Position-aware eXplanation: A Model-Agnostic Framework for Positional Attributions

Chaehyeon Kim[1*] **Gary E. Weissman**[2] **Eric Wong**[1]

[1]Department of Computer and Information Science, University of Pennsylvania
[2]Perelman School of Medicine, University of Pennsylvania

Abstract

Transformer predictions depend on positional embeddings, which are known to produce biases such as first- and last-token dominance. However, feature attribution methods implicitly entangle influence from positions with features into a single importance score per feature. To uncover these positional effects, we propose Position-aware eXplanation (PaX), a model-agnostic framework that transforms standard attribution methods to jointly produce feature and position attributions. PaX additionally produces counterfactual position explanations: actionable scores that quantify how the prediction would change if a feature were relocated. PaX generalizes perturbation-, gradient-, and boundary-based attribution methods without architectural changes, which we evaluate across vision, language, and clinical time-series benchmarks. We demonstrate that separating positional effects consistently improves both feature and position attribution faithfulness across insertion and deletion evaluations. In a case study on sepsis forecasting, we find that PaX recovers clinically validated bedside signals that standard attributions miss. [🌀]

1 Introduction

Modern transformer-based models are highly sensitive to input order [1–5]. Reordering elements in a sequence, from sentences in a document to patches in an image, can substantially alter a model's output, raising fundamental questions about their reliability and interpretability [6, 7]. This positional sensitivity has significant practical consequences: an image classifier may become overly reliant on an object's location, or a large language model may overweight information at the beginning or end of a prompt [8]. When sequence position encodes time, as in electronic health records, this blind spot becomes critical: in sepsis forecasting, an elevated temperature at admission may be unremarkable, whereas the same reading hours later signals deterioration. Standard explanation methods attribute importance to the measurement, obscuring that timing is the true causal driver [9, 10].

Explanation methods help interpret black-box model behavior and play an important role in building trust in AI systems [11]. However, existing attribution methods [12–16] are not well suited for identifying positional effects, since they entangle position and feature influences into a single importance score, failing to distinguish a feature's intrinsic value ('what') from its position ('when' or 'where'). Recent work confirms predictions can depend on feature location as well as feature values [17, 18], yet the magnitude of positional influence remains unknown in standard attributions.

Recent approaches have attempted to account for temporal or spatial effects [19–23], yet they remain domain- or method-specific. These methods are typically designed for a specific attribution framework (e.g., SHAP, Integrated Gradients) or a single domain (e.g., medical time-series), leaving a gap for a general, model-agnostic solution that can be applied across vision, language, and time-series settings.

[*] Correspondence to: chaenyk@engineering.upenn.edu
[🌀] Code is available at https://github.com/chaehyeon-kim/PaX.

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

To address this, we introduce *Position-aware eXplanation (PaX)*, a model-agnostic framework that enhances standard feature attribution by explicitly treating position as a "first-class feature." PaX decomposes a feature's total importance into three components: (1) a feature score, which quantifies the importance of its content; (2) a position score, which quantifies the influence of its location; and (3) a counterfactual position score, an actionable measure of how the prediction changes when the content moves from its position to another. Unlike prior positional explanations tied to specific methods or domains, PaX augments any feature attribution method and applies across modalities.

Our contributions are summarized as follows:

- We formalize PaX, a general framework that transforms any feature attribution method to jointly produce feature and position attributions, with no architectural changes to the underlying model.
- We propose counterfactual position attribution, an actionable explanation whose scores quantify the predictive impact of relocating feature content across positions.
- We develop position-aware faithfulness tests for both position and counterfactual position attributions by extending feature-based insertion and deletion metrics to positions and comparing counterfactual position attributions with actual output changes under content relocation.
- We validate PaX across vision (ImageNet-S, PASCAL VOC), language (GSM8K, MedMCQA), and clinical time-series (PhysioNet 2019, MC-MED) benchmarks, augmenting seven attribution methods to improve feature and position attribution faithfulness.
- In a sepsis forecasting case study, PaX surfaces bedside-relevant signals beyond formal sepsis criteria and exposes positional artifacts that standard attributions mistake for biological signal.

2 Position-aware eXplanation (PaX)

2.1 Preliminaries: Feature Attributions

Feature attribution methods interprets models by assigning importance scores to input features [24]. Existing methods typically answer the question: *How important are input features to the prediction?*

To formalize, let $f : \mathbb{R}^d \to \mathbb{R}^m$ be the predictive model being explained, which maps an input $x \in \mathbb{R}^d$ to an output $f(x)$, such as class probabilities or regression outputs. A general approach in feature attribution is to define a perturbation function $g : \mathbb{R}^d \times [0,1]^d \to \mathbb{R}^d$ that maps x and a reference vector $z \in [0,1]^d$ to a perturbed input $\tilde{x} = g(x,z) \in \mathbb{R}^d$. The reference vector z specifies feature perturbation: $z_i = 0$ retains feature i, and $z_i = 1$ perturbs it, for example by replacing it with a baseline value. Under this convention, $g(x, \mathbf{0}) = x$. Path-based methods such as Integrated Gradients use intermediate values of z to interpolate between the original and perturbed features, while perturbation-based methods such as LIME sample binary values of z to drop random features.

A feature attribution explainer, denoted by $\mathcal{E}$, uses the perturbation function g alongside the model f and input x to produce an attribution vector $\alpha = \mathcal{E}(f, g, x) \in \mathbb{R}^d$, where each attribution score $\alpha_i \in \mathbb{R}$ represents the contribution of feature x_i to an output coordinate c of $f(x)$, typically the predicted class. Let $v(z) \coloneqq f_c\big(g(x,z)\big)$ denote the value function. An explainer $\mathcal{E}$ evaluates v at different values of z to create an explanation from outputs of perturbed inputs. For example, LIME evaluates v at random binary z and fits a linear surrogate, while Integrated Gradients (IntGrad) integrates $\partial v / \partial z$ along the path from $z = \mathbf{1}$ to $z = \mathbf{0}$. The formulation of explainers $\mathcal{E}(f, g, x)$ probing the model on perturbed inputs via the value function $v(z)$ generalizes a wide range of feature attribution methods, from classic approaches [12–14, 16, 25] to newer ones [26, 27].

Attributions for Sequence-Structured Inputs. Sequence models such as Transformers [28] take an input of n ordered elements $x = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$, where each $x_i \in \mathbb{R}^d$ represents a language token, an image patch, or the measurements at one time step. These models typically associate position indices $p = (1, \ldots, n)$ with x, so that $p_i = i$ is the position of x_i. The model $f : \mathbb{R}^{n \times d} \to \mathbb{R}^m$ takes x as input and uses p internally. Depending on the architecture, it adds a positional embedding $e(p_i)$ to x_i, rotates queries and keys according to p_i, or uses p only through the order of x. With each element treated as one feature, the perturbation function $g : \mathbb{R}^{n \times d} \times [0,1]^n \to \mathbb{R}^{n \times d}$ perturbs x according to a perturbation vector $z \in [0,1]^n$, where z_i governs the perturbation of x_i. The explainer $\mathcal{E}$ produces an element-level attribution vector

$$\mathcal{E}(f, g, x) = \alpha \in \mathbb{R}^n, \tag{1}$$

where α_i quantifies the contribution of the i-th sequence element x_i to $f_c(x)$.

Limitations of Existing Explanations. Existing attribution methods, including perturbation-based [12, 14, 25], gradient-based [13, 16], and boundary-based [27] approaches, focus on explaining the effect of features, and in general do not distinguish whether the assigned importance arises from the semantic content of the input or from positional effects internal to the model.

A second line of work attempts to incorporate positional information by perturbing or ablating features at different positions and measuring the resulting change in the model's output. TimeSHAP [19] extends SHAP with time-level masking for recurrent models; WindowSHAP [20] partitions a time series into windows and computes Shapley values at the window level; TIMING [21] averages IntGrad over segment masks so that the integration paths respect temporal dependencies; DeltaSHAP [22] adapts SHAP to attribute changes between consecutive predictions; and OrdShap [23] measures how predictions change when features are permuted to different positions. However, these approaches mostly target time-series settings and each extends a specific feature attribution technique.

2.2 The PaX Framework

We propose *Position-aware eXplanation (PaX)*, a framework that captures positional effects by treating positions as first-class features through augmenting the input representation. In contrast to prior work, our framework brings three key benefits:

1. PaX provides position attributions for any model that ingests sequential data, extending beyond tabular settings to Transformers used in modern vision and language tasks.
2. PaX introduces counterfactual position attribution, a new type of attribution that yields actionable explanations of positional effects.
3. PaX generalizes a broad class of attribution methods, creating a family of position-aware attribution methods without requiring bespoke solutions for individual techniques.

We first show how PaX generalizes feature attribution methods to produce position attributions, then further extend the framework to counterfactual position explanations.

2.2.1 From Feature to Position Attributions

PaX framework generalizes the standard explanation formulation in Eq. (1) by incorporating position into the explanation. Given a sequence x with position indices $p = (1, \ldots, n)$, let $\mathcal{E}(f, g, x)$ be a general feature attribution method. The key idea of PaX is to construct position-aware versions of the input x', model f', and perturbation function g' such that $\mathcal{E}(f', g', x')$ produces both feature and position attributions.

PaX Inputs. The first step is to explicitly formulate an augmented input with position indices:

$$x' \coloneqq (x, p) \in \mathbb{R}^{n \times d} \times \mathrm{Sym}([n]),$$

where $\mathrm{Sym}([n])$ is the set of position indices that permute $(1, \ldots, n)$. The augmented input jointly represents feature content and positional information. We refer to x' as the PaX input. We write $\tilde{p} \in \mathrm{Sym}([n])$ for perturbed position indices. The PaX model below defines the output for any $\tilde{p}$.

PaX Models. The model f uses the position indices p internally. PaX formulates the position as an explicit input to the model. For any $\tilde{p}$, let $\Pi_{\tilde{p}}$ be the permutation matrix with $(\Pi_{\tilde{p}})_{i,\tilde{p}_i} = 1$. We define the PaX model f' by reordering the input sequence,

$$f'(x, \tilde{p}) \coloneqq f\big(\Pi_{\tilde{p}}^{\top} x\big), \tag{2}$$

where $\Pi_{\tilde{p}}^{\top} x$ places x_i at slot $\tilde{p}_i$. The default forward pass is $f'(x') = f'(x, p) = f(x)$, and any other $\tilde{p}$ perturbs the positions. In a bidirectional model, Eq. (2) is equivalent to adding $e(\tilde{p}_i)$ to x_i in place when the model uses additive positional embeddings, and rotating the queries and keys of x_i by $\tilde{p}_i$ when it uses rotary embeddings. Because f' only permutes the input and requires no access to explicit positional representations, PaX also applies to models with implicit positions.

PaX Perturbations. The next component is the perturbation function g', which perturbs both features x and positions p. Let $z' = (z^x, z^p) \in [0, 1]^n \times [0, 1]^n$ be a reference vector, where z_i^x governs feature x_i and z_i^p governs position p_i, both under the perturbation convention of Section 2.1.

The base method's perturbation function $g(x, z^x)$ perturbs the features. Positions, however, should not be masked the same way. Replacing them with a baseline or mean, a common practice for

features that is known to introduce missingness bias [29], either yields positional embeddings that match no index or assigns one index to several elements, neither of which occurs during training. A permutation instead keeps every index in $[n]$ and uses each exactly once, so the model only sees position values that occur during training. PaX therefore shuffles a random subset of positions. We define the distribution of the resulting permutations as

$$(b, \tilde{p}) \sim \textsc{pos-swap}(z^p): \quad b_i \sim \mathrm{Bern}(z_i^p) \text{ independently}, \quad \tilde{p} \mid b \sim \mathrm{Unif}\big(\mathrm{Sym}(S)\big), \tag{3}$$

where $S = \{i : b_i = 1\}$ is the pool of shuffled positions and $\mathrm{Sym}(S)$ is the set of $\tilde{p} \in \mathrm{Sym}([n])$ with $\tilde{p}_j = p_j$ for every $j \notin S$. We write $\mathbb{E}_{b,\tilde{p}}$ for the expectation over $(b, \tilde{p}) \sim \textsc{pos-swap}(z^p)$.

The joint PaX perturbation of x' is

$$g'(x', z') \coloneqq \big(g(x, z^x),\, \tilde{p}\big), \qquad (b, \tilde{p}) \sim \textsc{pos-swap}(z^p). \tag{4}$$

Here z^p affects $\tilde{p}$ only through the inclusion probabilities of b. In practice, PaX applies $\tilde{p}$ by shuffling the selected elements of the input, such as tokens, image patches, or time steps. Section 3.5 compares swapping with zero, mean, and random replacement on models with additive positional embeddings and finds that swapping yields more faithful explanations.

PaX Value Function. Recall that in standard feature attributions, the explainer $\mathcal{E}$ evaluates the value function $v(z) = f_c(g(x, z))$ at varying z. A natural generalization to the PaX setting would apply the same explainer to $f_c'(g'(x', z'))$. However, the draw $(b, \tilde{p})$ is random, so PaX defines the value function as the expected model output over the shuffles:

$$v(z') \coloneqq \mathbb{E}_{b,\tilde{p}}\big[f_c'\big(g'(x', z')\big)\big] = \mathbb{E}_{b,\tilde{p}}\big[f_c'\big(g(x, z^x), \tilde{p}\big)\big]. \tag{5}$$

Attribution methods interact with v through one of two mechanisms.

Mechanism 1: evaluation at binary z^p. Perturbation-based explainers such as LIME and SHAP sample binary $z' \in \{0, 1\}^{2n}$ from their own distributions and need $v(z')$ at each sample. At binary z^p, the indicators are deterministic with $b = z^p$, so $\tilde{p}$ is the only remaining randomness. A single forward pass $f_c'\big(g(x, z^x), \tilde{p}\big)$ then gives an unbiased estimate of $v(z')$. Each method aggregates these outputs linearly given the sampled z', so this randomness adds only Monte Carlo variance and the explainer runs unchanged.

Mechanism 2: derivatives at continuous z^p. Gradient-based explainers such as IntGrad need $\partial v / \partial z_i^p$ at interior points $z^p \in (0, 1)^n$. Each draw $(b, \tilde{p})$ yields hard position indices, and v depends on z^p only through the distribution of b. Conditioning on b_i gives

$$v(z') = z_i^p\, \mathbb{E}_{b,\tilde{p}}\big[f_c'\big(g(x, z^x), \tilde{p}\big) \mid b_i = 1\big] + (1 - z_i^p)\, \mathbb{E}_{b,\tilde{p}}\big[f_c'\big(g(x, z^x), \tilde{p}\big) \mid b_i = 0\big],$$

where each conditional expectation draws the remaining indicators and $\tilde{p}$ as in Eq. (3). Since both terms fix b_i, neither depends on z_i^p, so v is affine in each z_i^p separately, with exact partial derivative

$$\frac{\partial v}{\partial z_i^p}(z') = \mathbb{E}_{b,\tilde{p}}\big[f_c'\big(g(x, z^x), \tilde{p}\big) \mid b_i = 1\big] - \mathbb{E}_{b,\tilde{p}}\big[f_c'\big(g(x, z^x), \tilde{p}\big) \mid b_i = 0\big]. \tag{6}$$

Eq. (6) requires only forward passes, and we estimate both terms from one set of draws $(b, \tilde{p}) \sim \textsc{pos-swap}(z^p)$ by averaging the outputs separately over draws with $b_i = 1$ and with $b_i = 0$. Each draw records the full vector b, so the same draws serve all n positions.

PaX Explainer. Combining these components, we define the PaX explainer:

$$\mathcal{E}_{\text{PaX}}(f, g, x) \coloneqq \mathcal{E}(f', g', x') = \alpha' = (\alpha^x, \alpha^p) \in \mathbb{R}^{2n}. \tag{7}$$

Eq. (7) applies the original explainer $\mathcal{E}$ to the PaX components (f', g', x') through the value function v in Eq. (5). A single run of $\mathcal{E}$ thus produces both feature and position attributions. We interpret α_i^x as the importance of the content of x_i and α_i^p as the importance of its position p_i.

2.2.2 Counterfactual Position Attributions

We extend the core idea of PaX to measure how the prediction changes when we rearrange the input. Just as PaX treats position as an additional feature, we treat the relocation of an element from position i to position j as another feature and call the resulting explanation a counterfactual position attribution. These explanations can be more *actionable* than position attributions because they specify where an element moves (e.g., moving an object in an image from a corner to the center).

Table 1: Position faithfulness on PASCAL VOC (vision, 20-class object classification). Across ViT, CLIP, and Llama, PaX position attributions match or outperform existing explanations that account for position on both insertion (↑) and deletion (↓). Bold marks the best score and scores with error bars overlapping it. The trend holds across other datasets and all PaX variants (PaX-A for $A \in$ {LIME, SHAP, RISE, SHAP-IQ, IntGrad, FullGrad, MFABA}), as reported in Appendix C.

Model	Faithfulness Metric	PaX-LIME	PaX-IntGrad	TimeSHAP [19]	WindowSHAP [20]	TIMING [21]	DeltaSHAP [22]	OrdShap [23]
ViT	Insertion	0.400 ± 0.006	$\mathbf{0.413 \pm 0.007}$	0.325 ± 0.014	0.344 ± 0.012	0.400 ± 0.006	0.399 ± 0.006	0.386 ± 0.009
	Deletion	$\mathbf{0.008 \pm 0.001}$	0.012 ± 0.001	0.017 ± 0.002	0.015 ± 0.002	0.011 ± 0.001	0.010 ± 0.001	0.012 ± 0.002
CLIP	Insertion	$\mathbf{0.441 \pm 0.005}$	$\mathbf{0.437 \pm 0.008}$	0.354 ± 0.016	0.373 ± 0.013	0.424 ± 0.007	0.423 ± 0.007	0.407 ± 0.010
	Deletion	$\mathbf{0.012 \pm 0.001}$	0.015 ± 0.001	0.020 ± 0.002	0.017 ± 0.002	0.013 ± 0.001	$\mathbf{0.011 \pm 0.001}$	0.014 ± 0.002
Llama	Insertion	$\mathbf{0.437 \pm 0.004}$	$\mathbf{0.445 \pm 0.006}$	0.372 ± 0.015	0.391 ± 0.013	$\mathbf{0.439 \pm 0.006}$	$\mathbf{0.437 \pm 0.005}$	0.428 ± 0.009
	Deletion	$\mathbf{0.006 \pm 0.001}$	0.011 ± 0.001	0.021 ± 0.002	0.018 ± 0.002	0.012 ± 0.001	0.010 ± 0.001	0.013 ± 0.002

PaX Counterfactual Inputs. We augment the PaX input with relocation indicators, $x'' = (x', q)$, where $q = (q_{ij})_{i \neq j} \in \{0,1\}^{n(n-1)}$ and $q_{ij} = 1$ indicates that the element at position i moves to position j. The original input relocates nothing, so $q = \mathbf{0}$. Unlike features and positions, relocations are not independently controllable. Placing element i at position j displaces the element at j, so no perturbation can vary one relocation while holding all others fixed. PaX therefore does not control relocations directly. It samples $\tilde{p}$ from POS-SWAP and records the relocations it induces as $\tilde{q}$, with $\tilde{q}_{ij} = 1$ if $\tilde{p}_i = j$ and $\tilde{q}_{ij} = 0$ otherwise.

PaX Counterfactual Models. We define the counterfactual model as

$$f''(x, \tilde{p}, \tilde{q}) \coloneqq f'(x, \tilde{p}),$$

where $\tilde{q}$ records the relocations of $\tilde{p}$. Because $\tilde{p}$ and $\tilde{q}$ determine each other, the forward pass on $(x, \tilde{p})$ carries out the exact relocations in $\tilde{q}$.

PaX Counterfactual Perturbations. The counterfactual perturbation g'' applies g' and also returns the relocations induced by the drawn permutation,

$$g''(x'', z') \coloneqq \big(g(x, z^x),\, \tilde{p},\, \tilde{q}\big), \qquad (b, \tilde{p}) \sim \text{POS-SWAP}(z^p).$$

Following the perturbation convention of z^x and z^p, we record $z^q = \tilde{q}$, so $z^q_{ij} = 1$ marks a relocation that occurred. Unlike z^x and z^p, z^q is induced by POS-SWAP rather than sampled by the explainer. Since z^q and the indicators b come from the same draw, every recorded relocation is consistent with the pool: $z^q_{ij} = 1$ implies $b_i = b_j = 1$. Every output of g'' pairs $\tilde{p}$ with its own relocations, so f'' agrees with f' on every draw and the value function in Eq. (5) stays unchanged.

Under Mechanism 1, explainers append z^q to their interpretable features and run unchanged. Their coefficients or masked averages on z^q give the counterfactual position attributions α^q. Because each position holds exactly one element, these coordinates correlate with each other and with z^p, so we compare feature, position, and counterfactual position attributions only within their own block.

Under Mechanism 2, explainers need a derivative of v with respect to each relocation, but this derivative has no natural definition because relocations are not independently controllable. PaX instead supplies the conditional contrast

$$C_{ij}(z') \coloneqq \mathbb{E}_{b,\tilde{p}}\big[f'_c\big(g(x, z^x), \tilde{p}\big) \mid \tilde{p}_i = j\big] - \mathbb{E}_{b,\tilde{p}}\big[f'_c\big(g(x, z^x), \tilde{p}\big) \mid b_i = 0\big], \quad i \neq j, \tag{8}$$

the effect of moving the element at position i to position j relative to keeping it in place. The contrast C_{ij} is defined whenever $z^p_i, z^p_j > 0$, so that both i and j can enter the pool. We estimate C_{ij} from the same draws as Eq. (6) by averaging the forward-pass outputs separately over draws with $z^q_{ij} = 1$ and with $b_i = 0$. Appendix A.2.2 integrates Eq. (6) and Eq. (8) along the path from $z' = \mathbf{1}$ to $\mathbf{0}$ to obtain PaX-IntGrad.

The counterfactual PaX explainer applies $\mathcal{E}$ to the counterfactual components and returns

$$\mathcal{E}^{\text{cf}}_{\text{PaX}}(f, g, x) \coloneqq \mathcal{E}(f'', g'', x'') = \alpha'' = (\alpha^x, \alpha^p, \alpha^q) \in \mathbb{R}^{2n + n(n-1)}, \tag{9}$$

where a positive α^q_{ij} means that keeping the element at position i in place increases f_c relative to moving it to position j. This effect is averaged over perturbed backgrounds and differs from the single-transposition effect we use for evaluation.

Table 2: Feature faithfulness on ImageNet-S (vision, 919-class object classification) using ViT. PaX-Feature consistently improves over Feature-only explanations across all seven attribution methods on both insertion (↑) and deletion (↓). The trend holds across other datasets (PASCAL VOC, GSM8K, MedMCQA, PhysioNet 2019, MC-MED) and models (CLIP, Llama 3.2 Vision, Llama 3, GPT-2, Mamba, MedGemma), as shown in Appendix C.

Metric	Attribution Method	LIME [12]	SHAP [14]	RISE [25]	SHAP-IQ [26]	IntGrad [13]	FullGrad [16]	MFABA [27]
Insertion	Feature-only	0.354 ± 0.006	0.342 ± 0.005	0.349 ± 0.007	0.346 ± 0.006	0.361 ± 0.009	0.336 ± 0.012	0.351 ± 0.008
	PaX-Feature	$\mathbf{0.419 \pm 0.005}$	$\mathbf{0.403 \pm 0.004}$	$\mathbf{0.411 \pm 0.006}$	$\mathbf{0.407 \pm 0.005}$	$\mathbf{0.427 \pm 0.008}$	$\mathbf{0.384 \pm 0.010}$	$\mathbf{0.417 \pm 0.007}$
Deletion	Feature-only	0.020 ± 0.001	0.019 ± 0.001	0.019 ± 0.001	0.019 ± 0.001	0.018 ± 0.001	0.017 ± 0.001	0.018 ± 0.001
	PaX-Feature	$\mathbf{0.008 \pm 0.001}$	$\mathbf{0.007 \pm 0.001}$	$\mathbf{0.009 \pm 0.001}$	$\mathbf{0.008 \pm 0.001}$	$\mathbf{0.009 \pm 0.001}$	$\mathbf{0.010 \pm 0.001}$	$\mathbf{0.007 \pm 0.001}$

2.2.3 Example: PaX-LIME

A major advantage of PaX is that it turns a feature attribution method into one that also produces position attributions without bespoke design. As a demonstration, we instantiate PaX with LIME for sequence-structured inputs by treating positions as additional interpretable features. LIME [12] has three main steps: (1) generate perturbed inputs, (2) make predictions on perturbed inputs, and (3) fit a weighted linear model to the predictions.

1. **Perturbing inputs.** LIME samples binary feature masks z^x uniformly at random, where $z_i^x = 0$ retains feature i and $z_i^x = 1$ perturbs it. PaX-LIME additionally samples a binary position mask $z^p \in \{0,1\}^n$ uniformly at random under the same convention. The positions with $z_i^p = 1$ form the pool that POS-SWAP shuffles, and the drawn $\tilde{p}$ determines z^q.
2. **Making predictions.** LIME passes each perturbed input to the model $f_c(g(x, z^x))$. PaX-LIME passes one draw of the joint perturbation to the PaX model and records $f'_c(g(x, z^x), \tilde{p})$, which gives an unbiased single-sample estimate of $v(z')$ under Mechanism 1.
3. **Fitting a linear model.** PaX-LIME repeats steps (1) and (2) to collect N samples $\{(z''^{(k)}, y''^{(k)})\}_{k=1}^N$ and weights each by its proximity to the original input. LIME fits the weighted linear model from z^x to the predictions. PaX-LIME instead fits it from the augmented masks $z'' = (z^x, z^p, z^q)$ to y''. Since $z = 1$ marks perturbation, the negated coefficients give the joint attributions $\alpha'' = (\alpha^x, \alpha^p, \alpha^q)$.

The PaX family of attribution methods. Beyond PaX-LIME, PaX applies to a broad class of attribution methods and yields a family of position attribution methods. Treating position as an explicit model input reduces position attribution to feature attribution on the value function v. Perturbation-based methods such as SHAP, RISE, and SHAP-IQ run unchanged on z^x, z^p, and z^q through Mechanism 1. Gradient-based and boundary-based methods such as IntGrad, FullGrad, and MFABA compute content gradients as usual and, through Mechanism 2, use Eq. (6) and Eq. (8) in place of gradients for positions and relocations. For IntGrad, this amounts to integrating both along the path from $z' = \mathbf{1}$ to $\mathbf{0}$. Appendix A.2 gives additional instantiations.

3 Experiments

We show that PaX explanation yields more faithful feature and position attributions [30] than existing attribution methods, and that the counterfactual position attributions correlate with the prediction changes under relocation.

3.1 Experimental Setup

Datasets and Models. We evaluate PaX across three modalities: vision (ImageNet-S and PASCAL VOC with ViT-B/16, CLIP-ViT-L/14, and Llama 3.2 11B Vision Instruct), language (MedMCQA and GSM8K with Llama 3 8B Instruct), and clinical time series (sepsis forecasting on PhysioNet 2019 and MC-MED with fine-tuned GPT-2 Small, Mamba-130M, and MedGemma 4B). Appendix B provides additional details.

Explanation Methods and Baselines. We compare PaX against three groups of baselines: (1) standard feature attribution methods, including LIME [12], SHAP [14], RISE [25], SHAP-IQ [26], IntGrad [13], FullGrad [16], and MFABA [27], which perturb features alone; (2) a position-only ablation of each method in (1), which perturbs positions alone with features held fixed; and (3)

existing explanations that account for position, including TimeSHAP [19], WindowSHAP [20], TIMING [21], DeltaSHAP [22], and OrdShap [23]. Appendix A provides further details.

3.2 The PaX Family Yields More Faithful Position Attributions without Bespoke Solutions

A faithfulness metric for position attributions. We extend the standard insertion and deletion protocols [25] from features to positions. Instead of adding or removing content, these protocols keep the content fixed and relocate elements in the order given by the position attributions. Position insertion starts from a random configuration and sequentially places features at the highest-ranked positions. Position deletion starts from the original input and moves features from highly ranked to random positions. As in the feature-based case, we summarize each curve by the area under it (AUC). Higher insertion AUC and lower deletion AUC indicate more faithful position attributions.

PaX outperforms existing position-aware methods. On PASCAL VOC, PaX achieves higher insertion AUC and lower deletion AUC than existing methods that account for position as shown in Table 1. Appendix C confirms that these gains hold across additional datasets, explanation methods, and models, indicating that PaX improves position-attribution faithfulness while remaining both model- and method-agnostic.

3.3 Accounting for Positional Effects Improves the Faithfulness of Feature Attributions

Although PaX was designed to produce position attributions, we find that accounting for position also improves feature attributions. We evaluate this with the standard feature insertion and deletion tests [25, 31], which progressively insert or remove features according to an attribution ranking and report the AUC. Higher insertion AUC and lower deletion AUC indicate more faithful feature attributions. For all methods, we rank features by their attribution scores in descending order. Table 2 shows that on ImageNet-S with ViT-B/16, PaX achieves higher insertion AUC and lower deletion AUC than the corresponding base method for every attribution method, and Appendix C shows this pattern across additional datasets and models. Jointly modeling feature content and position thus improves feature attributions even when position is not the primary focus.

3.4 Counterfactual Position Attributions Correlate with Prediction Changes under Relocation

We test whether counterfactual position attributions track the effect of relocating elements. For each input and each pair $i < j$, we swap the elements at positions i and j and measure the resulting drop in f_c. A swap moves x_i to position j and x_j to position i, so if α^q is faithful, $\alpha^q_{ij} + \alpha^q_{ji}$ and this drop should covary. We quantify this with their Pearson correlation.

Table 3: Faithfulness of counterfactual position attributions. Pearson r between each α^q_{ij} from PaX-SHAP and the output drop caused by swapping the elements at positions i and j.

Dataset	Model	Pearson correlation r
ImageNet-S	ViT-B/16	0.282 ± 0.041
MedMCQA	Llama 3 8B	0.431 ± 0.029
MC-MED	Mamba-130M	0.358 ± 0.032

Table 3 reports positive correlations in all three modalities, indicating that counterfactual position attributions capture part of the effect of relocations. Appendix D reports results for the remaining datasets and models.

3.5 Analysis of Missingness Approximations for Position Perturbations

A known challenge in attribution and its evaluation is the choice of missingness approximation, which determines how to represent the absence of a feature. Prior work in vision replaces missing regions with a constant color, the mean color, random noise, or a blurred region [29]. These approximations target feature content, not position. We therefore compare four positional analogues inside the PaX perturbation while keeping the evaluation protocols fixed: zeroing out positional embeddings, replacing them with their mean, replacing them with positional embeddings of random positions, and swapping them across positions. Zeroing out and mean replacement produce embeddings that match no position, and random replacement can assign one position to several elements. Swapping instead uses each training-time positional embedding exactly once, which keeps the positional inputs closer to the training distribution. Table 4 confirms that, on PhysioNet with GPT-2 Small and LIME, swapping yields the highest insertion AUC under both feature and position faithfulness metric. Additional

Table 4: Ablation of missingness approximations for insertion faithfulness on PhysioNet using GPT-2 Small with LIME. PaX-Feature uses feature faithfulness, while PaX-Position uses position faithfulness. Positional swapping achieves the highest insertion AUC under both evaluation protocols.

Evaluation	Zeroing out	Replacing with mean	Replacing with random	Swapping
PaX-Feature	0.229 ± 0.010	0.205 ± 0.009	0.216 ± 0.010	$\mathbf{0.354 \pm 0.005}$
PaX-Position	0.210 ± 0.010	0.197 ± 0.009	0.208 ± 0.009	$\mathbf{0.341 \pm 0.006}$

results are provided in Appendix E. We therefore adopt positional swapping (POS-SWAP) as the default missingness approximation in our position-based experiments.

4 Case Study: Alignment with Clinical Expertise in Sepsis Forecasting

To examine PaX in a real-world medical setting, we collaborated with a medical expert to analyze whether PaX explanations are actionable for sepsis forecasting, where understanding which patient factors drive a prediction directly informs clinical decision making.

Toward actionable explanations. Recent work argues that explanations need to enable concrete decisions and interventions [32, 24, 33, 34]. Under this framing, an explanation is *actionable* to the extent that it supplies insights domain experts can act on [24, 35, 36]. Sepsis prediction is a particularly demanding setting for this standard: bedside clinicians cannot act on explanations that fail to map onto clinical expertise [37], and they need assurance that measurements flagged as high-importance reflect biological signal rather than artifacts of input layout.

We conducted an expert-guided clinical case study to assess the actionability of PaX across three axes most relevant to its deployment in clinical decision-support. First, *understandability*: do attributions align with clinical expertise [35, 36, 38]? Second, *task-enhancement*: do explanations surface diagnostically useful signals, including ones outside formal guidelines [39, 32]? Third, *reliability*: are attributions free of structural confounds, such as the well-documented positional biases of large language models [40], that would render them unsafe to act on [41, 37]?

We focus on PhysioNet with MedGemma and LIME, the best-performing combination, as shown in Tables 6 and 10. Appendix F describes the setup and annotation protocol.

4.1 Finding 1: PaX attributions align more closely with clinical reasoning than standard explanation methods.

To assess understandability, we test whether PaX assigns higher importance to measurements that the clinician identified as relevant to sepsis, and whether this alignment is stronger than standard LIME.

Each measurement was annotated along a sepsis-related axis with three labels: "yes" for direct indicators of sepsis risk, "partial" for indirect or contextual markers, and "no" for measurements that primarily indicate a different pathology. Appendix F.1 gives the full criteria. Examining the PaX-LIME feature component $\alpha^x_{\text{PaX-LIME}}$, average attributions follow the expected clinical hierarchy: 0.40, 0.31, and 0.22 for "yes," "partial," and "no" respectively.

To quantify the contribution of separating position from feature, we define $\Delta_f = \alpha^x_{\text{PaX-LIME}} - \alpha_{\text{LIME}}$, where $\Delta_f > 0$ indicates that a measurement's importance was underestimated by standard LIME due to its temporal position in the sequence. Of the 29 features the clinician labeled "yes" or "partial," 24 (83%) exhibited $\Delta_f > 0$ as shown in Figure 1, indicating that standard LIME systematically dilutes the importance of clinically critical biomarkers by conflating them with positional effects.

Error analysis: indirect indicators beyond formal labels. PaX-LIME assigned higher importance than Standard LIME to five features that the clinician labeled as not sepsis-related (orange and gray in Figure 1). On closer inspection, these were not failure cases but reflect higher-order physiological signals that lie outside the formal sepsis definition. Two of them, Chloride and Troponin I, function as second-order markers. Although Troponin I primarily marks myocardial infarction, its elevated attribution plausibly reflects cardiovascular involvement and organ dysfunction during severe septic shock. The remaining three, Phosphate, Magnesium, and Calcium, function as third-order markers. These electrolytes are non-specific to sepsis but indicate physiological derangement and metabolic

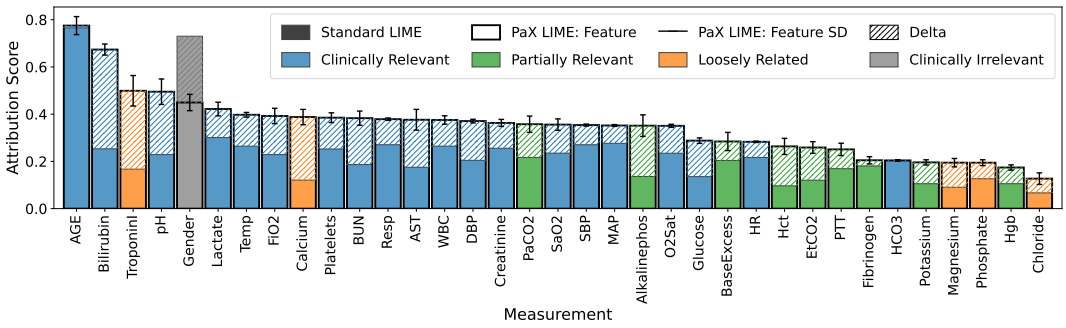


Figure 1: Feature attribution scores for sepsis-related clinical measurements on PhysioNet, stratified by clinician-assigned relevance (blue: clinically relevant, green: partially relevant, orange: loosely related, gray: irrelevant). Solid and white-filled bars show Standard LIME and PaX-LIME feature scores respectively; hatched bars show their difference (Δ). Clinically relevant features concentrate at the left with higher scores, loosely related features shift right, and the irrelevant feature Gender receives nearly half the attribution under PaX-LIME compared to Standard LIME, indicating that disentangling positional effects yields feature attributions more consistent with clinical reasoning.

instability. PaX-LIME thus surfaces secondary signals clinicians recognize as physiologically informative even outside first-tier sepsis criteria.

4.2 Finding 2: PaX surfaces bedside-relevant signals not codified in formal sepsis criteria.

Actionable explanations should surface diagnostically useful signals, including those beyond formal guidelines [32]. Standardized protocols (Sepsis-3, SOFA, qSOFA, NEWS, SOFA-2 [42]) offer structured criteria, but thresholds vary across institutions and bedside diagnosis depends on contextual cues no guideline fully captures.

Cross-referencing the clinician's annotations against established criteria, all 13 literature-flagged measurements were independently labeled "yes," and 16 of the 22 measurements absent from the literature (73%) nonetheless received a "yes" or "partial" label, as detailed inTable 5, suggesting that bedside clinical reasoning extends meaningfully beyond what formal sepsis criteria codify.

The pattern is clearer at PaX's top of the ranking. Ranking by the PaX-LIME feature component $\alpha^{x}_{\text{PaX-LIME}}$, the top ten features are Age, Bilirubin, pH, Lactate, Temp, FiO2, Platelets, BUN, Resp, and AST. Only 6 appear in the established literature, yet the clinician labeled all 10 as either "yes" (9) or "partial" (1). The four absent from formal guidelines, Age, pH, BUN, AST, were each recognized as clinically relevant, consistent with the task-enhancement criterion [32].

4.3 Finding 3: PaX exposes positional artifacts in the model that standard attribution would mistake for biological signal.

To assess reliability, we test whether PaX-LIME's decomposition into a feature component $\alpha^{x}_{\text{PaX-LIME}}$ and a position component $\alpha^{p}_{\text{PaX-LIME}}$ separates positional artifacts from genuine feature signal.

Each measurement was annotated along a Time-Dependent axis as either time-varying (e.g., vital signs) or time-stable (e.g., demographics), as described in Appendix F. As expected, $\alpha^{p}_{\text{PaX-LIME}}$ was higher for time-varying features (0.61) than for the two time-stable ones (0.33). These two features, Age and Gender, are the more revealing cases. Standard LIME assigns both high overall importance (α_{LIME} = 0.76 for Age, 0.73 for Gender). Once PaX decomposes attribution, a sharp divergence emerges: Age retains a feature score of 0.77, while Gender's drops to 0.33. Because PhysioNet places demographic variables at admission ($t = 0$), this pattern is consistent with a primacy bias. Age, despite appearing at the same time step, does not exhibit this divergence, consistent with its genuine clinical relevance.

Standard LIME would suggest that the model relies strongly on Gender for sepsis forecasting, a feature that the clinician considered irrelevant based on domain knowledge. PaX-LIME instead

Table 5: Clinician relevance labels for the 22 measurements absent from formal literature criteria. The clinician labeled a majority as sepsis-related ("yes" or "partial"), indicating that clinician reasoning captures physiological signals beyond what established sepsis definitions codify.

Clinician label	Count	Examples
yes	7	Age, AST, BUN, Glucose, HCO3, PaCO2, WBC
partial	9	Alkalinephos, BaseExcess, EtCO2, Fibrinogen, Hct, Hgb, PTT, Potassium, pH
no	6	Calcium, Chloride, Gender, Magnesium, Phosphate, Troponin I

attributes this effect to Gender's position rather than its value, a distinction clinicians need before trusting explanations in high-stakes settings.

5 Related Work

Model behavior often depends on where information appears in the input. Language models use information in the middle of long contexts less reliably than information at the beginning or end [3], and attention can be biased toward particular positions [43]. Post-hoc attributions also show position preferences [44]. These studies characterize positional effects but do not attribute an individual prediction to the content and position of each element.

Section 2.1 reviews temporal extensions of attribution methods [19–22], which target time-series settings and each extend a single base method. Closest to our work are methods that target position directly. PO-SHAP [45] builds on SHAP for sequence models, and PO-LRP [46] extends LRP to Transformers with positional encodings. OrdShap [23] also builds on SHAP and focuses on sensitivity to feature order. Each of these methods is tied to one attribution backbone, and PO-LRP further requires an explicit positional encoding. PaX instead wraps existing attribution methods. Perturbation-based methods run unchanged, and gradient-based methods replace only the positional derivative with an exact contrast computed from forward passes. Because PaX intervenes on position only through permutations, it also applies to models whose positions are implicit in the input order. As a result, PaX applies across attribution families, architectures, and input modalities, and it additionally yields counterfactual scores for moving an element to another position. Appendix G discusses these methods in more detail.

6 Conclusion

Standard attribution methods entangle *what* a feature is with *where* it appears, obscuring whether a prediction depends on feature content or on position. We introduce PaX, a model-agnostic framework that treats positions as explicit explanatory variables, letting existing feature attribution methods separate feature and position effects without modifying the model. PaX also produces counterfactual position scores that quantify how a prediction would change if identical content were relocated, an actionable form of explanation that standard attribution methods do not provide. To evaluate these explanations, we extend insertion-deletion protocols into positional faithfulness tests and introduce a counterfactual faithfulness test that measures how well counterfactual scores predict actual output changes under relocation.

Across seven attribution methods on vision, language, and clinical time-series benchmarks, PaX produced more faithful feature and position attributions than existing methods, alongside reliable counterfactual position scores. In a sepsis forecasting case study, it surfaced clinically validated bedside signals that standard attributions obscured and exposed positional artifacts otherwise misread as biological signal.

A limitation of PaX is that it scores individual positions, while some real-world signals reside in trajectories (e.g., slopes, peaks, plateaus) rather than in any single timestep. In sepsis forecasting, for example, PaX highlights the latest temperature reading, whereas a clinician would focus on the rising trend. Extending PaX to trajectory-level patterns is a natural next step.

More broadly, PaX can make explanations more reliable where position carries meaning, such as in clinical decision support, and can flag positional artifacts indicating data leakage or shortcuts. The principal risk is over-reliance: a confident-looking decomposition may mask model errors or distributional shift. PaX explanations should guide further analysis rather than replace it.

Acknowledgments and Disclosure of Funding

This research was partially supported by the ARPA-H program on Safe and Explainable AI under award D24AC00253-00, NIH grant R35GM155262, and NSF grant SHF 2442421.

References

[1] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=o2mbl-Hmfgd.

[2] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

[3] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

[4] Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

[5] Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. The Order Effect: Investigating Prompt Sensitivity to Input Order in LLMs, May 2025. URL http://arxiv.org/abs/2502.04134. arXiv:2502.04134 [cs].

[6] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

[7] Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models, 2024. URL https://arxiv.org/abs/2410.16983.

[8] Blerta Veseli, Julian Chibane, Mariya Toneva, and Alexander Koller. Positional biases shift as inputs approach context window limits, 2025. URL https://arxiv.org/abs/2508.07479.

[9] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17419-7. URL https://www.nature.com/articles/s41467-020-17419-7.

[10] Yubo Li, Xinyu Yao, and Rema Padman. No black box anymore: Demystifying clinical predictive modeling with temporal-feature cross attention mechanism, 2025. URL https://arxiv.org/abs/2503.19285.

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, July 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[14] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

[15] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017. doi: 10.1109/ICCV.2017.74. URL https://ieeexplore.ieee.org/document/8237336. ISSN: 2380-7504.

[16] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 371, pages 4124–4133. Curran Associates Inc., Red Hook, NY, USA, December 2019. URL https://dl.acm.org/doi/10.5555/3454287.3454658.

[17] Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations, 2021. URL https://arxiv.org/abs/2103.01890.

[18] Aahlad Puli, Nhi Nguyen, and Rajesh Ranganath. Explanations that reveal all through the definition of encoding, 2024. URL https://arxiv.org/abs/2411.02664.

[19] João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, and Pedro Bizarro. TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 2565–2573, New York, NY, USA, August 2021. Association for Computing Machinery. ISBN 978-1-4503-8332-5. doi: 10.1145/3447548.3467166. URL https://dl.acm.org/doi/10.1145/3447548.3467166.

[20] Amin Nayebi, Sindhu Tipirneni, Chandan K. Reddy, Brandon Foreman, and Vignesh Subbian. WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values. *J. of Biomedical Informatics*, 144(C), August 2023. ISSN 1532-0464. doi: 10.1016/j.jbi.2023.104438. URL https://doi.org/10.1016/j.jbi.2023.104438.

[21] Hyeongwon Jang, Changhun Kim, and Eunho Yang. TIMING: Temporality-Aware Integrated Gradients for Time Series Explanation, June 2025. URL http://arxiv.org/abs/2506.05035. arXiv:2506.05035 [cs].

[22] Changhun Kim, Yechan Mun, Sangchul Hahn, and Eunho Yang. DeltaSHAP: Explaining Prediction Evolutions in Online Patient Monitoring with Shapley Values, July 2025. URL http://arxiv.org/abs/2507.02342. arXiv:2507.02342 [cs].

[23] Davin Hill, Brian L. Hill, Aria Masoomi, Vijay S. Nori, Robert E. Tillman, and Jennifer Dy. Ordshap: Feature position importance for sequential black-box models, 2025. URL https://arxiv.org/abs/2507.11855.

[24] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017. URL http://arxiv.org/abs/1702.08608. arXiv:1702.08608 [stat].

[25] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, September 2018. URL http://arxiv.org/abs/1806.07421. arXiv:1806.07421 [cs].

[26] Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: unified approximation of any-order shapley interactions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 11515–11551, Red Hook, NY, USA, December 2023. Curran Associates Inc.

[27] Zhiyu Zhu, Huaming Chen, Jiayu Zhang, Xinyi Wang, Zhibo Jin, Minhui Xue, Dongxiao Zhu, and Kim-Kwang Raymond Choo. Mfaba: A more faithful and accelerated boundary-based attribution method for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17228–17236, 2024.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL https://dl.acm.org/doi/10.5555/3295222.3295349.

[29] Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness bias in model debugging. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Te5ytkqsnl.

[30] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7786–7795, Red Hook, NY, USA, 2018. Curran Associates Inc.

[31] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. Leveraging latent features for local explanations, 2021. URL https://arxiv.org/abs/1905.12698.

[32] Hadas Orgad, Fazl Barez, Tal Haklay, Isabelle Lee, Marius Mosbach, Anja Reusch, Naomi Saphra, Byron Wallace, Sarah Wiegreffe, Eric Wong, Ian Tenney, and Mor Geva. Interpretability can be actionable, 2026. URL https://arxiv.org/abs/2605.11161.

[33] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. URL https://arxiv.org/abs/1811.10154.

[34] Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310, 2021. doi: 10.1162/tacl_a_00367. URL https://aclanthology.org/2021.tacl-1.18/.

[35] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018. URL https://arxiv.org/abs/1706.07269.

[36] Helen Jin, Shreya Havaldar, Chaehyeon Kim, Anton Xue, Weiqiu You, Helen Qu, Marco Gatti, Daniel A Hashimoto, Bhuvnesh Jain, Amin Madani, Masao Sako, Lyle Ungar, and Eric Wong. The fix benchmark: Extracting features interpretable to experts, 2025. URL https://arxiv.org/abs/2409.13684.

[37] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, November 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00208-9. URL https://doi.org/10.1016/S2589-7500(21)00208-9.

[38] Shreya Havaldar, Weiqiu You, Chaehyeon Kim, Anton Xue, Helen Jin, Marco Gatti, Bhuvnesh Jain, Helen Qu, Amin Madani, Daniel A. Hashimoto, Gary E. Weissman, Rajat Deo, Sameed Khatana, Lyle Ungar, and Eric Wong. T-fix: Text-based explanations with features interpretable to experts, 2026. URL https://arxiv.org/abs/2511.04070.

[39] Laura Spillner, Rachel Ringe, Robert Porzel, and Rainer Malaka. Can ai explanations make you change your mind?, 2025. URL https://arxiv.org/abs/2508.08158.

[40] Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models, 2023. URL https://arxiv.org/abs/2310.01427.

[41] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods, 2017. URL https://arxiv.org/abs/1711.00867.

[42] Otavio T. Ranzani, Mervyn Singer, Jorge I. F. Salluh, Manu Shankar-Hari, David Pilcher, Joana Berger-Estilita, Craig M. Coopersmith, Nicole P. Juffermans, John Laffey, Matti Reinikainen, Ary Serpa Neto, Miguel Tavares, Jean-François Timsit, Maria Del Pilar Arias Lopez, Nish Arulkumaran, Diptesh Aryal, Elie Azoulay, Leo Anthony Celi, Dipayan Chaudhuri, Dylan De Lange, Jan De Waele, Claudia C. Dos Santos, Bin Du, Sharon Einav, Teresa Engelbrecht, Fathima Fazla, Ricard Ferrer, Stefano Finazzi, Tomoko Fujii, Hayley B. Gershengorn, John D. Greene, Rashan Haniffa, Sicheng Hao, Mohd Shahnaz Hasan, Steve Hollenberg, Mariachiara Ippolito, Christian Jung, Mikhail Kirov, Shigetaka Kobari, Inès Lakbar, Jeffrey Lipman, Vincent Liu, Xiaoli Liu, Suzana M. Lobo, Demetrio Magatti, Greg S. Martin, Barbara Metnitz, Philipp Metnitz, Sheila N. Myatra, Simon Oczkowski, José-Artur Paiva, Fathima Paruk, Pirkka T. Pekkarinen, Lise Piquilloud, Anssi Pölkki, Hallie C. Prescott, Annika Reintam Blaser, Ederlon Rezende, Chiara Robba, Bram Rochwerg, Stephane Ruckly, Rasoul Samei, Edward J. Schenck, Paul Secombe, Cornelius Sendagire, Moses Siaw-Frimpong, Andrew J. Simpkin, Márcio Soares, Charlotte Summers, Wojciech Szczeklik, Jukka Takala, Shiro Tanaka, Giovanni Tricella, Jean-Louis Vincent, Julia Wendon, Fernando G. Zampieri, Andrew Rhodes, and Rui Moreno. Development and validation of the sequential organ failure assessment (sofa)-2 score. *JAMA*, 334(23):2090–2103, 12 2025. ISSN 0098-7484. doi: 10.1001/jama.2025.20516. URL https://doi.org/10.1001/jama.2025.20516.

[43] Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *ICML*, 2025.

[44] Jonathan Kamp, Roos Bakker, and Dominique Blok. Explanation bias is a product: Revealing the hidden lexical and position preferences in post-hoc feature attribution, 2025. URL https://arxiv.org/abs/2512.11108.

[45] Quinn Dickinson and Jesse G. Meyer. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. *PLOS Computational Biology*, 18(1):1–24, January 2022. doi: 10.1371/journal.pcbi.1009736. URL https://doi.org/10.1371/journal.pcbi.1009736.

[46] Yarden Bakish, Itamar Zimerman, Hila Chefer, and Lior Wolf. Revisiting lrp: Positional attribution as the missing ingredient for transformer explainability, 2025. URL https://arxiv.org/abs/2506.02138.

[47] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-Scale Unsupervised Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7457–7476, June 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3218275. URL https://ieeexplore.ieee.org/document/9933726.

[48] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. ISSN 1573-1405. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

[49] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[50] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

[51] Matthew Reyna, Chris Josef, Russell Jeter, Supreeth Shashikumar, Benjamin Moody, M. Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D. Clifford. Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. *PhysioNet*, August 2019. doi: 10.13026/v64v-d857. URL https://doi.org/10.13026/v64v-d857. Version 1.0.0.

[52] Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, July 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05419-5. URL https://doi.org/10.1038/s41597-025-05419-5.

[53] Mayank Keoliya, Seewon Choi, Rajeev Alur, Mayur Naik, and Eric Wong. Stable prediction of adverse events in medical time-series data, 2025. URL https://arxiv.org/abs/2510.14286.

[54] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[56] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah,

Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo

Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

[57] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[58] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

[59] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026. URL `https://arxiv.org/abs/2507.05201`.

A Attribution Methods

A.1 Feature Attributions

This section briefly describes the feature attribution methods we instantiate with PaX. They fall into two groups: perturbation-based (LIME, SHAP, RISE, SHAP-IQ) and gradient-based (IntGrad, FullGrad, MFABA). We denote the instantiation of a method A as PaX-A, such as PaX-SHAP or PaX-IntGrad.

- **LIME** (Local Interpretable Model-agnostic Explanations) [12], in its sparse linear formulation, randomly masks interpretable components of the input (e.g., superpixels or words), queries the black-box model on the perturbed samples, and fits a sparse linear surrogate to the binary masks using proximity-weighted squared error. Each coefficient represents the corresponding component's contribution in this local surrogate.
- **SHAP** (SHapley Additive exPlanations) [14] assigns feature attributions using Shapley values. Its model-agnostic estimator, Kernel SHAP, fits an additive linear model to coalition values using binary feature indicators and the Shapley kernel. For a specified coalition value function, unregularized weighted least squares over all coalitions recovers the exact Shapley values when the empty- and full-coalition constraints are enforced; sampling coalitions provides an approximation.
- **IntGrad** (Integrated Gradients) [13] attributes a scalar model output to input features by integrating its gradients along the straight-line path from a baseline, typically chosen to represent the absence of signal, to the input, and multiplying each integrated gradient by the corresponding input–baseline difference. The exact attributions satisfy completeness, summing to the output difference between the input and the baseline, while numerical integration approximates this identity.
- **RISE** (Randomized Input Sampling for Explanation) [25] estimates pixel importance using random soft masks generated by bilinearly upsampling low-resolution binary grids and randomly shifting and cropping them. It queries the black-box model on each masked input and forms a weighted sum of the masks using the target-class scores, normalized by the number of samples and the expected mask value.
- **SHAP-IQ** (SHAPley Interaction Quantification) [26] estimates cardinal interaction indices, including Shapley interaction indices, by rewriting their weighted discrete derivatives as sums over coalition values. It reuses each coalition evaluation to update multiple interaction estimates, with weights determined by the chosen index, interaction order, coalition size, and overlap with the feature set being explained. Its estimator combines exact evaluation of selected coalition sizes with sampling-based estimation of the remaining terms, correcting for the sampling probabilities.
- **FullGrad** [16] constructs a saliency map for convolutional networks by aggregating the input-gradient-times-input term and spatial bias-contribution maps, obtained by multiplying gradients with respect to convolutional pre-activations by their corresponding broadcast bias values. It applies absolute values, rescales the maps to $[0, 1]$, and uses bilinear interpolation to match the input resolution before aggregation. The underlying full-gradient decomposition exactly reconstructs a ReLU network's scalar output when all bias contributions are included, whereas the processed saliency map does not generally preserve this equality.
- **MFABA** (More Faithful and Accelerated Boundary-based Attribution) [27] constructs an adversarial perturbation path from the input using gradient-based updates, stopping when the predicted class changes or a maximum iteration count is reached. It accumulates elementwise products of input displacements and the average attribution gradients at adjacent path points. This trapezoidal accumulation is motivated by a second-order Taylor expansion and uses a model-dependent perturbation path rather than a straight-line path to a manually specified baseline.

A.2 Position-aware eXplanation (PaX)

We provide additional examples of the PaX family beyond PaX-LIME introduced in Section 2.2.3, demonstrating that the framework generalizes across perturbation-based and gradient-based feature attribution methods.

A.2.1 Example: PaX-SHAP

We similarly instantiate PaX with SHAP [14] by treating positions as additional players in a coalition. We focus on KernelSHAP, which has three main steps: (1) sample coalitions, (2) evaluate the model on each coalition, and (3) fit a weighted linear regression under the Shapley kernel, whose solution estimates the Shapley values.

1. **Sampling coalitions.** Under our convention, a coalition is a binary feature mask $z^x \in \{0,1\}^n$ with $z^x_i = 0$ if feature i is present and $z^x_i = 1$ if it is missing. PaX-SHAP treats positions as additional players and samples a position mask $z^p \in \{0,1\}^n$ under the same convention. The positions with $z^p_i = 1$ form the pool that POS-SWAP shuffles. As in PaX-LIME, the explainer does not sample z^q. The drawn $\tilde{p}$ determines it, with $z^q_{ij} = 1$ if $\tilde{p}_i = j$ and $z^q_{ij} = 0$ otherwise.
2. **Making predictions.** KernelSHAP evaluates each coalition by mapping the mask back to input space and querying the model, $f_c(g(x, z^x))$. PaX-SHAP passes one draw of the joint perturbation to the PaX model and records $f'_c(g(x, z^x), \tilde{p})$, which gives an unbiased single-sample estimate of $v(z')$ under Mechanism 1. Since $\tilde{p}$ determines z^q, recording z^q requires no additional forward pass.
3. **Fitting the weighted regression.** PaX-SHAP repeats steps (1) and (2) to collect N samples $\{(z''^{(k)}, y''^{(k)})\}_{k=1}^N$ and weights each by the Shapley kernel over the $2n$ players (z^x, z^p), which assigns higher weight to very small and very large coalitions. KernelSHAP regresses the predictions onto z^x. PaX-SHAP instead regresses y'' onto the augmented masks $z'' = (z^x, z^p, z^q)$, and the negated coefficients give $\alpha'' = (\alpha^x, \alpha^p, \alpha^q)$. Because z^q correlates with z^p, these coefficients are kernel-weighted regression coefficients rather than exact Shapley values of the $2n$-player game.

A.2.2 Example: PaX-IntGrad

We similarly instantiate PaX within the Integrated Gradients paradigm [13] by applying its path integral to the value function v in Eq. (5). Standard IntGrad has three main steps: (1) choose a baseline and a straight-line path to the input, (2) compute gradients at points along the path, and (3) average the gradients and scale them by the difference between the input and the baseline.

1. **Choosing a baseline and path.** IntGrad interpolates along a straight line from the baseline $g(x, \mathbf{1})$ to the input x. PaX-IntGrad follows this path in the reference vector, from $z' = \mathbf{1}$ to $z' = \mathbf{0}$. For content, z^x traces the same path as IntGrad. For positions, z^p_i is the inclusion probability of b_i, so the path runs from a fully shuffled sequence at $z^p = \mathbf{1}$, where $\tilde{p}$ is uniform over $\mathrm{Sym}([n])$, to the original order at $z^p = \mathbf{0}$. The model therefore only receives position indices in $\mathrm{Sym}([n])$.
2. **Computing derivatives along the path.** IntGrad computes the gradient of the model output with respect to the input at each path point. PaX-IntGrad evaluates evenly spaced points on this path and, at each point z', draws samples $(b, \tilde{p}) \sim$ POS-SWAP(z^p), logging b, z^q, and the output $f'_c\big(g(x, z^x), \tilde{p}\big)$. For content, it computes the IntGrad gradient on each draw and averages over the draws. For positions, PaX-IntGrad does not use a model gradient, since implicit-position models expose no positional pathway and interpolated indices would fall outside $\mathrm{Sym}([n])$. PaX-IntGrad therefore estimates, under Mechanism 2, the exact partial in Eq. (6) and the contrast in Eq. (8) at each z' from the logged outputs, replacing each conditional expectation with the mean output over the draws that satisfy its condition.
3. **Integrating along the path.** IntGrad averages the gradients over the path and scales them by $x - g(x, \mathbf{1})$. PaX-IntGrad does the same for content to obtain α^x. For positions and relocations, it integrates the quantities from Step 2 along the same path,

$$\alpha^p_i = \int_1^0 \frac{\partial v}{\partial z^p_i}\, dz^p_i, \qquad \alpha^q_{ij} = \int_1^0 C_{ij}\, dz^p_i, \tag{10}$$

   where both integrands are evaluated at $z' = z^p_i \mathbf{1}$. We approximate each integral by the average of the Step 2 estimates over the evaluated points, multiplied by -1 because each reference coordinate moves from 1 to 0.

The position and counterfactual attributions reuse the same forward passes and need no backward pass. Only the content attribution uses backward passes, as in standard IntGrad. Since $v(\mathbf{0}) = f_c(x)$,

the feature and position attributions together sum to $f_c(x) - v(\mathbf{1})$, up to quadrature and sampling error.

B Experimental Setup

Our experiments evaluate four aspects of the proposed framework and analyze its design choices. First, we measure performance against existing attribution baselines under standard feature faithfulness metrics. Second, we assess positional faithfulness, comparing against existing baselines as well as methods explicitly designed to attribute positional importance. Third, we evaluate counterfactual positional explanations, a new explanation type enabled by our framework. Finally, we conduct ablation studies that isolate the contribution of each component, demonstrating that the observed gains arise from deliberate design choices rather than incidental effects.

B.1 Dataset Description

B.1.1 Vision, Language, and Clinical Time-Series Benchmarks

Our evaluation spans six benchmarks across three modalities: two vision datasets (ImageNet-S, PASCAL VOC), two language datasets (GSM8K, MedMCQA), and two clinical time-series datasets (PhysioNet 2019, MC-MED). All datasets are publicly available and used in accordance with their respective licenses and terms of use.

ImageNet-S [47]. A segmentation-oriented benchmark derived from ImageNet, designed to evaluate spatial localization and object-centric reasoning through pixel-level annotations. *License: subject to the ImageNet terms of access (research, non-commercial use).*

PASCAL VOC (Visual Object Classes) [48]. A widely used visual recognition benchmark with object segmentation annotations, supporting evaluation of positional sensitivity in visual explanations across diverse object categories. *License: "flickr" terms of use (research use).*

GSM8K (Grade School Math 8K) [49]. A collection of grade-school math word problems requiring multi-step numerical reasoning, where correctness depends critically on the order and position of intermediate statements. *License: MIT.*

MedMCQA (Multiple-Choice Question Answering) [50]. A multiple-choice medical question answering benchmark covering a broad range of clinical topics, supporting evaluation of attribution methods in medically grounded reasoning tasks. *License: Apache 2.0.*

PhysioNet 2019 [51]. An ICU sepsis prediction dataset comprising over 40,000 patients with up to 40 clinical variables recorded hourly, totaling approximately 2.5 million hourly time windows. *License: Creative Commons Attribution 4.0 International (CC BY 4.0).*

MC-MED (Multimodal Clinical Monitoring in the Emergency Department) [52]. An emergency department sepsis prediction dataset comprising 118,385 visits from 70,545 unique patients, combining minute-level vital signs and continuous physiological waveforms with comprehensive clinical data. *License: PhysioNet Credentialed Health Data License 1.5.0; access requires agreement to the specified license terms.*

B.1.2 Sepsis Prediction Task Curation

We follow the sepsis labeling procedure from CAREBench [53], with adaptations specific to each dataset's clinical context and data availability.

PhysioNet 2019. Sepsis labels are pre-defined using Sepsis-3 criteria, which require both clinical suspicion of infection (evidenced by blood culture or intravenous antibiotic orders) and a two-point increase in the SOFA score.

MC-MED. Because MC-MED captures emergency department encounters rather than ICU stays, we adopt a two-stage labeling procedure tailored to this setting, using a prediction horizon of $h = 1.5$ hours.

1. **At-Risk Cohort Selection.** We restrict the cohort to patients meeting all of the following criteria:

- Admission through the emergency department.
- Temperature $< 36°$C or $> 38.5°$C within 24 hours of admission (`Temp_time`).
- At least one abnormal vital sign or laboratory value within 24 hours:
  - WBC count > 12K or < 4K/μL (`WBC_time`),
  - Heart rate > 90 bpm (`HR_time`),
  - Respiratory rate > 20 (`RR_time`).
- At least one of `WBC_time`, `HR_time`, or `RR_time` occurs within 12 hours of `Temp_time`.
- No intravenous antibiotic was administered at or before the time the first criterion was met.

2. **Sepsis Labeling.** We assign a positive label when emergency SOFA (eSOFA) criteria are met within ± 2 days of blood culture collection. The eSOFA criteria require both:
   - Evidence of presumed serious infection:
     - Blood culture obtained, and
     - ≥ 4 qualifying antimicrobial days (QADs).
   - Evidence of acute organ dysfunction, defined as any of the following: vasopressor initiation, mechanical ventilation, renal dysfunction, hepatic dysfunction, thrombocytopenia, or elevated serum lactate.

B.2 Model Description

This section describes the models used in our experiments. We consider pre-trained foundation models for evaluating attribution faithfulness on the vision and language benchmarks, and sequence models fine-tuned for sepsis prediction on the clinical time-series benchmarks. All models are publicly available and used in accordance with their respective licenses and terms of use.

B.2.1 Models for Vision Benchmarks

ViT-B/16 [54]. A Vision Transformer that partitions images into fixed-size patches and processes them as a sequence via self-attention. Its patch-based representation makes it well suited for analyzing spatial positional effects in visual explanations. *License: Apache 2.0.*

CLIP-ViT-L/14 [55]. A large-scale vision–language model trained via contrastive learning on paired image–text data. We analyze its vision encoder to evaluate positional attribution under multimodal supervision. *License: MIT.*

Llama 3.2 11B Vision Instruct [56]. A multimodal large language model that integrates visual representations with text for joint visual–textual reasoning. *License: Llama 3.2 Community License.*

B.2.2 Models for Language Benchmarks

Llama 3 8B Instruct [56]. A decoder-only language model fine-tuned for instruction following. We use this model to evaluate attribution faithfulness on text-based reasoning tasks under input reordering. *License: Meta Llama 3 Community License.*

B.2.3 Models for Time-Series Benchmarks

For the clinical time-series benchmarks, we use three sequence models fine-tuned for sepsis prediction on the CAREBench-curated datasets: GPT-2 Small (124M), Mamba-130M, and MedGemma 4B.

GPT-2 Small [57]. A 124M-parameter decoder-only transformer with 12 layers, a hidden dimension of 768, and causal self-attention. Its autoregressive formulation naturally captures temporal dependencies in patient trajectories. *License: MIT.*

Mamba-130M [58]. A 130M-parameter state-space model designed for efficient long-sequence modeling with linear computational complexity in sequence length. Its state-space formulation provides inductive biases that align well with physiological dynamics. *License: Apache 2.0.*

MedGemma 4B [59]. A 4B-parameter instruction-tuned multimodal model based on the Gemma 3 decoder-only transformer architecture and adapted for medical text and image comprehension. Its medical-domain pre-training provides clinical priors well suited to modeling complex relationships among physiological measurements in patient trajectories. *License: Health AI Developer Foundations License.*

Table 6: Predictive performance of the fine-tuned GPT-2, Mamba, and MedGemma models on PhysioNet 2019 and MC-MED. We report standard classification metrics for comparability with prior work, together with the Brier score used for model selection during training.

Dataset	Fine-tuned Model	Accuracy	F1	AUROC	AUPRC	Brier
PhysioNet	GPT-2	0.8135	0.6705	0.7082	0.6831	0.081
	Mamba	0.7318	0.7368	0.8039	0.7664	0.141
	MedGemma	0.8836	0.8622	0.9051	0.9141	0.082
MC-MED	GPT-2	0.7850	0.7784	0.8608	0.8376	0.085
	Mamba	0.8135	0.6851	0.8655	0.5923	0.111
	MedGemma	0.8554	0.8474	0.8868	0.8634	0.085

B.3 Experimental Details

Data Splits. We follow the standard splits provided by each benchmark where available. For datasets without predefined splits—specifically PhysioNet 2019 and MC-MED—we partition the data into 80% training, 10% validation, and 10% testing.

Training Setup. All sepsis prediction models are trained from scratch following the CAREBench [53] protocol, with one clinically motivated modification. Whereas CAREBench primarily reports AUROC, we select hyperparameters using the **Brier score** as the primary validation metric, which we found to be more reliable given the extremely low event rate in sepsis prediction. The remaining configuration is as follows:

- **Tokenization:** Dataset-specific tokenizers for medical codes, following CAREBench.
- **Training duration:** 100 epochs.
- **Hyperparameter selection:** Learning rate $\in \{1 \times 10^{-5},\ 5 \times 10^{-5},\ 1 \times 10^{-4}\}$, selected by validation Brier score.

For completeness, Table 6 reports standard predictive performance metrics alongside the Brier score used for model selection.

Evaluation Protocols. Attribution methods are assessed along three complementary axes of faithfulness.

- *Feature faithfulness* is measured using standard feature insertion and deletion tests [25, 31]. Features are ranked by their attribution magnitude $|\alpha_i^x|$ and progressively inserted or removed according to this ranking. The area under the resulting performance curve (AUC) is reported; higher insertion AUC and lower deletion AUC indicate more faithful feature attributions.
- *Position faithfulness* is measured by extending the insertion and deletion protocol to operate over positions rather than feature content. Position insertion starts from a random configuration and sequentially places features at the highest-ranked positions; position deletion starts from the original input and relocates features from highly ranked positions to random ones. As in the feature-based case, each curve is summarized by its AUC, where higher insertion AUC and lower deletion AUC indicate more faithful position attributions.
- *Counterfactual faithfulness* is measured by testing whether counterfactual scores α_{ij}^q predict the effect of relocating content across positions. For each input and each ordered pair $i \neq j$, content is physically relocated from position i to position j and the resulting change in model output is recorded. Faithfulness is quantified as the Pearson correlation between α_{ij}^q and this observed output change; a higher positive correlation indicates that the counterfactual scores capture meaningful positional effects.

Compute Resources. All experiments were conducted on a single NVIDIA A100 80GB PCIe GPU. Runtime varied by configuration: each experiment—comprising one dataset, one model, and seven attribution methods—required approximately 3 to 18 hours, depending on the complexity of the dataset and the computational demands of the model architecture.

C Additional Feature and Position Faithfulness Results

This section presents comprehensive insertion and deletion results across all experimental configurations, complementing the main-paper results reported in Tables 1 and 2.

C.1 Experimental Setup

For each explanation method, we evaluate four attribution approaches:

- *Feature-only* is the traditional perturbation-based feature attributions, which serve as the baseline for feature faithfulness comparisons.
- *Position-only* is the positional explanations obtained by perturbing only the positions of tokens, serving as the baseline for position faithfulness comparisons.
- *PaX-Feature* is the feature attribution component extracted from our PaX decomposition.
- *PaX-Position* is the position attribution component extracted from our PaX decomposition.

Faithfulness is assessed using two complementary metrics: insertion AUC, where higher values indicate more faithful attributions, and deletion AUC, where lower values indicate more faithful attributions.

C.2 Results and Discussion

Tables 7–11 present the complete results across all datasets and models. The findings are consistent with those reported in the main paper and extend them to additional settings.

Insertion tests. PaX-Feature consistently achieves higher insertion AUC than Feature-only, and PaX-Position consistently surpasses Position-only, across all datasets and model architectures. This indicates that the PaX decomposition more faithfully identifies features and positions that are genuinely important to model predictions.

Deletion tests. The advantage of PaX is further confirmed under deletion. PaX-Feature achieves lower deletion AUC than Feature-only, and PaX-Position outperforms Position-only, indicating that removing PaX-attributed features causes greater degradation in model performance — the expected behavior when attributions correctly identify the most influential inputs.

Taken together, these results corroborate the findings of Section 3.2 and Section 3.3 and confirm that the benefits of decomposing attributions into feature and positional components are robust across datasets, explanation methods, and model architectures.

D Additional Counterfactual Faithfulness Results

Table 12 shows across all domains, the correlations remain positive, indicating that counterfactual positional attributions consistently reflect the effects of relocating content across positions.

E Additional Ablation of Missingness Approximations

Table 13 extends the missingness approximation analysis to datasets and models used in our insertion faithfulness experiments. Across vision, language, and clinical time-series benchmarks, positional swapping consistently yields the highest insertion AUC for both PaX-Feature and PaX-Position. This supports our use of POS-SWAP as the default missingness approximation throughout the main experiments.

F Additional Details on the Sepsis Forecasting Case Study

Experimental setup. The case study uses MedGemma on the PhysioNet sepsis prediction task with LIME as the backbone attribution technique—the highest-performing configuration in our quantitative evaluation (Tables 6 and 10). For each of the 35 input measurements, PaX decomposes the

Table 7: Faithfulness comparison across explanation methods on ImageNet-S. We report AUC for insertion (higher is better) and deletion (lower is better) tests. Across methods and models, separating feature and positional contributions yields more faithful explanations.

Model	Metric	Explanation Method	Feature-only (Original Method)	PaX-Feature	Feature Improvement	Position-only	PaX-Position	Position Improvement
ViT	Insertion	LIME	0.354 ± 0.006	$\mathbf{0.419} \pm 0.005$	18.4%	0.323 ± 0.007	$\mathbf{0.396} \pm 0.006$	22.6%
		SHAP	0.342 ± 0.005	$\mathbf{0.403} \pm 0.004$	17.8%	0.337 ± 0.006	$\mathbf{0.401} \pm 0.005$	19.0%
		RISE	0.349 ± 0.007	$\mathbf{0.411} \pm 0.006$	17.8%	0.330 ± 0.008	$\mathbf{0.398} \pm 0.006$	20.6%
		SHAP-IQ	0.346 ± 0.006	$\mathbf{0.407} \pm 0.005$	17.6%	0.334 ± 0.007	$\mathbf{0.399} \pm 0.006$	19.5%
		IntGrad	0.361 ± 0.009	$\mathbf{0.427} \pm 0.008$	18.3%	0.346 ± 0.010	$\mathbf{0.412} \pm 0.008$	19.1%
		FullGrad	0.336 ± 0.012	$\mathbf{0.384} \pm 0.010$	14.3%	0.314 ± 0.013	$\mathbf{0.393} \pm 0.011$	25.2%
		MFABA	0.351 ± 0.008	$\mathbf{0.417} \pm 0.007$	18.8%	0.325 ± 0.009	$\mathbf{0.402} \pm 0.007$	23.7%
	Deletion	LIME	0.020 ± 0.001	$\mathbf{0.008} \pm 0.001$	60.0%	0.016 ± 0.001	$\mathbf{0.007} \pm 0.001$	56.2%
		SHAP	0.019 ± 0.001	$\mathbf{0.007} \pm 0.001$	63.2%	0.018 ± 0.001	$\mathbf{0.008} \pm 0.001$	55.6%
		RISE	0.019 ± 0.001	$\mathbf{0.009} \pm 0.001$	52.6%	0.017 ± 0.001	$\mathbf{0.008} \pm 0.001$	52.9%
		SHAP-IQ	0.019 ± 0.001	$\mathbf{0.008} \pm 0.001$	57.9%	0.017 ± 0.001	$\mathbf{0.007} \pm 0.001$	58.8%
		IntGrad	0.018 ± 0.001	$\mathbf{0.009} \pm 0.001$	50.0%	0.021 ± 0.001	$\mathbf{0.011} \pm 0.001$	47.6%
		FullGrad	0.017 ± 0.001	$\mathbf{0.010} \pm 0.001$	41.2%	0.019 ± 0.001	$\mathbf{0.010} \pm 0.001$	47.4%
		MFABA	0.018 ± 0.001	$\mathbf{0.007} \pm 0.001$	61.1%	0.015 ± 0.001	$\mathbf{0.006} \pm 0.001$	60.0%
CLIP	Insertion	LIME	0.381 ± 0.007	$\mathbf{0.442} \pm 0.006$	16.0%	0.343 ± 0.008	$\mathbf{0.411} \pm 0.007$	19.8%
		SHAP	0.364 ± 0.006	$\mathbf{0.427} \pm 0.005$	17.3%	0.352 ± 0.007	$\mathbf{0.425} \pm 0.006$	20.7%
		RISE	0.372 ± 0.008	$\mathbf{0.435} \pm 0.007$	16.9%	0.347 ± 0.009	$\mathbf{0.418} \pm 0.007$	20.5%
		SHAP-IQ	0.368 ± 0.007	$\mathbf{0.431} \pm 0.006$	17.1%	0.349 ± 0.008	$\mathbf{0.421} \pm 0.006$	20.6%
		IntGrad	0.392 ± 0.010	$\mathbf{0.454} \pm 0.009$	15.8%	0.372 ± 0.011	$\mathbf{0.435} \pm 0.009$	16.9%
		FullGrad	0.361 ± 0.013	$\mathbf{0.407} \pm 0.011$	12.7%	0.332 ± 0.014	$\mathbf{0.419} \pm 0.012$	26.2%
		MFABA	0.383 ± 0.009	$\mathbf{0.445} \pm 0.008$	16.2%	0.351 ± 0.010	$\mathbf{0.431} \pm 0.008$	22.8%
	Deletion	LIME	0.021 ± 0.001	$\mathbf{0.010} \pm 0.001$	52.4%	0.019 ± 0.001	$\mathbf{0.009} \pm 0.001$	52.6%
		SHAP	0.022 ± 0.001	$\mathbf{0.009} \pm 0.001$	59.1%	0.020 ± 0.001	$\mathbf{0.010} \pm 0.001$	50.0%
		RISE	0.021 ± 0.001	$\mathbf{0.009} \pm 0.001$	57.1%	0.019 ± 0.001	$\mathbf{0.009} \pm 0.001$	52.6%
		SHAP-IQ	0.019 ± 0.001	$\mathbf{0.009} \pm 0.001$	52.6%	0.019 ± 0.001	$\mathbf{0.009} \pm 0.001$	52.6%
		IntGrad	0.020 ± 0.001	$\mathbf{0.011} \pm 0.001$	45.0%	0.022 ± 0.001	$\mathbf{0.013} \pm 0.001$	40.9%
		FullGrad	0.019 ± 0.001	$\mathbf{0.012} \pm 0.001$	36.8%	0.021 ± 0.001	$\mathbf{0.011} \pm 0.001$	47.6%
		MFABA	0.020 ± 0.001	$\mathbf{0.009} \pm 0.001$	55.0%	0.018 ± 0.001	$\mathbf{0.008} \pm 0.001$	55.6%
Llama	Insertion	LIME	0.386 ± 0.007	$\mathbf{0.456} \pm 0.006$	18.1%	0.348 ± 0.008	$\mathbf{0.433} \pm 0.007$	24.4%
		SHAP	0.369 ± 0.006	$\mathbf{0.444} \pm 0.005$	20.3%	0.357 ± 0.007	$\mathbf{0.441} \pm 0.006$	23.5%
		RISE	0.377 ± 0.008	$\mathbf{0.452} \pm 0.007$	19.9%	0.352 ± 0.009	$\mathbf{0.436} \pm 0.007$	23.9%
		SHAP-IQ	0.373 ± 0.007	$\mathbf{0.448} \pm 0.006$	20.1%	0.354 ± 0.008	$\mathbf{0.438} \pm 0.006$	23.7%
		IntGrad	0.397 ± 0.010	$\mathbf{0.472} \pm 0.009$	18.9%	0.377 ± 0.011	$\mathbf{0.456} \pm 0.009$	21.0%
		FullGrad	0.366 ± 0.013	$\mathbf{0.426} \pm 0.011$	16.4%	0.337 ± 0.014	$\mathbf{0.439} \pm 0.012$	30.3%
		MFABA	0.388 ± 0.009	$\mathbf{0.458} \pm 0.008$	18.0%	0.356 ± 0.010	$\mathbf{0.445} \pm 0.008$	25.0%
	Deletion	LIME	0.006 ± 0.001	$\mathbf{0.003} \pm 0.001$	50.0%	0.005 ± 0.001	$\mathbf{0.003} \pm 0.001$	40.0%
		SHAP	0.007 ± 0.001	$\mathbf{0.004} \pm 0.001$	42.9%	0.006 ± 0.001	$\mathbf{0.004} \pm 0.001$	33.3%
		RISE	0.006 ± 0.001	$\mathbf{0.004} \pm 0.001$	33.3%	0.005 ± 0.001	$\mathbf{0.003} \pm 0.001$	40.0%
		SHAP-IQ	0.005 ± 0.001	$\mathbf{0.003} \pm 0.001$	40.0%	0.005 ± 0.001	$\mathbf{0.003} \pm 0.001$	40.0%
		IntGrad	0.008 ± 0.001	$\mathbf{0.005} \pm 0.001$	37.5%	0.007 ± 0.001	$\mathbf{0.004} \pm 0.001$	42.9%
		FullGrad	0.006 ± 0.001	$\mathbf{0.004} \pm 0.001$	33.3%	0.006 ± 0.001	$\mathbf{0.004} \pm 0.001$	33.3%
		MFABA	0.007 ± 0.001	$\mathbf{0.004} \pm 0.001$	42.9%	0.006 ± 0.001	$\mathbf{0.003} \pm 0.001$	50.0%

LIME attribution into a feature component $\alpha^{(x)}_{\text{PaX-LIME}}$ and a position component $\alpha^{(p)}_{\text{PaX-LIME}}$ following Section 2.2. Clinician annotations along the Sepsis-Related and Time-Dependent axes (defined below) were collected from a medical expert blinded to the attribution scores.

F.1 Clinician annotation protocol

Each measurement was annotated along two axes by a medical expert.

Sepsis-Related axis. Three labels: "yes" for direct indicators of sepsis risk; "partial" for indirect or contextual markers; "no" for unrelated markers. The "no" label was reserved for features that primarily indicate a different pathology (e.g., lipase, which marks pancreatitis), even if they co-occur or appear on the differential diagnosis with sepsis.

Time-Dependent axis. Two labels: "yes" if a measurement's temporal trajectory carries diagnostic meaning (e.g., vital signs), or "no" if the marker is largely stable across a hospitalization (e.g., demographics). Of the 35 measurements, 33 were labeled "yes" and 2 (Age, Gender) "no."

Table 8: Faithfulness comparison across explanation methods on PASCAL VOC. We report AUC for insertion (higher is better) and deletion (lower is better) tests. Across methods and models, separating feature and positional contributions yields more faithful explanations.

Model	Metric	Explanation Method	Feature-only (Original Method)	PaX-Feature	Feature Improvement	Position-only	PaX-Position	Position Improvement
ViT	Insertion	LIME	0.359 ± 0.006	$\mathbf{0.424} \pm 0.005$	18.1%	0.326 ± 0.007	$\mathbf{0.400} \pm 0.006$	22.7%
		SHAP	0.345 ± 0.005	$\mathbf{0.408} \pm 0.004$	18.3%	0.341 ± 0.006	$\mathbf{0.405} \pm 0.005$	18.8%
		RISE	0.352 ± 0.007	$\mathbf{0.416} \pm 0.006$	18.2%	0.331 ± 0.008	$\mathbf{0.401} \pm 0.006$	21.1%
		SHAP-IQ	0.349 ± 0.006	$\mathbf{0.411} \pm 0.005$	17.8%	0.337 ± 0.007	$\mathbf{0.402} \pm 0.006$	19.3%
		IntGrad	0.365 ± 0.009	$\mathbf{0.431} \pm 0.008$	18.1%	0.350 ± 0.010	$\mathbf{0.413} \pm 0.007$	18.0%
		FullGrad	0.337 ± 0.010	$\mathbf{0.389} \pm 0.010$	15.4%	0.318 ± 0.013	$\mathbf{0.396} \pm 0.011$	24.5%
		MFABA	0.355 ± 0.010	$\mathbf{0.421} \pm 0.007$	18.6%	0.328 ± 0.009	$\mathbf{0.406} \pm 0.007$	23.8%
	Deletion	LIME	0.022 ± 0.001	$\mathbf{0.009} \pm 0.001$	59.1%	0.018 ± 0.001	$\mathbf{0.008} \pm 0.001$	55.6%
		SHAP	0.020 ± 0.001	$\mathbf{0.008} \pm 0.001$	60.0%	0.019 ± 0.001	$\mathbf{0.009} \pm 0.001$	52.6%
		RISE	0.015 ± 0.001	$\mathbf{0.010} \pm 0.001$	33.3%	0.018 ± 0.001	$\mathbf{0.009} \pm 0.001$	50.0%
		SHAP-IQ	0.020 ± 0.001	$\mathbf{0.009} \pm 0.001$	55.0%	0.018 ± 0.001	$\mathbf{0.008} \pm 0.001$	55.6%
		IntGrad	0.019 ± 0.001	$\mathbf{0.010} \pm 0.001$	47.4%	0.023 ± 0.001	$\mathbf{0.012} \pm 0.001$	47.8%
		FullGrad	0.018 ± 0.001	$\mathbf{0.011} \pm 0.001$	38.9%	0.021 ± 0.001	$\mathbf{0.011} \pm 0.001$	47.6%
		MFABA	0.019 ± 0.001	$\mathbf{0.008} \pm 0.001$	57.9%	0.016 ± 0.001	$\mathbf{0.007} \pm 0.001$	56.2%
CLIP	Insertion	LIME	0.412 ± 0.005	$\mathbf{0.432} \pm 0.004$	4.9%	0.368 ± 0.006	$\mathbf{0.441} \pm 0.005$	19.8%
		SHAP	0.385 ± 0.008	$\mathbf{0.456} \pm 0.007$	18.4%	0.372 ± 0.008	$\mathbf{0.449} \pm 0.006$	20.7%
		RISE	0.394 ± 0.007	$\mathbf{0.465} \pm 0.005$	18.0%	0.355 ± 0.009	$\mathbf{0.432} \pm 0.008$	21.7%
		SHAP-IQ	0.388 ± 0.006	$\mathbf{0.459} \pm 0.005$	18.3%	0.361 ± 0.007	$\mathbf{0.438} \pm 0.006$	21.3%
		IntGrad	0.403 ± 0.009	$\mathbf{0.471} \pm 0.008$	16.9%	0.382 ± 0.010	$\mathbf{0.437} \pm 0.008$	14.4%
		FullGrad	0.374 ± 0.012	$\mathbf{0.428} \pm 0.006$	14.4%	0.342 ± 0.013	$\mathbf{0.436} \pm 0.011$	27.5%
		MFABA	0.401 ± 0.008	$\mathbf{0.474} \pm 0.010$	18.2%	0.366 ± 0.009	$\mathbf{0.451} \pm 0.008$	23.2%
	Deletion	LIME	0.028 ± 0.001	$\mathbf{0.014} \pm 0.001$	50.0%	0.024 ± 0.001	$\mathbf{0.012} \pm 0.001$	50.0%
		SHAP	0.026 ± 0.001	$\mathbf{0.012} \pm 0.001$	53.8%	0.025 ± 0.001	$\mathbf{0.013} \pm 0.001$	48.0%
		RISE	0.027 ± 0.001	$\mathbf{0.013} \pm 0.001$	51.9%	0.023 ± 0.001	$\mathbf{0.011} \pm 0.001$	52.2%
		SHAP-IQ	0.025 ± 0.001	$\mathbf{0.012} \pm 0.001$	52.0%	0.022 ± 0.001	$\mathbf{0.011} \pm 0.001$	50.0%
		IntGrad	0.024 ± 0.001	$\mathbf{0.015} \pm 0.001$	37.5%	0.026 ± 0.001	$\mathbf{0.015} \pm 0.001$	42.3%
		FullGrad	0.022 ± 0.001	$\mathbf{0.016} \pm 0.001$	27.3%	0.024 ± 0.001	$\mathbf{0.014} \pm 0.001$	41.7%
		MFABA	0.025 ± 0.001	$\mathbf{0.011} \pm 0.001$	56.0%	0.021 ± 0.001	$\mathbf{0.010} \pm 0.001$	52.4%
Llama	Insertion	LIME	0.401 ± 0.007	$\mathbf{0.462} \pm 0.006$	15.2%	0.366 ± 0.008	$\mathbf{0.437} \pm 0.004$	19.4%
		SHAP	0.392 ± 0.006	$\mathbf{0.455} \pm 0.005$	16.1%	0.381 ± 0.007	$\mathbf{0.452} \pm 0.005$	18.6%
		RISE	0.398 ± 0.008	$\mathbf{0.459} \pm 0.007$	15.3%	0.372 ± 0.009	$\mathbf{0.441} \pm 0.006$	18.5%
		SHAP-IQ	0.395 ± 0.007	$\mathbf{0.457} \pm 0.006$	15.7%	0.374 ± 0.008	$\mathbf{0.444} \pm 0.006$	18.7%
		IntGrad	0.417 ± 0.010	$\mathbf{0.476} \pm 0.009$	14.1%	0.392 ± 0.011	$\mathbf{0.445} \pm 0.006$	13.5%
		FullGrad	0.386 ± 0.013	$\mathbf{0.431} \pm 0.011$	11.7%	0.358 ± 0.014	$\mathbf{0.445} \pm 0.009$	24.3%
		MFABA	0.403 ± 0.009	$\mathbf{0.464} \pm 0.008$	15.1%	0.369 ± 0.010	$\mathbf{0.448} \pm 0.007$	21.4%
	Deletion	LIME	0.019 ± 0.001	$\mathbf{0.007} \pm 0.001$	63.2%	0.017 ± 0.001	$\mathbf{0.006} \pm 0.001$	64.7%
		SHAP	0.020 ± 0.001	$\mathbf{0.006} \pm 0.001$	70.0%	0.018 ± 0.001	$\mathbf{0.007} \pm 0.001$	61.1%
		RISE	0.019 ± 0.001	$\mathbf{0.007} \pm 0.001$	63.2%	0.017 ± 0.001	$\mathbf{0.006} \pm 0.001$	64.7%
		SHAP-IQ	0.018 ± 0.001	$\mathbf{0.007} \pm 0.001$	61.1%	0.017 ± 0.001	$\mathbf{0.006} \pm 0.001$	64.7%
		IntGrad	0.019 ± 0.001	$\mathbf{0.008} \pm 0.001$	57.9%	0.021 ± 0.001	$\mathbf{0.011} \pm 0.001$	47.6%
		FullGrad	0.018 ± 0.001	$\mathbf{0.009} \pm 0.001$	50.0%	0.020 ± 0.001	$\mathbf{0.008} \pm 0.001$	60.0%
		MFABA	0.019 ± 0.001	$\mathbf{0.006} \pm 0.001$	68.4%	0.016 ± 0.002	$\mathbf{0.005} \pm 0.001$	68.8%

F.2 Bedside-relevant features beyond formal sepsis criteria

We cross-referenced clinician Sepsis-Related annotations against established sepsis criteria. The 13 literature-flagged measurements (Bilirubin, Creatinine, DBP[3], FiO2, HR, Lactate, MAP, O2Sat, Platelets, Resp, SBP, SaO2, Temp) were independently labeled "yes" by clinicians. Table 5 reports clinician labels for the 22 measurements absent from the literature: roughly two-thirds carry diagnostic value at the bedside.

G Extended Related Work

We situate PaX with respect to three threads of prior work: standard feature attribution, attribution methods adapted to temporal structure, and methods that explicitly target positional effects.

Positional effects and standard feature attribution. Recent work shows that model predictions depend not only on input feature values but also on their positions, with effects ranging from degraded

[3]DBP is included as it is used to compute MAP.

Table 9: Faithfulness comparison across explanation methods on Language benchmarks using Llama 3 8B Instruct. We report AUC for insertion (higher is better) and deletion (lower is better) tests. Across methods, PaX-Feature consistently outperforms Feature-only and PaX-Position outperforms Position-only, demonstrating that separating feature and positional contributions yields more faithful explanations.

Dataset	Metric	Explanation Method	Feature-only (Original Method)	PaX-Feature	Feature Improvement	Position-only	PaX-Position	Position Improvement
GSM8K	Insertion	LIME	0.347 ± 0.006	$\mathbf{0.355} \pm 0.005$	2.3%	0.337 ± 0.007	$\mathbf{0.343} \pm 0.006$	1.8%
		SHAP	0.392 ± 0.005	$\mathbf{0.405} \pm 0.004$	3.3%	0.387 ± 0.006	$\mathbf{0.393} \pm 0.005$	1.6%
		SHAP-IQ	0.446 ± 0.006	$\mathbf{0.458} \pm 0.005$	2.7%	0.437 ± 0.007	$\mathbf{0.443} \pm 0.006$	1.4%
		IntGrad	0.463 ± 0.008	$\mathbf{0.476} \pm 0.007$	2.8%	0.458 ± 0.009	$\mathbf{0.465} \pm 0.008$	1.5%
		FullGrad	0.431 ± 0.007	$\mathbf{0.443} \pm 0.006$	2.8%	0.422 ± 0.008	$\mathbf{0.430} \pm 0.007$	1.9%
		MFABA	0.454 ± 0.006	$\mathbf{0.467} \pm 0.005$	2.9%	0.445 ± 0.007	$\mathbf{0.452} \pm 0.006$	1.6%
	Deletion	LIME	0.284 ± 0.004	$\mathbf{0.270} \pm 0.003$	4.9%	0.283 ± 0.004	$\mathbf{0.276} \pm 0.003$	2.5%
		SHAP	0.257 ± 0.004	$\mathbf{0.238} \pm 0.003$	7.4%	0.253 ± 0.004	$\mathbf{0.244} \pm 0.003$	3.6%
		SHAP-IQ	0.234 ± 0.004	$\mathbf{0.215} \pm 0.003$	8.1%	0.232 ± 0.004	$\mathbf{0.224} \pm 0.003$	3.4%
		IntGrad	0.229 ± 0.005	$\mathbf{0.205} \pm 0.004$	10.5%	0.223 ± 0.005	$\mathbf{0.214} \pm 0.004$	4.0%
		FullGrad	0.246 ± 0.005	$\mathbf{0.225} \pm 0.004$	8.5%	0.241 ± 0.005	$\mathbf{0.232} \pm 0.004$	3.7%
		MFABA	0.236 ± 0.004	$\mathbf{0.214} \pm 0.003$	9.3%	0.230 ± 0.004	$\mathbf{0.221} \pm 0.003$	3.9%
MedMCQA	Insertion	LIME	0.368 ± 0.006	$\mathbf{0.378} \pm 0.005$	2.7%	0.357 ± 0.007	$\mathbf{0.363} \pm 0.006$	1.7%
		SHAP	0.416 ± 0.005	$\mathbf{0.429} \pm 0.004$	3.1%	0.407 ± 0.006	$\mathbf{0.414} \pm 0.005$	1.7%
		SHAP-IQ	0.462 ± 0.006	$\mathbf{0.479} \pm 0.005$	3.7%	0.457 ± 0.007	$\mathbf{0.463} \pm 0.006$	1.3%
		IntGrad	0.487 ± 0.008	$\mathbf{0.498} \pm 0.007$	2.3%	0.477 ± 0.009	$\mathbf{0.484} \pm 0.008$	1.5%
		FullGrad	0.438 ± 0.007	$\mathbf{0.451} \pm 0.006$	3.0%	0.429 ± 0.008	$\mathbf{0.436} \pm 0.007$	1.6%
		MFABA	0.471 ± 0.006	$\mathbf{0.486} \pm 0.005$	3.2%	0.462 ± 0.007	$\mathbf{0.469} \pm 0.006$	1.5%
	Deletion	LIME	0.259 ± 0.004	$\mathbf{0.238} \pm 0.003$	8.1%	0.253 ± 0.004	$\mathbf{0.244} \pm 0.003$	3.6%
		SHAP	0.226 ± 0.004	$\mathbf{0.205} \pm 0.003$	9.3%	0.223 ± 0.004	$\mathbf{0.214} \pm 0.003$	4.0%
		SHAP-IQ	0.207 ± 0.004	$\mathbf{0.185} \pm 0.003$	10.6%	0.202 ± 0.004	$\mathbf{0.194} \pm 0.003$	4.0%
		IntGrad	0.194 ± 0.005	$\mathbf{0.175} \pm 0.004$	9.8%	0.192 ± 0.005	$\mathbf{0.184} \pm 0.004$	4.2%
		FullGrad	0.218 ± 0.005	$\mathbf{0.197} \pm 0.004$	9.6%	0.213 ± 0.005	$\mathbf{0.204} \pm 0.004$	4.2%
		MFABA	0.203 ± 0.004	$\mathbf{0.183} \pm 0.003$	9.9%	0.199 ± 0.004	$\mathbf{0.191} \pm 0.003$	4.0%

Table 10: Faithfulness comparison across explanation methods on PhysioNet. We report AUC for insertion (higher is better) and deletion (lower is better). Across methods, PaX-Feature consistently outperforms Feature-only and PaX-Position outperforms Position-only, demonstrating that separating feature and positional contributions yields more faithful explanations.

Model	Metric	Explanation Method	Feature-only (Original Method)	PaX-Feature	Feature Improvement	Position-only	PaX-Position	Position Improvement
GPT-2	Insertion	LIME	0.332 ± 0.006	$\mathbf{0.354} \pm 0.005$	6.6%	0.325 ± 0.007	$\mathbf{0.341} \pm 0.006$	4.9%
		SHAP	0.341 ± 0.006	$\mathbf{0.362} \pm 0.005$	6.2%	0.336 ± 0.007	$\mathbf{0.348} \pm 0.006$	3.6%
		IntGrad	0.356 ± 0.008	$\mathbf{0.378} \pm 0.007$	6.2%	0.348 ± 0.008	$\mathbf{0.366} \pm 0.007$	5.2%
		FullGrad	0.324 ± 0.007	$\mathbf{0.346} \pm 0.006$	6.8%	0.318 ± 0.008	$\mathbf{0.334} \pm 0.007$	5.0%
		MFABA	0.339 ± 0.006	$\mathbf{0.361} \pm 0.005$	6.5%	0.327 ± 0.007	$\mathbf{0.349} \pm 0.006$	6.7%
	Deletion	LIME	0.152 ± 0.004	$\mathbf{0.138} \pm 0.003$	9.2%	0.146 ± 0.004	$\mathbf{0.141} \pm 0.003$	3.4%
		SHAP	0.149 ± 0.004	$\mathbf{0.134} \pm 0.003$	10.1%	0.148 ± 0.004	$\mathbf{0.139} \pm 0.003$	6.1%
		IntGrad	0.151 ± 0.005	$\mathbf{0.137} \pm 0.004$	9.3%	0.154 ± 0.005	$\mathbf{0.142} \pm 0.004$	7.8%
		FullGrad	0.147 ± 0.005	$\mathbf{0.135} \pm 0.004$	8.2%	0.150 ± 0.005	$\mathbf{0.138} \pm 0.004$	8.0%
		MFABA	0.150 ± 0.004	$\mathbf{0.136} \pm 0.003$	9.3%	0.145 ± 0.004	$\mathbf{0.140} \pm 0.003$	3.4%
Mamba	Insertion	LIME	0.335 ± 0.006	$\mathbf{0.357} \pm 0.005$	6.6%	0.331 ± 0.007	$\mathbf{0.344} \pm 0.006$	3.9%
		SHAP	0.343 ± 0.006	$\mathbf{0.366} \pm 0.005$	6.7%	0.323 ± 0.007	$\mathbf{0.341} \pm 0.006$	5.6%
		IntGrad	0.358 ± 0.008	$\mathbf{0.381} \pm 0.007$	6.4%	0.348 ± 0.008	$\mathbf{0.369} \pm 0.007$	6.0%
		FullGrad	0.329 ± 0.007	$\mathbf{0.351} \pm 0.006$	6.7%	0.312 ± 0.008	$\mathbf{0.333} \pm 0.007$	6.7%
		MFABA	0.346 ± 0.006	$\mathbf{0.369} \pm 0.005$	6.6%	0.334 ± 0.007	$\mathbf{0.352} \pm 0.006$	5.4%
	Deletion	LIME	0.153 ± 0.004	$\mathbf{0.139} \pm 0.003$	9.2%	0.149 ± 0.004	$\mathbf{0.142} \pm 0.003$	4.7%
		SHAP	0.154 ± 0.004	$\mathbf{0.141} \pm 0.003$	8.4%	0.148 ± 0.004	$\mathbf{0.143} \pm 0.003$	3.4%
		IntGrad	0.145 ± 0.005	$\mathbf{0.133} \pm 0.004$	8.3%	0.147 ± 0.005	$\mathbf{0.136} \pm 0.004$	7.5%
		FullGrad	0.144 ± 0.005	$\mathbf{0.131} \pm 0.004$	9.0%	0.146 ± 0.005	$\mathbf{0.134} \pm 0.004$	8.2%
		MFABA	0.152 ± 0.004	$\mathbf{0.138} \pm 0.003$	9.2%	0.147 ± 0.004	$\mathbf{0.141} \pm 0.003$	4.1%
MedGemma	Insertion	LIME	0.338 ± 0.005	$\mathbf{0.344} \pm 0.004$	1.8%	0.334 ± 0.006	$\mathbf{0.340} \pm 0.005$	1.8%
		SHAP	0.346 ± 0.005	$\mathbf{0.352} \pm 0.004$	1.7%	0.329 ± 0.006	$\mathbf{0.335} \pm 0.005$	1.8%
		IntGrad	0.361 ± 0.007	$\mathbf{0.368} \pm 0.006$	1.9%	0.351 ± 0.007	$\mathbf{0.358} \pm 0.006$	2.0%
		FullGrad	0.331 ± 0.006	$\mathbf{0.337} \pm 0.005$	1.8%	0.319 ± 0.007	$\mathbf{0.325} \pm 0.006$	1.9%
		MFABA	0.348 ± 0.005	$\mathbf{0.355} \pm 0.004$	2.0%	0.337 ± 0.006	$\mathbf{0.344} \pm 0.005$	2.1%
	Deletion	LIME	0.151 ± 0.004	$\mathbf{0.145} \pm 0.003$	4.0%	0.150 ± 0.004	$\mathbf{0.144} \pm 0.003$	4.0%
		SHAP	0.152 ± 0.004	$\mathbf{0.146} \pm 0.003$	3.9%	0.149 ± 0.004	$\mathbf{0.143} \pm 0.003$	4.0%
		IntGrad	0.148 ± 0.005	$\mathbf{0.142} \pm 0.004$	4.1%	0.151 ± 0.005	$\mathbf{0.145} \pm 0.004$	4.0%
		FullGrad	0.146 ± 0.005	$\mathbf{0.140} \pm 0.004$	4.1%	0.147 ± 0.005	$\mathbf{0.141} \pm 0.004$	4.1%
		MFABA	0.153 ± 0.004	$\mathbf{0.147} \pm 0.003$	3.9%	0.148 ± 0.004	$\mathbf{0.142} \pm 0.003$	4.1%

long-context retrieval to systematic biases in attention [3, 43, 44]. These findings motivate attribution methods that account for position rather than feature values alone. Standard feature attribution

Table 11: Faithfulness comparison across explanation methods on MC-MED. We report AUC for insertion (higher is better) and deletion (lower is better). Across methods, PaX-Feature consistently outperforms Feature-only and PaX-Position outperforms Position-only, demonstrating that separating feature and positional contributions yields more faithful explanations.

Model	Metric	Explanation Method	Feature-only (Original Method)	PaX-Feature	Feature Improvement	Position-only	PaX-Position	Position Improvement
GPT-2	Insertion	LIME	0.318 ± 0.006	$\mathbf{0.339} \pm 0.005$	6.6%	0.301 ± 0.007	$\mathbf{0.327} \pm 0.006$	8.6%
		SHAP	0.327 ± 0.006	$\mathbf{0.349} \pm 0.005$	6.7%	0.314 ± 0.007	$\mathbf{0.336} \pm 0.006$	7.0%
		IntGrad	0.338 ± 0.008	$\mathbf{0.361} \pm 0.007$	6.8%	0.322 ± 0.008	$\mathbf{0.348} \pm 0.007$	8.1%
		FullGrad	0.309 ± 0.007	$\mathbf{0.331} \pm 0.006$	7.1%	0.296 ± 0.008	$\mathbf{0.319} \pm 0.007$	7.8%
		MFABA	0.324 ± 0.006	$\mathbf{0.346} \pm 0.005$	6.8%	0.312 ± 0.007	$\mathbf{0.334} \pm 0.006$	7.1%
	Deletion	LIME	0.162 ± 0.004	$\mathbf{0.148} \pm 0.003$	8.6%	0.171 ± 0.004	$\mathbf{0.156} \pm 0.003$	8.8%
		SHAP	0.168 ± 0.004	$\mathbf{0.151} \pm 0.003$	10.1%	0.164 ± 0.004	$\mathbf{0.159} \pm 0.003$	3.0%
		IntGrad	0.182 ± 0.005	$\mathbf{0.164} \pm 0.004$	9.9%	0.191 ± 0.005	$\mathbf{0.173} \pm 0.004$	9.4%
		FullGrad	0.179 ± 0.005	$\mathbf{0.162} \pm 0.004$	9.5%	0.187 ± 0.005	$\mathbf{0.171} \pm 0.004$	8.6%
		MFABA	0.165 ± 0.004	$\mathbf{0.149} \pm 0.003$	9.7%	0.169 ± 0.004	$\mathbf{0.155} \pm 0.003$	8.3%
Mamba	Insertion	LIME	0.321 ± 0.006	$\mathbf{0.343} \pm 0.005$	6.9%	0.311 ± 0.007	$\mathbf{0.331} \pm 0.006$	6.4%
		SHAP	0.334 ± 0.006	$\mathbf{0.356} \pm 0.005$	6.6%	0.322 ± 0.007	$\mathbf{0.344} \pm 0.006$	6.8%
		IntGrad	0.341 ± 0.008	$\mathbf{0.364} \pm 0.007$	6.7%	0.336 ± 0.008	$\mathbf{0.352} \pm 0.007$	4.8%
		FullGrad	0.314 ± 0.007	$\mathbf{0.336} \pm 0.006$	7.0%	0.303 ± 0.008	$\mathbf{0.325} \pm 0.007$	7.3%
		MFABA	0.329 ± 0.006	$\mathbf{0.351} \pm 0.005$	6.7%	0.321 ± 0.007	$\mathbf{0.339} \pm 0.006$	5.6%
	Deletion	LIME	0.174 ± 0.004	$\mathbf{0.159} \pm 0.003$	8.6%	0.183 ± 0.004	$\mathbf{0.168} \pm 0.003$	8.2%
		SHAP	0.178 ± 0.004	$\mathbf{0.162} \pm 0.003$	9.0%	0.186 ± 0.004	$\mathbf{0.170} \pm 0.003$	8.6%
		IntGrad	0.181 ± 0.005	$\mathbf{0.165} \pm 0.004$	8.8%	0.189 ± 0.005	$\mathbf{0.173} \pm 0.004$	8.5%
		FullGrad	0.179 ± 0.005	$\mathbf{0.163} \pm 0.004$	8.9%	0.187 ± 0.005	$\mathbf{0.171} \pm 0.004$	8.6%
		MFABA	0.172 ± 0.004	$\mathbf{0.158} \pm 0.003$	8.1%	0.181 ± 0.004	$\mathbf{0.166} \pm 0.003$	8.3%
MedGemma	Insertion	LIME	0.324 ± 0.005	$\mathbf{0.330} \pm 0.004$	1.9%	0.314 ± 0.006	$\mathbf{0.321} \pm 0.005$	2.2%
		SHAP	0.335 ± 0.005	$\mathbf{0.342} \pm 0.004$	2.1%	0.325 ± 0.006	$\mathbf{0.332} \pm 0.005$	2.2%
		IntGrad	0.344 ± 0.007	$\mathbf{0.351} \pm 0.006$	2.0%	0.338 ± 0.007	$\mathbf{0.345} \pm 0.006$	2.1%
		FullGrad	0.317 ± 0.006	$\mathbf{0.323} \pm 0.005$	1.9%	0.306 ± 0.007	$\mathbf{0.313} \pm 0.006$	2.3%
		MFABA	0.331 ± 0.005	$\mathbf{0.338} \pm 0.004$	2.1%	0.324 ± 0.006	$\mathbf{0.335} \pm 0.005$	3.4%
	Deletion	LIME	0.171 ± 0.004	$\mathbf{0.165} \pm 0.003$	3.5%	0.179 ± 0.004	$\mathbf{0.172} \pm 0.003$	3.9%
		SHAP	0.176 ± 0.004	$\mathbf{0.169} \pm 0.003$	4.0%	0.182 ± 0.004	$\mathbf{0.175} \pm 0.003$	3.8%
		IntGrad	0.183 ± 0.005	$\mathbf{0.176} \pm 0.004$	3.8%	0.190 ± 0.005	$\mathbf{0.183} \pm 0.004$	3.7%
		FullGrad	0.180 ± 0.005	$\mathbf{0.173} \pm 0.004$	3.9%	0.186 ± 0.005	$\mathbf{0.179} \pm 0.004$	3.8%
		MFABA	0.172 ± 0.004	$\mathbf{0.166} \pm 0.003$	3.5%	0.178 ± 0.004	$\mathbf{0.171} \pm 0.003$	3.9%

Table 12: Additional counterfactual faithfulness results across datasets and models. We report Pearson correlation r between counterfactual positional scores and output changes induced by positional interventions.

Image			
Dataset	ViT	CLIP	Llama
ImageNet-S	0.282 ± 0.041	0.316 ± 0.037	0.344 ± 0.035
PascalVOC	0.291 ± 0.039	0.328 ± 0.036	0.352 ± 0.034
Text			
Dataset		Llama	
GSM8K		0.407 ± 0.031	
MedMCQA		0.431 ± 0.029	
Clinical time-series			
Dataset	GPT-2	Mamba	MedGemma
PhysioNet	0.337 ± 0.034	0.371 ± 0.031	0.392 ± 0.030
MC-MED	0.326 ± 0.035	0.358 ± 0.032	0.381 ± 0.031

techniques, however, were designed to assign importance to features in isolation. This includes perturbation-based methods [12, 14, 25], gradient-based methods [13, 16, 27], and decomposition-based methods [26]. Section 2.1 discusses these in detail; here we note only that none treats position as a first-class quantity. PaX builds on this family: it takes any such method as a black-box subroutine and lifts it to a positional attribution.

Temporal attribution. A line of work adapts standard attribution to time-series and healthcare settings, where temporal order is intrinsic to the input. TimeSHAP [19] extends SHAP to recurrent models via temporal masking. WindowSHAP [20] aggregates attributions over fixed windows to

Table 13: Ablation of missingness approximations for insertion faithfulness using LIME. We restrict this ablation to models with additive positional embeddings (ViT, CLIP, and GPT-2). PaX-Feature uses feature faithfulness, while PaX-Position uses position faithfulness. Across vision and clinical time-series benchmarks, positional swapping achieves the highest insertion AUC under both evaluation protocols.

Dataset	Model	Evaluation	Zeroing out	Replacing with mean	Replacing with random	Swapping
ImageNet-S	ViT	PaX-Feature	0.259 ± 0.009	0.240 ± 0.009	0.262 ± 0.009	$\mathbf{0.419 \pm 0.005}$
		PaX-Position	0.242 ± 0.009	0.238 ± 0.008	0.245 ± 0.009	$\mathbf{0.396 \pm 0.006}$
	CLIP	PaX-Feature	0.275 ± 0.010	0.269 ± 0.010	0.276 ± 0.011	$\mathbf{0.442 \pm 0.006}$
		PaX-Position	0.256 ± 0.010	0.248 ± 0.009	0.260 ± 0.010	$\mathbf{0.411 \pm 0.007}$
PASCAL VOC	ViT	PaX-Feature	0.273 ± 0.010	0.267 ± 0.010	0.269 ± 0.010	$\mathbf{0.424 \pm 0.005}$
		PaX-Position	0.251 ± 0.008	0.240 ± 0.009	0.246 ± 0.009	$\mathbf{0.400 \pm 0.006}$
	CLIP	PaX-Feature	0.281 ± 0.010	0.275 ± 0.010	0.285 ± 0.011	$\mathbf{0.432 \pm 0.004}$
		PaX-Position	0.268 ± 0.010	0.255 ± 0.009	0.265 ± 0.010	$\mathbf{0.441 \pm 0.005}$
PhysioNet	GPT-2	PaX-Feature	0.229 ± 0.010	0.205 ± 0.009	0.216 ± 0.010	$\mathbf{0.354 \pm 0.005}$
		PaX-Position	0.210 ± 0.010	0.197 ± 0.009	0.208 ± 0.009	$\mathbf{0.341 \pm 0.006}$
MC-MED	GPT-2	PaX-Feature	0.215 ± 0.010	0.196 ± 0.010	0.209 ± 0.011	$\mathbf{0.339 \pm 0.005}$
		PaX-Position	0.202 ± 0.010	0.189 ± 0.009	0.187 ± 0.010	$\mathbf{0.327 \pm 0.006}$

reduce computational cost. TIMING [21] attributes predictions across time steps under a stepwise framing, and DeltaSHAP [22] compares predictions across consecutive time points. These methods establish that structured inputs benefit from attribution mechanisms aware of that structure, but each is specialized to time as the structuring axis and to a particular attribution backbone. PaX retains the core insight that structure matters while removing both restrictions: it is not tied to time and not tied to a specific backbone.

Positional attribution. Three methods address position directly. PoSHAP [45] extends SHAP to sequence models by tying attributions to explicit sequence positions, but inherits SHAP's combinatorial cost and is restricted to the SHAP backbone. Position-LRP [46] introduces LRP rules that propagate relevance through Transformer positional encodings (Rotary, Learnable, and Absolute PE), reformulating the input space as position-token pairs; it is restricted, however, to the LRP backbone and to Transformer architectures with explicit positional encodings. OrdSHAP [23] measures sensitivity to input order via permutations, isolating order effects from value effects, but order sensitivity is a narrower question than positional attribution: it asks whether reordering changes predictions, not how much each position contributes.

Positioning of PaX. Existing positional attribution methods are each coupled to a specific attribution backbone (SHAP, LRP), constrained to particular architectures (Transformers with positional encodings), or address a narrower question (order sensitivity). PaX is, to our knowledge, the first framework that wraps an arbitrary feature attribution method and returns a positional attribution, requiring only black-box access to the underlying method. This makes it applicable across attribution families, architectures, and input modalities.

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 contributions claimed in the abstract and introduction (PaX framework, counterfactual position score, and positional faithfulness evaluation protocols) are directly supported by the formalization in Section 2 and the empirical results in Section 3.

   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: We discuss the limitations of our work in Section 6.

   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 is a use-inspired empirical paper and does not include formal theoretical results.

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: All experimental details (datasets, models, training configurations, and method specifications) are provided in the main text and Appendices B.1–B.2. Code is released at `https://github.com/chaehyeon-kim/PaX`.

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: Our code is released at `https://github.com/chaehyeon-kim/PaX`, with a README providing setup instructions, dependencies, and commands to reproduce our experiments. All datasets used are publicly available (see Appendix B.1).

   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: Experimental details (data splits, training, and evaluation protocols) are provided in Appendix B.3, with reproduction commands in the released in `https://github.com/chaehyeon-kim/PaX`.

   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: All main experimental results are reported with bootstrap-based error bars (or confidence intervals), as described in the corresponding figures and tables.

   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: Compute resources, including the type of GPUs used and overall hardware setup, are reported in Appendix B.3.

   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 reviewed the NeurIPS Code of Ethics and confirm that our research conforms to it in every respect.

   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: Both positive impacts (more reliable explanations in high-stakes domains) and potential risks (over-reliance on attribution outputs) are discussed in Section 6.

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 paper presents an algorithmic contribution and does not release any new pre-trained models, generative models, or scraped datasets that pose a high risk for misuse.

    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 datasets and pre-trained models used in this work are publicly available and properly credited, with citations to the original papers and license information in Appendix B.1 (datasets) and Appendix B.2 (models). All assets are used in accordance with their respective licenses and terms of use.

    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: Our code is released at `https://github.com/chaehyeon-kim/PaX` with documentation for reproducing all experiments. An open-source license will be specified upon publication.

    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: The paper does not involve crowdsourcing 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 research with human subjects requiring IRB review.

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 (drafting and editing) and for help implementing standard methods. They were not a core, original, or non-standard component of the methodology.

    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.