003	0.200 ± 0.007	0.798 ± 0.050	0.423 ± 0.026	0.334 ± 0.026	0.212 ± 0.018	0.417 ± 0.030
TimeX _{cf} ++	0.848 ± 0.104	0.847 ± 0.101	0.084 ± 0.026	0.251 ± 0.097	0.767 ± 0.185	0.946 ± 0.047	0.924 ± 0.045	0.139 ± 0.024	0.255 ± 0.028	0.698 ± 0.041

TABLE V: Untargeted counterfactual explanation performance on univariate and multivariate datasets.

METHOD	VALIDITY	CONFIDENCE	SPARSITY	PROX.-L1	PROX.-L2	VALIDITY	CONFIDENCE	SPARSITY	PROX.-L1	PROX.-L2
FREQSHAPES						SEQCOMB-UV				
CoMTE	1.000 ± 0.000	0.930 ± 0.017	1.000 ± 0.000	0.438 ± 0.016	0.641 ± 0.028	1.000 ± 0.000	0.790 ± 0.023	0.999 ± 0.000	0.196 ± 0.000	0.376 ± 0.001
AB-CF	1.000 ± 0.000	0.826 ± 0.021	0.672 ± 0.006	0.283 ± 0.010	0.486 ± 0.021	1.000 ± 0.000	0.890 ± 0.018	0.513 ± 0.040	0.114 ± 0.004	0.293 ± 0.004
M-CELS	0.612 ± 0.039	0.570 ± 0.034	0.063 ± 0.006	0.040 ± 0.001	0.157 ± 0.002	0.336 ± 0.026	0.319 ± 0.022	0.013 ± 0.001	0.011 ± 0.001	0.083 ± 0.004
CONFETTI	0.621 ± 0.017	0.514 ± 0.019	0.257 ± 0.008	0.117 ± 0.007	0.260 ± 0.017	0.962 ± 0.010	0.519 ± 0.008	0.200 ± 0.005	0.052 ± 0.002	0.208 ± 0.005
TimeX _{cf} ++	1.000 ± 0.000	0.999 ± 0.000	0.233 ± 0.009	0.420 ± 0.016	0.872 ± 0.020	1.000 ± 0.000	0.999 ± 0.001	0.624 ± 0.082	0.393 ± 0.069	0.637 ± 0.076
SEQCOMB-MV						LOWVAR				
CoMTE	1.000 ± 0.000	0.796 ± 0.022	0.367 ± 0.013	0.063 ± 0.001	0.208 ± 0.002	1.000 ± 0.000	0.978 ± 0.003	0.520 ± 0.002	0.552 ± 0.002	0.951 ± 0.002
AB-CF	1.000 ± 0.000	0.941 ± 0.013	0.514 ± 0.018	0.075 ± 0.002	0.202 ± 0.003	1.000 ± 0.000	0.866 ± 0.014	0.556 ± 0.020	0.579 ± 0.020	0.916 ± 0.019
M-CELS	0.158 ± 0.026	0.159 ± 0.026	0.005 ± 0.000	0.004 ± 0.000	0.051 ± 0.005	0.312 ± 0.045	0.296 ± 0.037	0.021 ± 0.001	0.030 ± 0.001	0.206 ± 0.006
CONFETTI	0.917 ± 0.037	0.483 ± 0.014	0.270 ± 0.010	0.044 ± 0.001	0.169 ± 0.004	0.998 ± 0.001	0.513 ± 0.001	0.176 ± 0.004	0.175 ± 0.004	0.506 ± 0.006
TimeX _{cf} ++	1.000 ± 0.000	1.000 ± 0.000	0.378 ± 0.092	0.118 ± 0.010	0.291 ± 0.008	1.000 ± 0.000	0.999 ± 0.000	0.221 ± 0.049	0.231 ± 0.014	0.552 ± 0.023

the bottom k -percentile of salient features to measure the change in prediction AUROC. Besides the standard baselines, we also include a random explainer reference to control for potential misinterpretations. The explanation results by occluding salient features are presented in Figure 3. Our results show that TimeX_a++ outperforms others across both univariate (Epilepsy) and multivariate (PAM and Boiler) time series. Particularly, TimeX_a++ maintains non-decreasing performance on Epilepsy due to the retention of only salient features, and outperforms baselines at any threshold k on PAM and Boiler datasets. Moreover, our method maintains excellent stability compared to the strongest baseline TIMEX, where the error bars of our method are noticeably narrower. This is because TimeX_a++ avoids the label leakage caused by re-optimizing a white-box predictive model. We also delete the top 10% of salient features and substitute them with an average of this feature or with zero perturbations [33], and further expand our experiments to five real-world time series datasets. The results (Table III) consistently show that the proposed method outperforms existing explainers under different perturbations, with substitution with zeros being the most applicable and yielding the highest average rank. The statistical significance of these improvements is firmly corroborated by the Friedman Test yielding $F_F = 25.80$ and $p < 0.001$.

Transitioning to **counterfactual** reasoning, we compare TimeX_{cf}++ with state-of-the-art baselines across six benchmark datasets under a targeted setting, rigorously evaluating all

possible target labels for each prediction, as presented in Table IV. Given the multi-objective nature of counterfactual explanations, directly comparing individual evaluation metrics is insufficient. Ideally, counterfactual explanations should exhibit low sparsity, low proximity, high validity, and high confidence. However, these objectives are often conflicting, resulting in an inherent trade-off among sparsity, validity, and proximity. Therefore, we adopt Pareto frontier analysis to demonstrate their overall effectiveness. Among the baselines, CONFETTI achieves high validity scores; however, its relatively low confidence indicates that the generated counterfactual explanations tend to lie close to the decision boundary. M-CELS generally produces counterfactual explanations with low sparsity but underperforms in terms of validity and confidence. AB-CF and COMTE achieve high validity and confidence, but they exhibit substantially higher sparsity, requiring extensive modifications to the input, which may generate target-labeled references. In contrast, TimeX_{cf}++ achieves a more favorable balance between validity/confidence and sparsity/proximity. Across these datasets, our method attains competitive validity and confidence compared with AB-CF and COMTE while maintaining substantially lower sparsity and proximity. For instance, on the Epilepsy dataset, TimeX_{cf}++ achieves the highest confidence score while requiring only 0.14 average sparsity, demonstrating the effectiveness of the proposed framework.

To further validate the effectiveness of TimeX_{cf}++, we

TABLE VI: Difference between the distribution of different explanation instances and the distribution of original data.

METHOD	KDE $\uparrow$	KL-DIVERGENCE $\downarrow$	MMD $\downarrow$	KDE $\uparrow$	KL-DIVERGENCE $\downarrow$	MMD $\downarrow$
FREQSHAPES			SEQCOMB-UV			
ZERO	-36.671 $\pm$ 0.275	0.144 $\pm$ 0.007	0.077 $\pm$ 0.004	-90.093 $\pm$ 0.312	0.399 $\pm$ 0.005	0.167 $\pm$ 0.003
MEAN	-36.539 $\pm$ 0.194	0.147 $\pm$ 0.003	0.078 $\pm$ 0.006	-83.038 $\pm$ 0.271	0.325 $\pm$ 0.009	0.111 $\pm$ 0.005
$b \sim \mathbb{B}_{\mathcal{X}}$	-53.855 $\pm$ 0.509	0.289 $\pm$ 0.003	0.024 $\pm$ 0.000	-100.050 $\pm$ 0.658	0.498 $\pm$ 0.005	0.009 $\pm$ 0.000
TimeX _a ++	-28.757$\pm$2.658	0.096$\pm$0.024	0.016$\pm$0.004	-57.932$\pm$1.684	0.060$\pm$0.018	0.078$\pm$0.019
SEQCOMB-MV			LOWVAR			
ZERO	-257.040 $\pm$ 1.656	0.739 $\pm$ 0.029	0.248 $\pm$ 0.006	-431.993 $\pm$ 1.131	0.597 $\pm$ 0.012	0.105 $\pm$ 0.003
MEAN	-230.250 $\pm$ 1.313	0.455 $\pm$ 0.019	0.074 $\pm$ 0.009	-429.331 $\pm$ 1.249	0.571 $\pm$ 0.012	0.095 $\pm$ 0.004
$b \sim \mathbb{B}_{\mathcal{X}}$	-260.182 $\pm$ 1.571	0.721 $\pm$ 0.019	0.026 $\pm$ 0.000	-474.482 $\pm$ 1.319	0.976 $\pm$ 0.017	0.104 $\pm$ 0.000
TimeX _a ++	-191.265$\pm$0.890	0.038$\pm$0.010	0.014$\pm$0.012	-426.431$\pm$2.765	0.538$\pm$0.025	0.073$\pm$0.017

TABLE VII: Explainer results with LSTM and CNN predictors on FreqShapes and SeqComb-MV synthetic datasets.

METHOD	FREQSHAPES			SEQCOMB-MV			ECG		
	AUPRC	AUP	AUR	AUPRC	AUP	AUR	AUPRC	AUP	AUR
LSTM									
IG	0.928 $\pm$ 0.002	0.778 $\pm$ 0.001	0.693 $\pm$ 0.002	0.237 $\pm$ 0.002	0.515 $\pm$ 0.005	0.321 $\pm$ 0.003	0.504 $\pm$ 0.002	0.613 $\pm$ 0.003	0.403 $\pm$ 0.002
DYNAMASK	0.229 $\pm$ 0.001	0.342 $\pm$ 0.004	0.517 $\pm$ 0.001	0.284 $\pm$ 0.002	0.637 $\pm$ 0.005	0.182 $\pm$ 0.002	0.373 $\pm$ 0.001	0.630 $\pm$ 0.003	0.110 $\pm$ 0.001
WINIT	0.417 $\pm$ 0.002	0.511 $\pm$ 0.003	0.391 $\pm$ 0.002	0.352 $\pm$ 0.001	0.655 $\pm$ 0.003	0.342 $\pm$ 0.002	0.363 $\pm$ 0.001	0.381 $\pm$ 0.002	0.406 $\pm$ 0.001
TIME X	0.990 $\pm$ 0.000	0.789 $\pm$ 0.001	0.796 $\pm$ 0.001	0.130 $\pm$ 0.002	0.131 $\pm$ 0.002	0.475 $\pm$ 0.002	0.606 $\pm$ 0.002	0.642 $\pm$ 0.002	0.444 $\pm$ 0.002
TimeX _a ++	0.994$\pm$0.000	0.741$\pm$0.001	0.843$\pm$0.001	0.405$\pm$0.004	0.680$\pm$0.005	0.352$\pm$0.002	0.651$\pm$0.001	0.743$\pm$0.001	0.445$\pm$0.001
CNN									
IG	0.991 $\pm$ 0.001	0.878 $\pm$ 0.001	0.706 $\pm$ 0.002	0.598 $\pm$ 0.003	0.886 $\pm$ 0.001	0.229 $\pm$ 0.001	0.495 $\pm$ 0.001	0.537 $\pm$ 0.001	0.531 $\pm$ 0.001
DYNAMASK	0.257 $\pm$ 0.001	0.443 $\pm$ 0.003	0.526 $\pm$ 0.002	0.455 $\pm$ 0.002	0.731 $\pm$ 0.003	0.314 $\pm$ 0.002	0.460 $\pm$ 0.001	0.722 $\pm$ 0.003	0.131 $\pm$ 0.001
WINIT	0.532 $\pm$ 0.002	0.602 $\pm$ 0.003	0.397 $\pm$ 0.002	0.533 $\pm$ 0.001	0.832 $\pm$ 0.002	0.226 $\pm$ 0.002	0.396 $\pm$ 0.001	0.329 $\pm$ 0.002	0.352 $\pm$ 0.001
TIME X	0.749 $\pm$ 0.005	0.497 $\pm$ 0.003	0.792 $\pm$ 0.002	0.702 $\pm$ 0.002	0.767 $\pm$ 0.001	0.469 $\pm$ 0.002	0.640 $\pm$ 0.001	0.746 $\pm$ 0.001	0.416 $\pm$ 0.001
TimeX _a ++	0.913$\pm$0.001	0.607$\pm$0.001	0.795$\pm$0.001	0.782$\pm$0.001	0.890$\pm$0.001	0.343$\pm$0.001	0.673$\pm$0.001	0.757$\pm$0.001	0.432$\pm$0.001

conduct experiments under the untargeted setting using four synthetic datasets. For implementation, multiple target labels are considered, and the counterfactual explanation with the highest confidence is selected as the final result. As shown in Table V, TimeX_{cf}++ demonstrates performance trends consistent with those observed in the targeted setting. Baseline methods that rely heavily on subsequence search, such as CONFETTI, exhibit limited robustness to the selection of reference data. In contrast, TimeX_{cf}++ consistently achieves the highest validity and confidence across all four datasets, highlighting its superior robustness compared with the baselines.

C. Resilience to Out-of-Distribution Shift

A critical vulnerability of perturbation-based attribution methods is the OOD shift, where explanation instances fall outside the original data manifold, yielding unreliable model gradients. We systematically quantify this phenomenon using Kernel Density Estimation (KDE) [41], KL divergence [27], and Maximum Mean Discrepancy (MMD) [16] to assess the distributional alignment between the explanation-embedded instances and the original temporal sequences. Specifically, KDE provides a non-parametric estimation of the probability density function of the original data manifold; by evaluating the log-likelihood of perturbed instances under this density, KDE effectively identifies OOD samples, as instances falling into low-density regions exhibit significantly lower likelihood scores [41, 48]. Furthermore, MMD serves as a robust kernel-based two-sample test that measures the distance between the

mean embeddings of two distributions within a Reproducing Kernel Hilbert Space (RKHS). A smaller MMD value formally indicates that the distribution of the generated explanation instances is statistically indistinguishable from the original empirical distribution, directly quantifying the mitigation of the OOD shift [16, 44]. Table VI reveals that explanations generated by TimeX_a++ exhibit significantly lower MMD and KL divergence compared to conventional masking strategies.

D. Robustness Against Noise

Beyond OOD issues in attribution, counterfactual explanations frequently suffer from adversarial degeneration, where imperceptible, high-frequency noise flips the model's prediction without providing meaningful semantic transitions. We demonstrate that TimeX_{cf}++ eradicates this vulnerability by strictly confining perturbations. The proximity metrics analyzed alongside sparsity indicate that our framework produces highly localized structural changes rather than distributed trivial noise. This structural integrity is governed by the causal structural anchor $\mathcal{L}_{bound}$, acting as a stringent penalty against any perturbation occurring in the background. By forcing the generator ψ_{cf} to manipulate only the critical temporal windows identified by the bottleneck extractor, the resulting counterfactuals are guaranteed to represent semantic shifts.

Furthermore, because TimeX_{cf}++ strictly confines perturbations to critical structural regions rather than relying on distributed adversarial shortcuts, the resulting counterfactual explanations also exhibit superior robustness to external noise.

TABLE VIII: Inference runtime of occlusion experiments for IG, Dynamask, TIMEX, and TimeX_a++ on all real-world datasets.

METHOD	PAM	EPILEPSY	BOILER
IG	5.231±0.247	15.467±0.363	123.556±0.807
DYNAMASK	105.166±0.610	371.092±4.257	3780.433±468.144
TIMEX	0.164±0.219	0.204±0.207	0.918±0.203
TimeX _a ++	0.161±0.218	0.201±0.208	0.910±0.177

METHOD	WAFER	FREEZERREGULAR
IG	15.395±0.305	6.745±0.502
DYNAMASK	403.400±31.664	189.285±1.731
TIMEX	0.163±0.022	0.269±0.411
TimeX _a ++	0.126±0.002	0.270±0.417

TABLE IX: Inference runtime of **counterfactual** explanation experiments for one split.

METHOD	FREQSHAPES	LOWVAR	EPILEPSY
CoMTE	42.245 ± 1.146	64.708 ± 8.603	22.992 ± 0.858
AB-CF	39.042 ± 1.139	37.944 ± 2.479	32.750 ± 2.522
M-CELS	4271.181 ± 55.677	5411.136 ± 63.176	1887.034 ± 69.215
CONFETTI	6628.330 ± 103.100	8397.309 ± 172.012	4519.880 ± 453.902
TimeX _{cf} ++	0.075 ± 0.035	0.287 ± 0.119	0.140 ± 0.012

To evaluate this property, we provide an analysis in Figure 4. We randomly select timesteps and inject standard Gaussian noise into the generated counterfactual explanations, then assess their effectiveness on the FreqShape and LowVar datasets. The validity and confidence of all counterfactual explanations decrease as the noise injection ratio increases. However, TimeX_{cf}++ demonstrates greater robustness than the baseline methods. Even with a 50% noise ratio, our counterfactual explanations maintain a validity above 0.8 on both datasets. Although CoMTE achieves comparable robustness, it exhibits much lower sparsity than our method. CONFETTI experiences a sharp drop in validity as noise increases due to its lower confidence. In summary, TimeX_{cf}++ prevents adversarial degeneration by design, demonstrating superior robustness while maintaining better sparsity than the baseline methods.

E. Runtime Results

We conduct a runtime analysis evaluating both the training and inference phases across the two explanation paradigms. Regarding the training phase, the runtime for each experiment approximated 3 to 15 minutes per fold, depending on the dataset volume. For attribution explanations, compared to TIMEX, which requires training both a white-box model for consistency and the explanation masks, TimeX_a++ directly perturbs the black-box, thereby requiring less overall training time. For counterfactual explanations, while TimeX_{cf}++ requires an upfront training stage unlike instance-specific optimization baselines, training time remains highly manageable. It is a constant multiple of the attribution training time, with the factor determined by the size of the classification label space.

During the inference phase, as shown in Table VIII and Table IX, TimeX++ emerges as the most expedient model across both tasks. For **attribution**, TimeX_a++ is significantly faster than Dynamask and IG, which necessitate recursive operations

TABLE X: Ablation of TimeX_a++ considering whether there are different losses in our component.

	AUPRC	AUP	AUR
FREQSHAPES			
TimeX _a ++	0.891±0.002	0.781±0.001	0.662±0.002
w/o $\mathcal{L}_{LC}$	0.225±0.001	0.196±0.001	0.768±0.002
w/o STE	0.779±0.003	0.732±0.003	0.656±0.002
w/o $\mathcal{L}_{KL}$	0.830±0.003	0.646±0.003	0.703±0.002
w/o $\mathcal{L}_{dr}$	0.194±0.002	0.139±0.002	0.415±0.002
SEQCOMB-MV			
TimeX _a ++	0.759±0.001	0.878±0.001	0.391±0.001
w/o $\mathcal{L}_{LC}$	0.095±0.003	0.057±0.002	0.330±0.008
w/o STE	0.727±0.003	0.873±0.001	0.348±0.002
w/o $\mathcal{L}_{KL}$	0.748±0.002	0.869±0.001	0.353±0.001
w/o $\mathcal{L}_{dr}$	0.069±0.002	0.053±0.001	0.628±0.008
ECG			
TimeX _a ++	0.660±0.001	0.726±0.001	0.460±0.001
w/o STE	0.615±0.001	0.747±0.001	0.402±0.001
w/o $\mathcal{L}_{LC}$	0.621±0.002	0.642±0.002	0.429±0.002
w/o $\mathcal{L}_{KL}$	0.642±0.002	0.698±0.001	0.442±0.001
w/o $\mathcal{L}_{dr}$	0.152±0.000	0.141±0.000	0.631±0.001

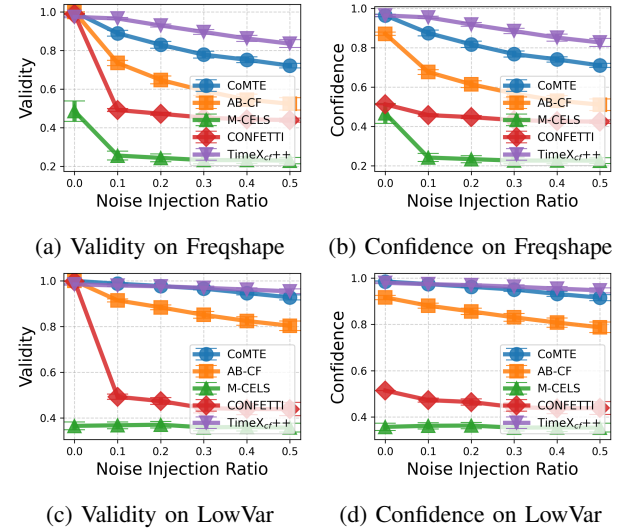


Fig. 4: The counterfactual explanations robustness analysis on Freqshape and LowVar.

for individual samples, and slightly faster than TIMEX, which requires Landmark calculations. For **counterfactual** generation, TimeX_{cf}++ demonstrates exceptional efficiency, being over 100 times faster than the fastest baseline (AB-CF). Traditional counterfactual baselines typically rely on costly iterative search procedures to identify suitable perturbations for each new sample. In contrast, because TimeX_{cf}++ is an inductive explainer, it requires only a single forward inference pass and fully benefits from GPU acceleration. Consequently, by requiring only a fixed amount of computation per instance during inference, our framework offers significant practical advantages for large-scale and real-time deployment scenarios.

F. Flexibility Across Different Black-box Classifiers

To explore the flexibility of TimeX++, we study two other time series classifiers and explore their explanatory role. We replace the original transformer-based black-box f with a long-short term memory (LSTM) and a convolutional neural network (CNN) as the underlying classifiers.

TABLE XI: Targeted counterfactual explanation performance on LSTM and CNN architectures.

Method	LSTM					CNN				
	Validity	Confidence	Sparsity	Prox.-L1	Prox.-L2	Validity	Confidence	Sparsity	Prox.-L1	Prox.-L2
FREQSHAPES										
CoMTE	1.000 ± 0.000	0.935 ± 0.014	1.000 ± 0.000	0.562 ± 0.003	0.821 ± 0.006	1.000 ± 0.000	0.922 ± 0.010	1.000 ± 0.000	0.578 ± 0.002	0.847 ± 0.003
AB-CF	1.000 ± 0.000	0.823 ± 0.017	0.664 ± 0.021	0.424 ± 0.007	0.723 ± 0.005	1.000 ± 0.000	0.790 ± 0.010	0.715 ± 0.015	0.451 ± 0.006	0.760 ± 0.005
M-CELS	0.294 ± 0.029	0.267 ± 0.026	0.057 ± 0.006	0.070 ± 0.008	0.192 ± 0.018	0.405 ± 0.025	0.363 ± 0.022	0.055 ± 0.003	0.072 ± 0.005	0.207 ± 0.013
CONFETTI	0.981 ± 0.008	0.506 ± 0.004	0.361 ± 0.005	0.260 ± 0.004	0.568 ± 0.007	0.984 ± 0.010	0.507 ± 0.005	0.360 ± 0.005	0.275 ± 0.004	0.612 ± 0.006
TimeX _{cf} ++	0.944 ± 0.036	0.926 ± 0.037	0.271 ± 0.011	0.421 ± 0.015	0.821 ± 0.022	0.976 ± 0.007	0.967 ± 0.007	0.283 ± 0.014	0.564 ± 0.031	1.120 ± 0.089
SEQCOMB-MV										
CoMTE	0.992 ± 0.003	0.932 ± 0.018	0.425 ± 0.020	0.078 ± 0.003	0.239 ± 0.007	0.999 ± 0.001	0.969 ± 0.008	0.454 ± 0.017	0.085 ± 0.003	0.255 ± 0.007
AB-CF	1.000 ± 0.000	0.862 ± 0.025	0.603 ± 0.033	0.094 ± 0.004	0.242 ± 0.007	1.000 ± 0.000	0.900 ± 0.015	0.578 ± 0.037	0.094 ± 0.005	0.249 ± 0.008
M-CELS	0.284 ± 0.039	0.277 ± 0.036	0.011 ± 0.001	0.007 ± 0.001	0.064 ± 0.006	0.549 ± 0.059	0.516 ± 0.056	0.012 ± 0.001	0.011 ± 0.001	0.089 ± 0.008
CONFETTI	0.911 ± 0.042	0.504 ± 0.021	0.236 ± 0.017	0.036 ± 0.003	0.140 ± 0.011	0.998 ± 0.001	0.512 ± 0.000	0.304 ± 0.011	0.051 ± 0.003	0.189 ± 0.010
TimeX _{cf} ++	0.941 ± 0.037	0.942 ± 0.034	0.129 ± 0.014	0.121 ± 0.013	0.364 ± 0.023	0.940 ± 0.035	0.941 ± 0.032	0.127 ± 0.013	0.124 ± 0.015	0.374 ± 0.028
ECG										
CoMTE	1.000 ± 0.000	0.852 ± 0.030	1.000 ± 0.000	0.279 ± 0.010	0.384 ± 0.016	1.000 ± 0.000	0.873 ± 0.016	1.000 ± 0.000	0.292 ± 0.002	0.403 ± 0.003
AB-CF	1.000 ± 0.000	0.779 ± 0.036	0.570 ± 0.057	0.174 ± 0.016	0.303 ± 0.019	0.999 ± 0.000	0.750 ± 0.015	0.464 ± 0.047	0.165 ± 0.006	0.317 ± 0.010
M-CELS	0.368 ± 0.057	0.358 ± 0.056	0.015 ± 0.001	0.007 ± 0.001	0.042 ± 0.004	0.486 ± 0.063	0.462 ± 0.062	0.023 ± 0.002	0.015 ± 0.002	0.077 ± 0.007
CONFETTI	0.858 ± 0.030	0.485 ± 0.020	0.250 ± 0.010	0.072 ± 0.005	0.180 ± 0.013	0.977 ± 0.007	0.503 ± 0.004	0.255 ± 0.010	0.076 ± 0.005	0.208 ± 0.012
TimeX _{cf} ++	0.772 ± 0.062	0.738 ± 0.057	0.138 ± 0.020	0.251 ± 0.037	0.709 ± 0.096	0.853 ± 0.047	0.849 ± 0.047	0.049 ± 0.016	0.104 ± 0.008	0.542 ± 0.033

TABLE XII: Ablation study Counterfactual explanation performance on LSTM and CNN architectures.

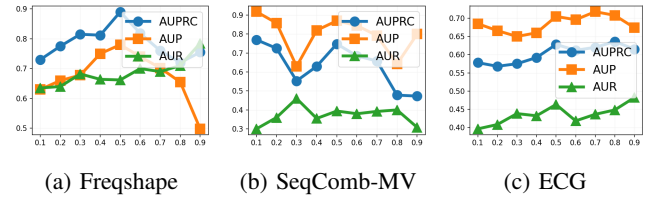
	VALIDITY	CONFIDENCE	SPARSITY	PROX.-L1	PROX.-L2
FREQSHAPES					
TimeX _{cf} ++	0.974 ± 0.004	0.964 ± 0.005	0.214 ± 0.005	0.339 ± 0.021	0.757 ± 0.030
w/o STE	0.769 ± 0.055	0.763 ± 0.055	0.174 ± 0.014	0.281 ± 0.031	0.616 ± 0.059
w/o $\mathcal{L}_{LC}$	0.000 ± 0.000	0.007 ± 0.001	0.091 ± 0.001	0.001 ± 0.000	0.002 ± 0.000
w/o $\mathcal{L}_{KL}$	0.936 ± 0.035	0.928 ± 0.034	0.220 ± 0.012	0.342 ± 0.030	0.730 ± 0.048
w/o $\mathcal{L}_{dr}$	0.955 ± 0.032	0.946 ± 0.032	0.216 ± 0.014	0.512 ± 0.070	1.107 ± 0.128
SEQCOMB-MV					
TimeX _{cf} ++	0.983 ± 0.007	0.979 ± 0.008	0.305 ± 0.041	0.152 ± 0.016	0.353 ± 0.033
w/o STE	0.804 ± 0.049	0.799 ± 0.049	0.114 ± 0.007	0.114 ± 0.016	0.368 ± 0.030
w/o $\mathcal{L}_{LC}$	0.000 ± 0.000	0.003 ± 0.001	0.079 ± 0.005	0.000 ± 0.000	0.001 ± 0.000
w/o $\mathcal{L}_{KL}$	0.940 ± 0.036	0.941 ± 0.033	0.150 ± 0.015	0.146 ± 0.016	0.395 ± 0.025
w/o $\mathcal{L}_{dr}$	0.907 ± 0.039	0.907 ± 0.038	0.129 ± 0.013	0.286 ± 0.062	0.771 ± 0.122
ECG					
TimeX _{cf} ++	0.848 ± 0.104	0.847 ± 0.101	0.084 ± 0.026	0.251 ± 0.097	0.767 ± 0.185
w/o STE	0.637 ± 0.128	0.638 ± 0.128	0.049 ± 0.015	0.111 ± 0.031	0.499 ± 0.080
w/o $\mathcal{L}_{LC}$	0.000 ± 0.000	0.015 ± 0.005	0.096 ± 0.003	0.000 ± 0.000	0.001 ± 0.000
w/o $\mathcal{L}_{KL}$	0.824 ± 0.124	0.831 ± 0.116	0.104 ± 0.037	0.311 ± 0.139	0.850 ± 0.214
w/o $\mathcal{L}_{dr}$	0.847 ± 0.110	0.853 ± 0.102	0.106 ± 0.013	0.539 ± 0.084	1.675 ± 0.221

For **attribution** explanations, TimeX_a++ retains the best AUPRC prediction on these datasets and is also slightly ahead for AUP and AUR overall, as shown in Table VII, while TIMEX fails to predict on SeqComb-MV due to non-convergence. Under the CNN predictor, our method achieves the highest AUPRC and AUP for both SeqComb-MV and ECG datasets.

For **counterfactual** explanations, Table XI presents the comparison results on three benchmarks. For the baselines, similar behaviors are observed across different architectures, except for M-CELS, whose performance exhibits large variance, performing better with CNN and worse with LSTM. We also observe results consistent with our main findings, where TimeX_{cf}++ achieves a balance among validity, confidence, sparsity, and proximity across both CNN and LSTM backbones, indicating strong generalizability across architectures.

G. Ablation Study and Hyperparameter Analysis

To disentangle the contributions of the proposed components, we conduct extensive ablation studies and hyperparameter analyses across both attribution and counterfactual tasks, evaluated on FreqShape, SeqComb-MV, and ECG datasets.

Fig. 5: **Attribution** explanation performance varying the sparsity parameter r .

a) *Effect of STE*: For **attribution** explanations as shown in Table X, TimeX_a++ outperforms leading baselines by 39.78% in AUPRC, 22.04% in AUP, and 3.10% in AUR on ECG arrhythmia detection, indicating that the use of the STE plays a critical role in optimizing the discrete explanation masks. On average, employing STE yields an average 8.87% increase in AUPRC across all datasets compared to a continuous masking strategy, alongside consistent enhancements in AUR, providing tangible evidence for utilizing hard masks. In the **counterfactual** domain, removing the STE results in a significant drop in validity and confidence. Although sparsity superficially improves, this indicates that the generation process has collapsed into a local optimum where perturbations are overly constrained and fail to effectively flip the model prediction, highlighting the necessity of STE in balancing the multi-objective trade-off.

b) *Effect of Different Losses*: We examine the effectiveness of the individual loss components constraining our framework. For **attribution** explanations, the absence of the Kullback-Leibler maintenance loss ($\mathcal{L}_{KL}$) widens the distributional gap between original and perturbed instances, dropping performance. The removal of the reference distance loss ($\mathcal{L}_{dr}$) naturally causes the explainer to fail, as it removes the primary target for bottleneck extraction. Furthermore, without consistency labeling ($\mathcal{L}_{LC}$), performance on synthetic datasets drops significantly.

In the **counterfactual** domain (Table XII), removing the label consistency loss ($\mathcal{L}_{LC}$) dramatically degrades validity, as the generator loses its primary signal for targeted label flipping.

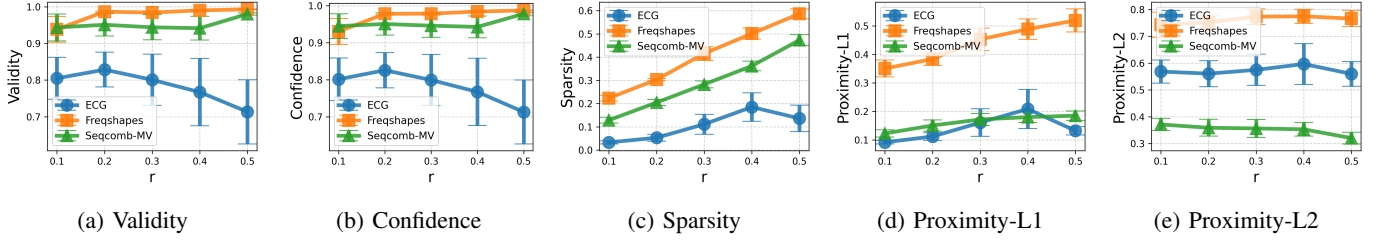


Fig. 6: The hyperparameter r analysis for **counterfactual** explanations on Freqshape (univariate), SeqComb-MV (multivariate), and ECG (real-world) datasets.

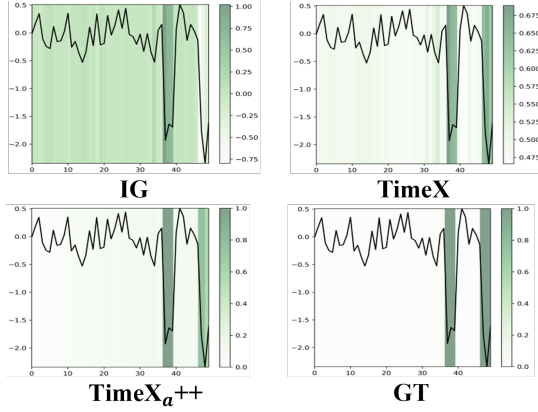


Fig. 7: Visualization of two attribution explainers and TimeX_a++ on the FreqShapes dataset.

Removing the reference distance loss ($\mathcal{L}_{dr}$), which acts as a structural regularization term, leads to a substantial decline in proximity. The maintenance loss ($\mathcal{L}_{KL}$) has a relatively smaller but consistent impact on preserving distributional consistency. Interestingly, the effect of the connectivity loss ($\mathcal{L}_{con}$) varies depending on the underlying data characteristics, which improves performance on FreqShapes by promoting continuous modifications but degrades performance on ECG, as critical ECG patterns are inherently distributed across multiple discrete segments rather than concentrated in a single contiguous window. Finally, as previously analyzed, excising the structural anchor ($\mathcal{L}_{bound}$) results in a dramatic deterioration of sparsity and proximity metrics; without this causal anchor, the generator scatters perturbations across the entire sequence.

c) Choosing the Hyperparameter r : The parameter r governs the sparsity of the masks learned during the process. For **attribution**, in Figure 5, lower values of the parameter are associated with a decrease in the performance (AUR). The performance stabilizes when the value is between 0.4 and 0.7, suggesting robustness in this range. In Figure 6, for **counterfactuals**, r balances sparsity, validity, and proximity by providing prior knowledge about the expected size of counterfactual explanations. As r increases, validity and confidence improve, whereas sparsity and proximity deteriorate, reflecting the inherent multi-objective trade-offs.

H. Case Study

Saliency maps are a potent tool for visualizing the significance of features. We demonstrate the attribution saliency maps of the benchmarks and TimeX_a++ on the FreqShapes

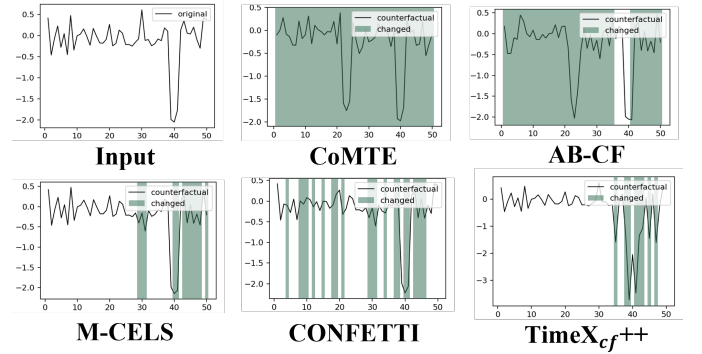


Fig. 8: The cases visualization of the FreqShapes dataset for counterfactual explanations. We demonstrate pairs of counterfactual cases with prediction 2 and target label 0.

dataset in Figure 7. IG identifies unnecessarily large areas as important, which can be untenable for noisy datasets. TIMEX struggles to describe important sub-instances with certainty due to data distribution bias from its constructed classifiers. In stark contrast, TimeX_a++ focuses accurately on the peaks for matching ground-truth explanations.

For counterfactual explanations in Figure 8, we compare TimeX_{cf}++ with baseline methods, with the modified parts highlighted. TimeX_{cf}++ requires fewer modifications, and these are highly related to the underlying patterns. In contrast, CoMTE and AB-CF modify a large portion of the original input, resulting in trivial solutions (e.g., generating highly similar instances). M-CELS fails to generate valid counterfactual explanations in most cases. CONFETTI produces sparse modifications but its results lack robustness to noise.

VII. CONCLUSION

In this work, we theoretically investigate an information-theoretic guided objective for time series explanations that ensures compactness and informativeness. We propose a novel approach, TimeX++, based on the IB principle, which allows a traceability computation to produce faithful attribution explanations and robust counterfactual explanations. Comprehensive studies on synthetic and real-world datasets have confirmed that TimeX++ surpasses existing approaches in performance and efficiency. Its effectiveness shows that TimeX++'s capability to reflect complex behaviors of pre-trained time series classifiers accurately. However, it may involve some hyperparameters in the learning objective to control the quantifiers of the explanation, especially when dealing with different datasets, which might be the key limitation.

REFERENCES

- [1] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, page 061907, 2001.
- [2] Emre Ates, Burak Aksar, Vitus J. Leung, and Ayse K. Coskun. Counterfactual explanations for multivariate time series. In *ICAPAI*, pages 1–8, 2021.
- [3] Omar Bahri, Soukaina Filali Boubrahimi, and Shah Muhammad Hamdi. Shapelet-based counterfactual explanations for multivariate time series. *arXiv preprint arXiv:2208.10462*, 2022.
- [4] João Bento, Pedro Saleiro, André F Cruz, Mário AT Figueiredo, and Pedro Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. In *SIGKDD*, pages 2565–2573, 2021.
- [5] Alan Gabriel Paredes Cetina, Kaouther Benguessoum, Raoni Lourenco, and Sylvain Kubler. Counterfactual explainable ai (xai) method for deep learning-based multivariate time series classification. *AAAI*, 40(21):17393–17400, Mar. 2026.
- [6] Zhuomin Chen, Gabriel Lucchesi, Qingkai Dong, Xu Zheng, Dongjin Song, Qingsong Wen, Wei Cheng, Jingchao Ni, and Dongsheng Luo. From signals to semantics: A survey on time series explainability through a human-cognitive lens. *Authorea Preprints*, 2026.
- [7] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*, pages 3504–3512, 2016.
- [8] Yu-Neng Chuang, Guanchu Wang, Fan Yang, Quan Zhou, Pushkar Tripathi, Xuanting Cai, and Xia Hu. CoRTX: Contrastive framework for real-time explanation. In *ICLR*, pages 1–23, 2023.
- [9] Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *ICML*, pages 2166–2177, 2021.
- [10] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.
- [11] Eoin Delaney, Derek Greene, and Mark T Keane. Instance-based counterfactual explanations for time series classification. In *International Conference on Case-Based Reasoning*, pages 32–47. Springer, 2021.
- [12] Joseph Enguehard. Learning perturbations to explain time series predictions. In *ICML*, pages 9329–9342, 2023.
- [13] Lukas Faber, Amin K. Moghaddam, and Roger Wattenhofer. When comparing to ground truth is wrong: On evaluating gnn explanation methods. In *SIGKDD*, pages 332–341, 2021.
- [14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.
- [15] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3(11):e745–e750, 2021.
- [16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.
- [17] Jacqueline Höllig, Cedric Kulbach, and Steffen Thoma. Tsevo: Evolutionary counterfactual explanations for time series classification. In *ICMLA*, pages 29–36. IEEE, 2022.
- [18] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, volume 32. Curran Associates, Inc., 2019.
- [19] Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. In *NeurIPS*, pages 26726–26739, 2021.
- [20] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 6441–6452. Curran Associates, Inc., 2020.
- [21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, pages 1–12, 2017.
- [22] Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In Silvia Chiappa and Roberto Calandra, editors, *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 26–28 Aug 2020.
- [23] Guillaume Jeanneret, Loic Simon, and Frederic Jurie. Diffusion models for counterfactual explanations. In *ACCV*, pages 858–876, December 2022.
- [24] Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A Pickett, and Varun Dutt. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data*, 3:4, 2020.
- [25] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AIES*, AIES '21, page 652–663, New York, NY, USA, 2021. Association for Computing Machinery.
- [26] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AIES*, AIES '21, page 652–663, New York, NY, USA, 2021. Association for Computing Machinery.

- [27] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [28] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: unjustified counterfactual explanations. In *IJCAI, IJCAI'19*, page 2801–2807. AAAI Press, 2019.
- [29] Kin Kwan Leung, Clayton Rooke, Jonathan Smith, Saba Zuberi, and Maksims Volkovs. Temporal dependencies in feature importance for time series prediction. In *ICLR*, pages 1–18, 2023.
- [30] Peiyu Li, Omar Bahri, Soukaïna Filali Boubrahimi, and Shah Muhammad Hamdi. Attention-based counterfactual explanation for multivariate time series. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 287–293. Springer, 2023.
- [31] Peiyu Li, Omar Bahri, Soukaïna Filali Boubrahimi, and Shah Muhammad Hamdi. M-cels: Counterfactual explanation for multivariate time series data guided by learned saliency maps. In *ICMLA*, pages 713–718, 2024.
- [32] Peiyu Li, Soukaïna Filali Boubrahimi, and Shah Muhammad Hamdi. Motif-guided time series counterfactual explanations. In Jean-Jacques Rousseau and Bill Kapralos, editors, *ICPR*, pages 203–215, Cham, 2023. Springer Nature Switzerland.
- [33] Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *ICLR*, pages 1–21, 2024.
- [34] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *NeurIPS*, 33:19620–19631, 2020.
- [35] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra, editors, *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR, 26–28 Aug 2020.
- [36] Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *ICML*, pages 15524–15543. PMLR, 2022.
- [37] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
- [38] George B. Moody and Roger G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20:45–50, 2001.
- [39] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.
- [40] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. CounterGAN: Generating counterfactuals for real-time recourse and interpretability using residual gans. In James Cussens and Kun Zhang, editors, *UAI*, volume 180 of *Proceedings of Machine Learning Research*, pages 1488–1497. PMLR, 01–05 Aug 2022.
- [41] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- [42] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In Gustavo Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pages 4574–4594. PMLR, 28–30 Mar 2022.
- [43] Owen Queen, Thomas Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency. In *NeurIPS*, 2023.
- [44] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, volume 32. Curran Associates, Inc., 2019.
- [45] Mario Refoyo and David Luengo. Sub-space: Subsequence-based sparse counterfactual explanations for time series classification problems. In *World Conference on Explainable Artificial Intelligence*, pages 3–17. Springer, 2024.
- [46] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *ISWC*, pages 108–109, 2012.
- [47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *SIGKDD, KDD '16*, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.
- [48] Karsten Roth et al. Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*, pages 9430–9440, 2022.
- [49] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
- [50] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AIES, AIES '20*, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery.
- [51] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328, 2017.
- [52] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *MLHC*, pages 322–337, 2017.
- [53] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE*

- Information Theory Workshop*, pages 1–5. IEEE, 2015.
- [54] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? Instance-wise feature importance for time-series black-box models. In *NeurIPS*, pages 799–809, 2020.
 - [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.
 - [56] Zhendong Wang, Isak Samsten, Ioanna Miliou, Rami Mochaourab, and Panagiotis Papapetrou. Glacier: guided locally constrained counterfactual explanations for time series classification. *Machine Learning*, 113(7):4639–4669, 2024.
 - [57] Zhendong Wang, Isak Samsten, Rami Mochaourab, and Panagiotis Papapetrou. Learning time series counterfactuals via latent space representations. In *International Conference on Discovery Science*, pages 369–384. Springer, 2021.
 - [58] Yu Zhang, Peter Tiño, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.
 - [59] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. OOD-CV: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *ECCV*, pages 163–180, 2022.
 - [60] Xu Zheng, Chaohao Lin, Sipeng Chen, Zhuomin Chen, Jimeng Shi, Jayantha Obeysekera, Jingchao Ni, Wei Cheng, Jason Liu, and Dongsheng Luo. Uncovering insights of compound flooding with data-driven AI. In *SIGKDD*, 2026.
 - [61] Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-fidelity: A robust framework for faithfulness evaluation of explainable ai. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *ICLR*, volume 2025, pages 12772–12804, 2025.