TRACE: TEMPORAL RELEVANCE ATTRIBUTION FOR CLINICAL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Irregularly sampled multivariate clinical time series are central to intensive care unit (ICU) risk prediction, yet deployable models must be both highly accurate and faithfully interpretable. Practitioners currently face a sharp trade-off: inherently interpretable architectures often rely on heuristic attention mechanisms without strict faithfulness guarantees, while rigorous post-hoc methods such as Integrated Gradients incur prohibitive computational costs. We introduce Temporal Relevance Attribution for Clinical Explanations (TRACE), a framework that resolves this tension through architecture-attribution co-design, achieving the strongest predictive performance together with single-pass, mathematically exact attribution. TRACE structurally reduces to an affine map for any fixed input, enabling a single backward pass to exactly decompose the prediction logit into signed per-variable, per-timestep contributions with a completeness gap at floating-point round-off. On PhysioNet-2012, PhysioNet-2019, and MIMIC-III, TRACE achieves the highest AUROC and AUPRC of all baselines. Under a graded remove-and-retrain evaluation (gROAR) that sweeps retention budgets with trend and equivalence testing, TRACE consistently outperforms existing inherently interpretable models. Its top-k degradation is comparable to that of Integrated Gradients, yet it runs two orders of magnitude faster and preserves machine-precision faithfulness, with a relative completeness gap on the order of 10^{-6}. Together, these results show that predictive accuracy, attribution faithfulness, and computational efficiency need not be traded off in clinical time-series modeling.

1 INTRODUCTION

Intensive care units (ICUs) produce continuous, high-frequency streams of physiological measurements that clinicians must integrate, under time pressure, into decisions with life-or-death consequences. Predicting adverse outcomes such as in-hospital mortality or the early onset of sepsis from these streams is of direct clinical value (Rajkomar et al., 2018; Komorowski et al., 2018; Moor et al., 2021; Sendak et al., 2020), and a series of public ICU benchmarks has standardized its study (Silva et al., 2012; Reyna et al., 2019; Johnson et al., 2016; Harutyunyan et al., 2019). Because deterioration usually manifests as a diffuse pattern across many weakly informative variables and time points, a risk score is actionable only if clinicians can identify *which* observations drive it and in *which direction*.

Modeling these signals is challenging for two reasons. Clinical time series are *irregularly sampled*, variables are measured at different, condition-dependent rates, so the data do not form a regular tensor. They are also *missing-not-at-random*, an absent laboratory value is not a normal one, and the decision to measure carries information. GRU-D first showed that modeling elapsed time and informative missingness explicitly is essential (Che et al., 2018), and much subsequent work has treated irregularity as a first-class modeling target (Cao et al., 2018; Shukla & Marlin, 2019; 2021; Rubanova et al., 2019; Sun et al., 2026).

Transparency is an equally important requirement. Predictions in high-stakes clinical settings must be inspectable and challengeable, and several authors have argued that such decisions should favor models that are interpretable by design rather than relying on post-hoc explanations (Rudin, 2019; Caruana et al., 2015). Clinicians, in particular, need an explanation that is *signed* (which direction an observation pushes the risk), *deterministic* (free of sampling and discretization variance), and *inexpensive* enough for bedside use (Tonekaboni et al., 2019); few existing methods satisfy all three.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches address only subsets of these requirements. Attention-based clinical models are inexpensive, but the attention weights themselves are non-negative and carry no direction (Guo et al., 2019), and their faithfulness is debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019); RETAIN obtains signed contributions by keeping its prediction linear in the input embeddings (Choi et al., 2016), at a cost in capacity (Table 1). Post-hoc methods are signed, but Integrated Gradients and SHAP need many forward passes for path integration or sampling (Sundararajan et al., 2017; Lundberg & Lee, 2017), DeepLIFT is single-pass yet approximate for nonlinear units (Shrikumar et al., 2017), and all are fragile to perturbations and to the evaluation protocol (Ghorbani et al., 2019; Adebayo et al., 2018; Hooker et al., 2019). Interpretable-by-construction networks such as SENN and B-cos (Alvarez Melis & Jaakkola, 2018; Böhle et al., 2022) are signed and inexpensive, but their linear coefficients are input-dependent outputs of the network itself, and neither addresses irregular sampling (Appendix K.3). To our knowledge, no existing method for irregularly sampled clinical time series is simultaneously signed, exact with completeness at machine precision, and single-pass.

We present TRACE (Temporal Relevance Attribution for Clinical Explanations), a clinical time-series framework co-designed for prediction and attribution. The design principle is deliberately minimal. Identify the few mechanisms that make exact credit assignment ill-posed, and replace only those. TRACE therefore keeps a high-capacity front end for irregular multivariate data (decay-aware imputation, per-variable encoding, and a learnable inter-variable graph) and replaces only the temporal core with a softmax-free retentive encoder (Sun et al., 2023), whose input-independent causal weighting makes the contribution of any past state to a future output a single closed-form coefficient (Section 3.3). For any fixed input, TRACE thus reduces to an affine map whose logit decomposes exactly into signed per-variable, per-timestep contributions in a single backward pass.

Our contributions are threefold:

- We co-design the architecture so that each prediction logit decomposes exactly into signed per-variable, per-timestep contributions in one backward pass.
- We instantiate this principle as TRACE, which pairs a graph-attention front end with a retentive temporal encoder and achieves the highest AUROC and AUPRC on three ICU benchmarks.
- We evaluate prediction and interpretability jointly, introducing graded ROAR (gROAR), a graded extension of remove-and-retrain with trend and equivalence testing, alongside faithfulness and efficiency benchmarks against gradient baselines and patient-level case studies.

2 Related Work

We summarize the four threads TRACE builds on; a fuller discussion of each is given in Appendix K.

Irregular clinical time series. GRU-D established that elapsed-time decay and missingness masks are essential for ICU prediction (Che et al., 2018); later work pursued learned imputation (Cao et al., 2018), interpolation and multi-time attention (Shukla & Marlin, 2019; 2021), and continuous-time dynamics (Rubanova et al., 2019; Kidger et al., 2020). TRACE adopts GRU-D's decay-aware imputation but factors it out as a standalone input-linear step whose coefficients depend only on elapsed time, so that for a fixed observation pattern the map from observed to imputed inputs is linear and does not entangle the attribution path.

Relational and Transformer models. Graph-based models learn dependencies among physiological variables (Zhang et al., 2021; Piao et al., 2025); GARLIC (Wang et al., 2026) combines graph message passing with sequential attention, but its explanations are unsigned attention weights. Transformers adapted to time series (Nie et al., 2022; Liu et al., 2024; Zhang et al., 2023) require an imputation or binning front-end for irregular data; we equip them with the same decay imputation for a paired comparison. TRACE keeps GARLIC's relational front-end and replaces only its temporal core.

Interpretability of clinical sequence models. Intrinsically interpretable predictors expose attention-style importances (Choi et al., 2016; Guo et al., 2019) whose faithfulness remains debated (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Post-hoc methods are signed: Integrated Gradients and SHAP (Sundararajan et al., 2017; Lundberg & Lee, 2017) and their sequence-specific variants (Tonekaboni et al., 2020; Crabbé & van der Schaar, 2021; Bento et al., 2021) rely on path integration, perturbation, or sampling, whose cost grows with sequence length, whereas DeepLIFT (Shrikumar et al., 2017) is single-pass but approximate for nonlinear units; their rankings are fragile and protocol-

dependent (Ghorbani et al., 2019; Adebayo et al., 2018; Hooker et al., 2019); attention-LRP rules (Chefer et al., 2021; Ali et al., 2022) conserve relevance only once attention is frozen. Interpretable-by-construction networks such as SENN and B-cos (Alvarez Melis & Jaakkola, 2018; Böhle et al., 2022) obtain a local linear summary through input-dependent coefficients; TRACE instead fixes its temporal coefficients by architecture and confines the remaining input dependence to attention scores frozen at evaluation, which makes the temporal coefficients unique and closed-form (Proposition 1).

Linear state-space and retentive models. S4, Mamba, and RetNet (Gu et al., 2021; Gu & Dao, 2023; Sun et al., 2023) replace gated recurrence with structured linear mixing. S4 and RetNet use input-independent temporal coefficients, whereas Mamba's selectivity re-introduces input-dependent gating and, with it, the non-unique decomposition that afflicts GRUs. TRACE therefore adopts RetNet's input-independent causal weighting; to our knowledge this interpretability property of the RetNet family has not been exploited. More broadly, the co-design principle is to diagnose which mechanisms make exact credit assignment ill-posed (input-dependent temporal mixing and softmax normalization on the credit path) and to replace only those; the diagnosis is not specific to retention or to healthcare.

3 METHODS

3.1 PROBLEM SETUP AND NOTATION

We consider binary outcome prediction, such as in-hospital mortality or early sepsis, from irregularly sampled, multivariate ICU time series. For each patient, we observe a time series $X \in \mathbb{R}^{K\times T}$ over K clinical variables and T (binned) time steps, together with a binary observation mask $M \in \{0,1\}^{K\times T}$, where $m_{k,t} = 1$ if variable k is observed at time t, the corresponding observation timestamps, and a binary outcome label $y \in \{0,1\}$. Our objectives are twofold: (i) estimate the outcome probability $\hat{y} = \sigma(\hat{z}) = \Pr(y = 1 \mid X, M)$ through a scalar logit $\hat{z}$, and (ii) produce an attribution matrix $C \in \mathbb{R}^{K\times T}$ that *exactly* reconstructs the logit, $\hat{z} = \sum_{k,t} C_{k,t} + b_{\text{global}}$, where each $C_{k,t}$ is the signed contribution of imputed input $\tilde{x}_{k,t}$ (Appendix A.1). Throughout, we use D to denote the per-variable embedding dimension, H the temporal hidden dimension, W the sliding-window length, and L the number of retention layers.

3.2 TRACE ARCHITECTURE

TRACE retains the strong irregular time-series feature extractor of recent graph-attention ICU models while replacing the components that prevent exact attribution. Figure 1 provides an overview of the architecture; full equations are in Appendix A.1. The architecture comprises seven stages: (1) input-linear decay imputation; (2) per-variable encoders; (3) per-variable time-lag attention for local temporal patterns; (4) linear graph message passing via a learnable adjacency matrix $A \in \mathbb{R}^{K\times K}$; (5) cross-variable signal attention that fuses K channels into one vector per position; (6) stacked multi-head softmax-free retention layers with input-independent causal weighting, GroupNorm, and residual FFN; (7) temporal attention pooling followed by a classifier.

TRACE is intentionally modular. The decay imputation follows GRU-D's input-decay principle (Che et al., 2018), factored into a standalone input-linear step (Appendix K.1); the learnable variable graph follows Raindrop (Zhang et al., 2021); and the relational front end as a whole (stages 1–5) is inherited from GARLIC (Wang et al., 2026). TRACE departs from GARLIC at the temporal core. The GRU is replaced by stacked *softmax-free* retention adapted from RetNet (Sun et al., 2023) (stage 6), with GroupNorm taking over the stabilizing role of the softmax. Section 3.3 explains why this single replacement is what makes exact attribution possible. Empirically it costs no accuracy. TRACE matches or exceeds GARLIC on all six dataset×metric cells (Table 1) while upgrading its attention-weight explanations to an exact, signed decomposition of the logit.

3.3 EXACT ATTRIBUTION VIA ONE-PASS DECOMPOSITION

For every prediction we seek signed, logit-scale contributions $C_{k,t}$ for each imputed input $\tilde{x}_{k,t}$ satisfying

$$\hat{z} \;=\; \sum_{k,t} C_{k,t} \;+\; b_{\text{global}}. \tag{1}$$

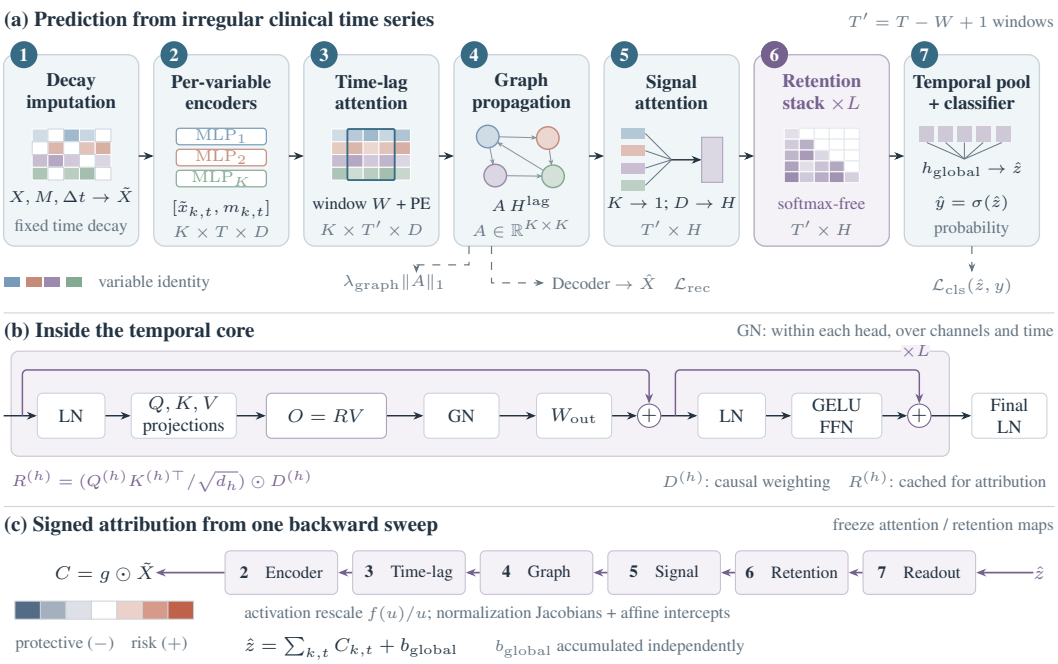


Figure 1: TRACE prediction and attribution. (a) Seven stages map irregular inputs, observation masks, and elapsed times to the logit $\hat{z}$ and probability $\hat{y}$. Colored rows identify variables; the dashed decoder branch supplies reconstruction supervision outside the prediction path. Time-lag, signal, and temporal attention retain softmax. (b) Each retention block contains pre-LayerNorm, multi-head softmax-free mixing, GroupNorm, an output projection, and a residual GELU FFN; a final LayerNorm follows the L blocks; $D^{(h)}$ is the input-independent causal weighting (Appendix A.1). (c) With attention and retention maps cached, one reverse sweep through stages 7–2 ends at the imputed input $\tilde{X}$, yielding signed contributions $C = g \odot \tilde{X}$. The independently accumulated affine terms complete $\hat{z} = \sum_{k,t} C_{k,t} + b_{\mathrm{global}}$.

Here, b_{global} collects the network's additive constants (layer biases, positional encodings, mask-path terms, and normalization intercepts), and the whole decomposition is computed in a single backward pass.

Once a forward pass is complete, all intermediate activations and dynamic coefficients, including attention weights and retention maps, are fixed at their operating-point values. The linearization tools themselves are classical; what is new is the architectural choice that makes them compose to an exact, unique decomposition end-to-end, and the falsifiable protocol that verifies it. We construct an equivalent local linear representation of the network by treating each attention or retention map as a fixed linear operator and replacing each elementwise nonlinearity f by its effective slope (Shrikumar et al., 2017; Ancona et al., 2017),

$$S_{\mathrm{eff}}(z) \;=\; f(z)/z, \qquad \text{so that} \quad f(z) \;=\; S_{\mathrm{eff}}(z)\, z \quad \textit{exactly} \text{ at the operating point.} \tag{2}$$

Unlike the ordinary derivative, which defines only a tangent approximation, this representation reproduces the activation exactly at the operating point. The primary obstacle in standard sequence models is the *temporal mixing*. A gated recurrence couples time steps through input-dependent products $z_t \odot h_{t-1}$: freezing either factor is an arbitrary convention, so the resulting coefficients are non-unique, and the influence of step j on step t becomes a product of $O(T)$ frozen matrices (Proposition 1).

TRACE constructs its temporal core so that the backward computation is a sequence of fixed linear operations. Throughout, g with a subscript denotes the effective gradient of the logit $\hat{z}$ with respect to the intermediate quantity indicated by the subscript. Retention mixes time steps as $o_t^{(h)} = \sum_{j \le t} R_{t,j}^{(h)}\, V_j^{(h)}$, in which $R_{t,j}^{(h)}$ is the retention coefficient from time step j to time step t and $V_j^{(h)}$ is the value representation of time step j for head h (Appendix A.1). Note that, without a softmax denominator, once the retention coefficients have been evaluated the corresponding backward

operation is a single matrix transpose:

$$g_{V^{(h)}} \;=\; \big(R^{(h)}\big)^{\top} g_{o^{(h)}}, \tag{3}$$

in which $g_{o^{(h)}}$ is the effective gradient with respect to the retention output $o^{(h)}$, $g_{V^{(h)}}$ is the effective gradient with respect to the values $V^{(h)}$ (both stacked over time steps). Because the causal weighting $D^{(h)}$ is input-independent and no gate multiplies the value stream, the coefficient of step j is a single closed-form scalar. The graph operator is likewise strictly linear, so its backward operation is exactly the transposed adjacency transformation, without approximation: $g_{H^{\text{lag}}} = A^{\top} g_{H^{\text{latent}}}$, in which $g_{H^{\text{latent}}}$ is the effective gradient with respect to the message-passing output H^{latent}, A is the learnable adjacency matrix, and $g_{H^{\text{lag}}}$ is the effective gradient with respect to its input H^{lag} (Appendix A.1). Chaining these frozen operators from the classifier back to the imputed input (stages $7 \to 2$) yields the effective gradients $g_{k,t}$ and the signed contributions $C_{k,t} = g_{k,t}\,\tilde{x}_{k,t}$ (positive = risk-increasing, negative = protective) at a computational cost of the same order as a forward evaluation (Appendix B); the full per-layer derivation is in Appendix A.2.

TRACE does *not* remove every nonlinearity. The three attention blocks (time-lag, signal, temporal) keep their softmax; at attribution time their weights are treated as constants and only the value path is attributed, as in attention-LRP (Chefer et al., 2021; Ali et al., 2022). We inherit the limitation of that convention: *why* the attention weights take their values is not itself attributed (Appendix H). What TRACE removes is the obstruction in the temporal core. In a gated recurrence, the influence of step j on step t is a product of up to T frozen gate matrices, and which factor to freeze is a convention, so the coefficients are neither unique nor closed-form. With gating replaced by softmax-free retention whose causal weighting is input-independent, the frozen network is, for a fixed input, an affine function of $\tilde{X}$, and the temporal coefficients of the decomposition in Eq. (1) are unique (Proposition 1; Appendix I).

Completeness is tested as a *falsifiable* identity: the constant b_{global} is accumulated independently by propagating every additive term of the frozen network (layer biases, positional encodings, normalization intercepts) through the same frozen operators, so a bug in any single operator produces a visible gap between the two sides of Eq. (1). Any approximation introduced by the decomposition would likewise manifest as a nonzero completeness gap between the two sides of the equation; during development this check caught two implementation bugs that a residual-recovery shortcut cannot see (Appendix A.2). In practice, on held-out test samples TRACE yields a relative completeness gap at single-precision round-off ($\sim 10^{-7}$–10^{-6}, Table 2) that falls below 10^{-13} when the same computation is run in double precision, i.e., within machine precision: the residual scales with machine epsilon, not with any approximation term, confirming that the decomposition is numerically exact. An approximation term, such as the Riemann discretization error of Integrated Gradients, is precision-independent and would remain at its $\sim 10^{-2}$ level in double precision.

3.4 Training Procedure

TRACE is trained with a three-term objective

$$\mathcal{L} \;=\; \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}} \;+\; \mathcal{L}_{\text{rec}} \;+\; \lambda_{\text{graph}}\,\|A\|_1, \tag{4}$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy on the outcome label, $\mathcal{L}_{\text{rec}}$ is the masked reconstruction MSE from the message-passing branch (its weight fixed to 1), and the ℓ_1 penalty on the learnable adjacency A encourages a sparse, interpretable variable graph. The reconstruction/graph and classification pathways are optimized by a alternating optimization schedule with separate Adam parameter groups; the warm-up and early-stopping schedule are detailed in Appendix D.

4 Experiments

4.1 Setup

We evaluate on three public ICU benchmarks: PhysioNet-2012 (P12, in-hospital mortality) (Silva et al., 2012), PhysioNet-2019 (P19, early sepsis) (Reyna et al., 2019), and MIMIC-III (ICU cohort, in-hospital mortality) (Johnson et al., 2016). The data are divided once into a stratified 80/10/10 train/validation/test split that is fixed and shared across every method; only the training seed varies, and each model is trained with five different seeds (Appendix D). We compare against nine baseline

models spanning several families: models for irregular series (GRU-D (Che et al., 2018), Raindrop (Zhang et al., 2021), mTAND (Shukla & Marlin, 2021), and MTSFormer (Zheng et al., 2024)), regular-series transformers with a decay front-end (PatchTST (Nie et al., 2022) and iTransformer (Liu et al., 2024)), and intrinsically interpretable models (IMV-LSTM (Guo et al., 2019), RETAIN (Choi et al., 2016), and GARLIC (Wang et al., 2026)). Reproduction details and selected configurations are in Appendix F. We report AUROC and AUPRC, given the strong positive-class imbalance; all values are 5-seed mean $\pm$ sample standard deviation. All runs use a single NVIDIA RTX 5090 (Appendix D).

4.2 Prediction

Table 1 demonstrates that TRACE achieves the strongest predictive performance overall, with the highest AUROC and AUPRC on all three datasets, while being the *only* model in the comparison that also provides an exact, complete attribution. Against the strongest interpretable baseline, TRACE improves AUPRC by 3.5 percentage points on P12, 1.2 points on P19, and 2.8 points on MIMIC-III; against non-interpretable models, it also remains superior. All models, baselines included, are tuned on the same split (Appendix F).

Table 1: Predictive performance on the three ICU benchmarks. Values are the mean $\pm$ sample standard deviation over five random seeds. The best result in each column is shown in bold.

	P12		P19		MIMIC-III	
Models	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
GRU-D	86.36 ± 0.22	56.21 ± 0.57	89.46 ± 0.31	54.75 ± 1.16	90.09 ± 0.29	65.77 ± 0.50
mTAND	87.33 ± 0.39	58.28 ± 1.62	83.97 ± 0.56	35.75 ± 1.26	90.03 ± 0.23	66.14 ± 0.66
RAINDROP	83.43 ± 0.80	46.91 ± 2.95	88.05 ± 1.06	51.33 ± 2.33	89.12 ± 0.34	62.29 ± 1.24
MTSFormer	85.80 ± 0.24	56.44 ± 0.79	90.37 ± 0.63	57.12 ± 1.74	90.15 ± 0.29	64.02 ± 1.41
PatchTST	86.24 ± 0.48	56.63 ± 1.89	86.93 ± 0.49	47.01 ± 1.60	89.82 ± 0.32	65.80 ± 0.53
iTransformer	86.60 ± 0.38	57.21 ± 1.35	87.60 ± 0.43	50.32 ± 1.43	90.33 ± 0.13	67.41 ± 0.41
IMV-LSTM	80.21 ± 1.24	43.58 ± 2.43	88.20 ± 0.83	48.34 ± 1.11	86.81 ± 1.31	58.11 ± 2.52
RETAIN	83.45 ± 1.28	48.40 ± 2.21	86.23 ± 0.74	45.45 ± 1.91	88.90 ± 0.43	62.67 ± 1.16
GARLIC	87.04 ± 0.62	56.76 ± 1.88	91.38 ± 0.36	57.93 ± 1.02	90.20 ± 0.35	66.36 ± 0.92
TRACE (Ours)	$\mathbf{87.52 \pm 0.14}$	$\mathbf{60.25 \pm 0.72}$	$\mathbf{91.46 \pm 0.41}$	$\mathbf{59.15 \pm 0.71}$	$\mathbf{91.12 \pm 0.33}$	$\mathbf{69.11 \pm 0.68}$

4.3 Attribution Fidelity and Efficiency

We compare against Gradient$\times$Input (Shrikumar et al., 2017; Ancona et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) (IG, n=50 steps). IG is a standard post-hoc method with a completeness guarantee in the continuous-path limit (Sundararajan et al., 2017). We omit SHAP: kernel-based SHAP is computationally intractable for inputs of dimensionality $K \times T$ in our test. The comparison thus spans the practical spectrum from fast-but-unfaithful to slow-but-faithful. We measure the relative completeness gap: $\text{gap}_{\text{rel}} = |\hat{z} - (\sum_{k,t} C_{k,t} + b_{\text{global}})| \,/\, |\hat{z}|$, following the infidelity framework of Yeh et al. (2019). In Table 2, Gradient$\times$Input leaves a large relative gap (≈ 4), reflecting its lack of any completeness guarantee; IG reduces this to $\sim 10^{-2}$ but requires 50 forward–backward passes; TRACE's exact decomposition leaves only floating-point round-off in a single backward pass, running 173–290$\times$ faster than IG.

4.4 Attribution Faithfulness: Comparison with Intrinsic Baselines

We assess attribution faithfulness with *graded ROAR* (gROAR), our graded extension of the remove-and-retrain protocol of Hooker et al. (2019): inputs are ranked by attribution magnitude, the top $k\%$ are retained, a *fresh model is retrained* on the masked data, and we measure the performance retained.[1] Rather than judging a single retention budget, gROAR sweeps $k \in \{10, 20, 30, 40, 50\}$

[1] Hooker et al. (2019) call the removal variant ROAR and the retention variant KAR; gROAR uses retention (Appendix G explains why) and keeps the ROAR name for the retrain-based family.

Table 2: Faithfulness (relative completeness gap, lower is better) and efficiency (ms per test sample; speedup vs. IG n=50) on the same TRACE checkpoint. Cross-seed mean over 5 seeds. TRACE's gap is measured under the falsifiable independent-bias protocol (Section 3.3) in single precision; in double precision it falls below 10^{-13}. Timings cover the one-pass computation of the contributions C; the independent-bias accumulation is a verification step and is not timed.

Dataset	Method	$\text{gap}_{\text{rel}} \downarrow$	ms/sample↓	speedup
P12	Gradient×Input	4.29	3.99	50×
P12	IG (n=50)	1.17×10^{-2}	200.9	1×
P12	**TRACE (exact)**	$\mathbf{4.1 \times 10^{-6}}$	**0.69**	**290×**
P19	Gradient×Input	3.67	0.89	50×
P19	IG (n=50)	1.00×10^{-2}	44.8	1×
P19	**TRACE (exact)**	$\mathbf{3.9 \times 10^{-7}}$	**0.26**	**173×**

and evaluates the sweep with two tests, Page's L for a monotone degradation trend and TOST for Top-50/Full equivalence. We compare against the other intrinsically interpretable models (IMV-LSTM, RETAIN, GARLIC), each using its own native importance. The retraining sweeps are run on P12 and P19; every cell retrains a model from scratch, over 200 runs per dataset across models, budgets, rankings, and seeds. This is computationally prohibitive on MIMIC-III's 103-variable, T=200 grid, so MIMIC-III is evaluated on prediction (Table 1) only. We use the full sweep rather than the prior single-point Top-/Bottom-50% protocol, which is confounded by "data starvation" (Appendix G). Table 3 shows that TRACE's exact decomposition degrades the most gracefully: at the 10% budget its AUROC/AUPRC are 6%/24% higher than the next-best method on P12 and 9%/67% higher on P19. Page's L test confirms a significant monotone trend for all methods. On P12 TRACE retains AUROC 84.43 / AUPRC 51.66 at a 10% budget and its Top-50% model is statistically equivalent to the full model by a paired TOST (Schuirmann, 1987). The same ordering holds on P19. TRACE is the only method achieving TOST equivalence on all four dataset×metric cells, confirming that its attribution captures genuinely informative inputs. Because each model is ranked by its own importance, these differences reflect both attribution quality and model capacity; Section 4.5 controls for model quality by comparing attribution strategies on the same checkpoint.

4.5 Same-Model Attribution Comparison

To isolate attribution quality from model quality, we run four attribution strategies on the *same* TRACE checkpoint and compare under the gROAR protocol: random, GARLIC-style attention-based attribution computed from TRACE's own attention weights, TRACE's exact attribution, and IG (n=32; the sweep's cost is dominated by retraining, and IG here serves only to rank inputs, so fewer steps suffice than in the faithfulness benchmark of Section 4.3). Table 4 reports the full results. Both TRACE exact and IG substantially outperform random and attention-based baselines at every budget: at the 10% budget TRACE's AUROC/AUPRC are 9%/26% higher than the attention-based ranking on P12 and 9%/67% higher on P19, confirming that gradient-informed attributions identify more informative subsets. TRACE's attribution matches or exceeds IG on P19 across all budgets and remains competitive on P12, while being over two orders of magnitude faster (Table 2). The attention-based method performs at or below random at low retention, consistent with findings that attention weights are unreliable importance indicators (Jain & Wallace, 2019).

4.6 Case Study

Because TRACE produces signed attributions, it distinguishes evidence that increases predicted risk from evidence that decreases it, a distinction unavailable from non-negative attention weights alone. Figure 2 illustrates four properties on representative patients. As the Glasgow Coma Scale (GCS) score deteriorates (a), attribution flips from negative (protective) to positive (risk-driving), conveying clinical direction. An isolated heart rate (HR) dip (b), a transient excursion of the kind typically produced by measurement artifacts, receives near-zero attribution at its minimum, whereas a sustained acute HR rise (c) is captured with strong positive attribution, consistent with sensitivity to real clinical events. The multi-variable view (d) for a 60-year-old medical-ICU patient who died (predicted mortality 0.992) shows how the attributions of GCS, lactate, HR, and creatinine jointly

Table 3: Interpretability assessment via gROAR on P12 and P19. Page's L tests the monotone trend full $>$ Top-50 $>$ $\ldots$ $>$ Top-10 (Page, 1963); TOST tests equivalence between Top-50 and Full with margins $\pm 4\%/\pm 8\%$ for AUROC/AUPRC (Schuirmann, 1987); margin rationale in Appendix J. (*: $p < 0.05$; **: $p < 0.005$)

(a) P12 AUROC/AUPRC (mean $\pm$ std)

	IMV-LSTM		RETAIN		GARLIC		Ours	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	64.50 ± 2.91	26.54 ± 3.75	79.93 ± 0.69	41.54 ± 1.67	74.13 ± 1.23	36.00 ± 3.88	**84.43 ± 0.83**	**51.66 ± 2.03**
Top 20	71.56 ± 6.89	32.37 ± 7.73	81.76 ± 1.46	42.82 ± 3.81	81.55 ± 0.81	45.58 ± 1.82	**85.96 ± 0.52**	**55.43 ± 1.51**
Top 30	76.04 ± 2.85	37.73 ± 3.35	82.00 ± 1.38	43.86 ± 1.56	84.34 ± 1.13	52.01 ± 2.27	**86.52 ± 0.83**	**56.32 ± 1.57**
Top 40	72.49 ± 5.22	33.22 ± 7.41	83.02 ± 0.83	46.05 ± 2.80	86.72 ± 0.55	57.43 ± 0.93	**86.87 ± 0.98**	**57.44 ± 2.18**
Top 50	74.88 ± 2.69	36.76 ± 3.07	83.71 ± 0.94	46.49 ± 3.80	86.29 ± 0.39	56.33 ± 1.31	**87.32 ± 0.37**	**58.75 ± 0.60**
Full	80.21 ± 1.24	43.58 ± 2.43	83.45 ± 1.28	48.40 ± 2.21	87.04 ± 0.62	56.76 ± 1.88	**87.52 ± 0.14**	**60.25 ± 0.72**
Page's L	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **
TOST	0.818 X	0.251 X	< 0.005 **	0.025 *	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **

(b) P19 AUROC/AUPRC (mean $\pm$ std)

	IMV-LSTM		RETAIN		GARLIC		Ours	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	73.19 ± 6.85	20.79 ± 5.69	76.54 ± 1.74	23.98 ± 1.96	78.62 ± 1.00	27.93 ± 1.09	**86.04 ± 0.49**	**46.60 ± 1.00**
Top 20	76.46 ± 7.52	26.50 ± 6.22	80.39 ± 0.61	28.53 ± 2.56	82.34 ± 0.89	31.95 ± 1.25	**88.67 ± 0.38**	**51.99 ± 1.88**
Top 30	80.42 ± 4.55	34.27 ± 7.99	80.62 ± 0.81	28.96 ± 4.80	86.33 ± 0.71	38.72 ± 0.97	**89.28 ± 0.66**	**53.59 ± 1.42**
Top 40	83.68 ± 2.95	39.83 ± 5.79	81.55 ± 0.40	31.15 ± 2.71	88.43 ± 0.79	46.72 ± 2.66	**90.77 ± 0.28**	**56.33 ± 1.30**
Top 50	82.05 ± 4.80	36.08 ± 8.99	81.70 ± 1.76	30.84 ± 3.54	89.22 ± 0.50	50.86 ± 1.15	**90.78 ± 0.51**	**57.52 ± 0.91**
Full	88.20 ± 0.83	48.34 ± 1.11	86.23 ± 0.74	45.45 ± 1.91	91.38 ± 0.36	57.93 ± 1.02	**91.46 ± 0.41**	**59.15 ± 0.71**
Page's L p	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **
TOST p	0.850 X	0.841 X	0.842 X	0.997 X	< 0.005 **	0.124 X	< 0.005 **	< 0.005 **

Table 4: Attribution comparison on the TRACE checkpoint. Four attribution methods rank the Top-$k\%$ and retrain on retained Top-$k\%$ data (the gROAR protocol). Page's L tests the monotone trend; TOST tests equivalence between Top-50 and Full, margins $\pm 4\%/\pm 8\%$ for AUROC/AUPRC (rationale: Appendix J). (*: $p < 0.05$; **: $p < 0.005$)

(a) P12 AUROC/AUPRC (mean $\pm$ std)

	Random		Attention-based		TRACE		IG (n=32)	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	78.07 ± 0.96	41.08 ± 1.39	77.46 ± 0.64	41.11 ± 0.65	84.43 ± 0.83	51.66 ± 2.03	**85.15 ± 1.53**	**55.77 ± 1.54**
Top 20	81.73 ± 1.31	48.30 ± 2.72	82.38 ± 1.61	49.32 ± 2.10	85.96 ± 0.52	55.43 ± 1.51	**86.62 ± 0.37**	**58.34 ± 1.19**
Top 30	83.84 ± 1.15	51.14 ± 2.19	85.22 ± 0.27	55.21 ± 0.88	86.52 ± 0.83	56.32 ± 1.57	**87.01 ± 0.40**	**58.73 ± 0.96**
Top 40	84.94 ± 0.90	52.30 ± 0.78	86.12 ± 0.55	56.52 ± 1.00	86.87 ± 0.98	57.44 ± 2.18	**87.27 ± 0.27**	**59.27 ± 0.71**
Top 50	85.41 ± 0.31	54.90 ± 1.22	86.84 ± 0.19	58.67 ± 0.64	**87.32 ± 0.37**	58.75 ± 0.60	87.26 ± 0.33	**59.56 ± 1.15**
Full	87.52 ± 0.14	60.25 ± 0.72	87.52 ± 0.14	60.25 ± 0.72	87.52 ± 0.14	60.25 ± 0.72	87.52 ± 0.14	60.25 ± 0.72
Page's L	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **
TOST	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **

(b) P19 AUROC/AUPRC (mean $\pm$ std)

	Random		Attention-based		TRACE		IG (n=32)	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	78.45 ± 1.43	31.18 ± 1.95	78.62 ± 1.08	27.96 ± 2.64	**86.04 ± 0.49**	**46.60 ± 1.00**	85.64 ± 1.73	43.45 ± 2.95
Top 20	83.61 ± 0.61	38.22 ± 2.29	83.45 ± 1.88	35.16 ± 2.44	**88.67 ± 0.38**	**51.99 ± 1.88**	87.48 ± 0.52	48.75 ± 1.56
Top 30	84.97 ± 1.10	41.39 ± 1.59	86.14 ± 1.43	40.60 ± 3.98	**89.28 ± 0.66**	**53.59 ± 1.42**	89.28 ± 1.00	53.52 ± 2.53
Top 40	86.91 ± 0.76	46.55 ± 1.15	88.46 ± 0.84	47.49 ± 3.44	**90.77 ± 0.28**	**56.33 ± 1.30**	90.08 ± 1.27	54.40 ± 3.17
Top 50	87.56 ± 0.43	47.80 ± 2.57	89.27 ± 0.20	51.34 ± 0.91	**90.78 ± 0.51**	**57.52 ± 0.91**	90.70 ± 0.53	56.99 ± 2.23
Full	91.46 ± 0.41	59.15 ± 0.71	91.46 ± 0.41	59.15 ± 0.71	91.46 ± 0.41	59.15 ± 0.71	91.46 ± 0.41	59.15 ± 0.71
Page's L	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **
TOST	0.372 X	0.985 X	< 0.005 **	0.363 X	< 0.005 **	< 0.005 **	< 0.005 **	< 0.005 **

describe the physiological trajectory: the neurological, perfusion, and cardiac signals accumulate positive (risk-driving) credit, while the recovering creatinine trajectory is scored as protective.

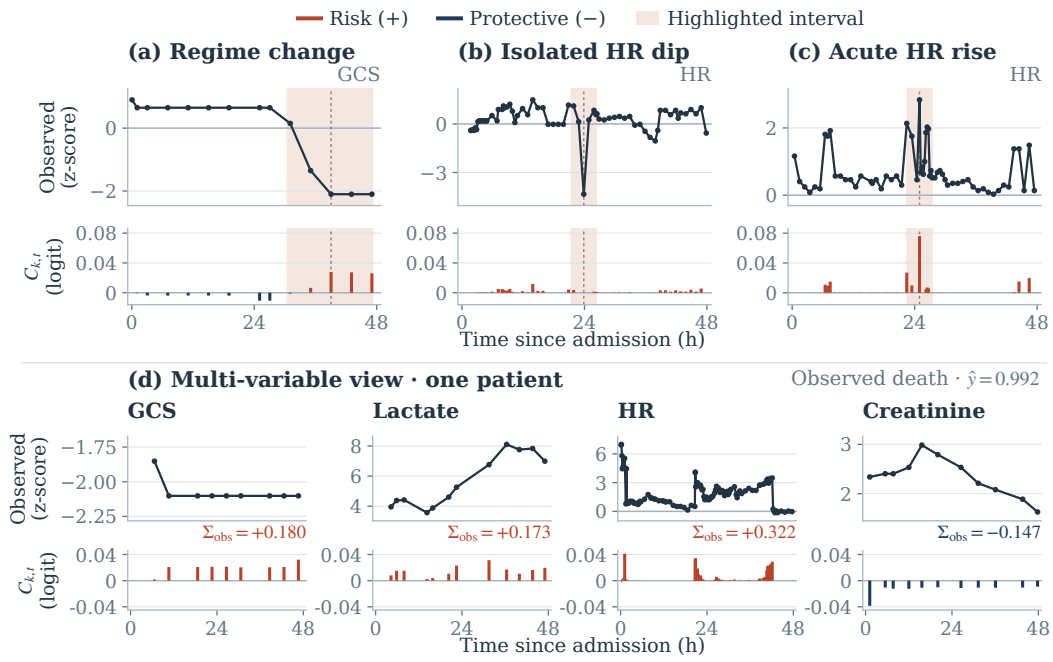


Figure 2: Signed attribution on representative P12 patients. (a) GCS regime change; (b) an isolated HR dip with little attributed credit; (c) an acute HR rise with positive credit; (d) four variables from one patient. Only observed entries ($m_{k,t} = 1$) are shown; Σ_{obs} in (d) sums the displayed contributions, whereas the full decomposition also includes imputed entries and b_{global}.

5 Conclusion

We presented TRACE, built on *architecture–attribution co-design*. The architecture is chosen so that, for any fixed input, one backward pass decomposes the logit exactly into signed per-variable, per-timestep contributions. On three ICU benchmarks TRACE achieves the highest AUROC and AUPRC; its completeness gap is at floating-point round-off; and its attribution runs two orders of magnitude faster than Integrated Gradients. We also introduced gROAR, a graded remove-and-retrain protocol that sweeps retention budgets and tests the sweep for a monotone trend and for equivalence to the full model. Under gROAR, TRACE outperforms every intrinsically interpretable baseline at every budget on both datasets, is the only method whose top-50% inputs are equivalent to the full model on all four dataset×metric cells, and still retains most of its accuracy when only 10% of the inputs are kept; on the same checkpoint, its exact attribution matches Integrated Gradients while attention-based rankings perform at random level. Faithful interpretability and strong prediction are therefore not mutually exclusive.

Some limitations remain. Attention weights are held fixed rather than attributed, so the attribution explains the value path only. The current experiments consider binary outcomes over 48-hour ICU windows, and the computationally intensive gROAR sweeps are limited to P12 and P19. Moreover, the evaluation is retrospective and based on public benchmarks; it does not establish clinical utility or safety in prospective practice.

Three directions follow. First, the same model can be applied more widely: multi-site cohorts, multi-class and continuous targets, and static covariates brought into the decomposition. Second, the co-design principle can be tested beyond this instance, by replacing retention with other sequence cores whose dynamic coefficients are input-independent, such as structured state-space models, and by moving to domains outside healthcare. Third, the clinical setting itself can be taken further. Attribution can be updated incrementally as new measurements arrive, and its effect on clinician decisions can be measured in human-in-the-loop studies. This would help connect mathematical faithfulness to clinician understanding and decision quality.

REPRODUCIBILITY STATEMENT

All datasets are public: PhysioNet-2012 and PhysioNet-2019 are freely available, and MIMIC-III is available under a credentialed data-use agreement; the preprocessing pipeline, stratified split (`random_state`$= 0$), and seeds are fixed and shared across every method (Appendix C). Architecture and training hyperparameters, including the selected configurations for every baseline, are listed in Appendices D and F. The full forward and backward derivations underlying the exact decomposition are given in Appendix A, and the independent-bias completeness check that validates them is released as a standalone script together with the training and attribution code (Appendix E).

ETHICS STATEMENT

This work uses de-identified, publicly available ICU datasets under their respective data-use agreements and involves no new data collection. TRACE is intended as decision support that makes a model's evidence inspectable; it is not a diagnostic device, and its attributions explain the model, not the patient's physiology. Faithful explanations can also invite over-trust, and we view prospective clinical validation and human-factors evaluation (Section 5) as prerequisites to any deployment.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used as writing and coding assistants: to polish prose, to refactor figure and evaluation scripts, and to check LaTeX consistency. All research ideas, model design, experiments, analyses, and claims are the authors' own, and every LLM-assisted edit was reviewed by the authors.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 435–451. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ali22a.html.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

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, pp. 2565–2573, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467166. URL https://doi.org/10.1145/3447548.3467166.

Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10319–10328, 2022. doi: 10.1109/CVPR52688.2022.01008.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791. IEEE, 2021.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2166–2177. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/crabbe21a.html.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Tian Guo, Tao Lin, and Nino Antulov-Fantulin. Exploring interpretable LSTM neural networks over multi-variable data. In *International conference on machine learning*, pp. 2494–2504. PMLR, 2019.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024. URL `https://arxiv.org/abs/2202.01602`.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

Michael Moor, Bastian Rieck, Max Horn, Catherine R Jutzeler, and Karsten Borgwardt. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Frontiers in medicine*, 8:607952, 2021.

Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased LSTM: Accelerating recurrent network training for long or event-based sequences. *Advances in neural information processing systems*, 29, 2016.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Ellis Batten Page. Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, 58(301):216–230, 1963.

Chengzhe Piao, Taiyu Zhu, Stephanie E Baldeweg, Paul Taylor, Pantelis Georgiou, Jiahao Sun, Jun Wang, and Kezhi Li. GARNN: an interpretable graph attentive recurrent neural network for predicting blood glucose levels via multivariate time series. *Neural Networks*, 185:107229, 2025.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

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.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif Cifci, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S Alkhawaldeh, Sadiq Hussain, et al. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, 2024.

Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680, 1987.

Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. "the human body is a black box": Supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 99–109, 2020.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of ICU patients: The PhysioNet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pp. 245–248. IEEE, 2012.

Chenxi Sun, Moxian Song, Derun Cai, Baofeng Zhang, Hongyan Li, and Shenda Hong. A review of deep learning methods for irregularly sampled medical time series data. *Health Data Science*, 6: 0456, 2026.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Andreas Theissler, Francesco Spinnato, Udo Schlegel, and Riccardo Guidotti. Explainable AI for time series classification: a review, taxonomy and research directions. *IEEE Access*, 10:100700–100724, 2022.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR, 2019.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 799–809. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/08fa43588c2571ade19bc0fa5936e028-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ruirui Wang, Yanke Li, Manuel Günther, and Diego Paez-Granados. GARLIC: Graph attention-based relational learning of multivariate time series in intensive care. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4ZAwmIaA9y.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3273–3285, 2023.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021.

Liangwei Nathan Zheng, Zhengyang Li, Chang George Dong, Wei Emma Zhang, Lin Yue, Miao Xu, Olaf Maennel, and Weitong Chen. Irregularity-informed time series analysis: Adaptive modelling of spatial and temporal dynamics. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pp. 3405–3414, 2024.

Xiao-Hua Zhou, Jiarui Sun, Gene A. Pennello, Nancy A. Obuchowski, and Donna K. McClish. *Statistical Methods in Diagnostic Medicine*. Wiley Series in Probability and Statistics. Wiley, 3rd edition, 2026. ISBN 9781394220212.

Appendix

A Full Derivation

A.1 Forward Pass

Notation and backward convention. Throughout the backward pass, g with a subscript denotes the gradient of the scalar logit $\hat{z}$ with respect to the intermediate quantity indicated by the subscript, written as a *column vector* (one such vector per time step and, where applicable, per variable). For a linear layer $y = Wx$ with W stored in the standard implementation layout $W \in \mathbb{R}^{\text{out}\times\text{in}}$, the gradient therefore back-propagates as $g_x = W^\top g_y$. For a mixing along the time or variable axis, $O = RV$, the gradient of each value v_j collects $g_{v_j} = \sum_t R_{t,j}\, g_{o_t}$, i.e., $G_V = R^\top G_O$ when the per-step gradients are stacked as rows. All transposes below follow these two rules, and the derivation matches the reference implementation.

Input Decay Imputation: Clinical time series in the ICU are inherently irregularly sampled and characterized by pervasive missing values. Naive imputation strategies (e.g. static zero-filling or pure forward-filling) often distort the underlying physiological dynamics. While GRU-D (Che et al., 2018) pioneered decay-based modeling of informative missingness, it applies elapsed-time decay both to the input and to the hidden state inside the gated recurrent transition. The decay rates themselves depend only on elapsed time, but the hidden-state decay path passes through the gates, so its effect on the prediction is not a closed-form per-feature quantity. TRACE keeps only the input decay, as a standalone pre-encoding step whose coefficients depend on the elapsed time and on input-independent parameters. As a result the imputation is a *linear* operator on the input values (for a fixed observation pattern), which is what keeps the attribution path closed-form and residual-free (Section 3.3).

Let $X \in \mathbb{R}^{K\times T}$ denote the multivariate input for a patient, with K clinical variables and T time steps, and $M \in \{0,1\}^{K\times T}$ the binary observation mask ($m_{k,t} = 1$ iff variable k is observed at time t). For missing entries we accumulate the normalized time elapsed since the last observation, $\Delta t_{k,t}$. As the gap grows, confidence in the last observed value decays and the estimate reverts to the empirical mean $\bar{x}_k$. We model this with a decay factor $\lambda_{k,t} \in (0,1]$ parameterized by variable-specific learnable scalars w_k, b_k:

$$\lambda_{k,t} = \exp\big(-\max(0,\, w_k\, \Delta t_{k,t} + b_k)\big). \tag{5}$$

Crucially $\lambda_{k,t}$ does *not* depend on the input values x; it is a fixed coefficient once Δt and the mask are given. (In the implementation the cumulative gaps are additionally normalized by $\Delta t \leftarrow \Delta t / \max \Delta t$, a sample-level constant that does not affect linearity.) The imputed value is the convex combination

$$\tilde{x}_{k,t} = m_{k,t}\, x_{k,t} + (1 - m_{k,t})\big[\lambda_{k,t}\, x_{k,t}^{(\text{last})} + (1-\lambda_{k,t})\, \bar{x}_k\big]. \tag{6}$$

Both $x_{k,t}^{(\text{last})}$ (a forward-fill) and $\bar{x}_k$ (a temporal mean) are linear in X, and $\lambda_{k,t}$ is input-independent; hence $X \mapsto \tilde{X}$ is a linear map. The completed tensor $\tilde{X} \in \mathbb{R}^{K\times T}$ is passed downstream and is the quantity to which we ultimately attribute.

Variable-Wise Feature Encoding: Each variable k is processed by its own 2-layer MLP, keeping the initial feature extraction strictly decoupled across variables. For each (k,t) we form the 2-vector $[\tilde{x}_{k,t},\, m_{k,t}]$ and compute

$$e_{k,t} = W_2^{(k)}\, \sigma_{\text{ReLU}}\Big(W_1^{(k)} \begin{bmatrix} \tilde{x}_{k,t} \\ m_{k,t} \end{bmatrix} + b_1^{(k)}\Big) + b_2^{(k)}, \tag{7}$$

with σ_{ReLU} the ReLU activation (we reserve plain σ for the logistic sigmoid; σ_{ReLU} and the GroupNorm scale σ^{GN} carry explicit marks), $W_1^{(k)} \in \mathbb{R}^{D\times 2}$, $W_2^{(k)} \in \mathbb{R}^{D\times D}$ with D denoting the hidden layer dimension and the parameters are not shared across variables. This yields $E \in \mathbb{R}^{K\times T\times D}$; each $e_{k,t}$ depends only on variable k and its imputed history.

Time-Lag Attention (intra-variable temporal alignment): We apply self-attention independently within each variable's sliding window of size W. Windows are indexed by their starting time step $\tau \in \{1,\dots,T'\}$, $T' = T - W + 1$; from this stage on, the sequence axis runs over windows rather than time steps. For window position $w \in \{0,\dots,W-1\}$ we add a sinusoidal positional encoding, $z_{\tau,w}^{(k)} = e_{k,\tau+w} + PE_w$. The resulting vectors are projected through variable-specific

query, key, and value transformations Q_k, K_k, and V_k. For window τ, the attention weights are $\beta_{i,i'}^{(k,\tau)} = \mathrm{Softmax}_{i'}(Q_k(z_{\tau,i}^{(k)})\, K_k(z_{\tau,i'}^{(k)})^\top/\sqrt{D})$; they are computed separately for every window and therefore depend on τ. The output for variable k and window τ is:

$$h_{k,\tau}^{\mathrm{lag}} = \sum_{i'=0}^{W-1} \left(\sum_{i=0}^{W-1} \beta_{i,i'}^{(k,\tau)} \right) V_k(z_{\tau,i'}^{(k)}), \tag{8}$$

i.e. each value $V_k(z_{\tau,i'}^{(k)})$ is weighted by the total attention it receives from all query positions, $\sum_i \beta_{i,i'}^{(k,\tau)}$. Collecting the window representations across variables and windows gives

$$H^{\mathrm{lag}} = \left\{ h_{k,\tau}^{\mathrm{lag}} \right\}_{k,\tau} \in \mathbb{R}^{K \times T' \times D}.$$

Inter-Variable Spatial Correlation: We model cross-variable dependencies with a learnable adjacency matrix $A \in \mathbb{R}^{K \times K}$. At each window τ, we apply a single linear message-passing operation along the variable axis:

$$H_{:,\tau,:}^{\mathrm{latent}} = A H_{:,\tau,:}^{\mathrm{lag}}, \qquad \tau = 1, \ldots, T'. \tag{9}$$

Equivalently, we write $H^{\mathrm{latent}} = A H^{\mathrm{lag}}$, with the multiplication understood to act on the variable dimension. The resulting tensor satisfies $H^{\mathrm{latent}} \in \mathbb{R}^{K \times T' \times D}$. We initialize

$$A = I + 0.2\,\mathbf{1},$$

to emphasize self-loops while allowing initial cross-talk. A separate 2-layer MLP branch (off the prediction path) decodes each latent vector $h_{k,\tau}^{\mathrm{latent}}$ (the entry of H^{latent} for variable k and window τ) into a reconstruction $\hat{x}_{k,\tau}$ of the observed value at the last time step of that window, $x_{k,\tau+W-1}$, which serves as a self-supervised regularizer:

$$\begin{aligned} \hat{x}_{k,\tau} &= W_2^{mp}\, \sigma_{\mathrm{ReLU}}\big(W_1^{mp}\, h_{k,\tau}^{\mathrm{latent}} + b_1^{mp}\big) + b_2^{mp}, \\ \mathcal{L}_{\mathrm{rec}} &= \frac{\sum_{k,\tau} m_{k,\tau+W-1} \left(x_{k,\tau+W-1} - \hat{x}_{k,\tau}\right)^2}{\sum_{k,\tau} m_{k,\tau+W-1}}. \end{aligned} \tag{10}$$

This reconstruction branch does not contribute to the prediction logit $\hat{z}$. Its ReLU therefore lies outside the attribution path. The predictive path uses only the linear graph operator A, whose exact backward operation is $A^\top$.

Signal Attention (cross-variable aggregation): Signal attention operates along the variable axis and compresses the K message-passed representations into a single vector. For query variable k and key/value variable k', define

$$\alpha_{k,k',\tau} = \mathrm{softmax}_{k'} \left(\frac{Q\left(h_{k,\tau}^{\mathrm{latent}}\right) K\left(h_{k',\tau}^{\mathrm{latent}}\right)^\top}{\sqrt{D}} \right). \tag{11}$$

The cross-variable representation is then

$$v_{\mathrm{out},\tau} = \sum_{k=1}^{K} \sum_{k'=1}^{K} \alpha_{k,k',\tau} V\left(h_{k',\tau}^{\mathrm{latent}}\right). \tag{12}$$

A linear projection $W_{\mathrm{proj}} \in \mathbb{R}^{H \times D}$ then maps $v_{\mathrm{out},\tau}$ to the hidden width H, giving the fused state $h_\tau = W_{\mathrm{proj}}\, v_{\mathrm{out},\tau} \in \mathbb{R}^H$. The sequence $\{h_\tau\}_{\tau=1}^{T'}$ is the input to the temporal core; stages 6–7 operate on this length-T' sequence, and we index its positions by $t, j \in \{1, \ldots, T'\}$ there (so $h_t \equiv h_\tau$ for $t = \tau$). *Implementation note.* The released implementation, inherited from the GARLIC front end (Wang et al., 2026), flattens the $K \times T'$ latent states in variable-major order and applies signal attention to consecutive groups of K tokens; the resulting group index serves as the position index of stages 6–7. This is a fixed permutation of the same tokens, i.e., a linear, input-independent re-indexing that the backward pass inverts exactly, so the decomposition and all reported results are unaffected; we keep the per-window indexing above for readability.

Explainable Temporal Dynamics via Stacked Retention: Conventional recurrent units (GRU/LSTM, Che et al., 2018) couple time steps through *input-dependent multiplicative gates* (e.g. $z_t \odot h_{t-1}$ with $z_t = \sigma(\cdot)$): a frozen linearization still exists, but the decomposition is then non-unique and requires freezing gates that are themselves functions of the input, so the resulting per-step coefficient is neither closed-form nor a stable importance. TRACE instead adapts RetNet (Sun et al., 2023), whose temporal coupling is governed by an input-independent causal weighting. The temporal contribution is then a single closed-form coefficient. For each of the n_h retention heads h we project $Q^{(h)}, K^{(h)}, V^{(h)} \in \mathbb{R}^{T' \times d_h}$, $d_h = H/n_h$, and form a causal weighting matrix $D^{(h)}$ (superscripted to distinguish it from the embedding width D) with $D^{(h)}_{t,j} = 0$ for $j > t$. RetNet's general form is $D^{(h)}_{t,j} = \gamma_h^{t-j}$ with a per-head rate $\gamma_h \in (0, 1]$; the configuration used in all experiments is the uniform causal weighting $D^{(h)}_{t,j} = \mathbf{1}[j \le t]$ ($\gamma_h = 1$), and any input-independent choice of $D^{(h)}$ leaves the derivation below unchanged. The retention coefficient is

$$R^{(h)}_{t,j} = \left(\frac{Q^{(h)}_t \, K^{(h)\top}_j}{\sqrt{d_h}} \right) \odot D^{(h)}_{t,j}. \tag{13}$$

Note that standard attention applies a softmax over the sequence, coupling all positions nonlinearly through a shared normalization denominator. TRACE removes this normalization from the retention operator. The output is therefore a direct linear combination of the value stream, weighted by retention coefficients cached at their forward-pass values:

$$o^{(h)}_t = \sum_{j \le t} R^{(h)}_{t,j} \, V^{(h)}_j. \tag{14}$$

Each retention sub-layer applies a LayerNorm to its input, computes the retention output above, applies Group Normalization within each head over its channels and all T' positions (which replaces the softmax for numerical stability) and an output projection W_{out}, and adds the residual; a residual FFN sub-layer with a GELU activation follows. After the L residual blocks, a final LayerNorm LN_f is applied to the stack output $U = \{u_t\}_{t=1}^{T'}$. We denote its output by $H_{\text{ret}} = \{h^{\text{ret}}_t\}_{t=1}^{T'} \in \mathbb{R}^{T' \times H}$. Replacing recurrent gates with this softmax-free linear causal mixing keeps long-range temporal modeling while enabling decomposition.

Temporal Attention (global aggregation): A global Temporal Attention condenses the sequence $H_{\text{ret}} \in \mathbb{R}^{T' \times H}$ into a single global representation. After adding temporal positional encodings, we construct query, key, and value representations and compute $\eta_{t,t'} = \mathrm{Softmax}_{t'}(Q_t K^\top_{t'} / \sqrt{H})$, then sum over the query positions t:

$$h_{\text{global}} = \sum_t \sum_{t'} \eta_{t,t'} \, V_{t'}, \qquad V_{t'} = W^{\text{ta}}_v \, \tilde{h}_{t'}, \tag{15}$$

where $\tilde{h}_{t'}$ is the retention output at step t' after positional encoding. During the backward pass η is held at its evaluated value (a frozen coefficient), so attribution propagates through the value path exactly.

Classification Head: The classifier produces the scalar logit

$$\hat{z} = W^{cls}_2 \, \sigma_{\mathrm{ReLU}}\big(W^{cls}_1 \, h_{\text{global}} + b^{cls}_1\big) + b^{cls}_2, \tag{16}$$

trained to minimize binary cross-entropy. The predicted probability is $\hat{y} = \sigma(\hat{z})$.

Summary. Every operator above is either (i) linear, (ii) an elementwise activation (ReLU, GELU, $\max$) with an exact rescale multiplier, (iii) a normalization with an exact Jacobian and affine intercept, or (iv) a dynamic coefficient (attention or retention map) frozen at its evaluated value. Consequently, for a fixed input the map $\tilde{X} \mapsto \hat{z}$ is affine and reproduces $\hat{z}$ exactly, which is the starting point for the decomposition of Appendix A.2.

A.2 Backward Pass (Proof of Exactness)

Exact additive decomposition. Fix an input X and cache every activation and dynamic weight at its evaluated operating point. Each element-wise activation f is replaced by its *rescale* multiplier (the DeepLIFT rule with a zero reference, Shrikumar et al., 2017), also known as the effective slope,

$$S_{\text{eff}}(z) = \frac{f(z)}{z}, \qquad \text{so that } f(z) = S_{\text{eff}}(z) \, z, \tag{17}$$

with the removable singularity at $z = 0$ taken as the limit ($S_{\text{eff}}(0) = \frac{1}{2}$ for GELU, 1 for the identity, and $\mathbf{1}[z > 0]$ for ReLU, which coincides with its derivative). Unlike the raw gradient, the rescale multiplier reproduces the activation exactly at the operating point, including for non-homogeneous activations such as GELU. For attention and retention operations, we hold the coefficient-generating branches fixed at their evaluated values and propagate attribution through the corresponding value paths. The composition of these frozen maps is affine, so there exist effective gradient coefficients $g_{k,t}$ and a constant b_{global} (absorbing all layer biases and positional-encoding contributions) with

$$\hat{z} = \sum_{k=1}^{K} \sum_{t=1}^{T} g_{k,t}\, \tilde{x}_{k,t} + b_{\text{global}}. \tag{18}$$

We define the signed contribution of input (k, t) as $C_{k,t} \coloneqq g_{k,t}\, \tilde{x}_{k,t}$, so that $\hat{z} = \sum_{k,t} C_{k,t} + b_{\text{global}}$. The coefficients $g_{k,t}$ are computed in two phases: (1) a forward pass caching all pre-activations and dynamic weights; (2) a backward pass that propagates g using the cached rescale multipliers and exact normalization Jacobians in place of autograd derivatives. The constant b_{global} is accumulated *independently* during the same backward sweep. Every additive constant of the frozen network (a layer bias, a positional encoding, the mask-path term of the per-variable encoder, or the affine intercept $f(x_0) - J_f\, x_0$ of a normalization layer) is a fixed vector $c_s \in \mathbb{R}^{d_s}$ injected at some site s of the computation graph. Because the frozen network is affine, the effect of c_s on the logit is the inner product of c_s with the gradient of $\hat{z}$ with respect to the quantity at that site, $g_s \in \mathbb{R}^{d_s}$, which the backward sweep has already computed when it reaches s. Hence

$$b_{\text{global}} \;=\; \sum_{\text{sites } s} g_s^{\top} c_s. \tag{19}$$

This choice is what makes Eq. (1) a testable identity. The left-hand side $\hat{z}$ comes from an ordinary forward pass. The right-hand side is assembled entirely from the backward sweep: $\sum_{k,t} C_{k,t}$ uses the gradients $g_{k,t}$ at the input, and b_{global} uses the gradients g_s at every bias site. If any backward operator were implemented incorrectly, the gradients downstream of it would be wrong in both terms, and the two sides would no longer agree. The alternative of defining the constant as the residual $b \coloneqq \hat{z} - \sum_{k,t} C_{k,t}$ would make the identity hold by construction, whether or not the gradients are correct; it is convenient at deployment but cannot detect an error. All reported completeness gaps use the independent accumulation of Eq. (19).

Classifier backward. Recall that the classification head is

$$\hat{z} = W_2^{\text{cls}} \sigma_{\text{ReLU}}\big(W_1^{\text{cls}} h_{\text{global}} + b_1^{\text{cls}}\big) + b_2^{\text{cls}}.$$

With ReLU mask $M_{\text{ReLU}} = \mathbf{1}[W_1^{cls} h_{\text{global}} + b_1^{cls} > 0]$, the gradient with respect to h_{global} is

$$g_{h_{\text{global}}} = \big(W_1^{cls}\big)^{\top} \operatorname{diag}(M_{\text{ReLU}}) \big(W_2^{cls}\big)^{\top} \in \mathbb{R}^{H}. \tag{20}$$

Temporal Attention backward. Holding η fixed, the value at step j receives the summed query attention $\sum_t \eta_{t,j}$, so

$$g_{h_j^{\text{ret}}} = \Big(\sum_t \eta_{t,j}\Big) \big(W_v^{\text{ta}}\big)^{\top} g_{h_{\text{global}}}, \tag{21}$$

the gradient with respect to the j-th output of the retention stack (the positional encoding is an additive constant and does not change the gradient).

Final LayerNorm backward. The retention stack ends with the LayerNorm LN_f whose input is u_j (Appendix A.1). With its exact Jacobian $J_{\text{LN}_f}(u_j)$,

$$g_{u_j} = J_{\text{LN}_f}(u_j)^{\top}\, g_{h_j^{\text{ret}}}, \tag{22}$$

and the affine intercept of LN_f is accumulated into b_{global}.

The stack of L retention layers is then traversed from the last layer to the first. Each layer consists of a retention sub-layer followed by an FFN sub-layer; for each sub-layer we write g_{out} and g_{in} for the gradients with respect to its output and its input. The traversal starts from $g_{\text{out}} = g_{u_j}$ at the top and ends with g_{h_τ}, the gradient with respect to the fused state h_τ (position $j = \tau$).

FFN sub-layer backward. For $x_{\text{out}} = x_{\text{ret}} + W_2^{\text{ffn}}\,\text{GELU}(W_1^{\text{ffn}}\,\text{LN}(x_{\text{ret}}) + b_1^{\text{ffn}}) + b_2^{\text{ffn}}$, with GELU rescale S_{GELU} and the exact LayerNorm Jacobian J_{LN},

$$g_{\text{in}} = g_{\text{out}} + J_{\text{LN}}^{\top}\left(W_1^{\text{ffn}}\right)^{\top}\text{diag}(S_{\text{GELU}})\left(W_2^{\text{ffn}}\right)^{\top} g_{\text{out}}. \tag{23}$$

Retention sub-layer backward (temporal mixing). The sub-layer computes $x_{\text{out}} = x_{\text{in}} + W_{\text{out}}\,\text{GN}\big(\text{Ret}(\text{LN}(x_{\text{in}}))\big)$. Project the incoming gradient through the output weight and the GroupNorm Jacobian J_{GN}, $J_{\text{GN}}^{\top}W_{\text{out}}^{\top}\,g_{\text{out}}$, and split the result into heads to obtain $g_{o^{(h)}}$, the gradient with respect to the retention output $o^{(h)}$ of Eq. (14) (stacked over time steps). Applying the transpose of the (frozen) retention matrix gives the gradient with respect to the values:

$$g_{V^{(h)}} = \left(R^{(h)}\right)^{\top} g_{o^{(h)}}, \tag{24}$$

i.e. step j collects $\sum_{t\geq j} R_{t,j}^{(h)}\, g_{o_t^{(h)}}$. Each head is then passed through the transpose of its value weight, $(W_v^{(h)})^{\top} g_{V^{(h)}}$, the heads are summed, the input LayerNorm Jacobian is applied, and the residual path is added: $g_{\text{in}} = g_{\text{out}} + J_{\text{LN}}^{\top}\sum_h (W_v^{(h)})^{\top} g_{V^{(h)}}$.

The GroupNorm operates on the $[H, T']$ layout, so its statistics pool over all $d_h \cdot T'$ entries of a head; its Jacobian therefore *couples time steps* and is block-diagonal per (sample, head) with blocks of size $(d_h T') \times (d_h T')$, not per-timestep. Writing $\bar{x} = (x - \mu_c^{\text{GN}})/\sigma_c^{\text{GN}}$ within a group c (mean and standard deviation of that group) and $n_g = d_h T'$, the exact VJP is applied in closed form without materializing the block:

$$g_{\text{in}} \;=\; \frac{1}{\sigma_c^{\text{GN}}}\Big(\tilde{g} \;-\; \overline{\tilde{g}} \;-\; \bar{x}\,\overline{\tilde{g}\odot\bar{x}}\Big), \qquad \tilde{g} = \gamma^{\text{GN}} \odot g_{\text{out}}, \tag{25}$$

where $\overline{(\cdot)}$ averages over the n_g entries of the group and γ^{GN} enters on the *output* index. (The per-position LayerNorm Jacobians take the analogous form with statistics over the feature axis only.)

Predictor projection backward. The fused state is $h_\tau = W_{\text{proj}}\, v_{\text{out},\tau}$, a strictly linear $D \to H$ map, so $g_{v_{\text{out},\tau}} = W_{\text{proj}}^{\top}\, g_{h_\tau} \in \mathbb{R}^D$.

Signal Attention backward. Undo the query-summation by expanding over the variable axis with the frozen cross-variable map α, where $V(\cdot) = W_v^{\text{sa}}(\cdot)$ is the value projection of Eq. (12):

$$g_{h_{k,\tau}^{\text{latent}}} = \Big(\sum_{k'} \alpha_{k',k,\tau}\Big)\left(W_v^{\text{sa}}\right)^{\top} g_{v_{\text{out},\tau}}, \tag{26}$$

one gradient in $\mathbb{R}^D$ for every variable k and window τ; stacking over k gives $g_{H_{:,\tau}^{\text{latent}}} \in \mathbb{R}^{K\times D}$.

Graph message passing backward. The forward step is $H^{\text{latent}} = A\,H^{\text{lag}}$, so the exact backward is the transpose of the (asymmetric) adjacency:

$$g_{H_{:,\tau}^{\text{lag}}} = A^{\top} g_{H_{:,\tau}^{\text{latent}}}, \qquad \tau = 1, \ldots, T', \tag{27}$$

whose rows are the gradients $g_{h_{k,\tau}^{\text{lag}}}$ with respect to the time-lag attention outputs. (Implementation note: this contracts the *row* index of A; using A in place of $A^{\top}$ is incorrect whenever $A \neq A^{\top}$, which holds in general.)

Time-Lag Attention backward. For each variable k and window τ, with frozen attention weights $\beta_{i,i'}^{(k,\tau)}$ (where i is the query position and i' the key/value position within the window),

$$g_{z_{\tau,w}^{(k)}} = \Big(\sum_{i} \beta_{i,w}^{(k,\tau)}\Big)\left(W_v^{(k)}\right)^{\top} g_{h_{k,\tau}^{\text{lag}}}. \tag{28}$$

Because the sliding windows overlap, the backward is a scatter-add that accumulates the window-position gradients onto the full timeline, expanding T' windows to T time steps: since $z_{\tau,w}^{(k)} = e_{k,\tau+w} + PE_w$ and the positional encoding is constant, $g_{e_{k,t}} = \sum_{\tau+w=t} g_{z_{\tau,w}^{(k)}} \in \mathbb{R}^D$ for every variable k and time step t.

Per-variable encoder backward. With ReLU mask $M_{\text{ReLU}}^{(k,t)} = \mathbf{1}\big[W_1^{(k)}[\tilde{x}_{k,t},\, m_{k,t}]^\top + b_1^{(k)} > 0\big]$,

$$\begin{bmatrix} g_{\tilde{x}_{k,t}} \\ g_{m_{k,t}} \end{bmatrix} = \big(W_1^{(k)}\big)^\top \operatorname{diag}(M_{\text{ReLU}}^{(k,t)})\, \big(W_2^{(k)}\big)^\top g_{e_{k,t}} \;\in \mathbb{R}^2. \tag{29}$$

The first entry is the gradient with respect to the physiological value and defines $g_{k,t} \coloneqq g_{\tilde{x}_{k,t}}$; the second multiplies the mask indicator, a constant, and is therefore booked into b_{global} as one of the sites in Eq. (19).

Final attribution coefficients. Collecting $g_{k,t}$ over all (k,t) gives the effective gradient of $\hat{z}$ with respect to the imputed input $\tilde{X}$. The signed contribution is

$$C_{k,t} = g_{k,t}\, \tilde{x}_{k,t}. \tag{30}$$

Completeness verification. Correctness of the backward is validated by two independent checks:

1. *Falsifiable completeness.* With b_{global} accumulated independently as above, we compare $\hat{z}_{\text{recon}} = \sum C + b_{\text{global}}$ against $\hat{z}$ from a standard forward pass. A mismatch between any backward operator and the frozen forward is directly visible as a gap; during development this check automatically caught two implementation bugs (an A-versus-$A^\top$ transpose in the graph backward and a mis-oriented normalization Jacobian) that the residual-recovery shortcut is structurally unable to detect.
2. *Precision scaling.* In single precision the relative gap is at the accumulated round-off level of the seven-stage chain (4.1×10^{-6} on P12, 3.9×10^{-7} on P19, 5-seed means; Table 2); re-running the identical computation in double precision drives it below 10^{-13} (1.1×10^{-14} on P12, 1.4×10^{-15} on P19). The gap scales with machine epsilon rather than with any approximation term, confirming that the decomposition is exact up to floating-point round-off.

B TIME COMPLEXITY

Notation. B: batch size; T: sequence length; $T' = T - W + 1$: sliding windows; K: clinical variables; D: embedding width; H: retention hidden width; W: window size; L: retention layers; n_h: heads.

Forward pass. We list each component's cost (omitting lower-order terms):

Component	Cost	Source
Encoder (K MLPs)	$B\,T\,K\,D^2$	$K \times \mathrm{Linear}(D \to D)$
Time-Lag Attention	$B\,K\,T'\,W(D^2 + W\,D)$	Q/K/V proj + $W \times W$ scores
Message Passing	$B\,T'\,K^2 D$	$A \in \mathbb{R}^{K \times K}$ matmul
Signal Attention	$B\,T'\,K(D^2 + K\,D)$	Q/K/V proj + $K \times K$ scores
Predictor Proj	$B\,T'\,DH$	$\mathrm{Linear}(D \to H)$
Retention ($\times L$)	$L\,B\,T'(H^2 + T'H)$	4 proj + scores + FFN
Temporal Attention	$B\,T'(H^2 + T'H)$	Q/K/V proj + $T' \times T'$ scores
Classifier	$B\,H^2$	$\mathrm{Linear}(H \to H)$

For our hyperparameters ($K = 36,\ D = 32,\ H = 128,\ W = 7,\ T = 215,\ L = 1$; the P12 configuration), the time-lag attention term dominates; for long T or deep L, the $\mathcal{O}(BLT'^2H)$ retention term dominates.

Attribution backward. The backward reuses all attention/retention weights cached during the forward and propagates a single effective gradient g. Crucially, Q/K projections and score computations are *not* re-executed. Only W_v projections and cached weight transposes are needed:

Component	Backward cost	Saving vs. forward
Encoder	$B\,T\,K\,D^2$	$1\times$ (same)
Time-Lag Attn	$B\,K\,T'\,D(D + W)$	$\sim W\times$ less expensive
Message Passing	$B\,T'\,K^2 D$	$1\times$ (same)
Signal Attn	$B\,T'\,D(D + K)$	$\sim K\times$ less expensive
Retention ($\times L$)	$L\,B\,T'(H^2 + T'H)$	$\sim 1\times$ (+ Jacobian overhead)
Temporal Attn	$B\,H^2$	$\sim T'\times$ less expensive

Comparison with IG. Integrated Gradients requires n forward–backward passes under autograd, giving a lower bound $\mathcal{T}_{\text{IG}}/\mathcal{T}_{\text{exact}} \geq n$. The speedup beyond n arises because TRACE's cached-weight backward is less expensive than a standard backward.

Memory. Peak activation memory is dominated by: retention coefficients ($L{\cdot}Bn_hT'^2$ floats), time-lag weights ($BT'W^2K$), signal/temporal attention maps ($BT'K^2 + BT'^2$), and the single largest tensor: one normalization Jacobian $[B, T', H, H]$.

C Dataset Details

PhysioNet-2012 (P12). The PhysioNet/Computing in Cardiology Challenge 2012 (Silva et al., 2012) targets in-hospital mortality prediction from the first 48 hours of an ICU stay. The three public sets (a, b, and c) contain 12,000 ICU stays; after excluding 12 corrupted records, as in prior work (Zhang et al., 2021; Wang et al., 2026), 11,988 patients remain, with 36 irregularly sampled clinical variables (vital signs, laboratory tests, and GCS sub-scores) and a binary mortality label (positive rate 14.2%). Each variable is observed at highly variable rates (median observation count per variable ranges from $\sim$2 to $\sim$40 across the 48-hour window), making missingness a dominant structural feature.

PhysioNet-2019 (P19). The PhysioNet/Computing in Cardiology Challenge 2019 (Reyna et al., 2019) targets early prediction of sepsis, defined by Sepsis-3 criteria. We use both public training sets (A and B; 40,336 patients, 34 variables) with hourly-binned observations over the first 48 hours; after excluding records without any valid observation in that window, 40,331 patients remain, each with a binary patient-level sepsis label (positive rate 7.3%). The extreme class imbalance and delayed label onset make AUPRC the primary metric.

MIMIC-III. The Medical Information Mart for Intensive Care III (Johnson et al., 2016) provides de-identified ICU records from Beth Israel Deaconess Medical Center. Following the preprocessing of Warpformer (Zhang et al., 2023), as adopted by GARLIC (Wang et al., 2026), we restrict the cohort to adult ICU stays, use the first 48 hours, and target in-hospital mortality (positive rate 11.6%) across 103 irregularly sampled variables. All models use the same pre-processed feature set.

Shared protocol. All three datasets are split into $80/10/10$ train/validation/test with a fixed stratified random seed (`random_state`$= 0$), shared across every method and seed. Class balance is preserved in all splits. Each variable is z-scored with per-variable statistics of its observed entries computed over the cohort, following the released GARLIC preprocessing (Wang et al., 2026); these statistics are label-free, and the identical preprocessed tensors are consumed by every method. No patient-level data augmentation is applied. Each benchmark also ships a small set of static covariates (age, gender, ICU type; row "Static dim" in Table 5); TRACE operates on the temporal channels and observation masks only, so static covariates are outside the attribution scope $K \times T$. Table 5 summarizes the key statistics.

Table 5: Summary statistics of the three clinical benchmarks.

	P12	**P19**	**MIMIC-III**
Samples	11,988	40,331	49,380
Positive	1,707 (14.2%)	2,931 (7.3%)	5,747 (11.6%)
Negative	10,281 (85.8%)	37,400 (92.7%)	43,633 (88.4%)
T (time steps)	215	49	200
K (variables)	36	34	103
$T \times K$	7,740	1,666	20,600
Static dim	9	6	2
Obs. rate	5.2%	14.2%	1.9%
Task	Mortality	Sepsis	Mortality

D TRAINING DETAILS AND HYPERPARAMETERS

Objective. TRACE is trained with a three-term objective: a classification loss, a self-supervised reconstruction regularizer (the message-passing branch of Section 3), and an ℓ_1 sparsity penalty on the learned graph adjacency A,

$$\mathcal{L} = \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{rec}} + \lambda_{\text{graph}}\,\|A\|_1, \tag{31}$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy on the mortality / sepsis label, $\mathcal{L}_{\text{rec}}$ is the masked MSE reconstruction of the observed inputs (its weight is fixed to 1), and $\|A\|_1$ encourages a sparse, clinically interpretable inter-variable graph. (An optional interpretability aux-head adds a further BCE term; it is disabled for the reported runs.)

Decoupled alternating optimization. The reconstruction/graph pathway and the classification pathway are optimized in an alternating curriculum with two Adam parameter groups (separate learning rates lr_{rec}, lr_{cls} and weight decays wd_{rec}, wd_{cls}). Training begins in a *reconstruction* stage minimizing $\mathcal{L}_{\text{rec}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{graph}}\|A\|_1$ with only the encoder/graph parameters trainable; when validation has not improved for `patience` epochs the model reloads its best checkpoint and switches to a *prediction* stage minimizing $\lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$ with only the temporal/classifier parameters trainable. The two stages alternate until early stopping.

Protocol. All models use a 5-seed protocol with seeds $\{100, 101, 102, 103, 104\}$ on a single fixed stratified split (`random_state`$= 0$, shared across every method and seed to enable paired comparison). Within a run, the retained checkpoint is the one with the lowest validation *classification* loss ($\lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$); the learning rate follows `ReduceLROnPlateau` (factor 0.5, patience 3), and early stopping triggers, after a warm-up, once the validation classification loss has not improved for `early_stopping` epochs.[2]

Hyperparameter search. For P12 and P19 we tune with Optuna (Akiba et al., 2019) maximizing validation AUPRC over the search space in Table 6. MIMIC-III uses a smaller architecture-only grid (memory pressure at $T{=}200 \times 103$ features). The selected configuration per dataset is in Table 7. GARLIC uses the per-dataset configurations reported in its paper (selected by its authors over the grid listed in Table 9) without retuning; the remaining baselines are tuned over the GARLIC Appendix C.1 ranges (Appendix F).

Table 6: TRACE hyperparameter search space

Hyperparameter	Range
hidden_size	$\{128, 256, 512\}$
feature_dim	$\{16, 32\}$
num_heads	$\{4, 8, 16\}$
num_layers	$\{1, 2, 3\}$
window_size	$\{3, 5, 7, 9\}$
batch_size	$\{64, 128, 256\}$
λ_{cls}	$\{1, 2, 5, 10, 20\}$
λ_{graph}	$\{0.001, 0.01, 0.1\}$
lr_{cls}	$[1\text{e}{-5},\ 5\text{e}{-4}]$ log-uniform
lr_{rec}	$[1\text{e}{-4},\ 5\text{e}{-3}]$ log-uniform
wd_{cls}	$[1\text{e}{-5},\ 1\text{e}{-2}]$ log-uniform
wd_{rec}	$[1\text{e}{-5},\ 5\text{e}{-3}]$ log-uniform

Table 7: Selected TRACE configuration

Hyperparameter	P12	P19	MIMIC-III
heads	4	8	8
layers	1	3	1
hidden_size	128	128	128
feature_dim	32	32	32
window_size	7	9	7
batch_size	64	64	64
λ_{cls}	5	20	20
λ_{graph}	0.1	0.01	0.01
epochs (cap)	150	150	100
lr_{cls}	$1.41\text{e}{-5}$	$3.16\text{e}{-4}$	$2.00\text{e}{-4}$
lr_{rec}	$2.15\text{e}{-4}$	$4.96\text{e}{-4}$	$3.00\text{e}{-4}$

[2] All five seeds reuse a single frozen configuration per dataset; no per-seed tuning is performed.

E REPRODUCTION DETAILS

Code and data. Code is provided in the supplementary material and will be released publicly upon acceptance. The repository is organized as follows:

```
trace/
|-- data/                # Preprocessing
|-- ours/
|   |-- model.py         # TRACE architecture
|   |-- train.py         # Alternating optimization training loop
|   |-- attribution.py   # Exact decomposition & Integrated Gradients
|   `-- scripts/train.sh # Single-seed training script
|-- requirements.txt
`-- README.md
```

We evaluate on PhysioNet-2012 (P12), PhysioNet-2019 (P19), and MIMIC-III. P12 and P19 are publicly available from `https://physionet.org`; MIMIC-III requires credentialed access (Johnson et al., 2016).

Reproducing results. To train TRACE on P12 with the paper's 5-seed protocol:

```
pip install -r requirements.txt
for seed in 100 101 102 103 104; do
  bash ours/scripts/train.sh "$seed"
done
```

Architecture hyperparameters are configured via environment variables (Table 7). Checkpoints are saved to `models/` and reused by all downstream evaluations. The gROAR sweep retrains from scratch on masked data as required by the protocol.

Environment. Python 3.10, PyTorch 2.1, CUDA 12.2, NVIDIA RTX 5090 (32 GB). Dependency versions are pinned in `requirements.txt`.

F Baseline Reproduction and Hyperparameter Search

Shared protocol. All non-GARLIC baselines are run on a single standardized scaffold to keep the comparison paired: cohort-level per-variable z-score normalization (`ours.load_data`), the same stratified $80/10/10$ split (`random_state=0`), Adam with weight decay, `ReduceLROnPlateau` (factor 0.5, patience 3) on validation AUPRC, early stopping (warm-up 10, patience 10, max 100 epochs), and a best-val-AUPRC checkpoint with a 2-logit cross-entropy head. Model selection is always by validation AUPRC (class-imbalanced setting). Table 8 lists the selected configuration per dataset, and Table 9 summarizes the corresponding search grids.

Table 8: Selected baseline configurations (val-AUPRC)

Model	P12	P19	MIMIC-III
GRU-D	hid 32, lat 64	hid 100, lat 64	hid 64, lat 20
Raindrop	d_ob 4, nh 2, lr 1e−4	d_ob 2, nh 2, lr 5e−4	d_ob 2, nh 2, lr 1e−4
mTAND	heads 4, H=20, lr 1e−3	heads 4, H=128, lr 5e−4	heads 2, H=64, lr 1e−3
MTSFormer	L=2, nh 4, lr 1e−4	L=4, nh 2, lr 1e−3	L=4, nh 4, lr 5e−4
PatchTST	d=128, L=2, pl 8, lr 1e−3	d=128, L=4, pl 8, lr 1e−3	d=128, L=4, pl 16, lr 1e−3
iTransformer	d=128, L=2, lr 1e−3, wd 1e−4	d=128, L=3, lr 1e−3, wd 1e−4	d=256, L=3, lr 5e−4, wd 1e−4
IMV-LSTM	d=64, lr 5e−3, bs 64	d=64, lr 1e−2, bs 64	d=64, lr 5e−3, bs 64
RETAIN	p=128, m=64, lr 1e−3	p=128, m=128, lr 1e−3	p=64, m=128, lr 5e−4
GARLIC	hid 128, feat 16, win 3	hid 512, feat 32, win 9	hid 128, feat 16, win 5
TRACE (ours)	heads 4, layers 1, win 7	heads 8, layers 3, win 9	heads 8, layers 1, win 7

Table 9: Baseline hyperparameter grids (selection: best val AUPRC)

Model	Axes (fixed settings omitted)	#cfg
GRU-D	hidden $\in \{32, 50, 64, 100\}$, latent $\in \{10, 20, 32, 64\}$; lr $= 1e{-3}$	16
Raindrop	d $\in \{2, 4\}$, nhead $\in \{1, 2, 4\}$, lr $\in \{1e{-4}, 5e{-4}, 1e{-3}\}$; nlayers $= 2$	18
mTAND	heads $\in \{1, 2, 4\}$, hidden $\in \{20, 32, 64, 128\}$, lr $\in \{1e{-4}, 5e{-4}, 1e{-3}\}$	36
MTSFormer	n_layers $\in \{2, 3, 4\}$, n_head $\in \{1, 2, 4\}$, lr $\in \{1e{-4}, 5e{-4}, 1e{-3}\}$	27
PatchTST	d $\in \{128, 256\}$, e $\in \{2, 3, 4\}$, patch_len $\in \{8, 16\}$, lr $\in \{1e{-4}, 5e{-4}, 1e{-3}\}$	36
iTransformer	d / e same as above, wd $\in \{1e{-5}, 1e{-4}\}$, lr $\in \{1e{-4}, 5e{-4}, 1e{-3}\}$	36
IMV-LSTM	h / b $\in \{64, 128\}$, lr $\in \{1e{-3}, 5e{-3}, 1e{-2}\}$, wd $\in \{1e{-5}, 1e{-4}, 1e{-3}\}$	36
RETAIN	$p \in \{64, 128, 256\}$, $m{=}q \in \{64, 128, 256\}$, lr $\in \{1e{-4}, 5e{-4}, 1e{-3}\}$	27
GARLIC	win $\in \{3, 5, 7, 9\}$, $D_f \in \{16, 32\}$, $D_h \in \{64, 128, 256, 512\}$	32
TRACE (ours)	Optuna over hidden $\in \{128, 256, 512\}$, feature_dim $\in \{16, 32\}$, heads $\in \{4, 8, 16\}$, layers $\in \{1, 2, 3\}$, window $\in \{3, 5, 7, 9\}$, batch $\in \{64, 128, 256\}$, $\lambda_{\text{cls}} \in \{1, 2, 5, 10, 20\}$, $\lambda_{\text{graph}} \in \{1e{-3}, 1e{-2}, 1e{-1}\}$, lr_{cls}, lr_{rec}, wd_{cls}, wd_{rec} (MIMIC: arch-only grid)	30 trials

G Ablation on Perturbation Strategies

Motivation. Prior interpretable-ICU work evaluates attribution with a single-point remove-and-retrain at a fixed budget (retain the Top-50%, Bottom-50%, or Random-50% most-attributed cells, retrain, and read the AUROC/AUPRC change) (Wang et al., 2026; Hooker et al., 2019). Table 10 shows our initial comparison between GARLIC's attention-based attribution and TRACE's attribution under this protocol. TRACE's Top-50 outperforms GARLIC's Top-50 on both datasets, yet GARLIC's Bottom-50 drops more, seemingly suggesting a more concentrated, faithful attribution. This paradox prompted a rigorous error analysis.

Table 10: Top-50 / Bottom-50 / Random-50 comparison on the same TRACE checkpoint (mean $\pm$ std, 5 seeds; an earlier run of the pipeline, so absolute values differ slightly from the final sweeps in Tables 3–4, which supersede them). Page's L tests the monotone trend All $>$ Top50 $>$ Random50 $>$ Bottom50; TOST tests equivalence between Top-50 and Full ($\pm 4\%/\pm 8\%$). (*: $p < 0.05$; **: $p < 0.005$)

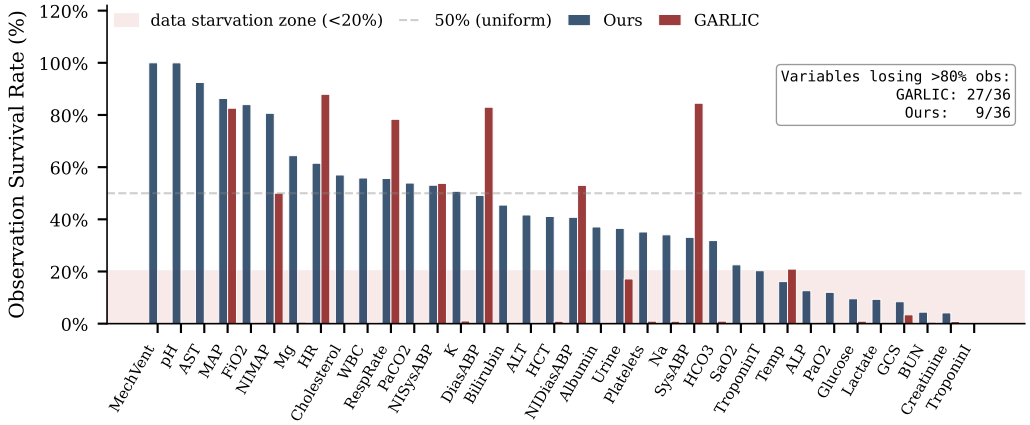
		P12		P19	
Method	Condition	AUROC	AUPRC	AUROC	AUPRC
GARLIC's	All	87.52 ± 0.11	60.32 ± 0.65	91.07 ± 0.22	58.95 ± 1.02
	Top50	86.84 ± 0.10	58.57 ± 0.69	89.67 ± 0.47	51.47 ± 0.62
	Random50	86.13 ± 0.42	56.92 ± 1.76	88.65 ± 0.59	51.28 ± 3.01
	Bottom50	$\mathbf{68.79 \pm 2.50}$	$\mathbf{30.20 \pm 2.16}$	$\mathbf{84.79 \pm 0.38}$	$\mathbf{41.92 \pm 0.93}$
	Equivalence	<0.005 **	<0.005 **	<0.005 **	0.012 *
	Monotone	<0.005 **	<0.005 **	<0.005 **	<0.005 **
Ours	All	87.52 ± 0.11	60.32 ± 0.65	91.07 ± 0.22	58.95 ± 1.02
	Top50	$\mathbf{87.44 \pm 0.27}$	$\mathbf{59.24 \pm 0.95}$	$\mathbf{90.28 \pm 0.83}$	$\mathbf{55.12 \pm 3.00}$
	Random50	86.13 ± 0.42	56.92 ± 1.76	88.65 ± 0.59	51.28 ± 3.01
	Bottom50	77.32 ± 0.70	40.98 ± 1.46	87.33 ± 0.90	47.92 ± 1.84
	Equivalence	<0.005 **	<0.005 **	<0.005 **	0.008 *
	Monotone	<0.005 **	<0.005 **	<0.005 **	<0.005 **

The paradox. Both methods pass the monotone trend and equivalence tests. However, GARLIC's Bottom-50 drops far more dramatically (P12 AUROC 68.79 vs. ours 77.32). Naïvely, this suggests GARLIC's attribution is "better concentrated". We show below that this interpretation is incorrect: the large Bottom-50 drop is an artifact of *data starvation*, not superior attribution quality.

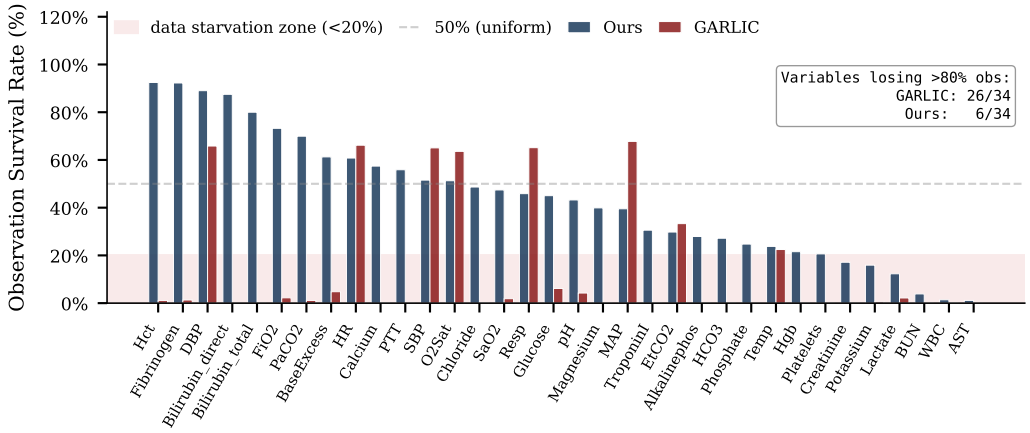
Data starvation (Figure 3). If we examine *which* cells are removed under Bottom-50%, we find that GARLIC's attention-based attribution pushes the vast majority of variables into near-total removal. On P12, 27 of 36 variables fall below 20% observation survival; on P19, 26 of 34 do. TRACE's more distributed attribution keeps only 9/36 (P12) and 6/34 (P19) below this threshold. The Bottom-50% protocol therefore conflates two effects: (i) the masked cells were genuinely unimportant, versus (ii) the masking destroyed the input distribution itself. GARLIC's large Bottom-50 drop is driven by (ii), not (i). A single-point Top/Bottom-50% comparison is confounded by the *shape* of the attribution distribution.

What makes a good attribution? Before proposing an alternative evaluation, we ask: what should a faithful clinical attribution look like? If an attribution method assigns importance to observations that are truly informative for predicting the outcome, then variables far from the population norm (high absolute z-score) should receive high attribution magnitude. In other words, we think anomalous values carry more predictive signal. We quantify this by computing the Spearman correlation $\rho(|\text{z-score}|, |\text{attribution}|)$ per variable per patient, then plotting the distribution of per-patient correlations (Figure 4).

On P12, TRACE achieves a mean Spearman $\rho = 0.51$ (median 0.62, 76% of patients above $\rho > 0.3$), indicating strong positive alignment between attribution magnitude and clinical severity. GARLIC's attention-based importance, by contrast, is essentially uncorrelated ($\mu = -0.03$, median -0.04, only 18% above 0.3). On P19 the pattern is similar: TRACE $\mu = 0.31$ (median 0.35, $55\% > 0.3$) vs. GARLIC $\mu = -0.02$ (median -0.01, $15\% > 0.3$). This shows that our model produces attributions



(a) P12 (36 variables)



(b) P19 (34 variables)

Figure 3: Per-variable observation survival under Bottom-50% masking. The dashed line at 50% marks uniform removal; the shaded red zone ($\leq 20\%$ survival) marks *data starvation*. GARLIC's concentrated attention pushes the vast majority of variables into starvation ($27/36$ on P12, $26/34$ on P19), whereas TRACE's attribution spreads importance more evenly (9/36 on P12, 6/34 on P19), preserving a usable input distribution.

that are physiologically coherent. Abnormal observations are highlighted, normal ones are not, whereas attention weights bear no systematic relationship to clinical severity.

Design principle of gROAR. Given that Bottom-50% is confounded and the z-score analysis confirms TRACE's attribution is clinically meaningful, we need a protocol that directly tests *ranking ability*: can the attribution correctly order inputs from most to least informative? Inspired by ROAR (Hooker et al., 2019), we retain only the Top-k% most-attributed cells for $k \in \{10, 20, 30, 40, 50\}$, mask the rest, retrain from scratch, and measure retained performance. This protocol always retains the same *number* of observations regardless of concentration. If an attribution is faithful, Top-k should substantially outperform a random mask of the same size, and the closer it is to the full model, the better its attribution is.

Results. Table 11 reports the Top-k sweep for each interpretable model alongside its random control. The key finding is that traditional attention-based models *cannot consistently outperform random masking*: TRACE's attribution is the only one that beats random at *every* k on *both* datasets and metrics.

Table 11: gROAR sweep (each model retrained on its own masked data; mean ± std %, 5 seeds). *attr* = built-in attribution; *rand* = random mask. Page's L tests the monotone trend; TOST tests equivalence between Top-50 and Full ($\pm 4\%/\pm 8\%$). (*: $p < 0.05$; **: $p < 0.005$; X: fail)

(a) RETAIN

	P12				P19			
	attr		rand		attr		rand	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	**79.93±0.69**	**41.54±1.67**	73.50±1.35	35.94±2.84	**76.54±1.74**	23.98±1.96	75.98±1.45	**24.03±4.18**
Top 20	**81.76±1.46**	**42.82±3.81**	77.81±3.04	38.13±2.68	80.39±0.61	28.53±2.56	**82.66±1.15**	**36.79±1.50**
Top 30	**82.00±1.38**	**43.86±1.56**	80.64±1.11	42.92±3.75	80.62±0.81	28.96±4.80	**84.62±0.66**	**39.86±2.44**
Top 40	**83.02±0.83**	**46.05±2.80**	81.72±1.15	43.46±1.35	81.55±0.40	31.15±2.71	**84.86±0.88**	**42.00±2.60**
Top 50	**83.71±0.94**	**46.49±3.80**	82.18±0.86	45.49±1.64	81.70±1.76	30.84±3.54	**84.83±0.85**	**42.05±2.36**
Full	83.45±1.28	48.40±2.21	83.45±1.28	48.40±2.21	86.23±0.74	45.45±1.91	86.23±0.74	45.45±1.91
Page's L	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**
TOST	0.004**	0.025*	<0.005**	<0.005**	0.842 X	0.997 X	<0.005**	<0.005**

(b) IMV-LSTM

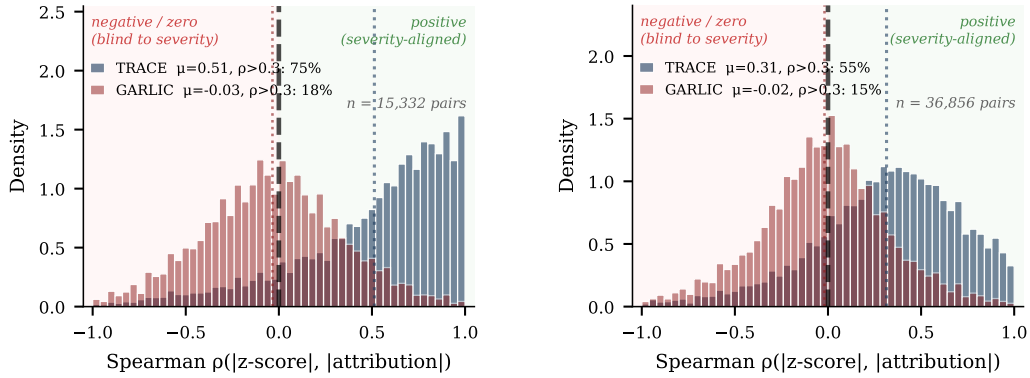
	P12				P19			
	attr		rand		attr		rand	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	**64.50±2.91**	**26.54±3.75**	62.67±3.82	25.49±4.88	**73.19±6.85**	**20.79±5.69**	55.20±0.84	9.26±0.38
Top 20	**71.56±6.89**	**32.37±7.73**	67.02±5.07	25.95±4.08	**76.46±7.52**	**26.50±6.22**	69.34±8.43	17.67±7.26
Top 30	**76.04±2.85**	**37.73±3.35**	65.97±4.10	24.87±3.93	**80.42±4.55**	**34.27±7.99**	76.56±6.40	27.31±8.27
Top 40	**72.49±5.22**	**33.22±7.41**	66.99±3.65	26.93±4.73	**83.68±2.95**	**39.83±5.79**	79.64±2.82	30.50±5.49
Top 50	**74.88±2.69**	**36.76±3.07**	69.13±6.92	32.19±9.32	82.05±4.80	36.08±8.99	**83.73±0.86**	**39.15±1.56**
Full	80.21±1.24	43.58±2.43	80.21±1.24	43.58±2.43	88.20±0.83	48.34±1.11	88.20±0.83	48.34±1.11
Page's L	<0.005**	<0.005**	<0.005**	0.006*	<0.005**	<0.005**	<0.005**	<0.005**
TOST	0.818 X	0.251 X	0.969 X	0.811 X	0.850 X	0.841 X	0.794 X	0.899 X

(c) GARLIC

	P12				P19			
	attr		rand		attr		rand	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	74.13±1.23	36.00±3.88	**77.89±1.76**	**41.85±2.46**	78.62±1.00	27.93±1.09	**78.77±0.69**	**28.25±1.40**
Top 20	81.55±0.81	45.58±1.82	**81.75±0.90**	**47.24±0.98**	**82.34±0.89**	31.95±1.25	82.10±1.05	**34.15±2.41**
Top 30	**84.34±1.13**	**52.01±2.27**	83.32±0.35	49.71±1.29	**86.33±0.71**	38.72±0.97	85.21±0.78	**41.87±1.62**
Top 40	**86.72±0.55**	**57.43±0.93**	84.73±0.58	51.39±1.65	**88.43±0.79**	**46.72±2.66**	87.24±0.37	46.17±1.69
Top 50	**86.29±0.39**	**56.33±1.31**	85.20±0.53	53.11±0.67	**89.22±0.50**	**50.86±1.15**	88.30±0.82	48.66±2.99
Full	87.04±0.62	56.76±1.88	87.04±0.62	56.76±1.88	91.38±0.36	57.93±1.02	91.38±0.36	57.93±1.02
Page's L	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**
TOST	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	0.124 X	0.033*	0.802 X

(d) TRACE (Ours)

	P12				P19			
	attr		rand		attr		rand	
Retain	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Top 10	**84.43±0.83**	**51.66±2.03**	78.07±0.96	41.08±1.39	**86.04±0.49**	**46.60±1.00**	78.45±1.43	31.18±1.95
Top 20	**85.96±0.52**	**55.43±1.51**	81.73±1.31	48.30±2.72	**88.67±0.38**	**51.99±1.88**	83.61±0.61	38.22±2.29
Top 30	**86.52±0.83**	**56.32±1.57**	83.84±1.15	51.14±2.19	**89.28±0.66**	**53.59±1.42**	84.97±1.10	41.39±1.59
Top 40	**86.87±0.98**	**57.44±2.18**	84.94±0.90	52.30±0.78	**90.77±0.28**	**56.33±1.30**	86.91±0.76	46.55±1.15
Top 50	**87.32±0.37**	**58.75±0.60**	85.41±0.31	54.90±1.22	**90.78±0.51**	**57.52±0.91**	87.56±0.43	47.80±2.57
Full	87.52±0.14	60.25±0.72	87.52±0.14	60.25±0.72	91.46±0.41	59.15±0.71	91.46±0.41	59.15±0.71
Page's L	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**
TOST	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	<0.005**	0.073 X



(a) P12 **(b)** P19

Figure 4: Distribution of per-patient Spearman correlation $\rho(|\text{z-score}|, |\text{attribution}|)$ for TRACE (blue) and GARLIC (red). TRACE's attributions are strongly positively correlated with deviation from the population norm, while GARLIC's attention weights are essentially uncorrelated.

Summary. The Top-k sweep reveals a clear hierarchy: TRACE's exact decomposition is the only attribution that *consistently* outperforms random masking across all conditions. Attention-based methods (RETAIN, IMV-LSTM, GARLIC) are unreliable. They can be anti-informative at specific budgets, particularly on P19. Combined with the z-score correlation analysis (Figure 4), we conclude that TRACE's attribution is both physiologically coherent and functionally faithful.

H Attribution Semantics and Post-hoc Baselines

What does a "good" attribution look like? Attribution methods differ not only in *how* they compute importance but in *what* the resulting scores mean. Consider three common paradigms:

- **Attention weights** $\alpha_{k,t} \in [0, 1]$ sum to one and are non-negative by construction. They indicate *relative routing* of information but not direction (positive vs. negative influence) or magnitude in output units (Jain & Wallace, 2019).
- **Gradient×input** $\nabla_x \hat{z} \odot x$ is a first-order Taylor approximation; it can be signed but does not sum to the output difference.
- **Integrated Gradients** $\mathrm{IG}(x) = (x - x') \odot \int_0^1 \nabla \hat{z}(x' + s(x - x'))\, ds$ satisfies completeness *in the limit* (Sundararajan et al., 2017), but the Riemann approximation introduces subtle discretization error ($\approx 10^{-2}$ at n=50 steps) and costs n forward–backward passes.

Discussion on attribution. We argue that an ideal attribution should express importance on the same scale as the model's output, i.e., the *actual contribution to the logit*. This is analogous to reporting exam scores in points rather than percentile ranks: both convey ordering, but points additionally convey magnitude and allow meaningful aggregation. Concretely, if $C_{k,t} = +0.3$ and $C_{k',t'} = -0.1$, a clinician knows that the first cell pushes the mortality logit up by 0.3 and the second pulls it down by 0.1. This is a statement that non-negative attention weights or rank-based importances cannot make.

What TRACE can achieve. TRACE's exact decomposition yields $\hat{z} = \sum_{k,t} C_{k,t} + b_{\text{global}}$, where each $C_{k,t} = g_{k,t}\, \tilde{x}_{k,t}$ is the signed, logit-scale contribution of input (k, t). The identity holds to machine precision (Table 2) in a single backward pass, with no Riemann discretization and no approximation. This is possible because, once frozen at the operating point, TRACE's softmax-free retention, its elementwise activations (via exact rescale multipliers), and its normalizations (via exact Jacobians) compose to an affine map, so the decomposition rule is *exact* rather than approximate.

Scope and caveats. Two remarks: (i) the decomposition is a *local* linearization, exact at the current input, not a global statement; moving to another input re-linearizes; (ii) attention/retention weights are frozen, so attribution flows through the value path and the question "why these attention weights" is not itself attributed, the same boundary drawn by transformer relevance propagation (Chefer et al., 2021; Ali et al., 2022).

Post-hoc baseline formulas. For completeness, the two gradient-based baselines are:

$$\text{Gradient} \times \text{Input:} \quad C_{k,t}^{\text{grad}} = \frac{\partial \hat{z}}{\partial x_{k,t}}\, x_{k,t}, \tag{32}$$

$$\text{Integrated Gradients:} \quad C_{k,t}^{\text{IG}} = (x_{k,t} - x'_{k,t}) \int_0^1 \frac{\partial \hat{z}(x' + s(x - x'))}{\partial x_{k,t}}\, ds, \tag{33}$$

where x' is the zero (masked) baseline. Gradient×input is a tangent linearization that leaks through saturated nonlinearities; IG satisfies completeness only in the continuous limit.

I RETENTIVE NETWORK AS THE TEMPORAL NETWORK

This appendix explains why we adopt the Retentive Network (Sun et al., 2023) as the temporal encoder. The key insight is that *input-dependent gating*, not nonlinearity per se, obstructs a well-defined exact decomposition.

The problem with gated recurrence. A frozen-activation linearization exists for almost any architecture. The relevant question is whether the resulting coefficients are (1) **unique**: independent of linearization conventions; (2) **closed-form**: not requiring iteration through frozen states; (3) **well-conditioned**: not a long product of frozen matrices. GRU/LSTM satisfy neither (1) nor (2), and (3) is not guaranteed. They couple time steps via input-dependent gates $z_t \odot h_{t-1}$ where $z_t = \sigma(W_z x_t + U_z h_{t-1})$. Linearizing this bilinear term requires freezing either z_t or h_{t-1} (both input-dependent) so the choice is arbitrary and yields different coefficients. The coefficient for input j reaching output t is a product $\prod_{s=j+1}^{t} \mathrm{diag}(z_s)$ of $O(T)$ frozen matrices: neither closed-form nor stable.

How retention resolves it. Retentive Networks replace the input-dependent gate with an input-independent causal weighting $D_{t,j}^{(h)}$ (RetNet's general form is γ_h^{t-j}; our configuration uses the uniform weighting $\gamma_h = 1$). No gate multiplies the value stream, so the temporal contribution of input j to output t is a single closed-form scalar; the only frozen quantities are the Q/K scores, exactly as in every attention attribution. Removing softmax further makes retention a *linear* mixing:

$$o_t^{(h)} = \sum_{j \le t} R_{t,j}^{(h)} \, V_j^{(h)}, \quad R_{t,j}^{(h)} := \left(\frac{Q_t^{(h)} K_j^{(h)\top}}{\sqrt{d_h}} \right) \cdot D_{t,j}^{(h)},$$

with no normalizing denominator (GroupNorm supplies the scale control instead; cf. Eq. (14)). Once $Q^{(h)}, K^{(h)}$ are evaluated, $R_{t,j}^{(h)}$ is fixed and the map $\{V_j^{(h)}\} \mapsto o_t^{(h)}$ is linear. No bilinear interaction to resolve in this case.

Proposition 1 (Uniqueness of temporal coefficients)**.** *Consider a temporal module mapping $\{v_j\}_{j \le t}$ to o_t.*

- *(a)* ***Retention:*** *If $o_t = \sum_j R_{t,j} v_j$ with $R_{t,j}$ held at its evaluated value (the value-path convention) and no gate multiplying v_j, the coefficient of v_j is uniquely $R_{t,j}$: no choice of which factor to freeze arises.*
- *(b)* ***GRU:*** *Linearizing $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t$ by freezing gates yields coefficients that depend on the convention (freeze z_t vs. h_{t-1}), hence non-unique.*

Proof. (a) Immediate from linearity. (b) Freezing z_t gives $\mathrm{diag}(z_t)$; freezing h_{t-1} gives $\mathrm{diag}(h_{t-1})$. These differ and propagate multiplicatively. Neither is canonical. □

Implication. Two valid "exact decomposition" implementations for GRU would yield different attribution maps, which is undesirable for reproducibility. Retention avoids this ambiguity. Importantly, this interpretability gain incurs no accuracy penalty: TRACE outperforms every other model on all three benchmarks (Table 1). The attention modules (time-lag, signal, temporal) use the standard weight-freezing convention shared by all attention attribution methods; retention's contribution is eliminating the *additional* non-uniqueness from gated recurrence.

J STATISTICAL TESTS

Page's trend test. To verify that the gROAR sweep produces a *monotone* degradation (full $>$ Top-50 $>$ Top-40 $> \ldots >$ Top-10), we apply Page's ordered alternatives test (Page, 1963), a non-parametric rank-based test for monotone trends across related samples. Each of the 5 seeds provides one paired observation at each retention level; the test statistic L summarizes whether the rank ordering is consistent with the hypothesized monotone sequence. All methods (TRACE exact, GARLIC, RETAIN, IMV-LSTM) show a significant monotone trend ($p < 0.005$ in every cell), confirming that removing more attributed inputs always degrades performance. Because the trend is significant for every method, the discriminative signal is the *degradation slope* and the absolute performance at small k, not the existence of a trend.

Two One-Sided Tests (TOST) for equivalence. To assess whether the Top-50%-reduced model is *statistically equivalent* to the full model, we apply the TOST procedure (Schuirmann, 1987). TOST tests whether the mean difference falls within a pre-specified equivalence margin $[-\Delta, +\Delta]$; rejection of both one-sided null hypotheses (at $\alpha = 0.05$) concludes equivalence. We set margins of $\pm 4\%$ (AUROC) and $\pm 8\%$ (AUPRC), fixed before running the sweep. The AUROC margin follows the 0.03–0.05 non-inferiority margins conventional in diagnostic-accuracy research (Zhou et al., 2026); the AUPRC margin is doubled because AUPRC's cross-seed dispersion is roughly twice that of AUROC in our runs (Table 1). The verdict pattern is not margin-critical: TRACE's observed Top-50 deficits (0.20/0.68 AUROC and 1.50/1.63 AUPRC points on P12/P19) lie 5–20$\times$ inside the margins, whereas the methods that fail TOST either exceed the margin outright (IMV-LSTM: 5.3/6.2 AUROC points; RETAIN on P19: 4.5) or exhibit seed variance too large to bound the difference. TRACE is the only model that achieves equivalence on all four dataset$\times$metric cells.

Interpretation. The combination of Page's L (confirms monotone degradation exists) and TOST (quantifies how close the top-50% reduced model is to the original full model) provides a two-part reading: monotonicity establishes that the attribution is informative, and equivalence at Top-50% establishes that the top half of the attribution captures essentially all predictive signal. A faithful attribution should satisfy both: significant Page trend *and* TOST equivalence at high retention budgets. TRACE is the only model that achieves this on all four cells.

K Extended Related Work

K.1 Modeling Irregular Clinical Time Series

Early deep learning models for sequential electronic health records used recurrent networks such as LSTMs and GRUs to carry a hidden state across visits or time steps, including attention-augmented recurrent predictors (Choi et al., 2016; Qin et al., 2017). A central limitation of these models in the ICU setting is that clinical measurements are inherently irregular and incomplete: different variables are observed at different frequencies, and missingness is often informative rather than random. GRU-D addressed this by incorporating elapsed-time decay and missingness masks directly into the recurrent transition, establishing explicit modeling of informative missingness as an important component of clinical time-series prediction (Che et al., 2018). Subsequent work pursued the same goal from complementary angles, including learned bidirectional imputation (Cao et al., 2018), interpolation networks that resample onto reference grids (Shukla & Marlin, 2019), multi-time attention over learned reference points (Shukla & Marlin, 2021), event-driven recurrence (Neil et al., 2016), and continuous-time formulations based on latent ODEs and controlled differential equations (Rubanova et al., 2019; Kidger et al., 2020). We adopt the decay-aware imputation philosophy of GRU-D but factor it out as a standalone *input-linear* step so that it does not entangle the attribution path. Specifically, because the decay coefficients depend only on elapsed time and learnable scalars, rather than on the physiological values themselves, for a fixed observation pattern the mapping from observed to imputed inputs remains linear, preserving a stable and closed-form decomposition of each variable's contribution throughout attribution. Recent surveys provide a broader overview of this rapidly expanding design space (Sun et al., 2026).

K.2 Relational and Transformer Models for Time Series

Because clinical deterioration in the ICU often reflects coordinated dysfunction across multiple physiological systems, modeling dependencies among variables is as important as capturing their temporal dynamics. Graph-based models treat physiological variables as nodes and learn their dependencies, as in Raindrop's graph-guided propagation (Zhang et al., 2021) and graph-attentive recurrent predictors for clinical signals (Piao et al., 2025). More recently, GARLIC (Wang et al., 2026) combines graph message passing and sequential attention for ICU prediction, but its explanations are unsigned attention weights. TRACE retains GARLIC's relational front-end and substitutes the temporal core, obtaining exact, signed per-variable, per-timestep attribution without compromising predictive capacity. In parallel, Transformer architectures (Vaswani et al., 2017) have been adapted from regular forecasting to clinical and long-horizon settings, including patch-based encoding (Nie et al., 2022), inverted variate-token attention (Liu et al., 2024), and multi-scale clinical modeling (Zhang et al., 2023), although their advantage over simpler baselines for time series remains debated (Zeng et al., 2023). These regular time-series models typically require an explicit imputation or binning front-end to process irregular clinical data. To ensure a fair and rigorous comparison, we equip these baselines with a decay imputation mechanism prior to evaluation.

K.3 Interpretability of Clinical Sequence Models

Interpretability has been a parallel concern in clinical sequence modeling. Intrinsically interpretable predictors such as RETAIN and IMV-LSTM expose attention-style importance over time and variables (Choi et al., 2016; Guo et al., 2019), in the same spirit as intelligible additive models for healthcare (Caruana et al., 2015). Whether attention weights constitute faithful explanations, however, remains debated. Some studies show that attention weights can be altered without changing predictions (Jain & Wallace, 2019), while others argue attention still carries explanatory signal under appropriate constraints (Wiegreffe & Pinter, 2019). Post-hoc attribution offers a model-agnostic alternative, such as Integrated Gradients with its completeness axiom (Sundararajan et al., 2017), DeepLIFT's reference-based propagation (Shrikumar et al., 2017; Montavon et al., 2018), and SHAP's additive game-theoretic credit (Lundberg & Lee, 2017). But these methods can be perturbation-fragile (Ghorbani et al., 2019), can fail saliency sanity checks (Adebayo et al., 2018), and yield rankings that depend heavily on the chosen benchmark such as remove-and-retrain (Hooker et al., 2019). Sequence-specific variants sharpen rather than remove these limitations: FIT scores each observation by the distribution shift it induces on the predictive posterior (Tonekaboni et al., 2020), Dynamask learns perturbation masks that respect temporal coherence (Crabbé & van der Schaar, 2021), and

TimeSHAP makes Shapley estimation tractable on sequences only through pruning and sampling (Bento et al., 2021). All remain perturbation- or sampling-based approximations whose cost grows with sequence length. For attention architectures, relevance-propagation rules extend LRP through attention and LayerNorm (Chefer et al., 2021), and conservation is preserved exactly when those modules are frozen as constants (Ali et al., 2022); on a softmax transformer, however, the propagated relevance remains a heuristic summary of a nonlinear computation. Compounding all of this, different post-hoc methods frequently disagree on the same prediction (Krishna et al., 2024). These concerns, together with broader cautions about over-claiming interpretability (Lipton, 2018; Rudin, 2019) and recent surveys of explainable AI in healthcare (Sadeghi et al., 2024) and time-series classification (Theissler et al., 2022), motivate an attribution that is signed, exact, and inexpensive rather than approximate and post hoc.

A complementary strategy makes the network itself carry its explanation. Self-explaining neural networks (SENN) generalize linear models by learning concept–relevance pairs whose combination is locally linear, with stability imposed through a gradient regularizer (Alvarez Melis & Jaakkola, 2018); B-cos networks replace each linear transform with an alignment-promoting one, so that the network computes an input-dependent linear map $W(x)\,x$ that summarizes the computation exactly (Böhle et al., 2022). TRACE belongs to this interpretable-by-construction family but differs in where the guarantee comes from. In SENN and B-cos the explanatory coefficients are themselves input-dependent functions computed by the network: the linear summary is faithful, but the mechanism producing the coefficients is not itself explained, and stability is encouraged by regularization (SENN) rather than structural. TRACE instead fixes the temporal coefficients *by architecture*, since its causal temporal weighting is input-independent, and confines the remaining input dependence to attention scores frozen at evaluation, which is what makes the temporal coefficients unique and closed-form (Proposition 1) and turns a local linear summary into a machine-precision identity on the logit. Moreover, both SENN and B-cos target densely observed vision or tabular inputs; neither addresses irregular sampling, informative missingness, or per-timestep credit assignment over hundreds of steps, the regime TRACE is designed for.

K.4 Linear State-Space and Retentive Sequence Models

Outside healthcare, a new class of sequence models replaces input-dependent recurrent gating and softmax attention with structured linear recurrences. Representative examples include structured state-space models such as S4 (Gu et al., 2021), selective state-space models such as Mamba (Gu & Dao, 2023), and retentive networks such as RetNet (Sun et al., 2023). S4 and RetNet share a key property: their temporal mixing is governed by *input-independent* coefficients, so the contribution of a past state to a future output becomes a stable, closed-form coefficient amenable to exact decomposition. Mamba, by contrast, re-introduces input-dependent gating (making its state-transition matrix a function of the current token) in order to improve content-based reasoning on language; this selectivity restores precisely the non-unique, input-entangled decomposition that afflicts GRUs, making Mamba unsuitable for the exact attribution we seek. TRACE therefore adopts RetNet's input-independent causal weighting rather than Mamba's selective mechanism. To our knowledge this interpretability property of the RetNet family has not been exploited. TRACE addresses this gap by showing that the temporal architecture can be chosen jointly with the attribution objective, such that exact decomposition emerges as a property of the model itself rather than as a post-hoc approximation. More broadly, this architecture–attribution co-design principle provides a potential direction for developing sequence models that combine strong predictive performance with built-in, mathematically faithful interpretability: diagnose which mechanisms make exact credit assignment ill-posed (here, input-dependent temporal mixing and softmax normalization on the credit path) and replace only those. The diagnosis is not specific to retention or to healthcare: any sequence encoder whose dynamic coefficients are either input-independent or can be frozen exactly admits the same one-pass decomposition.