Interpreting Sepsis Prediction through Positional Explanation

Anonymous Authors[1]

Abstract

Sepsis prediction models achieve high accuracy but go unused at the bedside. Their explanations cannot answer the questions clinicians actually ask: which measurement signaled deterioration, and when? Existing attribution methods score features but conflate a measurement's value with its timing, obscuring trajectories needed for early action. We introduce Position-aware eXplanation (PaX), a framework that decomposes each attribution into a "what" score for the measurement feature and a "when" score for its temporal position. Across Mamba, GPT-2, and MedGemma on PhysioNet and MC-MED, PaX surfaces second-order early-warning indicators that may signal early deterioration, improves alignment with clinical expertise, and exposes positional biases obscured by standard attribution methods.

1. Introduction

Sepsis is a leading cause of hospital mortality, often detected only after irreversible organ damage has occurred (Seymour et al., 2016). While modern machine learning models can predict its onset hours in advance, their lack of interpretability remains a barrier to clinical adoption: without understanding the underlying physiological drivers, clinicians cannot act on early warnings with confidence (Yuan et al., 2020; Bomrah et al., 2024). In high-stakes domain like medicine, explanations must support precise clinical reasoning and intervention timing, not merely build trust (Wong et al., 2021; Adams et al., 2022).

Sepsis is a disease of trajectory. A patient's physiological evolution over time is more informative than any single measurement (Zhu et al., 2023). An elevated heart rate, for instance, is clinically ambiguous in isolation; its significance depends on whether it is a sudden spike or a gradual rise coupled with declining blood pressure. Effective explanations must therefore reflect this longitudinal reasoning.

Yet widely used explanation methods such as LIME (Ribeiro et al., 2016) and IntGrad (Sundararajan et al., 2017) treat features independently of their temporal context, creating a fundamental mismatch with the way sepsis is diagnosed in practice. This mismatch is amplified by a key property of modern sequence models: they are highly sensitive to input order, and reordering a sequence can substantially change predictions (Liu et al., 2024). Existing feature attribution methods fail to expose this mechanism, and ignoring positional effects produces "false positive" attributions. We find that features measured at admission ($t = 0$), such as *Gender*, often appear highly important in standard LIME explanations simply because sequence models exhibit a known bias toward the beginning of an input string.

To bridge this gap, we propose Position-aware eXplanation (PaX), a framework that disentangles a measurement's clinical identity ("what") from its temporal position ("when"). By decomposing attribution into separate feature and positional scores, PaX reveals higher-order clinical factors—latent physiological relationships that emerge over time and are often masked by high-variance raw vitals, such as downstream biochemical responses to organ dysfunction. Across finetuned Mamba (Dao & Gu, 2024), GPT-2 (Radford et al., 2019), and MedGemma (Sellergren et al., 2025) models on PhysioNet (Reyna et al., 2019) and MC-MED (Kansal et al., 2025), this separation surfaces features that align significantly more closely with expert clinical annotations.

By isolating positional effects, PaX corrects architectural biases that mislead existing explanation methods, showing that separating position from feature identity is a requirement for faithfulness, not a detail. Our contributions are:

- We introduce PaX, a framework that separates attribution into clinical measurement and temporal position to improve faithfulness.
- We propose dual-axis faithfulness tests using insertion and deletion metrics on feature and position attributions.
- We show that PaX surfaces higher-order sepsis-related factors aligned with expert clinical reasoning but missed by existing methods.
- We show that PaX identifies and corrects positional biases in state-of-the-art sequence models.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Position-aware eXplanation (PaX)

2.1. Preliminaries: Feature Attributions

Feature attribution is a popular paradigm for interpreting model behavior, assigning an importance score to each input feature (Doshi-Velez & Kim, 2017). Existing methods typically answer the question: *How important are input features to the model's prediction?*

To formalize the problem, let $x \in \mathbb{R}^d$ denote the input to the model being explained, and let $f : \mathbb{R}^d \to \mathbb{R}^m$ map x to an output $f(x)$, such as class probabilities or regression values. Feature attribution methods typically define a perturbation function $g : \mathbb{R}^d \times [0,1]^d \to \mathbb{R}^d$ that produces a perturbed input $x_0 = g(x, z)$, where the reference vector $z \in [0,1]^d$ governs the degree of perturbation: $z_i = 1$ indicates presence of the i-th feature, while $z_i = 0$ indicates its absence or replacement by a baseline. While z is often binary, real-valued references can interpolate between the baseline and the original input. Attribution methods vary z to generate perturbations via g, which satisfies $g(x, \mathbf{1}) = x$.

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 an input feature x_i to the model prediction $f(x)$. This formulation generalizes a wide range of feature attribution methods, ranging from classic methods such as LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and IntGrad (Sundararajan et al., 2017) as well as newer and improved approaches (Fumagalli et al., 2023; Zhu et al., 2024; Kim et al., 2025).

Attributions for Sequence-Structured Inputs. Feature attributions extend naturally to sequence-structured inputs common in transformer architectures, where positional information is typically incorporated through positional embeddings. A sequential input consists of n elements (e.g., a sequence of measurements collected over time) $x = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d_1}$, where each $x_i \in \mathbb{R}^{d_1}$ is an element of the sequence. The sequence model is correspondingly defined as $f : \mathbb{R}^{n \times d_1} \to \mathbb{R}^m$, such as a transformer.

The perturbation function g now takes a reference vector $z \in [0,1]^n$ that specifies element-level perturbations, producing $x_0 = g(x, z)$. An explainer then produces an element-level attribution vector

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

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

Limitations of Existing Attribution Methods. Modern sequence models like Transformers and Mamba combine positional and feature embeddings to capture temporal context (Vaswani et al., 2017; Grattafiori et al., 2024). Classic feature attribution methods—including LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), Integrated Gradients (Sundararajan et al., 2017), and FullGrad (Srinivas & Fleuret, 2019)—explain feature effects but cannot distinguish whether importance arises from a feature's clinical identity or from positional biases internal to the model. Recent "time-aware" methods address temporal structure: TimeSHAP (Bento et al., 2021) masks at the time level for recurrent models, WindowSHAP (Nayebi et al., 2023) aggregates SHAP over fixed temporal windows, TIMING (Jang et al., 2025) extends IntGrad to low-dimensional time series, DeltaSHAP (Kim et al., 2025) examines differences between successive time steps, and OrdShap (Hill et al., 2025) quantifies positional importance via feature-order permutations. These approaches extend specific attribution methods rather than providing a unified treatment.

2.2. The PaX framework

We propose *Position-aware eXplanation (PaX)*, a framework that captures positional effects by treating positional embeddings as first-class features through an augmented input representation. Unlike prior work, PaX is general purpose, extending a broad class of attribution methods with positional attributions without requiring bespoke solutions.

PaX generalizes the standard explanation formulation in Eq. (1) by incorporating positional information as an implicit input to the model. Let $x \in \mathbb{R}^{n \times d_1}$ and $p \in \mathbb{R}^{n \times d_2}$ denote a sequence of features and positions, where the i-th element is represented by a feature vector x_i and an associated position p_i. Let $\mathcal{E}(f, g, x)$ be a general feature attribution method as defined in (1). The key idea 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 positional attributions jointly.

PaX Inputs. The first step is to explicitly formulate an augmented input with positional information:

$$x' := (x, p) \in \mathbb{R}^{n \times (d_1 + d_2)},$$

which jointly represents feature content and positional information. For simplicity, one can assume positions are represented as explicit vectors, such as positional embeddings in a transformer architecture. However, the proposed framework is compatible with implicit positions inferred from the architecture, which is expanded upon later.

PaX Models. The expanded input x' requires an adjustment to the formalization of the model being explained:

$$f' : \mathbb{R}^{n \times (d_1 + d_2)} \to \mathbb{R}^m, \quad f'(x') := f'(x, p) = f(x).$$

Here, positional information p is treated as explicit "feature" alongside the standard input x to the model f'. When p matches the positional information used internally by f, this reduces to a standard forward pass $f(x)$.

PaX Perturbations. The final component is the perturbation function g', which must perturb both the features x and positions p in the augmented input x'. Let $z' = (z_x, z_p) \in [0,1]^{d_1+d_2}$ be a reference vector specifying the degree to which features and positions are retained or perturbed. Then, the PaX perturbation function g' is defined as

$$g'(x', z') := (g(x, z_x), \text{pos-swap}(p, z_p)) .$$

In this perturbation function, features x are perturbed with respect to the reference z_x using the original perturbation function g from the feature attribution method. Positions p are perturbed by random swapping with respect to the reference z_p, $\text{pos-swap}(p, z_p)$: if $z_p = 0$, the position is randomly swapped with other perturbed positions; if $z_p = 1$, it stays put. For real-valued $z_p \in (0, 1)$, each position is retained with probability z_p, and attribution is aggregated in expectation over multiple samples.

With all of these components, the general PaX explanation operator for producing position attributions from a feature attribution method $\mathcal{E}$ is

$$\mathcal{E}_{\text{PaX}}(f, g, x) \;:=\; \mathcal{E}\big(f', g', x'\big) = \alpha'. \tag{2}$$

This calls the original feature attribution operator $\mathcal{E}$ on the position-aware inputs (f', g', x') to jointly produce attributions $\alpha' = (\alpha_x, \alpha_p)$ for features and positions. The feature attributions α_x are importance scores for feature contents; position attributions α_p are importance scores for positions.

A key advantage of PaX is that it transforms any feature attribution method into one that produces positional attributions, without bespoke design. Appendix A.2 instantiates PaX within LIME, SHAP, and IntGrad by treating positions as additional interpretable features.

3. Experiments

We conduct quantitative experiments showing that our framework yields more faithful explanations (Alvarez-Melis & Jaakkola, 2018) than existing methods and better aligns with clinician reasoning for sepsis prediction.

3.1. Experimental Setup

Datasets and Models. We finetune GPT-2 small (124M) (Radford et al., 2019), Mamba-130M (Dao & Gu, 2024), and MedGemma 4B (Sellergren et al., 2025) for sepsis prediction on MC-MED (Kansal et al., 2025) and PhysioNet (Reyna et al., 2019), and apply PaX to the finetuned models.

PhysioNet provides tabular EHR data, while MC-MED additionally includes waveforms, ventilator settings, medications, and per-minute vitals. We follow CareBench's sepsis labeling and cohort selection criteria for consistent preprocessing across both datasets. Details are in Appendix B.

Explanation Methods and Baselines. We instantiate PaX with five standard feature attribution methods, namely LIME, SHAP, IntGrad, FullGrad, and MFABA, with descriptions in Appendix A. We compare against two ablated variants of PaX that use only feature ($|\alpha^{(x)}|$) or only position ($|\alpha^{(p)}|$) attributions. The feature-only variant corresponds to the original feature attribution method.

3.2. Does Decomposing Attribution Improve Faithfulness of Explanations?

We evaluate whether separating feature and position attributions yields more faithful explanations than attributing them jointly. Faithfulness is measured by insertion and deletion AUC: higher insertion and lower deletion mean the attribution better identifies what the model relies on (Luss et al., 2021). We compare PaX-Feature and PaX-Position to Feature-only and Position-only baselines across PhysioNet and MC-MED, three backbones (GPT-2, Mamba, MedGemma), and five attribution methods (LIME, SHAP, IntGrad, FullGrad, MFABA). Full results are in Appendix C.

3.2.1. Faithful Feature Attributions

Standard feature insertion and deletion progressively add or remove features by attribution rank and report the AUC of the performance curve. PaX ranks features by $|\alpha_i^{(x)}|$.

PaX-Feature beats Feature-only on every configuration, with gains widening on MC-MED, whose sequences are longer and more positionally heterogeneous: insertion AUC rises by $\sim$0.02 on PhysioNet and 0.02–0.03 on MC-MED, while deletion AUC falls by 0.005–0.015 and 0.014–0.020, respectively. Gains are smallest on MedGemma ($\leq$ 0.007), consistent with stronger pretrained priors already absorbing part of the position–feature interaction. The ranking holds across perturbation-based (LIME, SHAP) and gradient-based (IntGrad, FullGrad, MFABA) methods, indicating the gain stems from decomposing attribution rather than from any single algorithm.

3.2.2. Faithful Position Attributions

We extend standard insertion and deletion protocols to evaluate position attributions. Rather than adding or removing feature content, these protocols measure how relocating identical content across positions affects predictions, guided by position attribution scores $|\alpha_i^{(p)}|$. Position insertion assigns features to top-ranked positions from a random configuration; position deletion moves features from top-ranked positions to random ones from the original input. AUC is computed as before.

PaX-Position beats Position-only on every configuration, and the same pattern as above holds: larger gains on MC-MED, smallest on MedGemma, and consistent across attribution methods.

3.3. Validating the actionability of positional explanations at the bedside

An explanation is only useful insofar as a stakeholder can act on it. Orgad et al. (2026) argue that interpretability methods should be judged by the decisions they enable, and propose criteria for doing so. Three apply here: understandability (can the audience make sense of the explanation?), task enhancement (does it improve their performance?), and mechanistic faithfulness (does it reflect what the model relies on, free of confounds?). We ask through them: does PaX, by decoupling position from feature attribution, produce explanations a bedside clinician can act on when treating a patient at risk of sepsis? We instantiate each criterion as a concrete bedside action and report one finding per action, all on the best-performing finetuned MedGemma model on the PhysioNet sepsis prediction task with clinician-annotated measurements. An understandable explanation is one a clinician can read as a statement about a measurement (Finding 1); a task-enhancing one surfaces measurements absent from formal guidelines (Finding 2); a mechanistically faithful one separates biological signal from structural artifacts of the input (Finding 3). Annotation protocol, full tables, and per-finding details are deferred to Appendix D.

Finding 1 (understandability). PaX-LIME feature attributions $\alpha^{(x)}_{\text{PaX-LIME}}$ track clinician judgments of sepsis relevance, with mean attributions of 0.40, 0.31, and 0.22 across direct, partial, and unrelated measurements. We define Δ_f as the difference between the PaX-LIME feature component and the standard LIME attribution; a positive value means standard LIME underestimated the measurement because some of its attribution was absorbed by position. We find $\Delta_f > 0$ for $24/29$ measurements labeled as direct or partial sepsis indicators. Standard LIME dilutes critical biomarkers by conflating them with their temporal placement; PaX-LIME recovers an attribution profile a clinician can read as a statement about the measurement itself.

Finding 2 (task enhancement). All 13 measurements named in formal sepsis guidelines (Sepsis-3, SOFA, qSOFA, NEWS, SOFA-2 (Ranzani et al., 2025)) are independently labeled relevant by clinicians, confirming PaX-LIME covers the canonical markers. The actionable surface, however, lies elsewhere: of the 22 off-guideline measurements, clinicians label 16 as relevant or partially relevant. The pattern sharpens at the top of the ranking, where four of the ten highest-attribution features (Age, pH, BUN, AST) are uncodified yet clinician-endorsed.

Finding 3 (mechanistic faithfulness). An attribution can mislead if what it labels "feature importance" is a structural artifact of the input. Standard LIME assigns Age and Gender comparable importance (0.76 and 0.73); once PaX decouples position from feature, Age retains a high feature score (0.77) while Gender drops to 0.33, with the remainder absorbed into the position component. PhysioNet places demographics at admission, so this pattern is consistent with a primacy bias rather than a biological link between gender and sepsis risk. Standard attribution would invite a clinician to act on the apparent link; PaX-LIME flags it as positional and tells the clinician to discount it.

4. Related Work

Section 2.1 motivates positional effects (Liu et al., 2024; Wu et al., 2025; Kamp et al., 2025) and reviews standard feature attribution methods, including perturbation-based, gradient-based, and decomposition-based approaches (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017; Srinivas & Fleuret, 2019; Zhu et al., 2024), as well as extensions to time-series and healthcare settings such as TimeSHAP (Bento et al., 2021), WindowSHAP (Nayebi et al., 2023), TIMING (Jang et al., 2025), and DeltaSHAP (Kim et al., 2025), which incorporate temporal structure through masking, windowing, or stepwise comparisons. A separate line of work targets positional structure directly. PoSHAP (Dickinson & Meyer, 2022) associates SHAP attributions with sequence positions, Position-Aware LRP (Bakish et al., 2025) distributes relevance across content and positional components, and OrdSHAP (Hill et al., 2025) disentangles value- from order-driven effects via input permutations. These methods are tied to specific attribution frameworks or focus narrowly on order sensitivity. In contrast, PaX provides a general, model-agnostic framework for positional attributions.

5. Conclusion

We address a core limitation of attribution methods for early sepsis prediction: positional information is implicitly absorbed into feature importance, obscuring how input order influences predictions. We propose Position-aware eXplanation (PaX), which separates a measurement's clinical identity from its temporal position. Across fine-tuned Mamba, GPT-2, and MedGemma, we show that standard attribution methods entangle clinical signals with architectural positional artifacts, and that separating the two yields more faithful explanations. Dual-axis faithfulness evaluations confirm that PaX corrects these artifacts and surfaces mediated prognostic factors aligned with clinical expertise yet missed by existing attribution methods. PaX is a step toward attributing sepsis predictions to clinical trajectories rather than isolated measurements, opening directions such as how physiological trends drive model risk estimates over time.

Impact Statement

This paper advances the field of Interpretable Machine Learning by addressing the critical role of positional sensitivity in modern sequence models, including Transformers and Large Language Models. In healthcare settings, our work separates feature and position importance to enable a more reliable and faithful understanding of model behavior. By identifying architectural biases and surfacing higher-order clinical factors, this research supports the development of safer decision-support systems for high-stakes medical interventions.

References

Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., A. Zell, K., S. L. Tan, C., N. Wiens, J., E. V. Barton, C., and A. Singh, K. Prospective, multi-site study of a deep learning model for early detection of sepsis. *Nature Medicine*, 28(8):1649–1654, 2022. doi: 10.1038/s41591-022-01894-0.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 7786–7795, Red Hook, NY, USA, 2018. Curran Associates Inc.

Bakish, Y., Zimerman, I., Chefer, H., and Wolf, L. Revisiting lrp: Positional attribution as the missing ingredient for transformer explainability, 2025. URL https://arxiv.org/abs/2506.02138.

Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A., and Bizarro, P. TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 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.

Bomrah, S., Uddin, M., Upadhyay, U., Priya, J., Dhar, E., Hsu, S.-C., and Syed-Abdul, S. A scoping review of machine learning for sepsis prediction- feature engineering strategies and model performance: a step towards explainability. *Critical Care*, 28:180, 2024.

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

Dickinson, Q. and Meyer, J. G. Positional shap (poshap) for interpretation of machine learning models trained from biological sequences. *PLOS Computational Biology*, 18(1):1–24, 01 2022. doi: 10.1371/journal.pcbi.1009736. URL https://doi.org/10.1371/journal.pcbi.1009736.

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

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. SHAP-IQ: unified approximation of any-order shapley interactions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 11515–11551, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S.,

Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hill, D., Hill, B. L., Masoomi, A., Nori, V. S., Tillman, R. E., and Dy, J. Ordshap: Feature position importance for sequential black-box models, 2025. URL https://arxiv.org/abs/2507.11855.

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

Kamp, J., Bakker, R., and Blok, D. 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.

Kansal, A., Chen, E., Jin, B. T., et al. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12:1094, 2025. doi: 10.1038/s41597-025-05419-5.

Keoliya, M., Choi, S., Alur, R., Naik, M., and Wong, E. Stable prediction of adverse events in medical time-series data, 2025. URL https://arxiv.org/abs/2510.14286.

Kim, C., Mun, Y., Hahn, S., and Yang, E. DeltaSHAP: Explaining Prediction Evolutions in Online Patient Monitoring with Shapley Values, July 2025. URL http://

arxiv.org/abs/2507.02342. arXiv:2507.02342 [cs].

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. 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/.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 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.

Luss, R., Chen, P.-Y., Dhurandhar, A., Sattigeri, P., Zhang, Y., Shanmugam, K., and Tu, C.-C. Leveraging latent features for local explanations, 2021. URL https://arxiv.org/abs/1905.12698.

Nayebi, A., Tipirneni, S., Reddy, C. K., Foreman, B., and Subbian, V. 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.

Orgad, H., Barez, F., Haklay, T., Lee, I., Mosbach, M., Reusch, A., Saphra, N., Wallace, B. C., Wiegreffe, S., Wong, E., Tenney, I., and Geva, M. Interpretability can be actionable. 2026. URL https://actionable-interpretability-guide.github.io.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Ranzani, O. T., Singer, M., Salluh, J. I. F., Shankar-Hari, M., Pilcher, D., Berger-Estilita, J., Coopersmith, C. M., Juffermans, N. P., Laffey, J., Reinikainen, M., Neto, A. S., Tavares, M., Timsit, J.-F., Arias Lopez, M. D. P., Arulkumaran, N., Aryal, D., Azoulay, E., Celi, L. A., Chaudhuri, D., De Lange, D., De Waele, J., Dos Santos, C. C., Du, B., Einav, S., Engelbrecht, T., Fazla, F., Ferrer, R., Finazzi, S., Fujii, T., Gershengorn, H. B., Greene, J. D., Haniffa, R., Hao, S., Hasan, M. S., Hollenberg, S., Ippolito, M., Jung, C., Kirov, M., Kobari, S., Lakbar, I., Lipman, J., Liu, V., Liu, X., Lobo, S. M., Magatti, D., Martin, G. S., Metnitz, B., Metnitz, P., Myatra, S. N., Oczkowski, S., Paiva, J.-A., Paruk, F., Pekkarinen, P. T., Piquilloud, L., Pölkki, A., Prescott, H. C., Blaser, A. R., Rezende, E., Robba, C., Rochwerg, B., Ruckly, S., Samei, R., Schenck, E. J., Secombe, P., Sendagire, C., Siaw-Frimpong, M., Simpkin, A. J., Soares, M., Summers, C., Szczeklik, W., Takala, J., Tanaka, S., Tricella, G., Vincent, J.-L., Wendon, J., Zampieri, F. G., Rhodes, A., and Moreno, R. 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.

Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., and Clifford, G. D. 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.

Ribeiro, M. T., Singh, S., and Guestrin, C. "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*, pp. 1135–1144, 2016.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Cían, H., Lau, C., et al. Medgemma technical report, 2025.

Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., Singer, M., Deutschman, C. S., Escobar, G. J., and Angus, D. C. Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):762–774, 2016. doi: 10.1001/jama.2016.0288.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems 32*, 2019.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 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.

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J. M., DeTroyer-Cooley, O., Pestrue, J., Phillips,

M. E., Konye, J., J. Schulte, P., A. Kora, M., A. Dligach, D., and Afshar, M. External validation of a widely implemented commercial sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021. doi: 10.1001/jamainternmed.2021.2626.

Wu, X., Wang, Y., Jegelka, S., and Jadbabaie, A. On the emergence of position bias in transformers. In *ICML*, 2025.

Yuan, K.-C., Tsai, L.-W., Lee, K.-H., Cheng, Y.-W., Hsu, S.-C., Lo, Y.-S., and Chen, R.-J. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International Journal of Medical Informatics*, 141:104176, 2020.

Zhu, J.-L., Yuan, S.-Q., Huang, T., Zhang, L.-M., Xu, X.-M., Yin, H.-Y., Wei, J.-R., and Lyu, J. Influence of systolic blood pressure trajectory on in-hospital mortality in patients with sepsis. *BMC Infectious Diseases*, 23(1): 90, 2023.

Zhu, Z., Chen, H., Zhang, J., Wang, X., Jin, Z., Xue, M., Zhu, D., and Choo, K.-K. R. 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, pp. 17228–17236, 2024.

A. Attribution Methods

A.1. Feature Attributions

This section briefly describes the explanation methods employed in conjunction with our PaX approach.

- **LIME (Local Interpretable Model-agnostic Explanations)** (Ribeiro et al., 2016) generates local explanations for individual predictions by fitting an interpretable surrogate model (typically linear) within the neighborhood of the target instance. The method creates perturbations around the input sample and trains the surrogate model on these variations, with samples weighted by their proximity to the original instance.
- **SHAP (SHapley Additive exPlanations)** (Lundberg & Lee, 2017) computes feature importance scores based on cooperative game theory principles. Each feature receives an attribution value representing its marginal contribution to the prediction relative to a baseline, with the property that all attribution values sum to the difference between the model's output and the baseline prediction.
- **Integrated Gradients (IntGrad)** (Sundararajan et al., 2017) computes feature attributions by integrating gradients along a linear path from a baseline input to the target input. This path integral approach ensures satisfaction of fundamental attribution axioms, including sensitivity and implementation invariance.
- **FullGrad** (Srinivas & Fleuret, 2019) extends standard gradient-based attribution by incorporating gradient information from all network layers. The method aggregates input gradients with bias gradients across all intermediate representations, providing more comprehensive attribution maps that capture multi-layer feature interactions.
- **MFABA (More Faithful and Accelerated Boundary-based Attribution)** (Zhu et al., 2024) computes attributions by constructing paths from input samples to adversarial examples that cross the model's decision boundary. The method employs second-order Taylor approximations to better model loss function changes during gradient ascent optimization.

A.2. Position-aware eXplanation (PaX)

Example illustrating PaX to LIME, SHAP, and IntGrad.

A.2.1. Example: PaX-LIME

A major advantage of our framework is that it can transform any standard feature attribution method into one that naturally produces positional attributions without having to design bespoke solutions. As a demonstration, we show how one can instantiate the PaX framework within the LIME paradigm for sequence-structured inputs by treating positions as additional interpretable features. Specifically, the traditional LIME (Ribeiro et al., 2016) algorithm has three main steps: (1) generate perturbed inputs, (2) make predictions on perturbed inputs, then (3) fit a weighted linear model to the predictions.

1. **Perturbing inputs.** LIME begins by creating feature masks z_x, which are sampled uniformly at random. Therefore, in PaX-LIME where positions are treated equivalently to features, both features and positions $z' = (z_x, z_p)$ are perturbed uniformly at random.
2. **Making predictions.** LIME then makes predictions on the perturbed inputs by passing them into the model, $y = f(g(x, z_x))$. The analogous step in PaX-LIME makes predictions on the augmented perturbed inputs by passing them into the position-explicit model, $y' = f'(g'(x, z'))$.
3. **Fitting a linear model.** The last step of LIME fit's a weighted linear model from the feature masks z_x to the predictions y. In PaX-LIME, this translates to fitting an expanded linear model from the joint position and feature masks $z' = (z_x, z_p)$ to the predictions y'.

The resulting coefficients of the linear model for PaX-LIME then consists of not only coefficients for features z_x, but also coefficients for positions z_p, which completes the positional attribution of PaX-LIME.

A.2.2. PaX-SHAP

PaX framework within the SHAP (SHapley Additive exPlanations) paradigm (Lundberg & Lee, 2017) for sequence-structured inputs by treating positions as additional participants in a coalition.

Specifically, the KernelSHAP algorithm—which is the most common model-agnostic implementation of SHAP—has three main steps: (1) sample coalitions of features, (2) evaluate the model on these coalitions, and (3) solve for Shapley values via

a weighted linear regression.

1. **Sampling coalitions.** SHAP begins by creating binary coalition masks $z_f \in \{0,1\}^n$, where each bit indicates whether a feature is "present" or "missing." Therefore, in PaX-SHAP where positions are treated as players alongside features, both feature and position masks $z' = (z_f, z_p)$ are sampled according to the SHAP kernel distribution.
2. **Making predictions.** SHAP makes predictions on the coalitions by mapping the masks back to the input space and passing them into the model, $y = f(g(x, z_f))$. The analogous step in PaX-SHAP evaluates these augmented coalitions by passing them into the position-explicit model: $y' = f'(g'(x, z'))$. This captures the interaction between feature content and its assigned position.
3. **Solving for Shapley values.** The last step of SHAP fits a weighted linear model from the masks z_f to the predictions y, where the weights are determined by the SHAP kernel. In PaX-SHAP, this translates to fitting the expanded linear model from the joint masks $z' = (z_f, z_p)$ to the predictions y'. Unlike LIME, the specific weighting of the SHAP kernel ensures the resulting coefficients satisfy the properties of Efficiency, Symmetry, and Linearity.

The resulting coefficients (Shapley values) for PaX-SHAP consist of attributions for both features ϕ_f and positions ϕ_p. Because SHAP is an additive attribution method, the sum of these values $\sum \phi_f + \sum \phi_p$ equals the difference between the model output $f'(x')$ and the base expectation, providing a theoretically grounded decomposition of positional importance.

A.2.3. PAX-INTGRAD

Standard Integrated Gradients (Sundararajan et al., 2017) computes attribution by integrating the gradients of the model's output with respect to the input along a straight-line path from a baseline to the input.

In PaX-IntGrad, we reconcile the continuous nature of path integrals with the discrete nature of positional swaps through a sampling-based aggregation:

1. **Generating Perturbed Paths.** For each sample k, we utilize the perturbation probability z_p. For each position i, we decide to either retain the original position or perform a swap based on z_p. This results in a perturbed positional configuration $p^{(k)}$. We then define the straight-line path between a baseline x'_0 and the perturbed augmented input $x'^{(k)} = (x, p^{(k)})$.
2. **Computing Path Gradients.** We compute the standard IG attribution for both the feature content and the specific positional embeddings actually used in the k-th forward pass:

$$\text{IG}(x'^{(k)}) = (x'^{(k)} - x'_0) \times \int_0^1 \frac{\partial f'(x'_0 + \gamma(x'^{(k)} - x'_0))}{\partial x'} d\gamma$$

   where γ is the interpolation parameter. This provides raw attribution scores for the features α_x and the positional embeddings α_p for that specific configuration.
3. **Indicator-based Aggregation.** To derive counterfactual attributions α_q, we cannot directly interpolate through a continuous z_q. Instead, we employ a post-hoc counterfactual indicator inferred from the sampling. By comparing the original positional embeddings p to the perturbed embeddings $p^{(k)}$, we define a binary mask $I_{ij}^{(k)} = 1$ if the feature originally at position i was moved to position j, and 0 otherwise. The final counterfactual attribution for moving a feature from i to j is aggregated across all samples where the movement was realized:

$$\alpha_{q,ij} = \text{aggregate}\left(\{\text{IG}_{p_j}(x'^{(k)}) \mid I_{ij}^{(k)} = 1\}\right)$$

The resulting α_q captures the effect of moving a feature from i to j by leveraging the model's gradient sensitivity to the positional embeddings actually present during the forward pass.

B. Experimental Setup

Our experiments are designed to evaluate three aspects of the proposed framework and to analyze its design choices. First, we evaluate performance relative to existing attribution baselines under standard feature faithfulness metrics. Second, we assess positional faithfulness, comparing against existing baselines, including methods specifically designed to attribute positional importance. Third, we present results evaluating counterfactual positional explanations, a new explanation type

enabled by our framework. Finally, we conduct ablation studies to assess the contribution of each component of the proposed method, demonstrating that observed gains arise from deliberate design choices rather than incidental effects.

B.1. Dataset Description

B.1.1. Clinical Time-Series Benchmarks

PhysioNet 2019 (Reyna et al., 2019). PhysioNet 2019 is 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.

MC-MED (Kansal et al., 2025). MC-MED is an emergency department sepsis prediction dataset containing 118,385 visits from 70,545 unique patients, combining minute-level vital signs and continuous physiological waveforms with comprehensive clinical data.

B.1.2. Sepsis Prediction Task Curation

We follow the sepsis labeling procedure described in CAREBench (Keoliya et al., 2025), with adaptations specific to each dataset's clinical context and data availability.

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

MC-MED. Sepsis labeling follows a two-stage procedure tailored to emergency department settings with a prediction horizon of $h = 1.5$ hours:

1. **At-Risk Cohort Selection.** Patients must meet 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 administered at or before the time the first criterion is met
2. **Sepsis Labeling.** Positive labels are assigned when emergency SOFA (eSOFA) criteria are met within ± 2 days of blood culture collection:
   - Evidence of presumed serious infection:
     - Blood culture obtained
     - ≥ 4 qualifying antimicrobial days (QADs)
   - Evidence of acute organ dysfunction, including 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 sequence models fine-tuned for clinical sepsis prediction.

We employ three sequence models fine-tuned for sepsis prediction on CAREBench-curated datasets: GPT-2 (124M parameters), Mamba-130M, and Medgemma 4B.

GPT-2 Small (Radford et al., 2019). A 124M-parameter decoder-only transformer with 12 layers, 768 hidden dimensions, and causal self-attention. Its autoregressive architecture naturally captures temporal dependencies in patient trajectories.

Mamba-130M (Dao & Gu, 2024). A 130M-parameter state-space model designed for efficient long-sequence modeling with linear computational complexity. Its state-space formulation provides inductive biases that align well with physiological dynamics.

Table 1. Performance of GPT-2, Mamba, and MedGemma on the MC-MED and PhysioNet datasets. We report standard classification metrics for comparability with prior work, along with the Brier score used for model selection during training.

Dataset	Finetuned 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

MedGemma 4B (Sellergren et al., 2025). 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 training provides clinical knowledge priors that make it suitable for modeling complex relationships among physiological measurements in patient trajectories.

Training Configuration. All sepsis prediction models are trained from scratch following the CAREBench (Keoliya et al., 2025) protocol, with one clinically motivated modification. Specifically, while CAREBench primarily reports AUROC, we select model hyperparameters using the **Brier score** as the primary validation metric to account for the extremely low event rate in sepsis prediction.

- **Custom Tokenization:** Dataset-specific tokenizers for medical codes
- **Training Duration:** 100 epochs
- **Hyperparameter Selection:** Learning rate $\in \{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$ selected based on validation Brier score

For completeness, we report standard predictive performance metrics, along with the Brier score used for model selection, in Table 1.

C. Additional Feature and Position Faithfulness Results

This section reports the full insertion and deletion results supporting the faithfulness claims in Section 3.2. The expanded grid spans two datasets (PhysioNet, MC-MED), three backbones (GPT-2, Mamba, MedGemma), and five attribution methods (LIME, SHAP, IntGrad, FullGrad, MFABA).

C.1. Setup

We compare four attributions per explanation method: **Feature-only** and **Position-only** (the conventional baselines that perturb either feature values or token positions in isolation), and **PaX-Feature** and **PaX-Position** (the corresponding components of our framework). We use insertion and deletion AUC as complementary faithfulness metrics; under insertion, higher is better, and under deletion, lower is better. We omit per-metric reminders in the discussion below and instead mark improvements directly in the tables.

C.2. Findings

PaX components dominate their baselines on every cell. Across all 60 (dataset, model, method) combinations in Tables 2 and 3, PaX-Feature beats Feature-only and PaX-Position beats Position-only on both insertion and deletion.

Gains are larger on MC-MED than on PhysioNet. Insertion AUC improves by roughly 0.02 on PhysioNet and 0.02–0.03 on MC-MED; deletion AUC drops by 0.005–0.015 on PhysioNet and 0.014–0.020 on MC-MED. The wider margin on MC-MED is consistent with its longer and more positionally heterogeneous sequences, where conflating feature and position attributions costs more.

Gains are smallest on MedGemma. Across both datasets and both metrics, MedGemma shows the narrowest PaX-vs.-baseline gap (often ≤ 0.007). We attribute this to MedGemma's stronger pretrained priors, which already absorb part of the position-feature interaction that PaX makes explicit in smaller models.

The pattern is method-agnostic. The ranking holds for perturbation-based methods (LIME, SHAP) and gradient-based

Table 2. Our Position-aware eXplanation (PaX) framework consistently outperforms traditional feature attribution methods. PaX-Feature achieve higher insertion AUC and lower deletion AUC than Feature-only counterparts, confirming more faithful identification of important features. The improvements hold across both GPT-2, Mamba, and MedGemma models and multiple explanation methods.

(a) Insertion AUC (↑).

Dataset	Model	Attribution	LIME	SHAP	IntGrad	FullGrad	MFABA
PhysioNet	GPT-2	Feature-only	0.332	0.341	0.356	0.324	0.339
		PaX-Feature	**0.354**	**0.362**	**0.378**	**0.346**	**0.361**
	Mamba	Feature-only	0.335	0.343	0.358	0.329	0.346
		PaX-Feature	**0.357**	**0.366**	**0.381**	**0.351**	**0.369**
	MedGemma	Feature-only	0.338	0.346	0.361	0.331	0.348
		PaX-Feature	**0.344**	**0.352**	**0.368**	**0.337**	**0.355**
MC-MED	GPT-2	Feature-only	0.318	0.327	0.338	0.309	0.324
		PaX-Feature	**0.339**	**0.349**	**0.361**	**0.331**	**0.346**
	Mamba	Feature-only	0.321	0.334	0.341	0.314	0.329
		PaX-Feature	**0.343**	**0.356**	**0.364**	**0.336**	**0.351**
	MedGemma	Feature-only	0.324	0.335	0.344	0.317	0.331
		PaX-Feature	**0.330**	**0.342**	**0.351**	**0.323**	**0.338**

(b) Deletion AUC (↓).

Dataset	Model	Attribution	LIME	SHAP	IntGrad	FullGrad	MFABA
PhysioNet	GPT-2	Feature-only	0.152	0.149	0.151	0.147	0.150
		PaX-Feature	**0.138**	**0.134**	**0.137**	**0.135**	**0.136**
	Mamba	Feature-only	0.153	0.154	0.145	0.144	0.152
		PaX-Feature	**0.139**	**0.141**	**0.133**	**0.131**	**0.138**
	MedGemma	Feature-only	0.151	0.152	0.148	0.146	0.153
		PaX-Feature	**0.145**	**0.146**	**0.142**	**0.140**	**0.147**
MC-MED	GPT-2	Feature-only	0.162	0.168	0.182	0.179	0.165
		PaX-Feature	**0.148**	**0.151**	**0.164**	**0.162**	**0.149**
	Mamba	Feature-only	0.174	0.178	0.181	0.179	0.172
		PaX-Feature	**0.159**	**0.162**	**0.165**	**0.163**	**0.158**
	MedGemma	Feature-only	0.171	0.176	0.183	0.180	0.172
		PaX-Feature	**0.165**	**0.169**	**0.176**	**0.173**	**0.166**

methods (IntGrad, FullGrad, MFABA) alike, indicating that the benefit comes from separating positional and feature attributions rather than from interactions with a particular attribution algorithm.

D. Validating actionability: protocol and per-finding details

This appendix expands Section 3.3: it specifies the annotation protocol, presents the per-finding tables, and develops the interpretation behind each finding. Throughout, α_{LIME} denotes the standard LIME attribution, while $\alpha^{(x)}_{\text{PaX-LIME}}$ and $\alpha^{(p)}_{\text{PaX-LIME}}$ denote the feature-content and positional components of the PaX-LIME attribution.

D.1. Annotation protocol

By "formal guidelines" we mean the established sepsis criteria: Sepsis-3, SOFA, qSOFA, NEWS, and the anticipated SOFA-2 update (Ranzani et al., 2025). These criteria vary across institutions and continue to evolve, so bedside diagnosis depends on clinician judgment that no single guideline fully captures. The protocol below operationalizes that judgment. Clinicians annotated each measurement on two orthogonal axes.

Sepsis-Related axis. Three labels: *yes* (direct indicator), *partial* (indirect or contextual), and *no* (unrelated). Labels are grounded in sepsis specifically rather than in co-occurring pathology. Lipase, for example, is labeled *no* because it primarily indicates pancreatitis, even when pancreatitis appears alongside sepsis on a differential.

Table 3. Position faithfulness: PaX-Position achieve higher insertion AUC and lower deletion AUC than Position-only attribution, confirming more faithful identification of important positions. The improvements hold across both GPT-2, Mamba, and MedGemma models and multiple explanation methods.

(a) Insertion AUC (↑).

Dataset	Model	Attribution	LIME	SHAP	IntGrad	FullGrad	MFABA
PhysioNet	GPT-2	Position-only	0.325	0.336	0.348	0.318	0.327
		PaX-Position	**0.341**	**0.348**	**0.366**	**0.334**	**0.349**
	Mamba	Position-only	0.331	0.323	0.348	0.312	0.334
		PaX-Position	**0.344**	**0.341**	**0.369**	**0.333**	**0.352**
	MedGemma	Position-only	0.334	0.329	0.351	0.319	0.337
		PaX-Position	**0.340**	**0.335**	**0.358**	**0.325**	**0.344**
MC-MED	GPT-2	Position-only	0.301	0.314	0.322	0.296	0.312
		PaX-Position	**0.327**	**0.336**	**0.348**	**0.319**	**0.334**
	Mamba	Position-only	0.311	0.322	0.336	0.303	0.321
		PaX-Position	**0.331**	**0.344**	**0.352**	**0.325**	**0.339**
	MedGemma	Position-only	0.314	0.325	0.338	0.306	0.324
		PaX-Position	**0.321**	**0.332**	**0.345**	**0.313**	**0.335**

(b) Deletion AUC (↓).

Dataset	Model	Attribution	LIME	SHAP	IntGrad	FullGrad	MFABA
PhysioNet	GPT-2	Position-only	0.146	0.148	0.154	0.150	0.145
		PaX-Position	**0.141**	**0.139**	**0.142**	**0.138**	**0.140**
	Mamba	Position-only	0.149	0.148	0.147	0.146	0.147
		PaX-Position	**0.142**	**0.143**	**0.136**	**0.134**	**0.141**
	MedGemma	Position-only	0.150	0.149	0.151	0.147	0.148
		PaX-Position	**0.144**	**0.143**	**0.145**	**0.141**	**0.142**
MC-MED	GPT-2	Position-only	0.171	0.164	0.191	0.187	0.169
		PaX-Position	**0.156**	**0.159**	**0.173**	**0.171**	**0.155**
	Mamba	Position-only	0.183	0.186	0.189	0.187	0.181
		PaX-Position	**0.168**	**0.170**	**0.173**	**0.171**	**0.166**
	MedGemma	Position-only	0.179	0.182	0.190	0.186	0.178
		PaX-Position	**0.172**	**0.175**	**0.183**	**0.179**	**0.171**

Temporal axis. Two labels: *time-dependent* (trajectory carries diagnostic meaning, e.g., vitals) and *time-invariant* (stable across a hospitalization, e.g., demographics).

D.2. Finding 1: attributions track clinical relevance

PaX-LIME assigns higher mean importance to features clinicians label *yes* than to those labeled *partial* or *no*, and raises $\alpha^{(x)}_{\text{PaX-LIME}}$ relative to standard LIME on 24 of the 29 clinically relevant features (Table 4).

Error analysis: indirect indicators rather than spurious attributions. Five features labeled *no* receive higher attribution under PaX-LIME than under LIME. These are not failure cases: they are higher-order physiological signals that sit outside the formal sepsis definition but carry diagnostic weight in practice.

Two act as second-order markers. Chloride and Troponin I, although primarily linked to other pathologies (notably myocardial infarction for Troponin I), plausibly reflect cardiovascular involvement during severe septic shock. Three act as third-order markers: Phosphate, Magnesium, and Calcium are non-specific electrolytes that signal general physiological derangement. These are exactly the kind of secondary cues a clinician wants surfaced, not suppressed.

Table 4. PaX-LIME feature attributions stratified by clinician Sepsis-Related label. $\Delta_f = \alpha^{(x)}_{\text{PaX-LIME}} - \alpha_{\text{LIME}}$ measures the per-feature shift from standard LIME to PaX-LIME; the rightmost column counts features with $\Delta_f > 0$ among the 29 clinically relevant ones (*yes* and *partial* combined).

Clinician label	# features	Mean $\alpha^{(x)}_{\text{PaX-LIME}}$	# with $\Delta_f > 0$
Yes (direct)	20	0.40	24 / 29
Partial	9	0.31	
No (unrelated)	6	0.22	—

D.3. Finding 2: PaX-LIME elevates the uncodified-but-used measurements

The actionable surface of an explanation is the set of measurements clinicians use but formal criteria do not codify. Table 5 crosses clinician labels with guideline membership; the top-right cell — clinically meaningful, not in guidelines — contains 7 of the 20 *yes*-labeled features and all 9 *partial*-labeled features.

Table 5. Clinician Sepsis-Related labels against guideline membership. The top-right cell, clinically meaningful measurements not codified in formal criteria, is the actionable surface.

Clinician label	In guidelines	Not in guidelines
Yes	13	7
Partial	0	9
No	0	6

PaX-LIME's top ten features by $\alpha^{(x)}_{\text{PaX-LIME}}$ are Age, Bilirubin, pH, Lactate, Temp, FiO_2, Platelets, BUN, Resp, and AST. Only six appear in formal guidelines, so at face value four of the top ten diverge from established criteria. Clinician annotations resolve the divergence: nine of the ten are labeled *yes* and one *partial*, and the four uncodified features (Age, pH, BUN, AST) are each independently endorsed. The headline signals PaX-LIME elevates are precisely the measurements an actionable bedside explanation should foreground.

D.4. Finding 3: separating positional artifact from intrinsic signal

For time-invariant markers the positional component $\alpha^{(p)}_{\text{PaX-LIME}}$ should be small: a feature whose value is fixed at admission cannot meaningfully owe its importance to position. As a consistency check, mean $\alpha^{(p)}_{\text{PaX-LIME}}$ is 0.61 for time-dependent features and 0.33 for time-invariant features, in the predicted direction.

The decomposition matters most when standard LIME conflates two features that PaX-LIME should distinguish. Age and Gender are a clean test case (Table 6).

Table 6. Time-invariant features under standard LIME vs. PaX-LIME. Standard LIME scores Age and Gender comparably; PaX-LIME shows that Gender's score is largely positional while Age's remains intrinsic.

Feature	α_{LIME}	$\alpha^{(x)}_{\text{PaX-LIME}}$
Age	0.76	0.77
Gender	0.73	0.33

Standard LIME scores the two features comparably. PaX-LIME preserves Age's intrinsic importance (0.77) but more than halves Gender's (0.33). Because PhysioNet places demographics at admission ($t = 0$), this divergence is consistent with a primacy bias: tokens at the start of a sequence draw disproportionate attention regardless of content, inflating the apparent importance of any feature that happens to sit there. Age survives the decomposition because its content carries genuine signal; Gender does not. Telling structural bias apart from learned biological correlation is exactly the audit a clinical explanation must support.