CoLaX: Context-Aware Local Explanations for Time Series Classification

Kexuan Zhang[1], **Paolo Frazzetto**[2], **Xiaobei Zou**[1],
Gustau Camps-Valls[2†], **Yang Tang**[1†]

[1]East China University of Science and Technology, China
[2]Image Processing Laboratory, Universitat de València, Spain

† Corresponding authors

Abstract

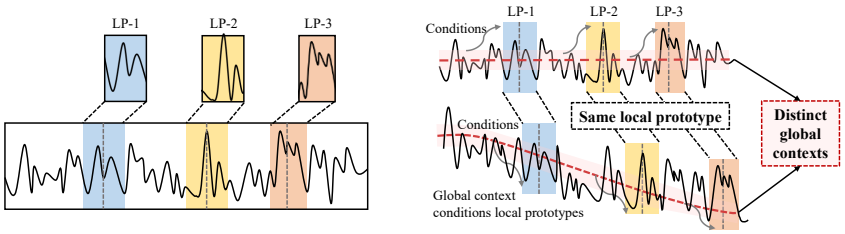
Post-hoc explanations for time series classifiers usually score local patterns, such as subsequences or shapelets, without regard to the rest of the sequence. Yet the same motif can carry different predictive meaning under different global temporal conditions, such as trends or regimes. We propose CoLaX, a context-aware post-hoc explanation framework that builds on shapelet-driven Shapley attribution and conditions local-pattern relevance on a learned representation of the full sequence. Global context acts at two levels: it produces sample-specific prototype weights and a temporally resolved signal that reweights the local activation map. The resulting context-conditioned segments are the players of a Shapley attribution game whose values come only from the frozen black-box classifier. On synthetic benchmarks where identical motifs change relevance across contexts, CoLaX improves AUPRC over the strongest baseline by 30% on average. On real-world data, it achieves the highest AUPRC on ECG. Additional ablations indicate that the gains come from input-dependent conditioning rather than from additional learnable weights. CoLaX is therefore most useful when recurring local motifs change relevance across sequence-level contexts.

1 Introduction

Time series are sequences of indexed data that follow a specific order, and their analysis is relevant across a wide range of domains, such as healthcare [24] or the environment [11]. Although Deep Learning methods specifically designed for sequential data have a long tradition [12], recent advances have demonstrated the capabilities of these technologies as highly accurate predictive tools across an increasing number of time series classification and regression tasks [20]. Such models are optimized to learn high-dimensional representations from rich, raw data; yet their complexity prevents an understanding of the logic behind their predictions. This hinders the adoption of Artificial Intelligence (AI) in real-world scenarios where users need to trust its predictions [6].

This lack of transparency necessitates a paradigm shift toward time series classifiers that are inherently interpretable or post-hoc explainable and, therefore, numerous solutions have been proposed [31, 23]. Most existing approaches attribute predictions to individual timesteps or short segments using perturbation-based techniques [33] or gradient-based saliency [14]. A growing line of work preserves temporal structure in the explanation itself: LIMESegment builds segment-level perturbation neighbourhoods [28], TIMING incorporates temporal dependence into integrated gradients [15], InteDisUX combines consecutive attribution scores into discriminative segments [32], and subsequence-level explainers such as Implet and ShapeX attribute predictions to meaningful motifs [19, 13]. Together with Shapley- and shapelet-based models [25, 16], these works reflect a consensus that short and discriminative patterns are central to both performance and interpretability.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Local-Prototype-only explanation **(b) Challenge: Ambiguity without global context**

Figure 1: Ambiguity of local prototype explanations without global temporal context. (a) **Local-pattern–only explanation.** Local prototypes (LP-1 to LP-3) capture discriminative motifs but implicitly assume that their contributions are invariant across contexts. (b) **Context-dependent ambiguity under distinct global temporal structures.** Identical local patterns may appear under distinct global temporal structures. Their contributions should be interpreted as conditional on the global context, motivating the need for context-aware explanation frameworks.

Despite these advances, existing explainers score each local pattern largely independently of the sequence in which it occurs. A local activation alone does not reveal whether a pattern is predictive in a given sequence, as identical motifs may contribute differently depending on the surrounding temporal context, such as long-term trends or regimes. Separating local and global temporal components is well established in forecasting [36, 34] and in explainable representation learning [27]; however, to the best of our knowledge, classification explainers do not model the relevance of a local motif as conditional on a representation of the full sequence. As illustrated in Figure 1, global temporal context acts as a conditioning variable that modulates the contribution of local patterns. For instance, a short-term fluctuation may signal an anomaly under a stable trend but represent normal variability within a volatile sequence (e.g., a heartbeat spike may mean something different during exercise vs. rest). While existing methods may correctly identify salient local patterns, they often fail to convey the overall temporal conditions under which the model interprets them. This gap substantially limits interpretability when predictions depend jointly on short-term dynamics and long-range structure.

In this work, we argue that faithful and informative explanations for time series classification models must jointly account for local pattern-level evidence and global temporal context. Therefore, we propose **CoLaX** (**Co**ntext-aware **L**oca**l** e**X**planations), a context-aware post-hoc explanation framework. that explicitly constructs distinct representations for short-term local temporal patterns and long-range temporal structures, integrating them through context-dependent modulation. The global temporal context summarizes holistic sequence-level characteristics and provides the conditioning background for interpreting local pattern contributions.

Specifically, CoLaX first identifies salient local temporal patterns that contribute to frozen black-box model predictions, and then contextualizes their contributions using a global temporal representation that captures diverse temporal regimes. In this manner, the global context modulates the influence of these local explanations, revealing not only *which patterns matter*, but also *under which temporal contexts they matter*. This design yields structured, instance-level explanations that aim to more closely reflect the model's inference process.

The main contributions can be summarized as follows:

- We frame time series explanation as a problem of context-dependent local relevance, where the contribution of a recurring motif may change with the sequence-level context in which it occurs;
- We introduce CoLaX,, which conditions local patterns on global temporal context at both the sequence level (prototype weights) and the temporal level (context signal), so that the units of segment-level Shapley attribution depend on each input, while attribution values are computed only from the frozen black-box classifier;
- Extensive experiments on diverse time series benchmarks demonstrate that CoLaX produces more faithful explanations than existing explainability methods, showcasing the validity of this approach when repeated local motifs shift in relevance across sequence-level contexts.

2 Method

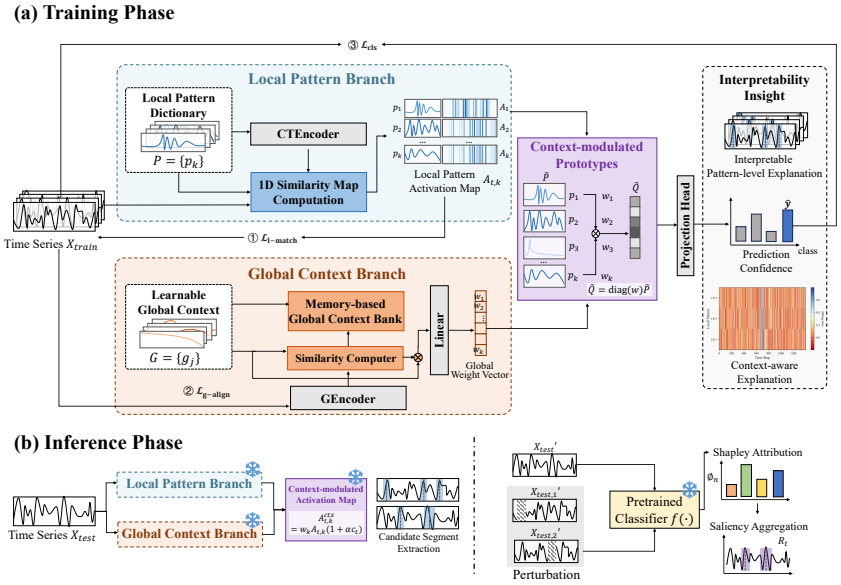


Figure 2: Overview of the proposed CoLaX framework for context-aware explanation in time series classification. The framework has two parts: a local pattern branch that builds a dictionary of short-term prototypes to detect important local subsequences, and a global context branch that uses a memory-based context bank to capture long-range background information. (a) During training, the branches are optimized end-to-end using three objectives: ① a local matching loss $L_{\text{l-match}}$ to align prototypes with salient subsequences; ② a global alignment loss $L_{\text{g-align}}$ to enforce coherent long-range structural representations; and ③ a classification loss L_{cls} for task-level supervision. (b) At inference time, the trained branches are frozen and serve as a post-hoc explainer. The framework explains a black-box time-series classifier by detecting context-dependent local patterns, generating pattern-based perturbations, and querying the classifier on original and perturbed inputs. It aggregates prediction changes into a saliency map, yielding instance-level, context-aware explanations.

This work focuses on the time series classification task, denoting a multivariate time series as $\boldsymbol{X} \in \mathbb{R}^{T \times D}$ of length T with D features. The associated label $\boldsymbol{y} \in \{0,1\}^C$ lies in a C-dimensional label space, where C denotes the number of classes. The objective is to learn a classifier $f(\cdot)$ that maps $\boldsymbol{X}$ to a predicted label $\hat{y} = f(\boldsymbol{X})$. Given a trained classifier, post-hoc explainability aims to identify which temporal components of the input sequence drive the prediction. To formalize this, we define an explanation function $\mathcal{E}(\cdot)$ that assigns a temporal saliency map $\boldsymbol{R} \in [0,1]^T$ to the input sequence, where each element R_t quantifies the importance of time step t to the predicted class probability (see also Appendix D).

Our approach is motivated by the observation that time series predictions are often driven by the interaction between short-term local dynamics and long-range temporal context [36, 34]. Local temporal patterns capture transient, event-level fluctuations, while global temporal context characterizes slowly evolving trends and structural regimes that condition the interpretation of these patterns. Explaining model behavior, therefore, requires modeling both components and their interactions.

To this end, we introduce CoLaX. It consists of three main components: *(i)* a local pattern branch that discovers discriminative subsequences, *(ii)* a global context branch that captures sequence-level structure, and *(iii)* a gating mechanism that conditions local contributions on global context. We train these jointly, then use the learned representations to guide post-hoc perturbations of any black-box classifier $f(\cdot)$. By using the discovered patterns as the basis for perturbation, CoLaX ensures that the interventions used to probe the black-box are temporally coherent and contextually grounded, leveraging the extracted dictionary of local temporal motifs and the memory bank of global contexts. In this fashion, the framework produces structured explanations that reveal both *which patterns are important* and *under which temporal conditions they are influential*. The overall architecture is illustrated in Figure 2, while the algorithmic description is provided in the Appendix F.

2.1 Local Pattern Branch

Short-term fluctuations in time series often encode salient high-frequency dynamics, such as transient events or specific local variations, that are crucial for classification. These fine-scale patterns provide localized evidence that, while distinct from global trends, contribute significantly to model predictions. To capture these behaviors, the Local Pattern Branch learns a dictionary of temporal prototypes that represent characteristic local dynamics.

Formally, let $P = \{p_1, p_2, \cdots, p_{K_{\text{local}}}\}$ denote a collection of K_{local} learnable local prototypes, where each $p_k \in \mathbb{R}^{L_p \times D}$ encodes a representative short-term pattern of length L_p. To enhance expressiveness and capture internal temporal dependencies, these prototypes are processed by a CNN-Transformer encoder (denoted as CTEncoder), yielding a set of refined feature representations $\widetilde{P} = \text{CTEncoder}(P) \in \mathbb{R}^{K_{\text{local}} \times D}$. This encoder integrates convolutional layers for local feature extraction with a Transformer module for modeling sequential dependencies [22].

Given the input $\boldsymbol{X}$, we compute its similarity with local prototypes via the convolution operator $I = \text{Conv}(\boldsymbol{X}, \widetilde{P})$, in which the prototype vectors act as the convolution filters. The activation map $A \in [0,1]^{T \times K_{\text{local}}}$ is then obtained by applying the softmax function across the prototype dimension, i.e., $A_{t,k} = \text{Softmax}(I_{t,:})_k$, where $A_{t,k}$ measures the alignment strength between the k-th prototype at time t.

Conceptually, the local pattern branch learns a dictionary of short-term motifs that summarize salient local dynamics. The activation $A_{t,k}$ reflects the relative matching strength of prototype k at time step t, while the local matching objective encourages the learned prototypes to remain aligned with observed subsequences.

2.2 Global Context Branch

While local prototypes capture short-term fluctuations, their semantic interpretation often relies on the broader and slowly varying temporal background. Thus, the Global Context Branch is introduced to learn a representation of this long-range structure, capturing information that is inaccessible to short observation windows and modulating local evidence accordingly.

2.2.1 Global Context Representation

To capture slowly evolving temporal dynamics, we employ a memory-based global context module. Let $\mathcal{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_{K_{\text{global}}}\}$ denote a set of K_{global} learnable context vectors in the shared feature space $\mathbb{R}^D$, each encoding a prototypical long-range structure.

First, global embeddings $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{T'}] \in \mathbb{R}^{T' \times D}$ are extracted from the input via a Transformer-based global encoder, $\boldsymbol{Z} = \text{GEncoder}(\boldsymbol{X})$, where T' denotes the length of the embedding sequence. The relevance between each time-step embedding $\boldsymbol{z}_t$ and the context vectors is computed via a soft attention mechanism:

$$a_{t,k} = \frac{\exp(\boldsymbol{z}_t^\top \boldsymbol{g}_k / \tau_1)}{\sum_{j=1}^{K_{\text{global}}} \exp(\boldsymbol{z}_t^\top \boldsymbol{g}_j / \tau_1)}, \tag{1}$$

where τ_1 is a temperature parameter. To derive a stable summary of the global structure, we aggregate the temporal embeddings into a single global representation: $\boldsymbol{E}_{\text{global}} = \text{Pool}(\boldsymbol{Z})$.

The central role of the global context is to make local-pattern relevance sample dependent. Given $\boldsymbol{E}_{\text{global}}$, we project the global representation into a query vector and the refined local prototypes into key representations:

$$\boldsymbol{q} = \boldsymbol{W}_q \boldsymbol{E}_{\text{global}}, \qquad \boldsymbol{K} = \widetilde{\boldsymbol{P}} \boldsymbol{W}_k, \tag{2}$$

where $\boldsymbol{W}_q \in \mathbb{R}^{D \times D}$ and $\boldsymbol{W}_k \in \mathbb{R}^{D \times D}$ are learnable projections.

Let $\boldsymbol{k}_k = \boldsymbol{K}_{k,:}^\top \in \mathbb{R}^D$ denote the column-vector form of the k-th row of $\boldsymbol{K}$. The sample-specific relevance of prototype k is:

$$w_k = \text{Sigmoid}\left(\frac{\boldsymbol{q}^\top \boldsymbol{k}_k}{\tau_2}\right), \tag{3}$$

The gate is applied only after its values have been computed:

$$\widetilde{\boldsymbol{Q}} = \mathrm{diag}(\boldsymbol{w})\widetilde{\boldsymbol{P}}. \tag{4}$$

Hence, $\mathrm{diag}(\cdot)$ is used for prototype reweighting rather than for gate computation. The gating weights $\boldsymbol{w}$ depend on the input-specific global representation, allowing the same local prototype to receive different weights across sequences. During training, these weights modulate the prototype representations used by the auxiliary prediction head. During post-hoc inference, they further condition local activation maps to construct context-dependent explanatory segments.

2.2.2 Memory-based Context Bank

To organize the global representation space around multiple prototypical temporal regimes, we maintain a memory bank preserving multiple prototypical regimes [29]. For each context prototype j, its assigned temporal embeddings are summarized through the normalized aggregation:

$$\bar{\boldsymbol{z}}_k = \frac{\sum_{t=1}^{T'} a_{t,k}\boldsymbol{z}_t}{\sum_{t=1}^{T'} a_{t,k} + \epsilon}, \tag{5}$$

where ϵ is a small constant for numerical stability. The context vectors are then updated through:

$$\boldsymbol{\psi}_j = \sigma\left(\boldsymbol{U}_\psi \boldsymbol{g}_j + \boldsymbol{W}_\psi \bar{\boldsymbol{z}}_j\right), \tag{6}$$

$$\boldsymbol{g}_j \leftarrow (1 - \boldsymbol{\psi}_j) \odot \boldsymbol{g}_j + \boldsymbol{\psi}_j \odot \bar{\boldsymbol{z}}_j, \tag{7}$$

where $\boldsymbol{U}_\psi, \boldsymbol{W}_\psi$ are learnable projections and $\sigma(\cdot)$ denotes the sigmoid function. The normalized aggregation prevents the scale of the memory update from growing with sequence length and allows the context bank to represent multiple recurring global regimes.

2.3 Prediction Head

The context-modulated representation $\widetilde{\boldsymbol{Q}}$ is fed into a multilayer perceptron (MLP) prediction head. Crucially, the prediction head explicitly incorporates a sigmoid activation to map the fused representations to the label space:

$$\hat{\boldsymbol{y}} = \sigma\left(\mathrm{MLP}(\widetilde{\boldsymbol{Q}})\right), \tag{8}$$

where $\hat{\boldsymbol{y}} \in [0,1]^C$ denotes the predicted probabilities.

2.4 Training Losses

To jointly optimize local pattern discovery, global context modeling, and task-level prediction, we employ a composite objective function comprising three complementary terms: (1) local semantic matching $\mathcal{L}_{\text{l-match}}$, (2) global semantic alignment $\mathcal{L}_{\text{g-align}}$, and (3) classification loss $\mathcal{L}_{\text{cls}}$.

(1) Local Semantic Matching. The local pattern branch is designed to capture representative short-term temporal motifs. We introduce a semantic matching loss, aligning each prototype with the most responsive subsequences in the input time series [13]. The local matching objective anchors the learned prototypes to representative subsequences identified by their activation responses, encouraging correspondence between prototype representations and observed temporal patterns.

Let $A_{t,k}$ denote the activation of prototype p_k at time t. The maximally activated position for prototype p_k is defined as

$$t_k^* = \arg\max_t A_{t,k}, \tag{9}$$

from which a local subsequence of length L_p is extracted:

$$\boldsymbol{s}_k = \boldsymbol{X}[t_k^* - \lfloor L_p/2 \rfloor : t_k^* + \lceil L_p/2 \rceil, :]. \tag{10}$$

Let $\mathcal{S}_k$ denote the set of subsequences assigned to prototype p_k across the batch. The local semantic matching loss is defined as

$$\mathcal{L}_{\text{l-match}} = \sum_{k=1}^{K_{\text{local}}} \sum_{x \in \mathcal{S}_k} \|x - p_k\|_F^2, \tag{11}$$

which encourages each prototype to align with the representative local dynamics in its activation neighborhood.

(2) Global Semantic Alignment. To encourage each global embedding to associate with the most relevant context representation while remaining discriminative from others, we introduce a global semantic alignment objective. For sample i, let $\boldsymbol{e}^{(i)} = \boldsymbol{E}^{(i)}_{\text{global}}$ denote its pooled global representation. The most strongly associated context prototype is $k_i^+ = \arg\max_k \left(\boldsymbol{e}^{(i)\top}\boldsymbol{g}_k\right)$. The global alignment objective is

$$\mathcal{L}_{\text{g-align}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\boldsymbol{e}^{(i)\top}\boldsymbol{g}_{k_i^+}/\tau_g\right)}{\sum_{k=1}^{K_{\text{global}}}\exp\left(\boldsymbol{e}^{(i)\top}\boldsymbol{g}_k/\tau_g\right)}, \tag{12}$$

where the context vectors are ℓ_2-normalized and τ_g is the temperature used for global alignment.

(3) Classification Objective. We adopt the standard cross-entropy loss $\mathcal{L}_{\text{cls}}$ for C classes, calculated using the predicted probabilities $\hat{\boldsymbol{y}}$.

Overall Objective. The final training loss is given by

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{l-match}} + \lambda_2\mathcal{L}_{\text{g-align}} + \mathcal{L}_{\text{cls}}, \tag{13}$$

where λ_1, and λ_2 balance the influence of the prototype-specific and context-specific objectives. Collectively, these terms drive the model to learn structured, interpretable local patterns and global temporal contexts without compromising task-level predictive performance.

2.5 Post-hoc Explainability

After training, CoLaX is frozen and used to construct context-aware explanatory segments for a separate frozen black-box classifier $f(\cdot)$. Following the shapelet-driven perturbation paradigm [13], CoLaX conditions segment construction on global temporal context, while their contributions are determined by black-box prediction differences.

Context-conditioned Segment Construction. CoLaX incorporates global context at two levels. The pooled representation $\boldsymbol{E}_{\text{global}}$ generates sample-specific prototype weights $\boldsymbol{w}$, while the temporal context associations $a_{p,j}$ in Eq. 1 provide a time-dependent conditioning signal:

$$c_t = \text{Norm}\left[\text{Interp}\left(\{\max_k a_{s,k}\}_{s=1}^{T'}\right)\right]_t, \tag{14}$$

where Interp restores the original temporal resolution and Norm scales the context-association confidence to $[0, 1]$.

The two levels of conditioning jointly modulate local activations:

$$A^{\text{ctx}}_{t,k} = w_k A_{t,k}(1 + \alpha c_t), \tag{15}$$

where $\alpha \geq 0$ controls the temporal modulation strength. Here, w_k adjusts prototype relevance across samples, while c_t modulates activation strengths across temporal locations. Contiguous regions satisfying $A^{\text{ctx}}_{t,k} > \Omega$ are extracted as explanatory segments, forming $\mathcal{P}(\boldsymbol{X}) = \{\mathcal{I}_1, \ldots, \mathcal{I}_N\}$, where Ω denotes the activation threshold.

Black-box Shapley Attribution. For each coalition $S \subseteq \mathcal{P}(\boldsymbol{X})$, we construct the perturbed input

$$\boldsymbol{X}_S = \boldsymbol{m}_S \odot \boldsymbol{X} + (1 - \boldsymbol{m}_S) \odot \boldsymbol{X}_{\text{base}}, \tag{16}$$

where $\boldsymbol{m}_S \in \{0, 1\}^T$ retains segments in S and masks candidate segments outside S, leaving non-candidate regions unchanged. The baseline $\boldsymbol{X}_{\text{base}}$ is generated by linear interpolation to reduce discontinuities.

Let $c^\star = \arg\max_c f_c(\boldsymbol{X})$ and $v(S) = f_{c^\star}(\boldsymbol{X}_S)$. The attribution of segment n is computed as

$$\phi_n = \sum_{S \subseteq \mathcal{P}(\boldsymbol{X})\setminus\{n\}} \frac{|S|!(N - |S| - 1)!}{N!}\left[v(S \cup \{n\}) - v(S)\right]. \tag{17}$$

Thus, CoLaX constructs context-dependent attribution players, while their Shapley values are determined solely by the frozen black-box classifier.

Saliency Aggregation. The segment-level contributions are projected onto the temporal domain using maximum absolute aggregation:

$$R_t = \frac{R_t^{\text{raw}}}{\max_{t'} R_{t'}^{\text{raw}} + \epsilon}, \quad R_t^{\text{raw}} = \max_{n:t\in\mathcal{I}_n} |\phi_n|, \tag{18}$$

where $R_t^{\text{raw}} = 0$ if no segment covers t. Consequently, CoLaX constructs an input-conditioned attribution game whose explanatory units depend on global temporal context, while their importance is evaluated through black-box prediction changes.

3 Experiments

3.1 Evaluation on Synthetic and Real-world Data

To assess the effectiveness of CoLaX, we evaluate explanation quality across both synthetic and real-world datasets. We employ saliency score evaluations on datasets with ground-truth saliency scores, while conducting occlusion experiments on datasets lacking them. All experiments use pretrained Transformer-based classifiers as frozen black-box models, with access restricted to input-output [1].

Datasets. We evaluate our model using synthetic data comprising four motif-based binary classification datasets: MCC-E, MTC-H, MCC-H, and MTC-E [9, 13]. Notably, each sample in these datasets is annotated with ground-truth saliency to facilitate the evaluation of explanation quality. Regarding real-world data, we utilize the ECG dataset [21] with ground-truth labels and the UCR Archive [5] without ground-truth saliency annotations. In this case, we assess faithfulness directly on the frozen black-box classifiers through saliency-guided removal tests (Appendix B.3).

Table 1: Saliency evaluation on the four synthetic datasets. Best results are highlighted in bold, second-best underlined.

	MCC-E			MCC-H		
Method	AUPRC	AUP	AUR	AUPRC	AUP	AUR
IG	0.363 ± 0.005	0.483 ± 0.008	0.467 ± 0.005	0.436 ± 0.006	0.510 ± 0.006	0.579 ± 0.005
DYNAMASK	0.325 ± 0.005	0.472 ± 0.010	0.071 ± 0.003	0.351 ± 0.003	0.183 ± 0.002	0.157 ± 0.003
WINIT	0.168 ± 0.004	0.158 ± 0.011	0.386 ± 0.007	0.148 ± 0.004	0.097 ± 0.005	0.444 ± 0.007
CoRTX	0.476 ± 0.006	0.524 ± 0.017	0.459 ± 0.016	0.620 ± 0.004	0.536 ± 0.016	0.543 ± 0.003
MILLET	0.173 ± 0.006	0.218 ± 0.007	0.164 ± 0.007	0.302 ± 0.005	0.274 ± 0.008	0.295 ± 0.003
TimeX	0.369 ± 0.004	0.321 ± 0.004	0.631 ± 0.003	0.444 ± 0.003	0.667 ± 0.006	0.365 ± 0.003
TimeX++	0.368 ± 0.004	0.500 ± 0.006	0.280 ± 0.004	0.639 ± 0.004	0.641 ± 0.004	0.488 ± 0.002
ShapeX	0.641 ± 0.004	0.561 ± 0.008	0.368 ± 0.005	0.811 ± 0.001	0.684 ± 0.005	0.743 ± 0.006
CoLaX	$\mathbf{0.933 \pm 0.003}$	$\mathbf{0.764 \pm 0.004}$	$\mathbf{0.853 \pm 0.003}$	$\mathbf{0.977 \pm 0.000}$	$\mathbf{0.904 \pm 0.002}$	$\mathbf{0.769 \pm 0.005}$

	MTC-E			MTC-H		
Method	AUPRC	AUP	AUR	AUPRC	AUP	AUR
IG	0.147 ± 0.001	0.147 ± 0.003	0.540 ± 0.004	0.310 ± 0.003	0.532 ± 0.006	0.496 ± 0.006
DYNAMASK	0.139 ± 0.001	0.101 ± 0.002	0.280 ± 0.004	0.260 ± 0.002	0.496 ± 0.007	0.189 ± 0.003
WINIT	0.141 ± 0.003	0.091 ± 0.005	0.005 ± 0.006	0.150 ± 0.004	0.091 ± 0.005	0.462 ± 0.006
CoRTX	0.188 ± 0.006	0.175 ± 0.017	0.526 ± 0.017	0.243 ± 0.010	0.241 ± 0.014	0.543 ± 0.006
MILLET	0.142 ± 0.006	0.110 ± 0.004	0.155 ± 0.015	0.207 ± 0.011	0.231 ± 0.003	0.352 ± 0.001
TimeX	0.248 ± 0.003	0.630 ± 0.007	0.094 ± 0.001	0.380 ± 0.002	0.889 ± 0.002	0.154 ± 0.001
TimeX++	0.242 ± 0.002	0.491 ± 0.006	0.263 ± 0.003	0.390 ± 0.003	$\mathbf{0.900 \pm 0.001}$	0.145 ± 0.001
ShapeX	0.610 ± 0.005	0.396 ± 0.007	0.547 ± 0.008	0.679 ± 0.001	0.425 ± 0.002	0.902 ± 0.004
CoLaX	$\mathbf{0.723 \pm 0.007}$	$\mathbf{0.638 \pm 0.007}$	$\mathbf{0.784 \pm 0.006}$	$\mathbf{0.920 \pm 0.001}$	0.686 ± 0.001	$\mathbf{0.977 \pm 0.000}$

Table 2: Saliency evaluation on the ECG dataset.

Method	AUPRC	AUP	AUR
IG	0.418 ± 0.001	0.595 ± 0.002	0.320 ± 0.001
DYNAMASK	0.328 ± 0.001	0.525 ± 0.003	0.108 ± 0.008
WINIT	0.305 ± 0.001	0.443 ± 0.003	0.347 ± 0.001
CoRTX	0.374 ± 0.001	0.497 ± 0.002	0.303 ± 0.001
MILLET	0.302 ± 0.002	0.472 ± 0.006	0.310 ± 0.001
TimeX	0.472 ± 0.002	0.566 ± 0.003	0.446 ± 0.001
TimeX++	0.660 ± 0.001	0.726 ± 0.001	0.460 ± 0.001
ShapeX	0.723 ± 0.003	$\mathbf{0.840 \pm 0.003}$	$\mathbf{0.696 \pm 0.003}$
CoLaX	$\mathbf{0.741 \pm 0.005}$	0.765 ± 0.005	0.674 ± 0.005

[1] The code is available at `https://github.com/zkkkkkkkkkkx/CoLaX`

Table 3: Ablation study of model components and training objectives. The full model is compared with variants that remove the global temporal context branch, or individual loss terms, as well as randomized control settings. All configurations are evaluated under the same experimental protocol. The results indicate that the interaction between local temporal patterns and global temporal context is critical for model performance and explanation quality.

	MCC-E			MCC-H		
Method	AUPRC	AUP	AUR	AUPRC	AUP	AUR
Full	$\mathbf{0.933 \pm 0.003}$	$\mathbf{0.764 \pm 0.004}$	$\mathbf{0.853 \pm 0.003}$	$\mathbf{0.977 \pm 0.000}$	$\mathbf{0.904 \pm 0.002}$	$\mathbf{0.769 \pm 0.005}$
w/o Global	0.290 ± 0.004	0.249 ± 0.005	0.320 ± 0.007	0.902 ± 0.001	0.814 ± 0.002	0.740 ± 0.004
Random Global Context	0.586 ± 0.004	0.529 ± 0.004	0.743 ± 0.005	0.936 ± 0.000	0.903 ± 0.001	0.768 ± 0.004
Random Local Pattern	0.888 ± 0.003	0.806 ± 0.004	0.641 ± 0.006	0.814 ± 0.001	0.699 ± 0.002	0.689 ± 0.005
w/o $\mathcal{L}_{\text{g-align}}$	0.586 ± 0.007	0.540 ± 0.007	0.493 ± 0.007	0.981 ± 0.001	0.863 ± 0.002	0.809 ± 0.004
w/o $\mathcal{L}_{\text{l-match}}$	0.576 ± 0.004	0.515 ± 0.004	0.750 ± 0.005	0.898 ± 0.002	0.792 ± 0.004	0.725 ± 0.005
	MTC-E			MTC-H		
Method	AUPRC	AUP	AUR	AUPRC	AUP	AUR
Full	$\mathbf{0.723 \pm 0.007}$	$\mathbf{0.638 \pm 0.007}$	$\mathbf{0.784 \pm 0.006}$	$\mathbf{0.920 \pm 0.001}$	$\mathbf{0.686 \pm 0.001}$	$\mathbf{0.977 \pm 0.000}$
w/o Global	0.237 ± 0.005	0.148 ± 0.004	0.359 ± 0.009	0.439 ± 0.007	0.304 ± 0.005	0.596 ± 0.009
Random Global Context	0.627 ± 0.007	0.611 ± 0.007	0.724 ± 0.006	0.761 ± 0.003	0.611 ± 0.001	0.919 ± 0.002
Random Local Pattern	0.560 ± 0.007	0.495 ± 0.008	0.543 ± 0.008	0.789 ± 0.002	0.605 ± 0.004	0.910 ± 0.003
w/o $\mathcal{L}_{\text{g-align}}$	0.238 ± 0.005	0.148 ± 0.004	0.359 ± 0.009	0.439 ± 0.007	0.303 ± 0.005	0.594 ± 0.009
w/o $\mathcal{L}_{\text{l-match}}$	0.285 ± 0.006	0.224 ± 0.006	0.423 ± 0.009	0.714 ± 0.005	0.537 ± 0.004	0.888 ± 0.004

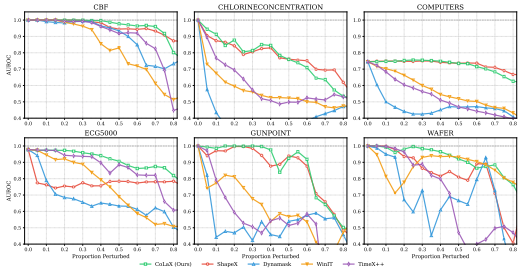


Figure 3: Least-salient-first removal on six of the 36 UCR datasets, shown as illustrative examples of different degradation behaviours.

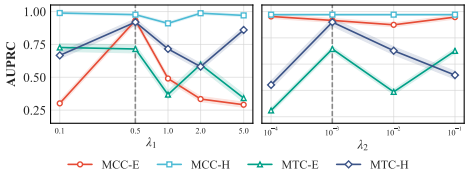


Figure 4: Sensitivity analysis on the selection of λ_1 and λ_2.

Baselines. Eight baselines, further described in Appendix B.3, are considered for comparison: IG [30], Dynamask [4], WinIT [17], CoRTX [3], MILLET [7], TIMEX [26], TIMEX++ [18], and ShapeX [13].

Results. Table 1 presents the evaluation results for the four synthetic datasets. CoLaX achieves the highest AUPRC across all four synthetic datasets and obtains the best performance on 11 of the 12 reported metrics. Regarding AUPRC, our method surpasses the second-best baseline by an average of 30.0%, with a peak of 45.6% improvement on MCC-E. Notably, TimeX [26] and TimeX++ [18] exhibit degraded performance on MCC-E and MTC-E, where identical local patterns occur under different temporal contexts. In contrast, CoLaX maintains stability across both settings by conditioning local pattern contributions on the global temporal context. Results for the ECG dataset are summarized in Table 2. CoLaX achieves the highest AUPRC, outperforming ShapeX by 2.49%, which indicates a better overall balance between precision and recall in identifying salient regions. However, ShapeX obtains higher AUP and AUR scores, showing that CoLaX does not uniformly dominate across all saliency metrics. This indicates a favorable precision–recall trade-off rather than uniform superiority across metrics. For the heterogeneous UCR archive, we selected a representative subset spanning distinct domains and exhibiting widely varying statistical properties, as shown in Figure 3. The analysis reveals that CoLaX demonstrates superior performance under the least-saliency removal (bottom-k) protocol, characterized by a more graceful degradation in AUROC scores.

Further details regarding the datasets, baselines, implementation, hyperparameters, metrics, and additional results are provided in the Appendices B and C.

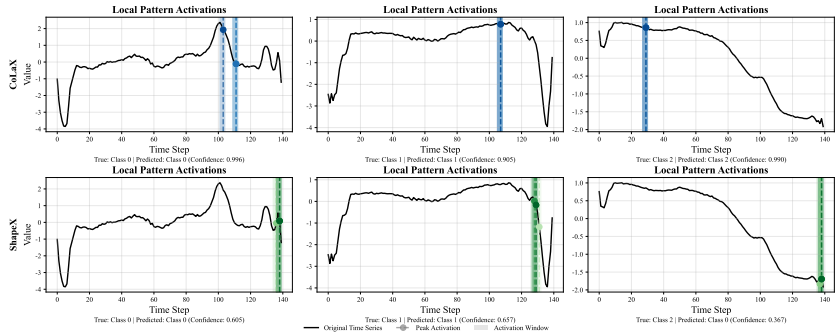


Figure 5: Representation-quality diagnostic on ECG5000: prototype activations of CoLaX (top) and ShapeX (bottom). Predictions and confidences are outputs of each method's internal surrogate head, which is used only during training; they are not predictions of the frozen black-box classifier and do not measure explanation faithfulness.

3.2 Ablation Study

To assess the contribution of individual components within CoLaX, we conduct a comprehensive ablation study by systematically removing or perturbing key architectural modules and training objectives. Experiments are performed on four synthetic datasets, with the full model as the reference. Quantitative results are reported in Table 3. Removing the global branch (*w/o Global*) causes severe degradation, with AUPRC drops of 68.9% and 66.8% on MCC-E and MTC-E, respectively, and an average reduction of 48.9%. This highlights the insufficiency of relying solely on local patterns. Replacing learned context with random signals (*Random Global Context*) still yields a 17.7% AUPRC drop, confirming that performance relies on semantically meaningful global context rather than mere architectural complexity. Randomizing local prototypes (*Random Local Pattern*) degrades performance by 14.4% on average. This asymmetry indicates that while local patterns are necessary, their interpretability depends heavily on conditioning by a coherent global context. Regarding training objectives, removing the global alignment loss (*w/o* $\mathcal{L}_{\text{g-align}}$) reduces performance on several datasets, particularly MCC-E and MTC-E, although modest improvements are observed for some metrics on MCC-H, suggesting that global representations must be semantically aligned. Similarly, excluding the local matching loss (*w/o* $\mathcal{L}_{\text{l-match}}$) consistently lowers performance, supporting its necessity in enforcing correspondence between prototypes and salient subsequences. To further determine whether these gains arise from sample-specific global conditioning or merely from introducing additional learnable weights over local prototypes, we replace the input-dependent prototype weights in Eq. 3 with a single learnable vector $\boldsymbol{w}_{\text{static}} \in \mathbb{R}^{K_{\text{local}}}$ shared across all samples. The global context branch and its associated alignment objective are removed, while the local pattern branch and black-box attribution procedure are retained. CoLaX improves the average AUPRC from 0.637 to 0.888, outperforming the static variant on 11 of the 12 reported metrics. These results show that a fixed dataset-level weighting of prototypes is insufficient: the main benefit arises from adapting prototype relevance to the global state of each input.

3.3 Parameter Analysis

Regarding the loss function hyperparameters, Figure 4 exhibits distinct peaks at our selected values ($\lambda_1 = 0.5, \lambda_2 = 10^{-3}$). Specifically, MCC-E is highly sensitive to λ_1, while the MTC datasets display an inverted U-shaped trend for λ_2, suggesting an effective balance between global alignment and regularization.

We also conducted investigations into the implementation of local patterns and global context, along with the selection of pattern length. More details can be found in Appendix C.2. Our analysis infers that increasing local patterns enhances performance only in the presence of sufficient global context, confirming the synergistic relationship between global context and local patterns.

3.4 Evaluation on Multivariate Time Series

Although the main experiments focus on univariate time series, CoLaX is formulated for multivariate inputs $\boldsymbol{X} \in \mathbb{R}^{T \times D}$. We therefore construct a controlled two-variable benchmark by combining MCC-E and MTC-H, where four classes are defined by different cross-variable motif combinations.

We extend CoLaX to **CoLaX-M**, which learns variable-specific local prototypes through parallel branches while conditioning them on the global multivariate context. Variable-specific temporal segments are treated as candidate explanation units and jointly attributed using segment-level Shapley values on the frozen black-box classifier. For comparison, we similarly extend ShapeX to **ShapeX-M** under the same multivariate attribution protocol, together with TimeX and TimeX++.

As shown in Table 4, CoLaX-M achieves the highest AUPRC and AUP, with scores of 0.535 and 0.663, respectively, providing initial evidence that context-dependent modulation extends effectively to multivariate attribution.

Table 4: Saliency evaluation on the controlled multivariate benchmark. Best results are shown in bold and second-best results are underlined.

Method	AUPRC ↑	AUP ↑	AUR ↑
TimeX	0.231 ± 0.002	0.210 ± 0.004	$\underline{0.531 \pm 0.004}$
TimeX++	0.299 ± 0.001	0.267 ± 0.002	$\mathbf{0.613 \pm 0.006}$
ShapeX-M	$\underline{0.488 \pm 0.004}$	$\underline{0.581 \pm 0.008}$	0.411 ± 0.007
CoLaX-M	$\mathbf{0.535 \pm 0.005}$	$\mathbf{0.663 \pm 0.004}$	0.408 ± 0.006

3.5 Showcase of Explanations

To assess whether the models have learned genuine electrophysiological features, we visualize the internal surrogate representations of CoLaX and ShapeX on the ECG5000 dataset in Figure 5. The visualization reveals a fundamental divergence in the underlying representation logic.

CoLaX (Top) demonstrates superior representation fidelity by consistently attending to clinically relevant physiological features, such as the T-wave and QRS complex, yielding high surrogate confidence. In contrast, ShapeX (Bottom) exhibits shortcut learning by fixating on tail artifacts. This reliance on boundary noise leads to a critical failure in the PVC case, where the arrhythmia is categorized as Normal because the model ignores the actual pathological deflections. Such a discrepancy underscores the critical importance of learning global morphological contexts rather than relying on discriminative local fragments. Details are provided in Appendix C.6.

4 Conclusion

This paper addressed a limitation of current time series explanations: local patterns are scored independently of the sequence in which they occur, although the same motif may contribute differently under different global temporal conditions. We introduced CoLaX, a post-hoc explanation framework that builds on shapelet-driven Shapley attribution and conditions local-pattern relevance on a learned representation of the full sequence, at both the sequence and the temporal level. Global context thus determines which local occurrences are evaluated as explanatory units, whereas their importance is computed solely from the frozen black-box classifier.

While CoLaX addresses this local-global ambiguity, it also presents certain limitations. As a post-hoc surrogate framework, it provides correlational rather than strictly causal explanations. In the future, we will focus on deepening the causal investigation to disentangle how global context drives local pattern importance, while simultaneously extending the framework to support decisions in practical applications such as meteorological analysis and health monitoring.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62293502, 62233005, and U2441245), Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX026D01), and the European Union's Horizon Europe research and innovation programme (Grant Agreement Nos. 101130544, ThinkingEarth, and 101120237, ELIAS). The computations in this research were performed using the CFFF platform of Fudan University. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Pierre-Daniel Arsenault, Shengrui Wang, and Jean-Marc Patenaude. A survey of explainable artificial intelligence (XAI) in financial time series forecasting. *ACM Computing Surveys*, 57(10):1–37, 2025.

[2] Sohee Cho, Wonjoon Chang, Ginkyeng Lee, and Jaesik Choi. Interpreting internal activation patterns in deep temporal neural networks by finding prototypes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 158–166, 2021.

[3] Yu-Neng Chuang, Guanchu Wang, Fan Yang, Quan Zhou, Pushkar Tripathi, Xuanting Cai, and Xia Hu. CoRTX: Contrastive framework for real-time explanation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[4] Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *International Conference on Machine Learning (ICML)*, pages 2166–2177. PMLR, 2021.

[5] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series classification archive, 2018. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/`.

[6] Jesper Sören Dramsch, Monique M Kuglitsch, Miguel-Ángel Fernández-Torres, Andrea Toreti, Rustem Arif Albayrak, Lorenzo Nava, Saman Ghaffarian, Ximeng Cheng, Jackie Ma, Wojciech Samek, et al. Explainability can foster trust in artificial intelligence in geoscience. *Nature Geoscience*, 18(2):112–114, 2025.

[7] Joseph Early, Gavin Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[8] Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. SOM-VAE: interpretable discrete representation learning on time series. In *7th International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=rygjcsR9Y7`.

[9] Thibaut Germain, Charles Truong, and Laurent Oudre. Persistence-based motif discovery in time series. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6814–6827, 2024.

[10] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[11] Luis Gómez-Chova, Julia Amorós-López, Gonzalo Mateo-García, Jordi Muñoz-Marí, and Gustau Camps-Valls. Cloud masking and removal in remote sensing image time series. *Journal of Applied Remote Sensing*, 11(1):015005–015005, 2017.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Bosong Huang, Ming Jin, Yuxuan Liang, Johan Barthelemy, Debo Cheng, Qingsong Wen, Chenghao Liu, and Shirui Pan. ShapeX: Shapelet-driven post hoc explanations for time series classification models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[14] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *Advances in Neural Information Processing Systems*, volume 33, pages 6441–6452, 2020.

[15] Hyeongwon Jang, Changhun Kim, and Eunho Yang. TIMING: Temporality-aware integrated gradients for time series explanation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 26877–26895. PMLR, 13–19 Jul 2025.

[16] Xuan-May Le, Ling Luo, Uwe Aickelin, and Minh-Tuan Tran. Shapeformer: Shapelet transformer for multivariate time series classification. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1484–1494, 2024.

[17] Kin Kwan Leung, Clayton Rooke, Jonathan Smith, Saba Zuberi, and Maksims Volkovs. Temporal dependencies in feature importance for time series prediction. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[18] Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. TimeX++: Learning time-series explanations with information bottleneck. In *International Conference on Machine Learning (ICLR)*, 2024.

[19] Fanyu Meng, Ziwen Kan, Shahbaz Rezaei, Zhaodan Kong, Xin Chen, and Xin Liu. Implet: A post-hoc subsequence explainer for time series models. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 2217–2226. IEEE, 2025.

[20] Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I. Webb, Germain Forestier, and Mahsa Salehi. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9):217:1–217:45, 2024. doi: 10.1145/3649448.

[21] George B. Moody and Roger G. Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[22] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[23] Oscar J Pellicer-Valero, Miguel-Ángel Fernández-Torres, Chaonan Ji, Miguel D Mahecha, and Gustau Camps-Valls. Explainable earth surface forecasting under extreme events. *Earth's Future*, 13(9): e2024EF005446, 2025.

[24] Chetanya Puri, Gerben Kooijman, Bart Vanrumste, and Stijn Luca. Forecasting time series in healthcare with Gaussian processes and dynamic time warping based subset selection. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6126–6137, 2022.

[25] Eric Qu, Yansen Wang, Xufang Luo, Wenqiang He, Kan Ren, and Dongsheng Li. CNN kernels can be the best shapelets. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[26] Owen Queen, Tom Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency. *Advances in Neural Information Processing Systems*, 36:32129–32159, 2023.

[27] Michele Ronco and Gustau Camps-Valls. Role of locality, fidelity and symmetry regularization in learning explainable representations. *Neurocomputing*, 562:126884, 2023.

[28] Torty Sivill and Peter Flach. Limesegment: Meaningful, realistic time series explanations. In *International Conference on Artificial Intelligence and Statistics*, pages 3418–3433. PMLR, 2022.

[29] Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: Memory-guided transformer for multivariate time series anomaly detection. *Advances in Neural Information Processing Systems*, 36: 57947–57963, 2023.

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, 2017.

[31] Andreas Theissler, Francesco Spinnato, Udo Schlegel, and Riccardo Guidotti. Explainable AI for time series classification: A review, taxonomy and research directions. *IEEE Access*, 2022.

[32] Viet-Hung Tran, Zichi Zhang, Tuan Dung Pham, Ngoc Phu Doan, Anh-Tuan Hoang, Peixin Li, Hans Vandierendonck, Ira Assent, and Son T Mai. Intedisux: intepretation-guided discriminative user-centric explanation for time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20921–20928, 2025.

[33] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3):250–260, 2023. doi: 10.1038/s42256-023-00620-w.

[34] Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, and Jianxin Liao. Rethinking the power of timestamps for robust time series forecasting: A global-local fusion perspective. *Advances in Neural Information Processing Systems*, 37:22206–22232, 2024.

[35] Yi Yang, Qilin Deng, Furao Shen, Jinxi Zhao, and Chaomin Luo. A shapelet learning method for time series classification. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 423–430. IEEE, 2016.

[36] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 27268–27286. PMLR, 2022.

A Related Work

The field of time series explainability has attracted significant attention for its role in interpreting the decision-making processes of black-box models applied to temporal data [31]. The inherent opacity of pure black-box models obscures their internal mechanisms, rendering decision logic difficult for users to comprehend.

Explainability is defined as the ability to elucidate a model's internal functioning and provide meaningful insights into how predictions are made [1]. Explainable Artificial Intelligence (XAI) addresses this challenge through two primary categories: intrinsic transparency and post-hoc approaches. The former utilizes simple structures or inherently interpretable formulations, achieving transparency by design. Conversely, post-hoc approaches are applied after model training to interpret what a model has learned, without altering its underlying predictive mechanism [33]. Within this class of methods, gradient-based techniques assess input sensitivity but often fail to yield temporally coherent or stable explanations in time series settings [30, 8]. Perturbation-based methods offer an alternative by estimating feature importance through the measurement of prediction changes under systematic masking strategies [26, 18]. Extending this paradigm, Shapley-based approaches approximate cooperative game-theoretic attributions to provide both local and global explanations for complex temporal models [13]. As time series models grow in scale and complexity, research has increasingly shifted from point-wise attributions to structured explanations that operate at the level of subsequences or dynamical patterns [35, 2, 13]. However, many existing approaches do not explicitly model how sequence-level temporal context conditions the selection and attribution of local explanatory patterns.

This perspective motivates the development of frameworks that transcend isolated saliency scores to capture the interplay between local transient patterns and global system evolution.

B Experimental Details

B.1 Dataset

We evaluate our CoLaX on both synthetic and real-world datasets.

Synthetic Datasets As noted in [13], class-discriminative features in real-world applications are often defined by specific temporal patterns rather than simple amplitude differences. Consequently, we utilize the four motif-based binary classification datasets proposed in [13]: MCC-E, MTC-H, MCC-H, and MTC-E. The statistical details are provided in Table 6. These datasets align more closely with real-world data-generation mechanisms, enabling us to more accurately assess whether an explanation method can effectively identify the time-series patterns crucial for classification.

Table 5: Overview of Synthetic Datasets

Category	Dataset Name	Classes	Testing Samples	Training Samples	Length of Series	Description	True Saliency Score
Synthesis	MCC-E	2	2,000	10,000	800	Shapelet count for classification, equal amplitude	Yes
	MCC-H	2	2,000	10,000	800	Shapelet count for classification, higher amplitude	Yes
	MTC-E	2	2,000	10,000	800	Shapelet type for classification, equal amplitude	Yes
	MTC-H	2	2,000	10,000	800	Shapelet type for classification, higher amplitude	Yes

Real-world dataset. We evaluate CoLaX on the ECG dataset following the experimental protocol of ShapeX [13]. This benchmark is used for binary atrial fibrillation (AF) classification and consists of 55,507 training samples and 37,004 test samples, each containing 360 time steps. Following ShapeX, we use the same ECG data and reference saliency annotations for explanation evaluation. These reference regions serve as domain-informed proxies rather than direct human explanations of the black-box predictions. We additionally evaluate CoLaX on selected datasets from the UCR Time Series Classification Archive.

B.2 Data Assets and Licenses

All datasets used in this study are publicly accessible research benchmarks. The four synthetic motif-based datasets (MCC-E, MCC-H, MTC-E, and MTC-H) follow the data generation protocol of ShapeX [13], and are used with appropriate attribution to the original authors. The MIT-BIH

Table 6: Overview of Real-world Datasets

Category	Dataset Name	Classes	Testing Samples	Training Samples	Length of Series	Description	True Saliency Score
Real-world Dataset	ECG	2	37004	55507	360	Biomedical ECG signals with expert annotation	No
	ACSF1	10	100	100	1460	Device	No
	ARROWHEAD	3	175	36	251	Image	No
	BME	3	150	30	128	Simulated	No
	CAR	4	60	60	577	Sensor	No
	CBF	3	900	30	128	Simulated	No
	ChlorineConcentration	3	3840	467	166	Simulated	No
	CinCECGTorso	4	1380	40	1639	ECG	No
	Coffee	2	28	28	286	Spectro	No
	Computers	2	250	250	720	Device	No
	CricketX	12	390	390	300	Har	No
	CricketY	12	390	390	300	Har	No
	CricketZ	12	390	390	300	Har	No
	Crop	24	16800	7200	46	Image	No
	DiatomSizeReduction	4	306	16	345	Image	No
	DistalPhalanxOutlineAgeGroup	3	139	400	80	Image	No
	DistalPhalanxOutlineCorrect	2	276	600	80	Image	No
	ECG200	2	100	100	96	ECG	No
	ECG5000	5	4500	500	140	ECG	No
	FordA	2	1320	3601	500	Sensor	No
	FreezerRegularTrain	2	2850	150	301	Device	No
	GunPoint	2	150	50	150	Har	No
	ItalyPowerDemand	2	1029	67	24	Sensor	No
	Lightning7	7	73	70	319	Sensor	No
	Mallat	8	2345	55	1024	Simulated	No
	MedicalImages	10	760	381	99	Image	No
	ShapesAll	60	600	600	512	Image	No
	SmoothSubspace	3	150	150	15	Simulated	No
	SonyAIBORobotSurface1	2	601	20	70	Sensor	No
	Strawberry	2	370	613	235	Spectro	No
	Symbols	6	995	25	398	Image	No
	SyntheticControl	6	300	300	60	Simulated	No
	TwoPatterns	4	4000	1000	128	Simulated	No
	UMD	3	144	36	150	Simulated	No
	UWaveGestureLibraryAll	8	896	3582	945	Har	No
	WAFER	2	6164	1000	152	Sensor	No
	WORMS	5	77	181	900	Motion	No

Arrhythmia Database is available through PhysioNet under the Open Data Commons Attribution License (ODC-By 1.0). The UCR Time Series Classification Archive is publicly accessible, with usage and redistribution conditions varying across individual datasets. We acknowledge the original sources and follow their applicable terms of use.

B.3 Benchmarks

We evaluate explanation quality using a diverse set of benchmark methods developed for interpreting deep learning models on time series data, covering gradient-based, perturbation-based, surrogate-model, and inherently interpretable approaches.

Integrated Gradients (IG) [30]. Integrated Gradients is a classical gradient-based attribution method that explains model predictions by integrating gradients along a straight path from a baseline input to the original input. By accumulating the sensitivity of the output with respect to each input dimension, IG assigns attribution scores that satisfy desirable axiomatic properties such as completeness and sensitivity.

Dynamask [4]. Dynamask is a perturbation-based explainer specifically designed for time series data. It learns a continuous mask that selectively occludes parts of the input while minimizing distortion to the model's prediction. The mask gradually deforms the input toward a carefully chosen baseline, yielding temporally coherent saliency patterns.

WinIT [17]. WinIT is a feature-removal-based explanation method that evaluates the importance of temporal segments by measuring the change in prediction induced by their removal. By eliminating specific time intervals and modeling dependencies across time steps, WinIT captures both local relevance and temporal interactions.

CoRTX [3]. CoRTX adopts an augmentation-based contrastive learning strategy to efficiently approximate Shapley values. By leveraging a small number of contrastive perturbations, it estimates feature attributions with reduced computational overhead while retaining the theoretical grounding of SHAP.

MILLET [7]. MILLET provides inherent interpretability for time series classifiers through a Multiple Instance Learning (MIL) formulation. It produces sparse and instance-level explanations directly from the model architecture, achieving interpretability without relying on post-hoc perturbations or sacrificing predictive performance.

TIMEX [26]. TIMEX generates explanations by training an interpretable surrogate model to mimic a pretrained black-box classifier. By preserving latent space relationships between inputs and predictions, TIMEX delivers structured and faithful explanations that reflect the behavior of the original model.

TIMEX++ [18]. TIMEX++ extends TIMEX by incorporating an information bottleneck principle. It produces label-preserving, in-distribution explanations through constrained input perturbations, explicitly mitigating trivial solutions and distributional shift issues commonly encountered in perturbation-based methods.

ShapeX [13]. ShapeX explains time series classifiers by segmenting inputs into semantically meaningful, shapelet-driven subsequences and assigning saliency scores via Shapley value attribution. This segment-level formulation enables explanations that align with the model's underlying decision logic and capture higher-level temporal structures.

Metrics Following [26], we adopt AUP (Area Under the Precision Curve), AUR (Area Under Recall Curve), and AUPRC (Area Under the Precision–Recall Curve) to quantitatively evaluate saliency maps in experiments where ground-truth explanations are available. These metrics assess the discrepancy between the generated saliency scores and the true salient regions. In all cases, metric values closer to 1 indicate higher agreement with the ground truth and thus higher explanation quality.

In occlusion-based experiments, where ground-truth saliency annotations are unavailable, we evaluate explanation faithfulness using the AUROC (Area Under the Receiver Operating Characteristic Curve) of the underlying black-box classifier. Specifically, AUROC measures the degradation in classification performance as progressively least salient (bottom-k) subsequences identified by the explanation method are removed. In this setting, a faithful explainer should result in a slower and smoother performance drop, as it successfully prioritizes the removal of non-informative background regions while preserving key motifs.

Let $x \in \mathbb{R}^{T \times D}$ denote an input time series, where T is the temporal length and D the feature dimension. Let Q be the ground-truth saliency matrix, where $Q_t = 1$ indicates that time step t is truly salient, and 0 otherwise. Let $R \in [0,1]^T$ denote the saliency map produced by the explanation method. Given a detection threshold $\tau \in (0,1)$, the binary saliency estimator $\hat{Q}_t(\tau)$ is defined as:

$$\hat{Q}_t(\tau) = \begin{cases} 1 & \text{if} \quad R_t \geq \tau \\ 0 & \text{else.} \end{cases} \tag{19}$$

We define the set of truly salient indices and the set predicted as salient by the explanation method as:

$$A = \{t \in [1:T] \mid Q_t = 1\}$$
$$\hat{A}(\tau) = \{t \in [1:T] \mid \hat{Q}_t(\tau) = 1\}.$$

For each threshold τ, the precision and recall are defined as:

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

The Area Under the Precision Curve (AUP) and Area Under Recall Curve (AUR) are computed as:

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

To jointly assess the trade-off between precision and recall, we additionally compute the Area Under the Precision–Recall Curve (AUPRC):

$$\text{AUPRC} = \int_0^1 \mathbf{P}(\mathbf{R}) d\mathbf{R}, \tag{20}$$

where precision is treated as a function of recall as the threshold τ varies.

Table 7: Training parameters across different datasets.

Parameter	MCCE	MCCH	MTCE	MTCH	ECG
Length of pattern	100	100	100	100	40
Number of local pattern	2	2	2	2	2
Number of global context	2	2	2	2	2
Number of CTEncoder	6	6	6	1	1
Number of GEncoder	1	6	1	1	1

Table 8: Comparison of wall-clock time (in seconds) on different datasets.

Dataset \ Method	Ours	ShapeX	DYNAMASK	WINIT	TimeX++
MCC-E	68.1	58.3	385.2	43.4	20.9
MTC-E	72.1	63.1	383.8	30.7	27.2
MCC-H	70.3	58.6	384.3	39.7	23.0
MTC-H	55.4	51.4	383.1	39.7	21.2
ECG5000	35.3	33.1	181.8	8.7	17.6

B.3.1 Experimental Settings

All experiments were conducted on an NVIDIA A100 GPU with all models (our framework and baselines) implemented in PyTorch. For each dataset, we use the standard train/test split from the corresponding benchmark and reserve 20% of the training set for validation to tune hyperparameters and perform early stopping. Black-box classifiers are trained with cross-entropy loss until convergence, using validation accuracy for early stopping. Unless otherwise specified, we use each baseline's default hyperparameters; the main settings for all datasets are summarized in Table 7 and kept fixed to ensure fair comparison. Results are averaged over five random seeds (controlling initialization and data shuffling), and we report the mean and standard deviation (shown as error bars where applicable).

C Further Analysis Experiments

C.1 Computational Cost

To assess the computational cost of each baseline and our CoLaX, we compare the inference time required for explanation generation across several datasets, as shown in Table 8. As observed, our method maintains a competitive computational profile. It is significantly more efficient than DYNAMASK and demonstrates efficiency comparable to ShapeX. While it is well known that computing Shapley values is highly time-consuming, and CoLaX may not be the fastest method among all baselines, it strikes a superior balance between efficiency and performance. It keeps the computational overhead within an acceptable range while achieving significantly higher accuracy.

C.2 Parameter Analysis

We investigate model capacity by varying the counts of local patterns and global context, as shown in Figure 6, 7, and 8. Results reveal a non-additive interaction: increasing local patterns improves performance only when there is sufficient global context, confirming that global context modulates local pattern interpretation.

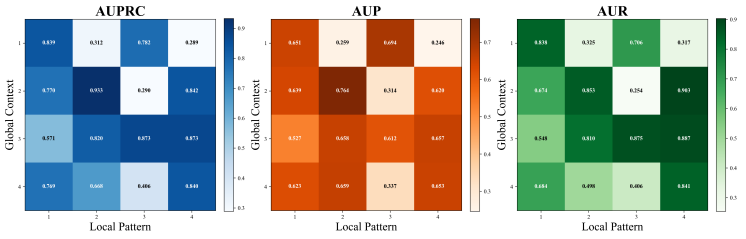


Figure 6: Metric under different local–global interaction configurations on MCC-E datasets.

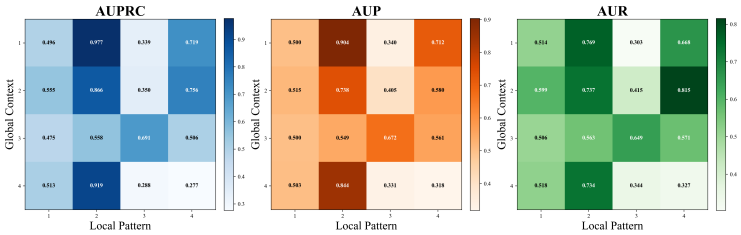


Figure 7: Metric under different local–global interaction configurations on MCC-H datasets.

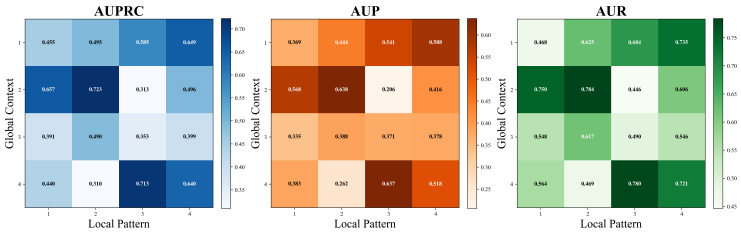


Figure 8: Metric under different local–global interaction configurations on MTC-E datasets.

Figure 9 demonstrates that intermediate local pattern lengths also yield optimal performance. Crucially, this optimum depends on the global context, necessitating their joint optimization rather than isolated tuning.

Regarding the loss function hyperparameters, Figure 10 exhibits distinct peaks at our selected values ($\lambda_1 = 0.5, \lambda_2 = 10^{-3}$). Specifically, MCC-E is highly sensitive to λ_1, while the MTC datasets display an inverted U-shaped trend for λ_2, suggesting an effective balance between global alignment and regularization.

Overall, these results indicate that the numbers of local patterns and global contexts, and the pattern length, should be tuned jointly rather than in isolation.

C.3 Robustness Evaluation

We evaluate the local stability of CoLaX explanations under three types of input perturbations: Gaussian noise ($\sigma \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$), temporal shifts ($s \in \{1, 2, 3, 5\}$), and amplitude scaling ($a \in \{0.8, 0.9, 1.1, 1.2\}$). Only perturbations preserving the frozen black-box classifier's predicted label are retained. Explanation consistency is measured using Spearman rank correlation (ρ) and top-k intersection-over-union (IoU), with higher values indicating greater stability.

As shown in Table 9, CoLaX exhibits high explanation consistency across all four datasets. Gaussian noise and amplitude scaling yield average Spearman correlations of 0.9885 and 0.9853, respectively, indicating stable saliency rankings. Temporal shifts produce relatively lower overlap (IoU = 0.7225), potentially reflecting changes in the positions of salient patterns. Nevertheless, the average rank correlation remains high (0.9607), suggesting that CoLaX maintains substantial explanation consistency under the evaluated prediction-preserving perturbations.

C.4 Effect of Context-dependent Prototype Weighting

To examine whether the benefits of CoLaX arise from context-dependent weighting rather than merely introducing learnable prototype weights, we introduce a *Static Learnable Weights* variant. This variant removes the global encoder, context bank, global alignment loss, and context-to-prototype interaction, replacing them with an input-independent learnable weight vector $\boldsymbol{w}_{\text{static}} \in \mathbb{R}^{K_{\text{local}}}$. The local prototype learning, segment extraction protocol, perturbation operator, and black-box Shapley estimator are retained. However, the context-dependent modulation is replaced by static prototype weighting.

As shown in Table 10, CoLaX outperforms the static-weight variant on 11 of 12 metrics, improving average AUPRC from 0.637 to 0.888, AUP from 0.577 to 0.748, and AUR from 0.672 to 0.846.

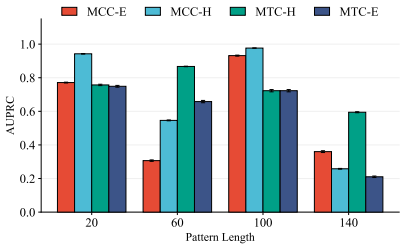


Figure 9: AUPRC under different pattern length configurations on synthetic datasets.

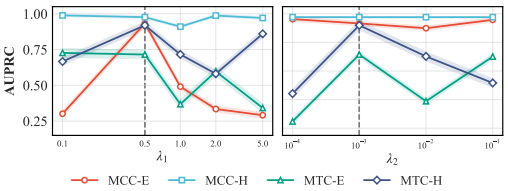


Figure 10: Analysis on the selection of λ_1 and λ_2.

Table 9: Robustness of CoLaX explanations under prediction-preserving perturbations. Results are averaged across perturbation strengths. Higher is better.

Dataset	Gaussian Noise		Temporal Shift		Amplitude Scaling	
	$\rho \uparrow$	IoU $\uparrow$	$\rho \uparrow$	IoU $\uparrow$	$\rho \uparrow$	IoU $\uparrow$
MCC-E	0.9839	0.8618	0.9436	0.6635	0.9827	0.8402
MCC-H	0.9935	0.9056	0.9691	0.7797	0.9849	0.8093
MTC-E	0.9824	0.8461	0.9468	0.6493	0.9805	0.8226
MTC-H	0.9943	0.9085	0.9831	0.7976	0.9929	0.8948
Average	**0.9885**	**0.8805**	**0.9607**	**0.7225**	**0.9853**	**0.8417**

Unlike static weights, which assign identical prototype relevance across samples, CoLaX adapts prototype weights to the global temporal context of each input. These results highlight the importance of global context conditioning for constructing more informative explanatory segments, beyond simply introducing learnable prototype weights.

C.5 More experiments on Real-world dataset

The UCR Archive is widely adopted for comparison given their inherent diversity, covering various domains, sequence lengths, and class cardinalities. The results for occlusion-based evaluation are illustrated in Figure 11 and 12.

It can be inferred that no single explanation model dominates across all datasets. Our CoLaX appears to degrade more gracefully on challenging cases, as reflected in the smoother drop of AUROC scores across datasets, illustrating its robustness.

C.6 Interpretable comparisons

To evaluate whether the models capture genuine electrophysiological features during their initial training phase, we visualize the internal surrogate activations of CoLaX versus ShapeX on the ECG5000 dataset (Figure 13). This comparison reveals a fundamental divergence in the internal representation logic learned by the two frameworks.

The ECG5000 benchmark, derived from the BIDMC Congestive Heart Failure Database (CHFDB) [10], comprises 5,000 heartbeat segments. Originally sampled at 250 Hz, each segment was length-normalized to 140 time steps via interpolation; thus, the temporal axis reflects the relative cardiac phase rather than absolute duration. The dataset includes five heartbeat classes: Normal (N), R-on-T PVC, PVC, SP/EB, and Unclassified (UB). We report the surrogate performance of both models across these classes to validate the quality of their learned prototype dictionaries.

It can be clearly inferred from Figure 13 that CoLaX demonstrates dynamic adaptability, with activation regions (blue) shifting distinctively according to the morphological characteristics of each class. Conversely, ShapeX exhibits a static bias, with its activations (green) consistently fixated on the tail of the time series (Time Step > 130), regardless of the input pathology. Specifically, for Normal and R-on-T classes, CoLaX concentrates on the T-wave repolarization phase (Time Step $\sim$90–110), whereas for PVCs, it shifts attention to the onset QRS complex. However, ShapeX exhibits a static bias, focusing more on the tail of the signal. As previously discussed, the dataset underwent interpolation and length normalization, enforcing a fixed resolution of 140 time steps for all heartbeats. A critical artifact of this process lies at the boundaries of the time series, particularly the tail. These

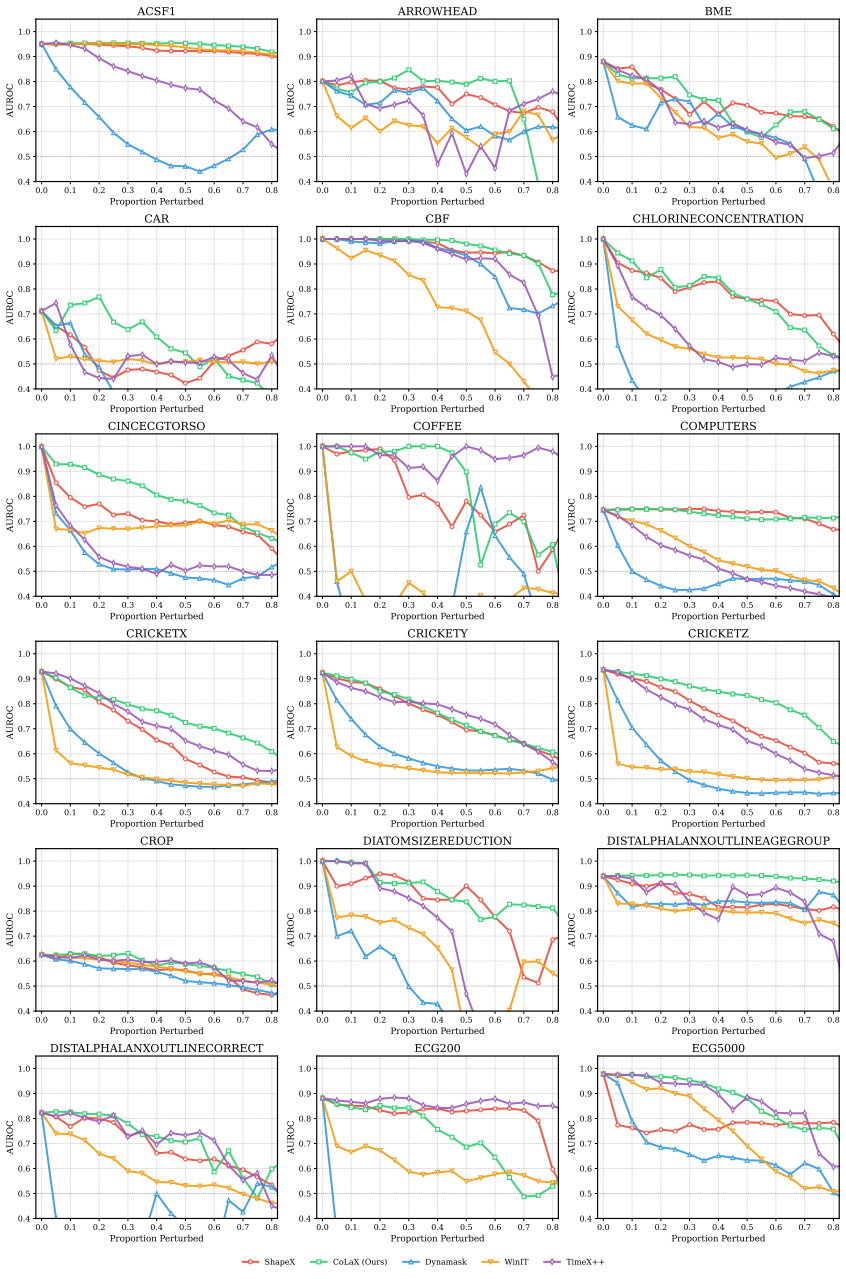


Figure 11: The occlusion experimental results on UCR archive (1/2).

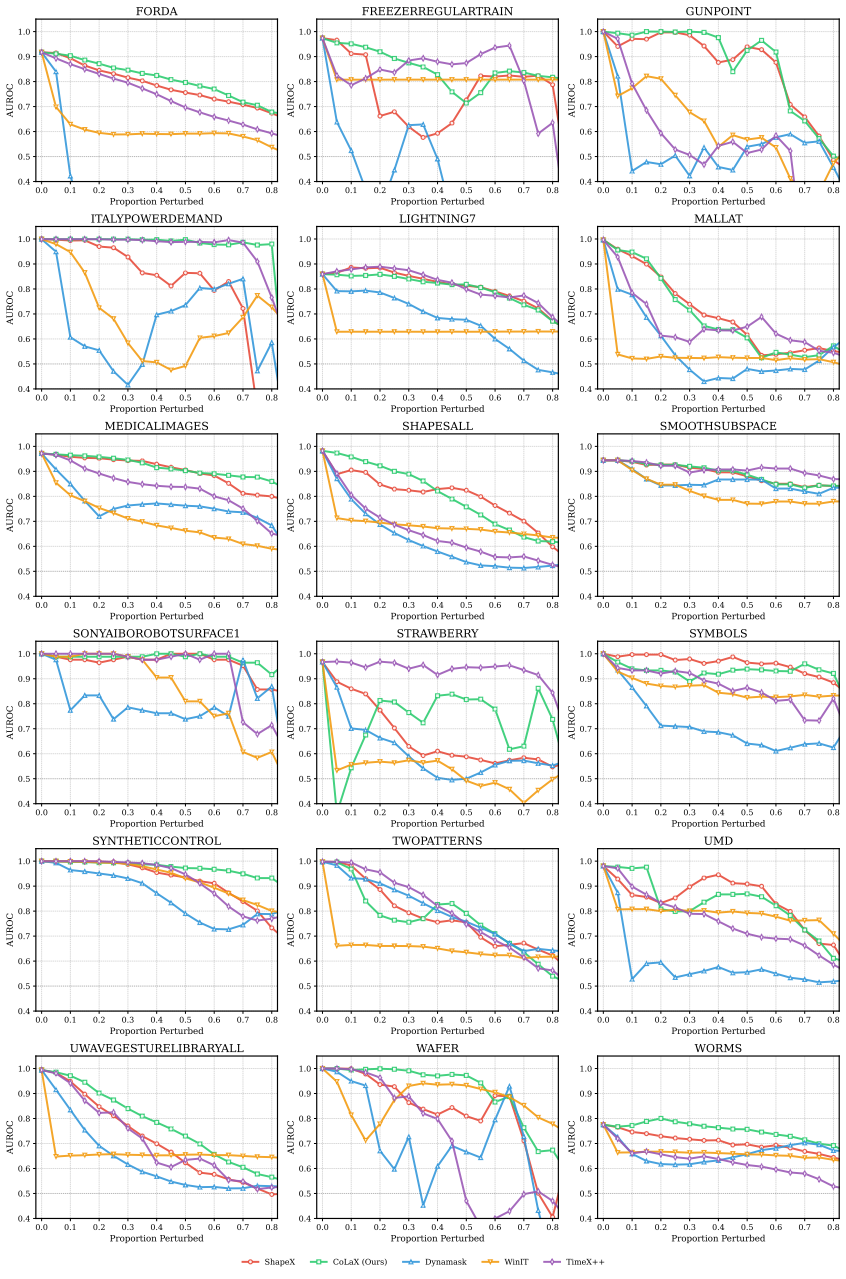


Figure 12: The occlusion experimental results on UCR archive (2/2).

Table 10: Comparison between CoLaX and static learnable prototype weights. Best results are highlighted in bold.

Dataset	Method	AUPRC ↑	AUP ↑	AUR ↑
MCC-E	CoLaX	**0.933±0.003**	**0.764±0.004**	**0.853±0.003**
	Static Weights	0.284±0.003	0.293±0.004	0.268±0.004
MCC-H	CoLaX	**0.977±0.000**	**0.904±0.002**	**0.769±0.005**
	Static Weights	0.785±0.002	0.798±0.003	0.671±0.005
MTC-E	CoLaX	**0.723±0.007**	**0.638±0.007**	**0.784±0.006**
	Static Weights	0.660±0.006	0.595±0.005	0.762±0.005
MTC-H	CoLaX	**0.920±0.001**	**0.686±0.001**	0.977±0.000
	Static Weights	0.819±0.004	0.621±0.003	**0.986±0.001**

Table 11: Additional comparisons with Implet and TIMING on four synthetic datasets. The best results are highlighted in bold.

Dataset	Method	AUPRC	AUP	AUR
MCC-E	CoLaX	$\mathbf{0.933 \pm 0.003}$	$\mathbf{0.764 \pm 0.004}$	$\mathbf{0.853 \pm 0.003}$
	Implet	0.382 ± 0.005	0.522 ± 0.003	0.501 ± 0.003
	TIMING	0.364 ± 0.005	0.472 ± 0.008	0.469 ± 0.005
MCC-H	CoLaX	$\mathbf{0.977 \pm 0.000}$	$\mathbf{0.904 \pm 0.002}$	$\mathbf{0.769 \pm 0.005}$
	Implet	0.525 ± 0.006	0.488 ± 0.004	0.556 ± 0.004
	TIMING	0.446 ± 0.007	0.582 ± 0.009	0.433 ± 0.008
MTC-E	CoLaX	$\mathbf{0.723 \pm 0.007}$	$\mathbf{0.638 \pm 0.007}$	$\mathbf{0.784 \pm 0.006}$
	Implet	0.174 ± 0.002	0.180 ± 0.001	0.508 ± 0.001
	TIMING	0.188 ± 0.002	0.449 ± 0.007	0.412 ± 0.005
MTC-H	CoLaX	$\mathbf{0.920 \pm 0.001}$	0.686 ± 0.001	$\mathbf{0.977 \pm 0.000}$
	Implet	0.431 ± 0.006	0.523 ± 0.001	0.458 ± 0.005
	TIMING	0.402 ± 0.005	$\mathbf{0.799 \pm 0.007}$	0.206 ± 0.005

regions often accumulate significant non-biological information, such as edge-smoothing artifacts introduced by interpolation algorithms or artificial zero-padding effects resulting from segmentation. These artifacts act as spurious features. The fact that CoLaX avoids these regions suggests that its representation learning is more robust and physiologically grounded.

In particular, with proper accuracy and confidence, CoLaX focuses more on the characteristic of the given signal. This difference is particularly evident in the Class 2 sample. CoLaX's internal surrogate head captures physiologically consistent features, whereas ShapeX's representation may reflect boundary-related artifacts. Consequently, ShapeX's surrogate prediction fails to align with the actual class, categorizing the severe arrhythmia as 'Normal' with low surrogate confidence (0.367). We emphasize that this does not directly measure post-hoc faithfulness to a black-box; rather, it demonstrates that CoLaX learns a higher-fidelity dictionary.

It is worth noting that both surrogates struggle with Class 4 identification. Unlike ventricular arrhythmias (e.g., Class 2), which exhibit distinct broad QRS complexes, Class 4 waveforms often present with subtle morphological deviations that significantly overlap with normal beats. Consequently, improving the sensitivity to Class 4 through fine-grained feature extraction and multi-scale attention mechanisms remains a priority for our future research.

C.7 Additional Comparisons with Subsequence-Based Explainers

To further assess the fairness of baseline comparisons, we include two additional methods: Implet [19], a post-hoc subsequence-based explainer, and TIMING [15], a temporality-aware Integrated Gradients method. These methods complement the existing baselines by providing more structurally relevant comparisons beyond conventional point-wise attribution.

Table 11 reports the results on four synthetic benchmarks. CoLaX consistently outperforms Implet across all datasets and evaluation metrics. Compared with TIMING, CoLaX achieves higher AUPRC and AUR on all four datasets, while TIMING obtains a higher AUP on MTC-H. These results suggest that the improvements of CoLaX cannot be attributed solely to comparisons with conventional point-wise explanation methods.

Nevertheless, these synthetic datasets are deliberately constructed around discriminative temporal motifs and may favor methods that produce coherent subsequence-level explanations. We therefore

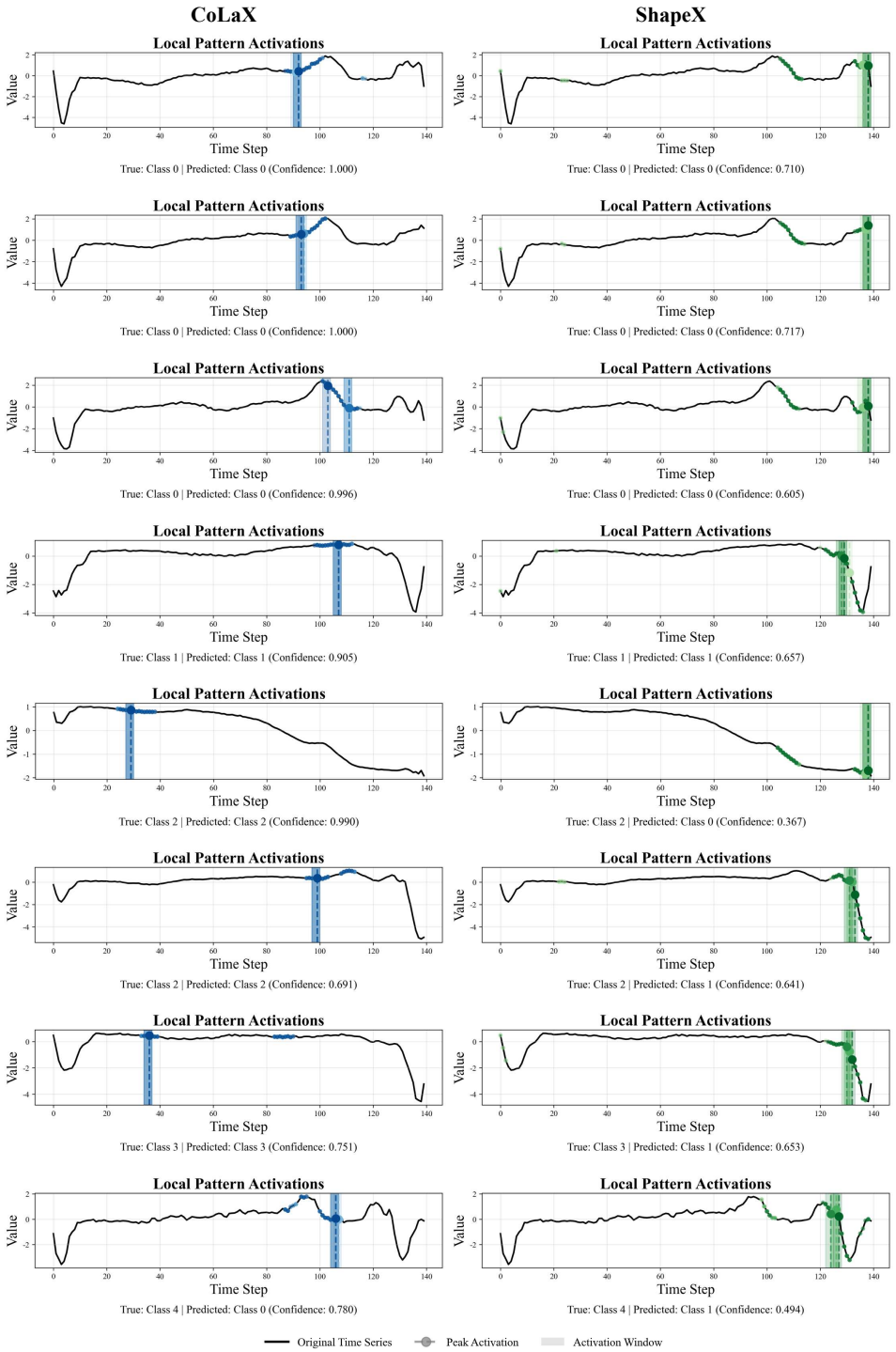
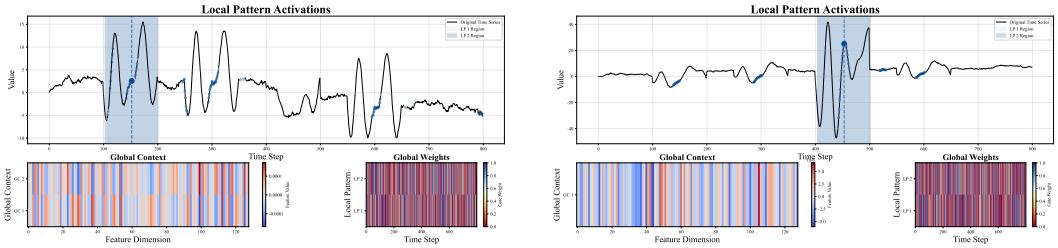


Figure 13: Comparison of Class-Discriminative Regions learned by CoLaX and ShapeX on the ECG5000 Dataset. Predictions and confidence scores refer to the outputs of the internal surrogate heads used to validate the representation fidelity of the learned prototype dictionary.



(a) MCCE (b) MCCH

Figure 14: Visualization of model explanations on synthetic datasets. Saliency maps highlight the localized discriminative segments identified by the model under different motif–context configurations.

place particular emphasis on comparisons with structurally related methods, especially ShapeX and Implet. The results support the effectiveness of CoLaX in motif-rich settings, particularly when the relevance of recurring local patterns depends on the broader temporal context.

C.8 Visualizations for local pattern and global context

Figure 14 provides qualitative insights into how local pattern activations, global context representations, and their context-dependent gating jointly contribute to the model's decision-making process.

The top panel displays the original time series together with multiple detected local pattern activation regions. These regions correspond to temporally localized fluctuations with distinct morphological characteristics. Notably, similar local patterns appear at different time intervals, including both pronounced transient events and more subtle variations. This observation highlights that the mere occurrence of a local pattern does not uniquely determine its predictive relevance.

The bottom-left panel visualizes the learned global context representations, which capture slowly varying, high-level temporal structures across feature dimensions. These global context embeddings evolve smoothly over time and are not temporally aligned with individual local activations, indicating that they encode background system conditions rather than instantaneous dynamics.

Crucially, the bottom-right panel illustrates the resulting global-local gating weights, which modulate the contribution of each local pattern over time. The sample-specific gating weights adjust the overall relevance of each local prototype, while the temporal context signal modulates its activation strength across time. Consequently, context-conditioned activations can vary across temporal locations even when the underlying prototype remains unchanged. In particular, strong local transients receive high gating weights only when aligned with compatible global context states, whereas similar local patterns occurring under different global conditions are down-weighted.

Overall, these visualizations provide intuitive evidence that our CoLaX does not rely on local patterns in isolation. Instead, it implements a context-aware weighting mechanism in which global temporal context conditions the contribution of local patterns to the final prediction. This qualitative behavior is consistent with the non-additive and interaction-driven trends observed in the quantitative parameter analysis and ablation studies, and supports the interpretability of the proposed global-to-local modulation design.

D The Saliency Score for CoLaX

D.1 Definition of Saliency Score

Given a trained classification model f that maps an input time series $\boldsymbol{X}$ to a predicted label $\hat{Y}$, CoLaX aims to identify the contribution of shapelet-aligned subsequences to the model's decision. Unlike timestep-level attribution, we focus on identifying meaningful segments driven by learned shapelets. A saliency map R is generated to quantify this importance, adopting the methodology proposed in ShapeX [13].

Let $\boldsymbol{X} = [x_1, x_2, \cdots, x_T]$ denote a time series of length T. The saliency map R is defined as a sequence of importance values, where each element R_t quantifies the contribution of x_t to the model's

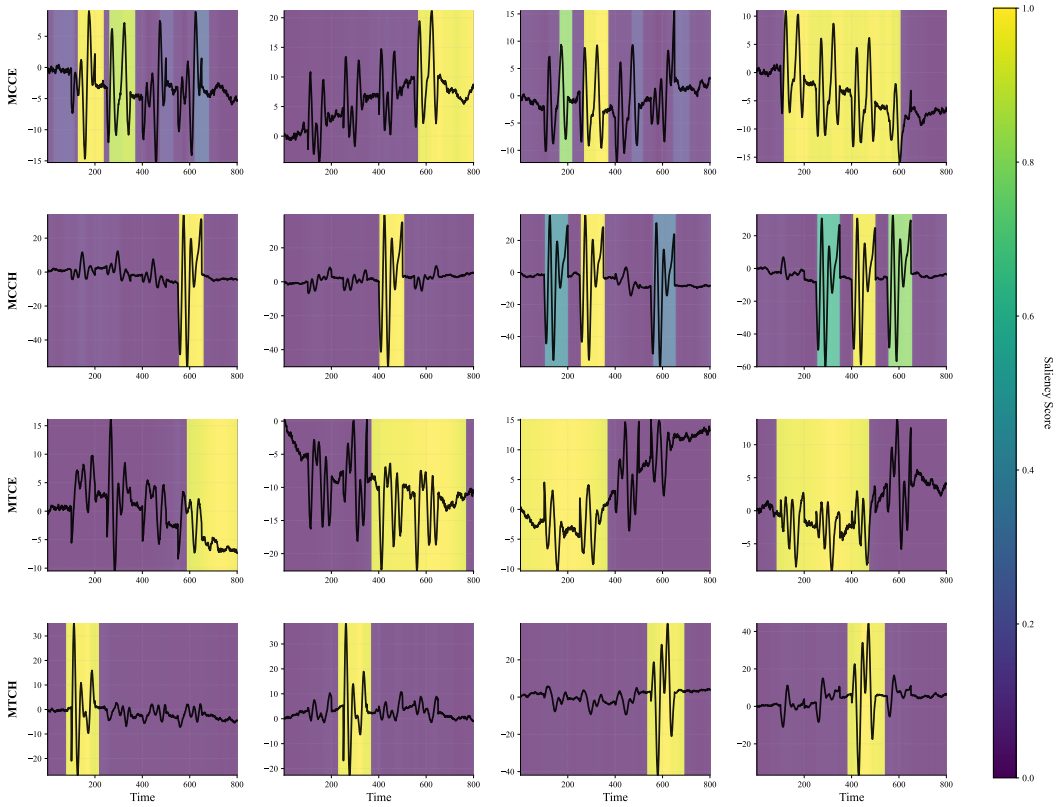


Figure 15: The visualization of saliency score in MCC-E, MCC-H, MTC-E, and MTC-H dataset.

prediction $f(\boldsymbol{X})$. Formally, Shapley values are employed to quantify the marginal contribution of each segment to the prediction $f(\boldsymbol{X})$. The final saliency score R_t for each time step corresponds to the importance of the shapelet-driven segment to which it belongs. The normalized saliency score R_t represents the magnitude of the aggregated segment-level attribution at time step t. Higher values indicate greater attribution magnitude, without distinguishing positive from negative contributions. For visualization and comparison purposes, saliency scores are normalized to the range $[0, 1]$.

D.2 Perturbation-based Explanation Methods

Adopting the perturbation-based paradigm, the saliency score is interpreted as the expected variation in the model's output under input perturbation. Let $P(\boldsymbol{X}, \boldsymbol{M})$ denote a perturbation function that modifies the input $\boldsymbol{X}$ based on a binary mask $\boldsymbol{M} \in \{0, 1\}^T$. The perturbed input $\boldsymbol{X}'$ is defined as:

$$\boldsymbol{X}' = \boldsymbol{M} \odot \boldsymbol{X} + (1 - \boldsymbol{M}) \odot \boldsymbol{X}_{\text{base}}, \tag{21}$$

where $\odot$ denotes the Hadamard product with $\boldsymbol{M}$ broadcasted across the D feature dimensions. Here, $\boldsymbol{X}_{\text{base}}$ represents a context-consistent baseline signal generated via linear interpolation to maintain temporal continuity and avoid artificial spikes at segment boundaries.

The temporal saliency map is obtained by aggregating the absolute segment-level Shapley values:

$$\boldsymbol{R} = \text{Aggregate}\left(\{\phi_n, \mathcal{I}_n\}_{n=1}^{N}\right), \tag{22}$$

where $\text{Aggregate}(\cdot)$ denotes the maximum absolute aggregation and normalization defined in Eq. 18.

D.3 Visualization of Saliency Maps

Figure 15 presents the saliency maps generated by CoLaX on the four synthetic datasets.

E Conceptual Motivation for Context-Aware Attribution

This section provides an intuitive motivation for context-aware attribution rather than a formal theoretical guarantee. From a conditional-attribution perspective, the relevance of a local temporal pattern may depend on the broader sequence-level context in which it occurs.

Conceptual Analogy to Conditional Attribution. Consider a time series characterized by local temporal patterns X_{local} and global context X_{global}. The marginal relationship $P(Y \mid X_{\text{local}})$ does not explicitly distinguish the different contexts in which a local pattern may appear. In contrast, $P(Y \mid X_{\text{local}}, X_{\text{global}})$ illustrates how the predictive relevance of the same local pattern may vary across global temporal conditions. This distinction motivates incorporating global context when constructing local explanations, without implying that these conditional distributions are explicitly estimated or that the resulting attributions have a causal interpretation.

Context-aware Representation and Attribution. Motivated by this perspective, CoLaX learns a global representation $\boldsymbol{E}_{\text{global}}$ to generate sample-specific prototype weights $\boldsymbol{w}$. During training, these weights modulate the learned prototype representations through $\widetilde{\boldsymbol{Q}} = \text{diag}(\boldsymbol{w})\widetilde{\boldsymbol{P}}$. During post-hoc inference, the same weights, together with the temporal context signal c_t, modulate local activations as $A^{\text{ctx}}_{t,k} = w_k A_{t,k}(1 + \alpha c_t)$. The resulting activations determine the candidate segments used for black-box Shapley attribution. Thus, global context influences which local patterns are selected as explanatory units, while their contributions are evaluated through the frozen classifier's prediction changes. This design provides an explicit mechanism for context-dependent attribution, rather than a theoretical guarantee of explanation faithfulness.

F Algorithm for CoLaX

The training and inference procedures of CoLaX are summarized in Algorithm 1:

G Limitations

While CoLaX addresses local–global ambiguity in time-series explanations, several limitations remain.

First, as a post-hoc surrogate-based framework, CoLaX inherits the limitations of perturbation-based explainability. Its explanations characterize the responses of a frozen black-box classifier to structured input perturbations rather than revealing causal mechanisms of the underlying data-generating process. Moreover, the resulting attributions remain sensitive to the perturbation strategy, as segment-level masking may generate out-of-distribution samples. Since CoLaX is trained independently of the black-box classifier, its learned prototypes may also fail to capture classifier-specific or spurious features, limiting explanation faithfulness.

Second, the empirical advantages of CoLaX are most pronounced on synthetic, motif-rich datasets designed to exhibit local–global ambiguity. Improvements on real-world datasets are generally less consistent. Although CoLaX achieves the highest AUPRC on ECG, it does not outperform ShapeX on AUP and AUR, and no single method consistently dominates across the heterogeneous UCR datasets. Thus, the current results support its effectiveness primarily in settings where recurring local patterns have context-dependent relevance, rather than establishing universal superiority.

Third, although CoLaX supports multivariate inputs and has been evaluated on several multivariate benchmarks, the experimental evaluation remains predominantly univariate. Its ability to capture complex cross-variable interactions in high-dimensional settings requires further investigation.

Fourth, the dual-branch architecture and representation-learning stage introduce additional computational overhead compared with simpler attribution methods. While computationally practical on the evaluated benchmarks, scalability to extremely long sequences and high-dimensional multivariate data remains an open challenge.

Finally, CoLaX remains a post-hoc explanation framework without formal guarantees of attribution faithfulness or stability. Extending it toward inherently interpretable architectures and establishing theoretical guarantees under non-stationary conditions are promising directions for future research.

Algorithm 1 Training and Inference of **CoLaX**

Require: Multivariate time series $\boldsymbol{X} \in \mathbb{R}^{T \times D}$, labels $\boldsymbol{y} \in \{0,1\}^C$, frozen black-box classifier $f(\cdot)$; local prototype length L_p, number of prototypes K_{local}, number of global contexts K_{global}, temperatures τ_1, τ_2, modulation strength α, activation threshold Ω, loss weights λ_1, λ_2, learning rate η.
Require: CoLaX parameters θ, local prototypes $\boldsymbol{P}$, and global context bank $\mathcal{G}$.

Training Phase (CoLaX modules trainable)
for epoch $= 1$ to max epoch **do**
  for each batch of input data $(\boldsymbol{X}, \boldsymbol{y})$ **do**
    Local Pattern Branch:
    Refine local prototypes: $\widetilde{\boldsymbol{P}} = \text{CTEncoder}(\boldsymbol{P})$
    Compute activation map: $\boldsymbol{A} = \text{Softmax}(\text{Conv}(\boldsymbol{X}, \widetilde{\boldsymbol{P}}))$
    Global Context Branch:
    Extract global embeddings: $\boldsymbol{Z} = \text{GEncoder}(\boldsymbol{X})$
    Compute context attention: $a_{t,k} = \text{Softmax}(\boldsymbol{z}_t^\top \boldsymbol{g}_k / \tau_1)$
    Aggregate global representation: $\boldsymbol{E}_{\text{global}} = \text{Pool}(\boldsymbol{Z})$
    Context-Aware Modulation:
    Query $\boldsymbol{q} = \boldsymbol{W}_q \boldsymbol{E}_{\text{global}}$, Key $\boldsymbol{K} = \widetilde{\boldsymbol{P}} \boldsymbol{W}_k$
    Compute gating: $\boldsymbol{w} = \text{Sigmoid}(\boldsymbol{K}\boldsymbol{q}/\tau_2)$
    Modulate local prototypes: $\widetilde{\boldsymbol{Q}} = \text{diag}(\boldsymbol{w})\widetilde{\boldsymbol{P}}$
    Prediction: $\hat{\boldsymbol{y}} = \sigma(\text{MLP}(\widetilde{\boldsymbol{Q}}))$
    Loss Computation:
    $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{l-match}} + \lambda_2 \mathcal{L}_{\text{g-align}}$
    Optimization: Update θ via gradient descent
    Memory Update:
    Compute normalized $\bar{\boldsymbol{z}}_k = \frac{\sum_{t=1}^{T'} a_{t,k} \boldsymbol{z}_t}{\sum_{j=1}^{T'} a_{j,k} + \epsilon}$
    Update $\mathcal{G}$ via gated memory rule.
  end for
end for
Inference Phase (CoLaX modules frozen)
for each test sample $\boldsymbol{X}_{\text{test}}$ **do**
  Compute local activation map $\boldsymbol{A}$, temporal embeddings $\boldsymbol{Z}$, pooled global context $\boldsymbol{E}_{\text{global}}$, sample-specific prototype weights $\boldsymbol{w}$, and context attention $a_{t,k}$.
  Compute temporal context signal $\boldsymbol{c}$ from the context-attention matrix:
  $c_t \leftarrow \text{Norm}\left[\text{Interp}\left(\{\max_k a_{s,k}\}_{s=1}^{T'}\right)\right]_t$
  Construct context-conditioned activation $A_{t,k}^{\text{ctx}} = w_k A_{t,k}(1 + \alpha c_t)$.
  Extract contiguous intervals satisfying $A_{t,k}^{\text{ctx}} > \Omega$.
  Form the candidate segment set $\mathcal{P}(\boldsymbol{X}_{\text{test}}) = \{\mathcal{I}_1, \ldots, \mathcal{I}_N\}$.
  Construct coalition-based perturbations using the detected segments.
  Black-box Shapley Attribution: $c^\star \leftarrow \arg\max_c f_c(\boldsymbol{X}_{\text{test}})$
  Construct coalition perturbations $\boldsymbol{X}_S$ using interpolation masking.
  Evaluate $v(S) \leftarrow f_{c^\star}(\boldsymbol{X}_S)$ and compute segment Shapley values $\{\phi_n\}$.
  Aggregate $|\phi_n|$ over overlapping segments and normalize to obtain $\boldsymbol{R}$.
end for
return Trained CoLaX parameters θ and saliency maps $\{\boldsymbol{R}\}$.

H Broader Impacts

CoLaX helps practitioners inspect time-series classifiers by highlighting which local patterns a frozen model responds to and in what sequence-level contexts, supporting debugging, spurious-cue detection, and expert auditing (e.g., healthcare or environmental monitoring).

CoLaX provides post-hoc, correlational approximations of a model's response to structured perturbations; results depend on the intervention strategy and prototype dictionary, and they do not establish causality or clinical validity. Benefits are strongest on synthetic benchmarks, and real-world utility needs further study. The main risk is over-reliance (treating explanations as evidence a prediction is correct), so CoLaX should be used only as decision support with domain-specific validation and human oversight, not for autonomous high-stakes decisions.

All experiments use public, de-identified benchmarks; we collected no new data and conducted no user study or crowd annotation. Since CoLaX explains existing models and generates no content, we see no meaningful additional dual-use risk beyond that of the underlying classifiers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction are well-aligned with the paper's actual contributions. The abstract accurately outlines the core method Co-LaX for time series classification, highlighting its ability to capture both local discriminative patterns and global contextual information. These core ideas are consistently developed throughout the paper (Section 2), and the experimental results on UCR datasets (Section 3) and visualizations (ECG5000, synthetic data) empirically support the stated claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated section in Appendix G discussing the limitations of the proposed method.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not contain formal theoretical results such as theorems or lemmas. The contributions are primarily algorithmic and empirical, presenting the CoLaX method with experimental validation on benchmark datasets. Therefore, this question is not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have disclosed all necessary information to reproduce our main results. This includes the complete mathematical derivation of the algorithm, detailed architecture diagrams, pseudocode in Appendix F, and comprehensive hyperparameter settings. Furthermore, the source code and reproduction scripts provided at `https://github.com/zkkkkkkkkkkx/CoLaX` facilitate full replication.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The complete code for our proposed method, along with scripts to reproduce the main experimental results, is provided in the supplementary material and at `https://github.com/zkkkkkkkkkkx/CoLaX`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all essential experimental details, including data splits, hyperparameter configurations in Appendix B.3.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the mean and standard deviation for all main experimental results across five independent runs with different random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information regarding the hardware specifications in Appendixes B.3.1 and C.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and confirm that our research, including data collection, algorithmic development, and experimental evaluation, fully adheres to all its principles and guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H discusses potential benefits (model auditing and debugging) and risks, chiefly over-reliance on correlational, perturbation-based explanations, and recommends use as decision support under human oversight with domain validation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The models and methods developed in this research are intended for time series analysis and do not pose a high risk for misuse. Therefore, specific safeguards for responsible release are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets (datasets and codebases) used in this work are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide the full implementation of our proposed method, including the core algorithm, experimental scripts, and environment configurations, in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This research does not involve any crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any human subjects or participant-based studies. Therefore, no Institutional Review Board (IRB) approval is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing assistance and did not play any role in the core methodology, algorithm design, experimentation, or scientific decision-making process.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.