Temporal Island Sparse Autoencoders for Interpreting Clinical Time-Series Models

Muhan Yeo[1,2] **Bong Gyun Kang**[1] **HyunGi Kim**[1,3] **Sungroh Yoon**[1]*

[1]Department of Electrical and Computer Engineering and Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea
[2]McWilliams School of Biomedical Informatics,
University of Texas Health Science Center at Houston, Houston, TX, USA
[3]UNIST InnoCORE AI-Space Solar Initiative,
Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

Muhan.Yeo@uth.tmc.edu
{luckypanda,rlagusrl0128,sryoon}@snu.ac.kr

Abstract

Deep learning models for electronic health records achieve strong predictive performance but provide little insight into *what clinical concepts* they represent internally. Post-hoc attribution methods produce per-cell importance scores that lack temporal coherence, cross-variable structure, and reusability across patients. We introduce the **Temporal Island Sparse Autoencoder** (TI-SAE), which learns a backbone-specific, patient-shared *clinical concept dictionary* from a frozen predictive model's representations. Each dictionary entry (latent) is defined by learned temporal Gaussian islands—specifying *when* the concept is active—and a sparse variable-composition vector—specifying *which variables* are involved. A decoupled selector then identifies which dictionary entries drive each patient's prediction, separating representation learning from attribution. We evaluate TI-SAE across five backbone architectures spanning different representation strategies on MIMIC-III and MIMIC-IV (three prediction tasks each). TI-SAE produces structured clinical concepts that go beyond transient attribution heatmaps while achieving strong performance on standard faithfulness benchmarks.

1 Introduction

Deep learning models for electronic health records (EHR) have achieved strong predictive performance in clinical tasks including mortality, decompensation, and readmission [Rajkomar et al., 2018, Harutyunyan et al., 2019, Tipirneni and Reddy, 2022]. However, these models remain opaque. For clinical models to move toward actual insight, clinicians need to understand a model's behavior in terms they already use—clinical patterns with temporal extent and cross-variable structure [Tonekaboni et al., 2019, Ghassemi et al., 2021].

Post-hoc explanation methods such as Integrated Gradients [Sundararajan et al., 2017], SHAP [Lundberg and Lee, 2017], Dynamask [Crabbé and Van Der Schaar, 2021], and TimeX [Queen et al., 2023] produce importance scores over the time×variable input grid. These attributions answer *which individual inputs matter* but not *what clinical concepts the model uses*. Point attributions lack (a) temporal and multivariable structure, which are prerequisites for clinical interpretation, and (b) compositionality into concepts that generalize across patients. Moreover, most time-series

*Corresponding author.

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

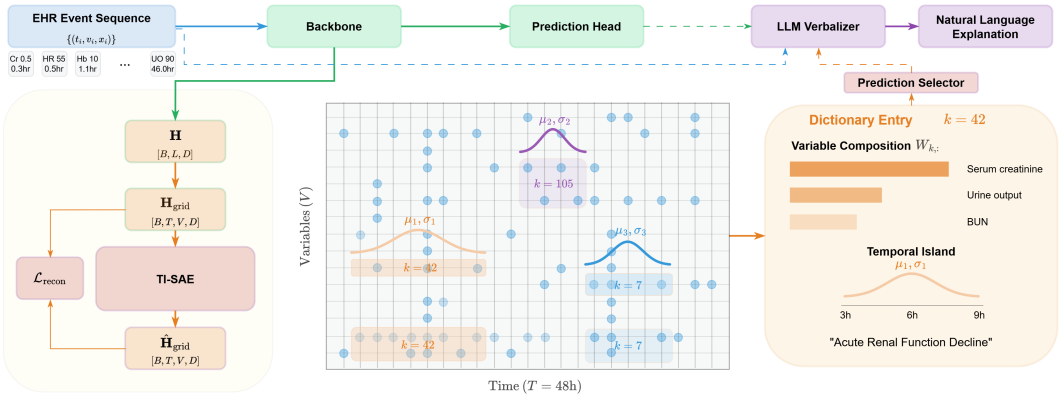


Figure 1: **Overview of the Temporal Island SAE (TI-SAE).** A frozen backbone transforms an EHR event sequence into a hidden representation $\mathbf{H}[B, L, D]$, which is then scattered into a dense representation grid $\mathbf{H}_{\text{grid}}[B, T, V, D]$. TI-SAE learns a dictionary of K latents, each with M temporal Gaussian islands (colored curves) defining when the concept is active, and sparse variable-composition weights $\mathbf{w}_k$ defining which variables are involved. Each island corresponds to an interpretable clinical concept with explicit temporal extent and variable structure.

explanation methods assume dense, regularly-sampled inputs, yet clinical event sequences are sparse and irregular, forcing ad-hoc adaptations.

Sparse autoencoders (SAEs) offer a principled approach to this problem by decomposing a model's internal representations into sparse, interpretable features via dictionary learning [Bricken et al., 2023, Templeton et al., 2024]. However, standard SAEs often recover token-specific or highly local features, which are not well-aligned with temporally extended clinical concepts [Bhalla et al., 2026]. Moreover, clinical episodes span hundreds of variables over extended time windows, and extracting a coherent concept from raw patient trajectories is infeasible without knowing which variables and what time window to examine. This motivates an SAE architecture that structurally encodes spatiotemporal composition.

We propose the **Temporal Island Sparse Autoencoder** (TI-SAE), which learns a backbone-specific, patient-shared clinical concept dictionary—a finite vocabulary of the latent patterns reused across patients (Figure 1). Each dictionary entry (latent k) has two components: (a) M learned *temporal Gaussian islands* with centers $\mu_{k,m}$ and widths $\sigma_{k,m}$ that define *when* the concept is active, and (b) a sparse variable-composition vector $\mathbf{w}_k \in \mathbb{R}^V$ that defines *which variables* are involved. The *island* structure, our key architectural novelty, encodes spatiotemporal organization directly in the latent space via temporally localized Gaussian windows and sparse variable composition.

We evaluate TI-SAE across five backbone architectures on MIMIC-III and MIMIC-IV with three prediction tasks each. TI-SAE achieves the highest faithfulness (AOPC) on four of five backbones, while simultaneously providing a decomposable clinical concept dictionary that no baseline method can produce. The learned dictionary is judged clinically plausible by four clinicians (mean Likert 4.07–4.40), transfers to unseen data, and produces superior concept-level explanations in both LLM-judge and clinician evaluations. Code is available at `https://github.com/MuhanY/TI-SAE`.

2 Related work

2.1 EHR model architectures

Each clinical event carries a variable identity, a measured value, and a timestamp. Code-level models such as BEHRT [Li et al., 2020], CLMBR [Steinberg et al., 2021], and MOTOR [Steinberg et al., 2024] encode only discrete medical concept codes without measurement values, often targeting long outpatient trajectories. For inpatient settings where measurement magnitudes and their temporal evolution are central to prediction, models must explicitly represent all three components.

Table 1: **Comparison of explanation methods along structural properties.** *Temporal structure*: attributions are organized in time, not independent per cell. *Decomposable*: explanations decompose into components that generalize across patients. *Post-hoc*: no backbone modification. *No concept-label supervision*: does not require external concept-label annotation.

Method family	Temporal struct.	Decomposable	Post-hoc	No concept sup.
Attention			✓	✓
Gradient-based	(✓)		✓	✓
Perturbation-based	(✓)		✓	✓
Learned selection	(✓)		✓	✓
Concept bottleneck		✓		
TI-SAE (ours)	✓	✓	✓	✓

Event representation. Early approaches to irregular clinical time series operated on dense time×variable grids with imputation or learned decay for unobserved cells [Che et al., 2018]. Modern architectures instead process variable-length event sequences directly, adopting different strategies for representing each event. Continuous value embedding (CVE) methods such as STraTS [Tipirneni and Reddy, 2022] embed raw scalar values and timestamps through learned MLPs, producing a single token per event; HyMaTE [Mottalib et al., 2025] extends this approach with a hybrid Mamba/transformer encoder. Discrete methods quantize values into bins: ETHOS [Renc et al., 2024] serializes each event into up to seven tokens, while EHRMamba [Fallahpour et al., 2025] inserts time-interval tokens between events. Grid-native architectures such as DuETT [Labach et al., 2023] and BAT [DeVries et al., 2025] operate directly on the time×variable matrix with factored bi-axial attention.

2.2 Post-hoc explainability for time series

Gradient-based methods. Integrated Gradients [Sundararajan et al., 2017], DeepLIFT [Shrikumar et al., 2017], and GradSHAP assign per-feature importance scores via backpropagation. Applied to time series, they produce $(T \times V)$ attribution maps with no temporal structure—each cell is scored independently. TSR [Ismail et al., 2020] combines gradient methods via two-step saliency rescaling but the output remains a flat attribution map.

Perturbation-based methods. KernelSHAP [Lundberg and Lee, 2017] and TimeSHAP [Bento et al., 2021] estimate feature importance by sampling input subsets and observing output changes; TimeSHAP adapts this for sequences by grouping time steps, but produces flat cell-level attributions. Dynamask [Crabbé and Van Der Schaar, 2021] learns continuous perturbation masks with temporal smoothness regularization. FIT [Tonekaboni et al., 2020] uses KL-divergence shifts in the predictive distribution; WinIT [Leung et al., 2023] aggregates feature-removal scores over time windows. ContraLSP [Liu et al., 2024b] adds contrastive regularization to mask optimization. These methods improve *how* attributions are computed but do not decompose explanations into structured concepts.

Learned selection. TimeX [Queen et al., 2023] trains an interpretable surrogate model via model behavior consistency. TimeX++ [Liu et al., 2024a] reformulates selection through an information bottleneck. L2X [Chen et al., 2018] introduces Gumbel-softmax selection for instance-wise feature importance. They operate at the cell level without concept decomposition.

2.3 Sparse autoencoders for interpretability

Dictionary learning for mechanistic interpretability has been transformative in understanding language models. Bricken et al. [2023] showed that SAEs decompose MLP activations into monosemantic features, with Templeton et al. [2024] scaling this to production models. Gated SAEs [Rajamanoharan et al., 2024] improve reconstruction–sparsity tradeoffs via gated encoding. TopK SAEs [Gao et al., 2025] enforce fixed sparsity via top-k activation selection. Zhang et al. [2025] extended SAEs to vision-language models, validating the approach beyond text. Most recently, Mishra [2026] applied SAEs to time-series foundation models, revealing causal feature hierarchies. No prior work applies SAEs to clinical time-series models with explicit spatiotemporal structure.

2.4 Concept-based explanations

TCAV [Kim et al., 2018] tests whether models use human-defined concepts via directional derivatives. ACE [Ghorbani et al., 2019] automates concept extraction via superpixel clustering. Concept Bottleneck Models (CBMs) [Koh et al., 2020] enforce a concept layer during training, with post-hoc extensions removing the training requirement [Yuksekgonul et al., 2023].

TI-SAE differs from CBMs in several ways: (a) it is fully post-hoc on a frozen backbone (no architectural modification or retraining), (b) concepts are *discovered* from data via dictionary learning rather than defined a priori, (c) each concept has learned temporal extent (Gaussian islands) rather than being a static attribute, and (d) concept labels are post-hoc validation, not the concept definition itself. Table 1 summarizes these structural distinctions.

3 Method

3.1 Problem setup and notation

We consider inpatient EHR episodes consisting of irregular multivariate observations $\{(t_i, v_i, x_i)\}_{i=1}^{L}$, where $t_i \in [0, T)$ is the timestamp, $v_i \in \{1, \ldots, V\}$ is the variable index, and $x_i \in \mathbb{R}$ is the observed value. The clinical outcome is a binary label $y \in \{0, 1\}$. Variables are organized into C clinical systems (e.g., cardiovascular, renal, respiratory) via a known ontology mapping $\phi : \{1, \ldots, V\} \to \{1, \ldots, C\}$, where some variables are shared across two channels.

A pretrained backbone f_θ transforms events into contextual embeddings $\mathbf{H} \in \mathbb{R}^{B \times L \times D}$, which are pooled into a patient-level summary $\mathbf{q} \in \mathbb{R}^D$ and passed to a prediction head $\hat{y} = \sigma(g_\psi(\mathbf{q}))$. The embeddings are organized into a dense time$\times$variable grid $\mathbf{H}_{\text{grid}} \in \mathbb{R}^{B \times T \times V \times D}$, with an observation mask $\mathbf{O} \in \{0, 1\}^{B \times T \times V}$ indicating which cells contain real observations. The backbone is frozen throughout; TI-SAE operates entirely on $\mathbf{H}_{\text{grid}}$.

3.2 TI-SAE architecture

TI-SAE decomposes $\mathbf{H}_{\text{grid}}$ into K latents, each with M temporal islands. The architecture has four components: encoder, island parameter network, island selector, and decoder (Figure 2).

3.2.1 Encoder

The encoder maps $\mathbf{H}_{\text{grid}}$ to pre-selection activation traces $\mathbf{A}_0 \in \mathbb{R}^{B \times K \times T}$. Each latent k has a sparse variable-routing vector $\mathbf{w}_k^{\text{enc}} \in \mathbb{R}^V$ (top-κ_{enc}, row-normalized, non-negative) and a projection vector $\mathbf{d}_k \in \mathbb{R}^D$:

$$A_0[b, k, t] = \text{ReLU}\left(\sum_{v=1}^{V} O[b, t, v] \cdot W^{\text{enc}}[k, v] \cdot (\mathbf{H}_{\text{grid}}[b, t, v, :] \cdot \mathbf{d}_k) + b_k\right) \tag{1}$$

Each latent attends to a sparse subset of variables (via $\mathbf{w}_k^{\text{enc}}$), projected onto $\mathbf{d}_k$ to produce a scalar activation.

3.2.2 Temporal Gaussian islands

The **Island Parameter Network** transforms activation traces into per-latent temporal parameters. For each latent k and island $m \in \{1, \ldots, M\}$, it produces a center $\mu_{k,m}$ and width $\sigma_{k,m}$:

$$\mathbf{h}_k(t) = \text{Conv1D}(\mathbf{A}_0[b, k, :]) \in \mathbb{R}^{H \times T} \tag{2}$$

$$\mu_{k,m} = \sum_{t=0}^{T-1} \text{softmax}_t\big((\boldsymbol{\ell}_\mu^{(m)})^\top \mathbf{h}_k(t)\big) \cdot t \tag{3}$$

$$g_{k,m}(t) = \exp\left(-\frac{(t - \mu_{k,m})^2}{2\sigma_{k,m}^2}\right) \tag{4}$$

A 1D convolutional feature extractor maps each activation trace to a temporal feature map $\mathbf{h}_k(t)$ (Eq. 2). Each island head m applies a pointwise projection $\boldsymbol{\ell}_\mu^{(m)}$ to produce attention logits over

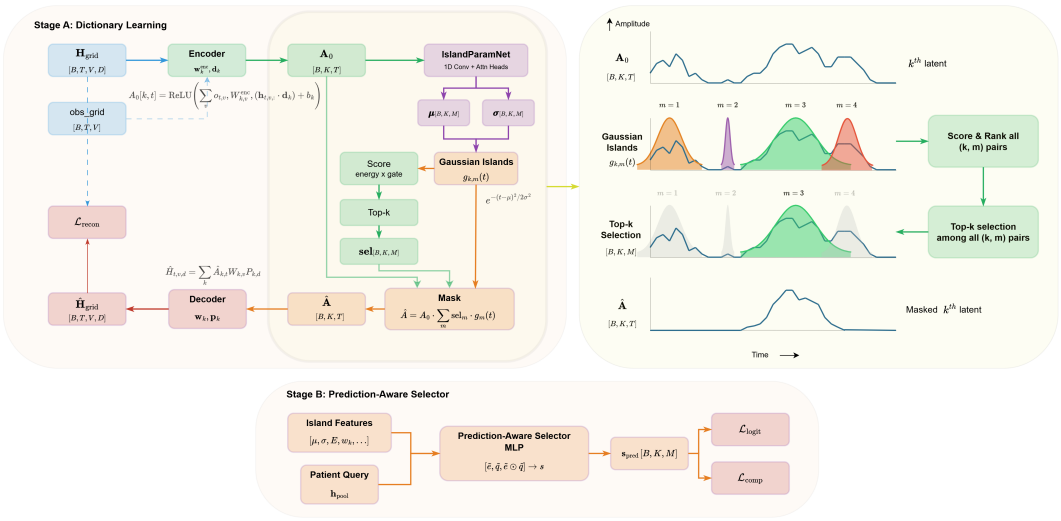


Figure 2: **TI-SAE architecture.** The encoder maps $\mathbf{H}_{\text{grid}}$ to activation traces $\mathbf{A}_0[B, K, T]$. The IslandParamNet produces M Gaussian islands per latent with learned centers μ and widths σ. Energy-gated selection picks top-k islands, defining temporal masks. The decoder reconstructs $\hat{\mathbf{H}}_{\text{grid}}$ via sparse routing weights $\mathbf{W}$ and atoms $\mathbf{P}$. The prediction-aware selector is trained separately on the frozen latent basis to identify prediction-relevant islands.

time, and $\mu_{k,m}$ is the softmax-weighted average time position (Eq. 3). The width $\sigma_{k,m}$ is estimated via EM: iterating between soft assignment of time steps to islands and variance estimation, with σ clamped to $[\sigma_{\min}, \sigma_{\max}]$. Each Gaussian island $g_{k,m}(t)$ defines a soft temporal window. The window is a temporal support gate indicating *when* the concept is active. The concept's activation trajectory is carried by the encoder activation $\mathbf{A}_0$ within the window.

3.2.3 Island selection

From $K \times M$ candidate islands per sample, we select a budget of n_{global} islands via differentiable top-k with straight-through estimation (STE) [Bengio et al., 2013]. Each island (k, m) is scored by its energy-gated activation:

$$s_{k,m} = \frac{\sum_t A_0[b, k, t] \cdot g_{k,m}(t)}{\sqrt{\sum_t g_{k,m}(t)}} \cdot \text{gate}_{k,m} \tag{5}$$

where the numerator measures total activation energy within the Gaussian window, the denominator normalizes by effective window length, and $\text{gate}_{k,m} \in [0, 1]$ is a learned budget-control signal.

The selected islands define a temporal mask for each latent:

$$\text{mask}[b, k, t] = \text{clamp}\left(\sum_{m=1}^{M} \text{sel}_{k,m} \cdot g_{k,m}(t),\ 0,\ 1\right) \tag{6}$$

and the masked activation is $\hat{\mathbf{A}}[b, k, t] = A_0[b, k, t] \cdot \text{mask}[b, k, t]$.

3.2.4 Decoder

The decoder reconstructs $\mathbf{H}_{\text{grid}}$ from the masked activations. Each latent k has an atom $\mathbf{p}_k \in \mathbb{R}^D$ and reuses the same underlying routing parameters as the encoder, with top-κ_{dec} sparsification ($\kappa_{\text{dec}} < \kappa_{\text{enc}}$):

$$\hat{\mathbf{H}}_{\text{grid}}[b, t, v, d] = \sum_{k=1}^{K} \hat{\mathbf{A}}[b, k, t] \cdot W[k, v] \cdot P[k, d] + b_d \tag{7}$$

where $\mathbf{b} \in \mathbb{R}^D$ is a shared bias.

3.3 Stage A: dictionary learning

TI-SAE is trained to reconstruct frozen backbone representations. Before encoding, we subtract a population-level per-variable mean $\bar{\mathbf{H}} \in \mathbb{R}^{V \times D}$ computed from training data, so the SAE operates on $\mathbf{H}_{\text{grid}} - \bar{\mathbf{H}}$; the mean is added back after decoding. The primary loss is per-cell normalized mean squared error (NMSE) on observed grid cells:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{O}|} \sum_{(b,t,v) \in \mathcal{O}} \frac{\|\hat{\mathbf{H}}_{\text{grid}}[b,t,v,:] - \mathbf{H}_{\text{grid}}[b,t,v,:]\|_c^2}{\|\mathbf{H}_{\text{grid}}[b,t,v,:]\|_c^2 + \epsilon} \tag{8}$$

where $\|\cdot\|_c$ denotes centered (mean-subtracted) norm and $\mathcal{O} = \{(b,t,v) : O[b,t,v] = 1\}$.

Dead latent revival. Latents that are not selected for $>\delta$ batches are declared "dead." An auxiliary loss $\mathcal{L}_{\text{auxk}}$ trains dead latents to reconstruct the residual $\mathbf{H}_{\text{grid}} - \hat{\mathbf{H}}_{\text{grid}}$, reviving unused capacity [Gao et al., 2025].

Regularizers. μ-repulsion (λ_{repul}) pushes island centers apart within each latent, preventing multiple islands converge to the same time position. Decoder diversity (λ_{div}) penalizes cosine similarity between $\mathbf{W}$ rows, encouraging variable-composition diversity across latents. Mask coverage (λ_{mcov}) penalizes the total Gaussian mass of selected islands, preventing TI-SAE from increasing temporal coverage to favor reconstruction. Channel purity ($\lambda_{\text{ch_purity}}$) encourages each latent's routing weights to concentrate on variables from a single clinical channel. Full regularizer formulas are in Appendix C.

The Stage A total loss is:

$$\mathcal{L}_A = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{auxk}}\mathcal{L}_{\text{auxk}} + \lambda_{\text{repul}}\mathcal{L}_{\text{repul}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{mcov}}\mathcal{L}_{\text{mcov}} + \lambda_{\text{ch}}\mathcal{L}_{\text{ch}} \tag{9}$$

3.4 Stage B: prediction-aware selector

The dictionary from Stage A captures what the model represents; a prediction-aware concept selector (hereafter, the prediction selector) identifies which concepts drive each prediction. The entire TI-SAE is frozen during training.

Each candidate island (k, m) is described by a feature vector $\mathbf{e}_{k,m}$ combining its temporal parameters (μ, σ), activation statistics, the routing weights $\mathbf{w}_k$, and observation coverage within the island window (full specification in Appendix E). A patient-level query $\mathbf{q}$ is obtained from the backbone's attention pooling. Island features and patient query are projected into a shared space, combined via element-wise interaction, and scored:

$$s_{k,m}^{\text{pred}} = \text{MLP}([\mathbf{U}_e\mathbf{e}_{k,m};\ \mathbf{U}_q\mathbf{q};\ \mathbf{U}_e\mathbf{e}_{k,m} \odot \mathbf{U}_q\mathbf{q}]) \tag{10}$$

where $\mathbf{U}_e$ and $\mathbf{U}_q$ are learned projection matrices.

Islands are selected via softmax-weighted scoring with temperature τ, producing continuous selection weights that sum to $n_{\text{global}}^{\text{pred}}$.

Losses. Let z_{keep}, z_{drop}, and z_{full} denote the prediction logit when using only selected islands, only the complement, and all islands respectively, with teacher logit z_{teacher} from the frozen backbone.

$$\mathcal{L}_{\text{logit}} = \|z_{\text{keep}} - z_{\text{teacher}}\|^2 \tag{11}$$

$$\mathcal{L}_{\text{comp}} = \text{ReLU}(\delta - c \cdot (z_{\text{full}} - z_{\text{drop}})) \tag{12}$$

where $c = \text{sign}(z_{\text{teacher}})$ is the teacher's predicted class direction and δ is a hinge margin. The total prediction selector loss is $\mathcal{L}_B = \lambda_{\text{logit}}\mathcal{L}_{\text{logit}} + \lambda_{\text{comp}}\mathcal{L}_{\text{comp}}$.

3.5 Latent characterization

The learned dictionary is characterized post-hoc. Each latent k is described by actual raw-value patterns of its variable composition ($\mathbf{w}_k$) and temporal extents (island μ, σ) from top-N exemplar episodes ranked by activation and the prediction-aware selection. In this study, an LLM labeler produces per-latent clinical concept labels and descriptions from these summaries (Appendix L).

4 Experimental setup

4.1 Dataset

We use MIMIC-III [Johnson et al., 2016] and MIMIC-IV [Johnson et al., 2023], publicly available critical care databases. We extract 48-hour ICU episodes with $V=135$ clinical variables (MIMIC-III) and $V=132$ variables (MIMIC-IV). We evaluate on three binary prediction tasks: in-hospital mortality, 30-day readmission, and length-of-stay >7 days. MIMIC-III contains $N=28{,}626$ episodes (14.1% mortality); MIMIC-IV contains $N=46{,}140$ episodes (14.4% mortality). Both are split into train/validation/test (64/16/20%) by patient.

4.2 Model training

We evaluate TI-SAE across five backbone configurations (Section 2.1): a discrete transformer (with and without pretraining), STraTS [Tipirneni and Reddy, 2022], HyMaTE [Mottalib et al., 2025], and BAT [DeVries et al., 2025]. TI-SAE ($K=1280$, $M=4$, $n_{\text{global}}=128$) and the prediction selector ($n_{\text{global}}^{\text{pred}}=64$) are trained on each frozen backbone. Full architecture specifications and training details are in Appendix A.

Baselines. We compare against backbone attention, Integrated Gradients [Sundararajan et al., 2017], TimeSHAP [Bento et al., 2021], Dynamask [Crabbé and Van Der Schaar, 2021], TSR [Ismail et al., 2020], Cluster-TCAV [Ghorbani et al., 2019, Kim et al., 2018], FIT [Tonekaboni et al., 2020], WinIT [Leung et al., 2023], and random sanity checks, all evaluated at matched top-f fractions.

4.3 Evaluation metrics

4.3.1 Faithfulness

Comprehensiveness measures the prediction change when the top-f most important events are removed; sufficiency measures the change when only those events are kept. We evaluate under three perturbation strategies that define how removed events are treated: *removal* (event deleted entirely), *marginal* (replaced with the per-variable median from the training set), and *neighborhood-based* (replaced with the patient's temporally nearest observed value of the same variable, approximating conditional sampling $P(x_v \mid x_{\text{observed}})$). We report neighborhood-based sampling as the primary strategy, as it best approximates realistic clinical counterfactuals. We summarize the comprehensiveness curve with AOPC (Area Over the Perturbation Curve).

4.3.2 Clinical plausibility

We evaluate whether attributions align with established clinical severity indicators, using SOFA organ-system severity [Vincent et al., 1996] as soft ground truth generated on both MIMIC-III and MIMIC-IV datasets. We measure whether each method's top-attributed cells correspond to clinically severe observations (AUROC and AUPRC). Ground truth generation details are in Appendix I.

4.3.3 Verbalized explanation quality

For patient-level explanations, the top prediction selector-selected latents are combined with raw clinical values and the model's prediction, then passed to an LLM verbalizer to produce natural language explanations. We evaluate verbalization quality by comparing explanations under two conditions: *attribution-augmented* (top-K attributed cells) and *TI-SAE-augmented* (top prediction selector-selected concepts with dictionary labels), both receiving the same patient data table with matched information budget. This comparison mimics a clinician reading each method's explanation, so the inputs intentionally differ in the two conditions. A pairwise LLM judge and four tertiary-hospital clinicians independently evaluate each dimension: *concept-level reasoning*, *clinical actionability*, and *faithfulness*. Full protocol in Appendix M.

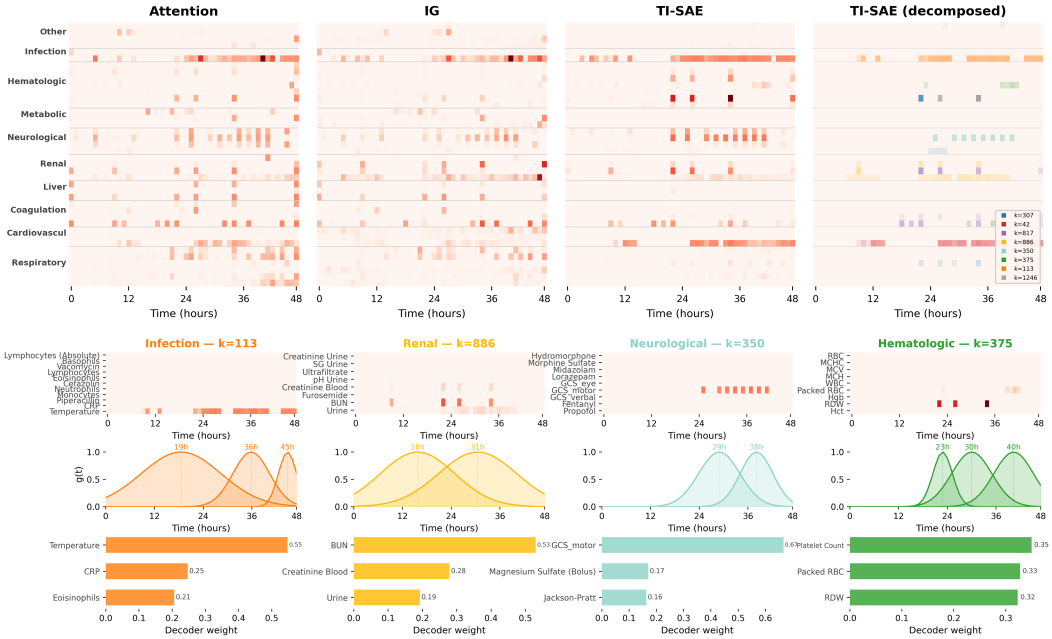


Figure 3: **Qualitative comparison of attribution methods and TI-SAE decomposition.** (a) Attribution heatmaps for an example mortality episode. TI-SAE produces structured attribution that decomposes into per-channel latent contributions (rightmost column: each color represents one latent). (b) Four selected latents from the decomposition. Top row: attribution restricted to the latent. Middle row: learned Gaussian islands defining when the concept is active. Bottom row: decoder variable-composition weights $\mathbf{w}_k$.

Table 2: **Faithfulness and clinical plausibility across backbone architectures.** AOPC: neighborhood-based perturbation. AUROC/AUPRC: plausibility against SOFA-derived clinical prior. Averaged over tasks and datasets. Best per column in **bold**, second-best underlined.

	Faithfulness (AOPC ↑)					Clinical Plausibility (AUROC/AUPRC ↑)				
	Disc.	D+PT	HyM.	BAT	STraTS	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	**.086**	**.076**	**.097**	**.043**	.016	**.593/.286**	.589/.255	.614/**.276**	.588/.251	.536/.227
TI-SAE (hard)	.079	.068	.090	.041	.016	.569/.267	.568/.243	.581/.257	.567/.243	.535/.228
Attention	.053	.019	.051	–	.016	.570/.242	.539/.219	.554/.218	–	.551/.217
IG	.060	.051	.049	.031	.067	.545/.225	.549/.229	.596/.244	.584/.239	.577/.223
TSR	.063	.054	.062	.037	**.071**	.566/.243	.573/.241	**.625**/.272	**.608/.261**	.595/.234
DynaMask	.019	.032	.012	.028	.034	.510/.201	.503/.200	.494/.187	.493/.187	.494/.187
TimeSHAP	.009	.011	.012	.016	.029	.512/.202	.517/.206	.516/.205	.519/.209	.513/.201
WinIT	.020	.030	.032	.021	.032	.578/.252	**.611/.280**	.611/.273	.565/.247	**.621/.280**
Random	.002	.003	.002	.006	.007	.502/.190	.502/.190	.502/.190	.502/.190	.502/.190

5 Results

We evaluate TI-SAE across five backbone architectures on MIMIC-III and MIMIC-IV (three prediction tasks each). Figure 3 illustrates that TI-SAE's attribution decomposes into per-channel latent contributions with learned temporal windows and variable composition.

5.1 Faithfulness and clinical plausibility

Table 2 presents faithfulness (AOPC under neighborhood-based perturbation) and clinical plausibility (AUROC and AUPRC against SOFA severity) across backbones, averaged over tasks and datasets. TI-SAE achieves the highest AOPC on four of five backbones (Discrete, Disc.+PT, HyMaTE, BAT), exceeding baselines. On STraTS, TI-SAE's advantage is reduced; we find that STraTS's residual-

averaging architecture does not preserve the per-variable locality that TI-SAE's routing requires (Appendix J). On clinical plausibility, TI-SAE achieves the highest AUPRC and AUROC on the discrete backbone, and competitive scores on the others.

5.2 Dictionary validation

The learned dictionary captures clinically coherent patterns: high-mortality latents correspond to vasopressor-dependent shock and respiratory failure, while protective latents capture hemodynamic stability and normal acid-base status (Appendix G). Four independent clinicians rated a sample of 30 dictionary labels on a 5-point Likert scale, yielding mean ratings of 4.07–4.40 (Appendix G.5).

The learned concepts are reproducible across independent training runs, both when retraining the full pipeline (decoder cosine .724, variable overlap 77%) and per stage (dictionary alone: decoder cosine .735, variable overlap 82%; selector on a fixed dictionary: top-64 Jaccard .708, rank correlation .87; Appendix G.7), and zero-ablating importance-ranked concepts degrades predictions monotonically and more than random-order removal, confirming that individual concepts carry non-redundant predictive contributions (Appendix G.6). The learned dictionary shows cross-dataset transfer: faithfulness and plausibility preserved when the MIMIC-III–trained model is applied to disjoint MIMIC-IV without retraining (Table 3, Appendix G.8).

Table 3: **Cross-dataset transfer**. MIMIC-III→IV. MIMIC-IV was restricted to admissions from 2014 to remove patient and temporal overlap with MIMIC-III. For the Discrete backbone trained on MIMIC-III mortality, TI-SAE and prediction selector are applied directly to MIMIC-IV. Baselines recomputed on disjoint MIMIC-IV. Neighborhood-based perturbation. Backbone AUROC: .861→.864. ID = in-distribution; OOD = out-of-distribution.

	AOPC ↑		AUROC/AUPRC ↑	
	ID (MIMIC-III)	OOD (MIMIC-IV)	ID (MIMIC-III)	OOD (MIMIC-IV)
TI-SAE (soft)	.149	.130	.756/.386	.756/.370
Attention	.100	.097	.616/.259	.631/.261
IG	.086	.045	.567/.255	.552/.224
TSR	.088	.046	.593/.280	.587/.245

5.3 Verbalized explanation quality

Table 4 reports pairwise win counts from the LLM judge (N=50 per baseline) and the clinician panel (N=50 on the IG comparison). TI-SAE wins concept-level reasoning and actionability across all baselines. Clinician majority votes match the LLM judge's ranking direction for concept-level reasoning and actionability, with faithfulness near-tied for both. Per-rater details and inter-rater agreement analysis are reported in Appendix M.2.

Table 4: **Verbalized explanation quality.** Pairwise win counts: TI-SAE / baseline (ties). LLM judge (N=50 per baseline) and clinician majority vote (four independent raters) on the IG comparison.

	LLM judge			Clinician
	vs. Att	vs. IG	vs. TSR	vs. IG
Concept	**30**–17 (3)	**33**–15 (2)	**33**–16 (1)	**39**–4 (7)
Actionability	**28**–22 (0)	**32**–18 (0)	**27**–23 (0)	**23**–9 (18)
Faithfulness	**19**–15 (16)	**15**–11 (24)	**28**–9 (13)	4–4 (42)

An ablation isolating the role of the dictionary labels (Appendix M.3) shows that much of this verbalized advantage derives from the cross-patient concept labels: with these labels removed, TI-SAE's concepts no longer clearly beat per-cell attribution. However, the concept labels are generated from cross-patient exemplar evidence, and only TI-SAE's shared dictionary provides them with spatiotemporal grounding.

5.4 Ablation study

We ablate TI-SAE's structural mechanisms, dictionary capacity, and loss components (Appendix J). Within TI-SAE, sparse variable routing is the critical mechanism for faithfulness (removing it costs 58–75% of AOPC), while the temporal islands and channel organization have minimal impact on faithfulness but provide the concepts' clinical readability: without them, channel coherence drops from 91.9% to 36.9% and temporal localization collapses (Appendix J.3). Isolating the dictionary from the prediction selector shows that the same selection objective applied to raw hidden states collapses (neighborhood-based AOPC .038; Appendix J.3). Composition over multiple windows matters, as a single island per concept ($M{=}1$) degrades both faithfulness and clinical plausibility (Appendix J.5). The symmetric Gaussian window shape is not restrictive: a learned split-normal window changes neither metric (Appendix J.4).

6 Limitations

First, TI-SAE operates on a fixed-length time$\times$variable grid; variable-length sequences or observation windows beyond the 48-hour setting used here would require adaptation, which we have not validated. Second, TI-SAE's variable-routing encoder requires hidden representations that balance per-variable locality with contextual concept formation. Not all backbone layers achieve this. Third, while TI-SAE's latents are structurally defined by sparse variable composition and temporal extent, their identification as *clinical concepts* relies on post-hoc validation. Fourth, SOFA-derived severity is a coarse, organ-independent proxy that cannot capture the full complexity of clinical reasoning. Fifth, TI-SAE assumes a single-token-per-event representation that maps to one cell in the time$\times$variable grid; multi-token strategies may require architectural adaptation. Sixth, our verbalization evaluation is mediated by an LLM verbalizer whose strong clinical priors can produce fluent explanations from any source: high quality does not guarantee the explanation reflects model reasoning, and although concept-based inputs constrain the verbalizer more tightly, this asymmetry is not directly measured.

7 Conclusion

We introduced TI-SAE, a sparse autoencoder that decomposes frozen clinical time-series model representations into a dictionary of temporal clinical concepts. Each concept is defined by learned Gaussian temporal islands and sparse variable composition, enabling structured explanations that go beyond cell-level attribution scores. Evaluated across five backbone architectures on MIMIC-III and MIMIC-IV with three different tasks, TI-SAE achieves the highest faithfulness on four of five backbones while providing a decomposable, clinically labeled dictionary that no baseline method can produce.

Acknowledgments and Disclosure of Funding

This work was supported by the National Institutes of Health and National Institute on Aging (project number R01AG084637 and R01AG082721); the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2026; the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Nos. 2022R1A3B1077720 and 2022R1A5A7083908); the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program, Seoul National University); and the InnoCORE program of the Ministry of Science and ICT(1.260007.01).

The authors declare no competing interests.

References

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

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*, page 2565–2573, 2021. doi: 10.1145/3447548.3467166.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Himabindu Lakkaraju, and Flavio Calmon. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bojVI4l9Kn.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892, 2018.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2166–2177, 2021.

Rachael DeVries, Casper Christensen, Marie Lisandra Zepeda Mendoza, and Ole Winther. Biaxial transformers: Addressing the increasing complexity of EHR classification, 2025. URL https://arxiv.org/abs/2508.12418.

Joseph Enguehard. Learning perturbations to explain time series predictions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 9329–9342, 23–29 Jul 2023.

Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. EHRMamba: Towards generalizable and scalable foundation models for electronic health records. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259, pages 291–307, 2025.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

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

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 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. doi: 10.1038/s41597-019-0103-9.

Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H Lehman, Leo A Celi, and Roger G Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10 (1):1, 2023.

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.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348, 2020.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL `https://arxiv.org/abs/2009.07896`.

Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G. Krishnan. DuETT: Dual event time transformer for electronic health records. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219, 2023.

Kin Kwan Leung, Clayton Rooke, Jonathan Smith, Saba Zuberi, and Maksims Volkovs. Temporal dependencies in feature importance for time series prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=C0q9oBc3n4`.

Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020.

Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. Timex++: learning time-series explanations with information bottleneck. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Zichuan Liu, Yingying ZHANG, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *The Twelfth International Conference on Learning Representations*, 2024b. URL `https://openreview.net/forum?id=qDdSRaOiyb`.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, 2017.

Anurag Mishra. Dissecting chronos: Sparse autoencoders reveal causal feature hierarchies in time series foundation models. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL `https://openreview.net/forum?id=Q5ox302DLN`.

Md Mozaharul Mottalib, Thao-Ly T. Phan, and Rahmatollah Beheshti. HyMaTE: A hybrid mamba and transformer model for EHR representation learning. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2025.

Owen Queen, Thomas Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=zLBlin2zvW.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1.

Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7 (1):256, 2024.

Jon Saad-Falcon, Rajan Pathe Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. LMUNIT: Fine-grained evaluation with natural language unit tests. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3303–3324, 2025. doi: 10.18653/v1/2025.findings-emnlp.176.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 3145–3153, 2017.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. doi: https://doi.org/10.1016/j.jbi.2020.103637.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 3319–3328, 2017.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data*, 16(6), 2022. doi: 10.1145/3516367.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, pages 359–380, 2019. URL https://proceedings.mlr.press/v106/tonekaboni19a.html.

Sana Tonekaboni, Shalmali Joshi, Kieran R Campbell, David Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Jean Louis Vincent, Rui Paolo Jinó Moreno, Jukka Takala, Sheila M. Willatts, Arnaldo de Mendonça, Hajo A. Bruining, Camelia Reinhart, Peter M. Suter, and L. G. Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22 (7):707–710, 1996.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=nA5AZ8CEyow`.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3650–3661, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

A Training and hyperparameter details

All models are trained on NVIDIA A100-SXM4-80GB GPUs. Deterministic seeding (seed 42) is used throughout.

Backbones. Table 5 summarizes the five backbone configurations. All backbones are trained for binary prediction via cross-entropy with inverse-prevalence class weighting. The discrete transformer and its pretrained variant share the same architecture (D=128, 4 layers, 8 heads); the pretrained variant additionally undergoes 2-hour value-bin forecast pretraining before finetuning. STraTS and HyMaTE use continuous value embedding (CVE) with D=64; STraTS uses 4 transformer layers with 16 heads, while HyMaTE uses 2 Mamba layers iterated over 3 blocks. BAT uses 1 bi-axial layer with embed_out=44 (total D=88 after track concatenation), 1 attention head, and input masking for unobserved cells. All backbones are frozen for subsequent stages.

Table 5: Backbone configurations.

Backbone	D	Layers	Heads	Dropout	LR	Tokenization
Discrete	128	4	8	0.3	10^{-4}	Binned (single-token)
Discrete+PT	128	4	8	0.3	10^{-4}	Binned (single-token)
STraTS	64	4	16	0.2	5×10^{-5}	CVE (single-token)
HyMaTE	64	2×3	4	0.2	5×10^{-5}	CVE (single-token)
BAT	88	1	1	0.2	5×10^{-4}	Grid-native

TI-SAE (stage A). Table 6 lists the full TI-SAE hyperparameter configuration, applied identically across all backbones. The model has ≈856K trainable parameters (encoder: 502K, IslandParamNet: 27K, decoder: 328K). The variable routing matrix $\mathbf{W}$ is shared between encoder and decoder, with independent sparsification (κ_{enc}=5 and κ_{dec}=3), along with row-normalization and non-negativity. Before encoding, a per-variable population mean $\bar{\mathbf{H}} \in \mathbb{R}^{V\times D}$ is subtracted from $\mathbf{H}_{\text{grid}}$ (computed from 100 training batches) and added back after decoding.

Training uses AdamW (lr 10^{-4}, weight decay 10^{-2}), cosine annealing, batch size 32, gradient norm clipping at 1.0, and early stopping on total validation loss (patience 15). No prediction loss is used in Stage A (λ_{pred}=0); prediction relevance is handled by the prediction selector in Stage B.

Prediction selector (stage B). The prediction selector is trained on the frozen TI-SAE dictionary with the backbone also frozen (≈38K trainable parameters). Each island is represented by a $6{+}2V$-dimensional feature vector (Appendix E). Candidates are pre-filtered to the top-2560 islands by energy score, from which the prediction selector selects $n^{\text{pred}}_{\text{global}}$=64 islands per sample via soft (softmax-weighted) selection with temperature τ=1.0. The reference logit z_{full} uses a separate pool of 256 top-energy islands, selected as the pool size that minimized $|z_{\text{full}} - z_{\text{teacher}}|^2$ on the validation set.

Table 6: TI-SAE hyperparameters.

Category	Parameter	Value
Architecture	Latents K	1280
	Islands per latent M	4
	Selection budget n_{global}	128
Encoder / Decoder	Encoder Variable Routing sparsity κ_{enc}	5
	Decoder Variable Routing sparsity κ_{dec}	3
	Atom normalization	L2-normalize
IslandParamNet	Hidden dimension H	32
	Conv kernel size	25
	EM iterations n_{iter}	3
	σ bounds $[\sigma_{\min}, \sigma_{\max}]$	$[0.5,\ 12.0]$
	Score mode	energy-gated
Selection	Gumbel noise scale	0.05
	Dead boost scale	0.5
	Dead threshold (batches)	5
Loss weights	λ_{recon}	1.0
	λ_{auxk} (dead latent revival, k_{aux}=256)	0.3
	λ_{repul} (island center repulsion, δ=2.0)	0.01
	λ_{mcov} (mask coverage)	0.005
	λ_{div} (decoder diversity)	0.01
	$\lambda_{\text{ch_purity}}$ (channel purity)	0.1
Optimization	Optimizer	AdamW
	Learning rate	10^{-4}
	Weight decay	10^{-2}
	Batch size	32
	LR schedule	Cosine annealing
	Max gradient norm	1.0
	Max epochs / patience	100 / 15

Training uses AdamW (lr 10^{-3}, no weight decay), batch size 52, gradient norm clipping at 1.0, and early stopping on validation $\mathcal{L}_B$ (patience 20, up to 100 epochs). Loss weights are λ_{logit}=1.0, λ_{comp}=0.1, and δ_{comp}=0.5. Projection dimension d_{sel}=64, MLP hidden dimension 64.

Pretraining details. STraTS, HyMaTE, and Discrete+PT use a 48-hour sliding window with 2-hour value forecast pretraining objective before finetuning. Pretraining uses a higher learning rate (5×10^{-4}) than finetuning (5×10^{-5}); architecture hyperparameters (dimensions, layers, dropout) are shared. The pretrained checkpoint is loaded and all layers are finetuned end-to-end on the downstream task.

Implementation notes. Our STraTS uses 4 transformer layers (vs. 2 in the original) to provide comparable model depth across backbones. For HyMaTE, we correct a tensor dimension mismatch in the original supervised forward path. BAT uses input masking (setting unobserved grid cell embeddings to zero before attention) to produce representations compatible with TI-SAE's grid-based encoding.

B Baseline method adaptation

Our backbone consumes irregular event sequences $(v_j, t_j, x_j)_{j=1}^{L}$, not a regular $[T, V]$ grid. Some baseline attribution methods require adaptation to this input format.

Gradient-based methods (IG, GradSHAP, TSR). The backbone embeds discrete token indices into continuous vectors $\mathbf{e}_j = \mathbf{e}_{\text{val}}(x_j) + \mathbf{e}_{\text{var}}(v_j) + \mathbf{e}_{\text{time}}(t_j)$, so we attribute in the continuous embedding space: `captum` [Kokhlikyan et al., 2020] treats the embedding tensor as the input and everything downstream (transformer layers, pooling, prediction head) as the forward function. This yields per-event, per-dimension attributions $a_{j,d}$. We first sum over embedding dimensions to obtain

a signed group attribution $\bar{a}_j = \sum_d a_{j,d}$ for each event. For ranking-based faithfulness evaluation, we use its magnitude $a_j = |\bar{a}_j|$ as the event-importance score, and place each a_j at its grid position (t_j, v_j) to obtain a $[T, V]$ attribution map.

The choice of reference embedding (baseline) determines what the attribution measures. Ideally, a null baseline representing "no event" would be used, but unobserved cells have no corresponding token in the sequence. A zero embedding is the closest approximation, attributing importance to the observation's existence. A per-variable median baseline $\mathbf{e}_j^{\text{ref}} = \mathbf{e}_{\text{val}}(b_{\text{med}}(v_j)) + \mathbf{e}_{\text{var}}(v_j) + \mathbf{e}_{\text{time}}(t_j)$ instead attributes importance to the event's *value* deviating from a typical value for that variable. We evaluated both and selected the median baseline, for yielding better faithfulness for each method.

Dynamask. Dynamask optimizes a per-sample soft mask $\mathbf{m} \in [0,1]^{T \times V}$ to identify features whose removal most changes the prediction. Since the backbone takes discrete token indices, the optimization requires a differentiable approximation. For removal masking, we multiply event embeddings by the soft mask value: $\tilde{\mathbf{e}}_j = m_{t_j,v_j} \cdot \mathbf{e}_j$. For marginal and neighborhood masking, we pre-compute a replacement embedding per event (by re-embedding a sampled or nearest-neighbor value) and interpolate: $\tilde{\mathbf{e}}_j = m_{t_j,v_j} \cdot \mathbf{e}_j + (1 - m_{t_j,v_j}) \cdot \mathbf{e}_j^{\text{ref}}$. At intermediate mask values these blended embeddings have no token-level equivalent, but the final attribution is binary-thresholded during faithfulness evaluation, where embedding interpolation and token-level replacement coincide. We fit masks at four sparsity levels (10%, 20%, 30%, 50%) and select the one with lowest prediction error.

TimeSHAP. TimeSHAP [Bento et al., 2021] computes cell-level Shapley values via hierarchical KernelSHAP: event-level pruning identifies the relevant time window, then event-level and feature-level SHAP values are computed, and finally cell-level values are estimated for the top events $\times$ features. The package requires a callable from $[T, V]$ observation grids to predictions. Each coalition determines which observed cells are included; excluded cells are handled according to the masking strategy (removed, marginally replaced, or neighborhood-replaced).

Cluster-TCAV. We construct a TCAV [Kim et al., 2018]-inspired prototype baseline using ACE-style unsupervised concept discovery [Ghorbani et al., 2019]. Contextual event embeddings $\mathbf{h}_j \in \mathbb{R}^D$ from all training samples are clustered into $K{=}50$ concepts via Mini-Batch K-means. For each concept, we use the normalized cluster center as a CAV proxy and compute the average directional derivative of the prediction logit along this direction over training samples; concepts whose importance is not statistically significant (one-sample t-test against 20 random directions, $p \geq 0.05$) are zeroed out. At inference, each event is assigned to its nearest concept and receives that concept's importance as its attribution score, placed at its grid position (t_j, v_j).

FIT and WinIT. FIT [Tonekaboni et al., 2020] and WinIT [Leung et al., 2023] require a conditional generator that models $P(\mathbf{x}_t \mid \mathbf{x}_{<t})$. The standard implementations assume dense, regularly-sampled time series data. We evaluated several adaptations: mask-aware generator training, mask-concatenated input, and event-removal counterfactual without generator training. We report results from the event-removal variant, which performed best; instead of replacing other features' values with generator samples, it removes their events by zeroing the observation mask. For FIT, the counterfactual at time t retains only feature i's event; for WinIT, feature f's events are removed in windows $[t, t+n]$ with finite-difference decomposition. Both methods' attributions are computed only at observed (t, v) cells.

TimeX, TimeX++, and ExtremalMask. TimeX [Queen et al., 2023] trains an interpretable surrogate that produces discrete attribution maps while preserving the behavior of a pretrained time-series model, and TimeX++ [Liu et al., 2024a] reformulates time-series explanation through an information-bottleneck objective. ExtremalMask [Enguehard, 2023] learns a per-sample mask and perturbation for time-series attribution. All three methods were designed for dense, regularly-sampled multivariate time series. Their default architectures fail on sparse clinical data (most grid cells unobserved); we adapted TimeX/TimeX++ by replacing their internal encoder with the frozen backbone (using only the `MaskGenerator` component conditioned on backbone-derived grid features), and ExtremalMask by wrapping the backbone in a grid-to-embedding adapter with tuned regularization ($\lambda_1{=}0.01$, lr$=0.01$). Despite these adaptations, all three produced near-random faithfulness, suggesting that the adaptations

did not yield reliable implementations for sparse event-sequence data. We therefore exclude them from the main comparison and report the adaptation details here for completeness.

Observation masking. All methods' final attribution maps are masked to observed cells. Without this, methods that produce full-grid attributions (Dynamask, TimeSHAP) waste top-f budget on unobserved cells that have no causal effect when perturbed.

CVE backbone adaptation. For CVE backbones (STraTS, HyMaTE), attributions can additionally be computed at the value level by treating the raw scalar input to the CVE MLP as the attributed input, yielding per-event value-sensitivity scores. We report the variant (embedding-level or value-level) with higher faithfulness per method.

C Regularizer formulas

Island center repulsion. Prevents islands within each latent from collapsing to the same center:

$$\mathcal{L}_{\text{repul}} = \left\langle \exp\left(-\frac{(\mu_{k,m} - \mu_{k,m'})^2}{2\delta^2}\right) \cdot \mathbf{1}[m \neq m'] \right\rangle_{b,k,m,m'} \tag{13}$$

averaged over all latents, island pairs, and samples, with δ=2.0.

Mask coverage. Penalizes total temporal mass of selected islands:

$$\mathcal{L}_{\text{mcov}} = \frac{1}{n_{\text{global}}} \sum_{k,m} \text{sel}_{k,m} \cdot \overline{g_{k,m}} \tag{14}$$

where $\overline{g_{k,m}} = \frac{1}{T}\sum_t g_{k,m}(t)$ is the temporal average of the Gaussian envelope, averaged over samples.

Decoder diversity. Penalizes cosine similarity between decoder weight rows:

$$\mathcal{L}_{\text{div}} = \left\langle \left(\frac{\mathbf{W}_k \cdot \mathbf{W}_{k'}}{\|\mathbf{W}_k\|\|\mathbf{W}_{k'}\|} - \mathbf{1}[k{=}k'] \right)^2 \right\rangle_{k,k'} \tag{15}$$

When channel organization is active, diversity is computed per-channel to avoid penalizing cross-channel differences.

Channel purity. Encourages each latent to concentrate its variable weight mass on a single clinical channel:

$$\mathcal{L}_{\text{ch_purity}} = 1 - \frac{1}{K}\sum_k \frac{\max_j \sum_v W[k,v] \cdot M_{\text{ch}}[j,v]}{\sum_v W[k,v]} \tag{16}$$

D Architecture details of IslandParamNet

Temporal feature extraction. The IslandParamNet transforms per-latent activation traces $A_0[b,k,:] \in \mathbb{R}^T$ into island parameters. A two-layer 1D convolutional network extracts temporal features:

$$\mathbf{h}_k = \text{GELU}(\text{Conv1D}_2(\text{GELU}(\text{Conv1D}_1(A_0[b,k,:])))) \in \mathbb{R}^{H\times T} \tag{17}$$

where both layers have hidden dimension H=32 and kernel size 25 (with zero-padding to preserve length).

Island center localization. For each island m, a pointwise projection $\boldsymbol{\ell}_\mu^{(m)} \in \mathbb{R}^H$ produces attention logits over time. The center $\mu_{k,m}$ is the softmax-weighted average time position (Eq. 3 in the main text).

Table 7: Components of the island feature vector $\mathbf{e}_{k,m} \in \mathbb{R}^{6+2V}$.

Component	Dim	Range	Description
$\mu_{k,m}/T$	1	$[0, 1]$	Normalized island center position
$\sigma_{k,m}/T$	1	$[0, 0.25]$	Normalized island width
$\widetilde{E}_{k,m}$	1	$[0, 1]$	Gaussian-weighted activation energy (batch-normalized)
$\widetilde{P}_{k,m}$	1	$[0, 1]$	Peak activation at $\mu_{k,m}$ (batch-normalized)
$\text{gate}_{k,m}$	1	$[0, 1]$	Budget gate value from IslandParamNet
$\widetilde{G}_{k,m}$	1	$[0, 1]$	Gradient$\times$activation prior score (batch-normalized)
$\mathbf{W}[k, :]$	V	$[0, 1]$	Variable-routing row (top-κ_{dec} sparse, row-normalized)
$\mathbf{o}_{k,m}$	V	$[0, 1]$	Observation profile within island window

EM sigma estimation. Island widths $\sigma_{k,m}$ are estimated via expectation-maximization, treating time steps as data points and islands as mixture components:

Initialize $\sigma_{k,m}^{(0)} = \min(T/2M,\ 3.0)$
for iter $= 1, \ldots, n_{\text{iter}}$ (default 3) **do**
  Gaussian kernel: $g_{m,t}^{\text{em}} = \exp\left(-\frac{(t-\mu_{k,m})^2}{2(\sigma_{k,m}^{(\text{iter}-1)})^2}\right)$
  Responsibilities: $r_{m,t} = g_{m,t}^{\text{em}} \big/ \sum_{m'} g_{m',t}^{\text{em}}$ (soft assignment)
  Weights: $w_{m,t} = A_0[b, k, t] \cdot r_{m,t}$ (activation-weighted)
  Update: $\sigma_{k,m}^{(\text{iter})} = \text{clamp}\left(\sqrt{\frac{\sum_t w_{m,t}(t-\mu_{k,m})^2}{\sum_t w_{m,t}}},\ \sigma_{\min},\ \sigma_{\max}\right)$
end for

The EM procedure has no learnable parameters; σ emerges from the shape of A_0. We use $\sigma_{\min}$=0.5, $\sigma_{\max}$=12.0, and n_{iter}=3.

Energy-gated scoring. Each island (k, m) is scored by its activation energy within the Gaussian window, normalized by effective support length:

$$s_{k,m} = \frac{\sum_t A_0[b, k, t] \cdot g_{k,m}(t)}{\sqrt{\sum_t g_{k,m}(t)}} \cdot \text{gate}_{k,m} \tag{18}$$

The $\sqrt{\cdot}$ normalization prevents wider islands from scoring higher purely due to coverage. The gate $\text{gate}_{k,m} \in [0, 1]$ is produced by a separate linear layer on *detached* temporal features (gradient from $\mathcal{L}_{\text{mcov}}$ reaches only the gate weights, not μ or σ), providing learned budget control.

Dead latent revival (AuxK). Latents not selected for $> \delta$ consecutive batches are declared dead (δ chosen so that approximately 150 training episodes elapse before revival). The top k_{aux}=256 dead latents (by energy) are trained to reconstruct the residual $\mathbf{H}_{\text{grid}} - \hat{\mathbf{H}}_{\text{grid}}$ via an auxiliary loss:

$$\mathcal{L}_{\text{auxk}} = \frac{\|(\hat{\mathbf{H}}_{\text{aux}} - (\mathbf{H}_{\text{grid}} - \hat{\mathbf{H}}_{\text{grid}})) \odot \mathbf{O}\|^2}{\|(\mathbf{H}_{\text{grid}} - \hat{\mathbf{H}}_{\text{grid}}) \odot \mathbf{O}\|^2} \tag{19}$$

where $\hat{\mathbf{H}}_{\text{aux}}$ is the dead-latent reconstruction and $\mathbf{O}$ is the observation mask. Latents with low gate values (< 0.1) are excluded from the dead set, as they are budget-paused rather than truly dead.

E Feature vector in the prediction selector

For each candidate island (k, m), we assemble a feature vector $\mathbf{e}_{k,m} \in \mathbb{R}^{6+2V}$:

$$\mathbf{e}_{k,m} = \left[\frac{\mu_{k,m}}{T},\ \frac{\sigma_{k,m}}{T},\ \widetilde{E}_{k,m},\ \widetilde{P}_{k,m},\ \text{gate}_{k,m},\ \widetilde{G}_{k,m},\ \mathbf{W}[k, :],\ \mathbf{o}_{k,m}\right] \tag{20}$$

The observation profile $\mathbf{o}_{k,m} \in \mathbb{R}^V$ measures what fraction of each variable's observations fall within the island's temporal window:

$$o_{k,m}[v] = \frac{\sum_t g_{k,m}(t) \cdot O[b, t, v]}{\sum_t g_{k,m}(t)} \tag{21}$$

Table 8: Clinical ontology mapping: 10 channels with all constituent variables (V=135). Counts include shared variables. Variables marked with * are assigned to two channels. Variables marked with † are absent in MIMIC-IV.

Ch	Name	Variables
0	Respiratory (12)	O_2 Saturation, RR, FiO_2, Intubated, Chest Tube, PO_2^*, PCO_2^*, pH Blood*, Total CO_2^*, Base Excess*, Propofol*, Fentanyl*
1	Cardiovascular (27)	MBP, HR, SBP, DBP, Vasopressin, Neosynephrine, Nitroglycerine, Amiodarone, Norepinephrine, Epinephrine, Milrinone, Nitroprusside, Troponin T†, Troponin I, NTproBNP, Hydralazine, Metoprolol, CRR, Lactate*, Dopamine*, Furosemide*, Heparin*, Normal Saline*, Lactated Ringers*, Colloid*, OR/PACU Crystalloid*, Weight*
2	Coagulation (7)	PTT, Fibrinogen, Heparin*, Platelet Count*, Fresh Frozen Plasma*, INR*, PT*
3	Liver (13)	Bilirubin (Total), Bilirubin (Direct), Bilirubin (Indirect), ALP, ALT, AST, Albumin, Ammonia, LDH*, Albumin 25%*, Albumin 5%*, INR*, PT*
4	Renal (16)	Creatinine Blood, Creatinine Urine, BUN, Urine, SG Urine, pH Urine, Ultrafiltrate†, Furosemide*, Dopamine*, Normal Saline*, Lactated Ringers*, Colloid*, OR/PACU Crystalloid*, Weight*, Albumin 25%*, Albumin 5%*
5	Neurological (9)	GCS_eye, GCS_verbal, GCS_motor, Midazolam, Lorazepam, Hydromorphone, Morphine Sulfate, Propofol*, Fentanyl*
6	Metabolic (27)	Anion Gap, Glucose (Serum), Glucose (Whole Blood), Glucose (Blood), Sodium, Calcium Total, Chloride, Potassium, Bicarbonate, Phosphate, Magnesium, Calcium Free, Calcium Gluconate, Magnesium Sulphate, Magnesium Sulfate (Bolus), KCl, KCl (Bolus), Insulin Regular, Insulin Humalog, Insulin Glargine, Insulin NPH, Lactate*, PO_2^*, PCO_2^*, pH Blood*, Total CO_2^*, Base Excess*
7	Hematologic (12)	Hgb, Hct, RBC, MCV, MCH, MCHC, RDW, Packed RBC, WBC*, Platelet Count*, Fresh Frozen Plasma*, LDH*
8	Infection (12)	Temperature, Monocytes, Neutrophils, Basophils, Lymphocytes, Lymphocytes (Absolute), Eosinophils, Cefazolin, Vancomycin, Piperacillin, CRP, WBC*
9	Other (24)	Unknown†, Solution, D5W, Dextrose Other, PO Intake, Gastric, Gastric Meds, GT Flush, Sterile Water, Stool, Jackson-Pratt, Fiber, TPN, Half Normal Saline, Pantoprazole, Free Water, Pre-admission Intake, Piggyback, Famotidine, Pre-admission Output, Height, Emesis, Residual, EBL

This allows the selector to distinguish islands with dense observations from those covering sparsely observed regions.

F Clinical ontology mapping

Table 8 lists the 10 clinical channels and their constituent variables for MIMIC-III (V=135). MIMIC-IV uses V=132 (three CareVue-era variables absent: Troponin I, Ultrafiltrate, Unknown); channel assignments are otherwise identical.

Of the V=135 variables, 24 are assigned to two channels (e.g., blood gases $\rightarrow$ Respiratory + Metabolic; IV fluids $\rightarrow$ Cardiovascular + Renal). The hard mask $\mathbf{M}_{\text{ch}} \in \{0,1\}^{C\times V}$ used for channel purity and decoder masking reflects these shared assignments.

G Dictionary quality

G.1 Reconstruction and island stability

TI-SAE achieves cosine similarity $\cos_H$ of 0.91 (Discrete), 0.83 (Disc.+PT), 0.88 (HyMaTE), 0.87 (BAT), and 0.87 (STraTS) between reconstructed and original backbone representations. Island identity stability under 5% event dropout yields Jaccard similarity 0.837–0.903 across backbones (vs. random baseline $\sim$0.01).

G.2 Latent utilization

Table 9 reports latent utilization for the mortality prediction task on both cohorts. A dead latent here is defined as a latent that never selected across the full evaluation cohort.

Table 9: Latent utilization across backbones (K=1280, mortality task). "Dead" latents are never selected across the full cohort. "$>1\%$" counts latents selected in more than 1% of episodes.

	MIMIC-III (N=28,626)			MIMIC-IV (N=46,140)		
Backbone	Dead	Dead%	$>1\%$	Dead	Dead%	$>1\%$
Discrete	26	2.0%	471	40	3.1%	554
Discrete+PT	54	4.2%	524	21	1.6%	537
HyMaTE	104	8.1%	413	36	2.8%	458
BAT	154	12.0%	381	251	19.6%	365
STraTS	295	23.0%	467	311	24.3%	404

G.3 Dictionary overview

Figure 4 summarizes the 479 latents with activation frequency $>1\%$ from the primary dictionary (Discrete backbone, MIMIC-III mortality).

High-risk latents capture clinically coherent critical-care states (vasopressor dependence, metabolic acidosis, acute kidney injury), while protective latents reflect hemodynamic stability and normal laboratory values.

Concept separability and groupings. We measure concept separability by how correlated two concepts' activations are across patients (decoder cosines are uninformative because the sparse atoms are near-orthogonal by construction). The mean absolute pairwise activation correlation is .125, and only 1.4% of concept pairs exceed .5, indicating that concepts fire in distinct contexts rather than duplicating one another. The dictionary is fine-grained rather than a few coarse clusters (k-means silhouette .073). There is a mild organ-level tendency (concepts within the same organ channel have more similar variable composition, .107 vs. .008 across organs), but this partly follows from the channel-purity regularizer.

Statistical significance of outcome association. We tested each active concept for association with the outcome (active vs. inactive mortality odds ratio with a 95% confidence interval; significant if the interval excludes 1). Of 479 concepts, 385 (80%) are statistically significant: 223 high-risk (OR > 1) and 162 protective (OR < 1). The remaining 20% we report as non-significant.

G.4 Exemplar latent concepts

Figures 5 and 6 present exemplar latents from the high-risk and protective groups of the dictionary. Each panel shows 4 episodes per latent with the island temporal window ($\mu \pm \sigma$, top track) and raw variable traces on clinical y-axes.

Mortality-predictive concepts (Figure 5). (a) **Vasopressor-Dependent Hyperlactatemia** (OR=6.63): abnormal capillary refill (CRR=1) co-occurs with elevated lactate and vasopressor (neosynephrine) administration, capturing a circulatory shock state with impaired tissue perfusion. (b) **Severe Metabolic Acidosis** (OR=5.50): persistently low total CO_2 and deeply negative base excess capture progressive acid–base derangement. The third variable (Insulin NPH) may reflect statistical co-occurrence in diabetic patients, rather than a mechanistic relationship. (c) **Acute Kidney Injury with Oliguria** (OR=5.09): rising BUN and creatinine with declining urine output form a coherent renal failure trajectory across all four episodes. (d) **Mechanical Ventilation with Respiratory Failure** (OR=4.85): sustained intubation with monitored oxygen saturation and total CO_2, capturing the ventilated respiratory failure state.

Protective concepts (Figure 6). (a) **Stable Acid–Base Balance** (OR=0.14): total CO_2 and base excess remain within normal ranges—a direct contrast to the acidosis concept above. (b) **Sustained Hemodynamic Stability** (OR=0.14): dense SBP and DBP readings in normal ranges across the full 48 hours. The third decoder variable (NTproBNP, w=0.18) has zero observations in all selected episodes, illustrating sparse variables can contribute decoder weight without contributing observations. (c) **Effective Mechanical Ventilation with Adequate Oxygenation** (OR=0.16): low FiO_2 (≤ 0.5) with adequate O_2 saturation and non-acidotic base excess, indicating controlled ventilation without

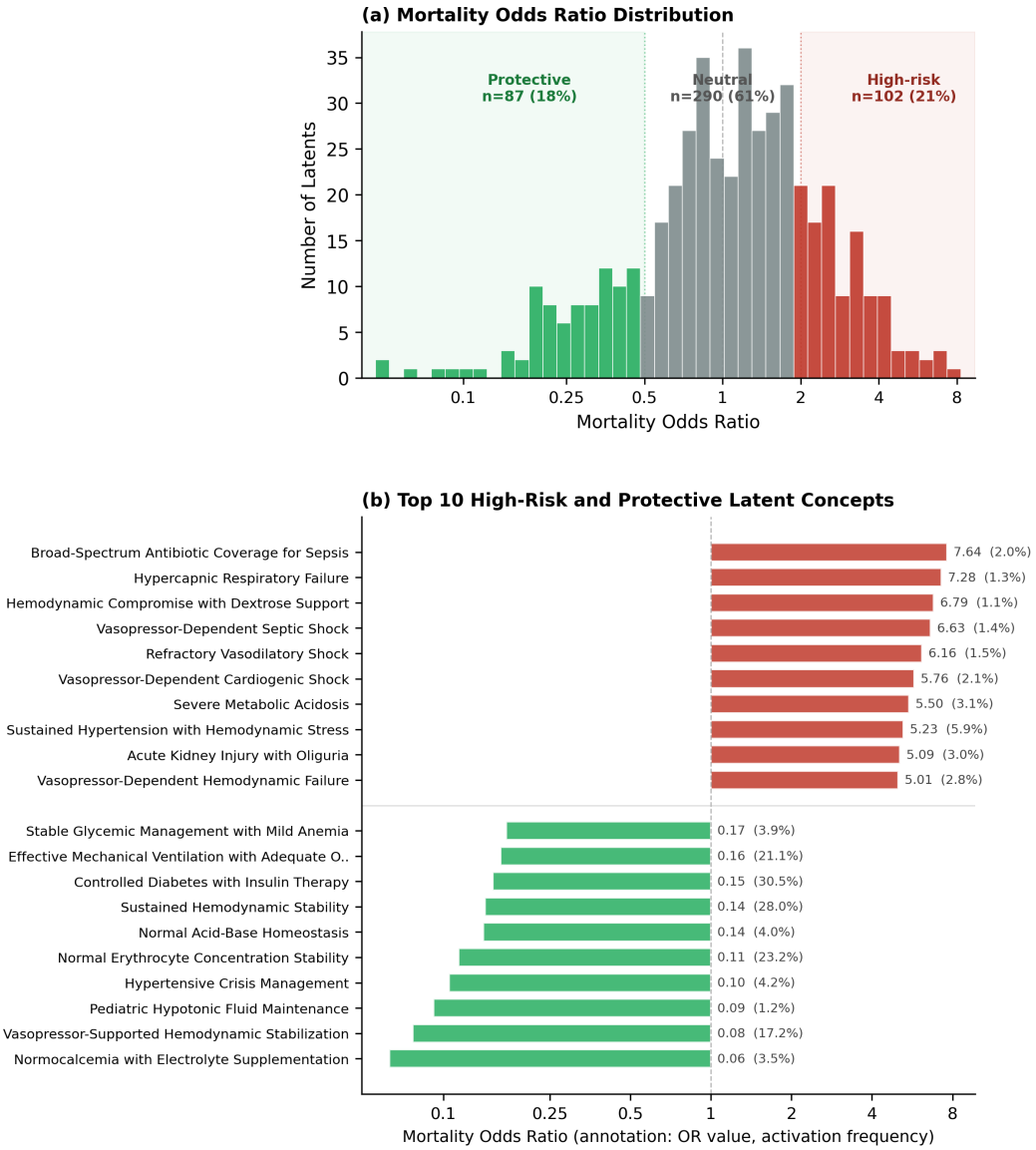


Figure 4: Dictionary overview (Discrete backbone, MIMIC-III mortality). (a) Mortality odds ratio distribution colored by risk tier: high-risk (OR $>$ 2, red), neutral (grey), protective (OR $<$ 0.5, green). (b) Top 10 high-risk and protective latent concepts with LLM-generated clinical labels. Annotations show OR and activation frequency.

deterioration. (d) **Sustained Normal Temperature** (OR=0.19): temperature remains in the normal range ($\sim$97–99°F) throughout the stay. Cefazolin and eosinophils are extremely sparse, making this latent effectively a single-variable detector.

Structural observations. These exemplars illustrate three properties of unsupervised concept discovery that merit discussion: (1) *Spurious variable grouping*: decoder weights capture statistical co-occurrence, sometimes pairing clinically unrelated variables (e.g., Insulin NPH in the acidosis concept). Whether such groupings reflect genuine latent confounders (e.g., diabetic patients presenting with more severe acidosis) or artifacts of the training distribution is unknown. (2) *Observation-frequency imbalance*: when constituent variables differ substantially in measurement frequency, the concept's observable behavior is dominated by the densely sampled variable (e.g., temperature in concept (d)), even though the decoder assigns meaningful weight to rarer variables. The learned

composition remains multi-variable, but its empirical grounding becomes uneven. (3) *Latent-space variables without direct observations*: a decoder variable can carry substantial weight yet have no raw measurements in a given episode (e.g., NTproBNP in concept (b)). Because the SAE operates on the backbone's contextual representations—which encode cross-variable dependencies even for unmeasured variables—this may reflect genuine imputed signal rather than a vacuous weight. However, such contributions are not directly verifiable from the clinical record, complicating interpretability.

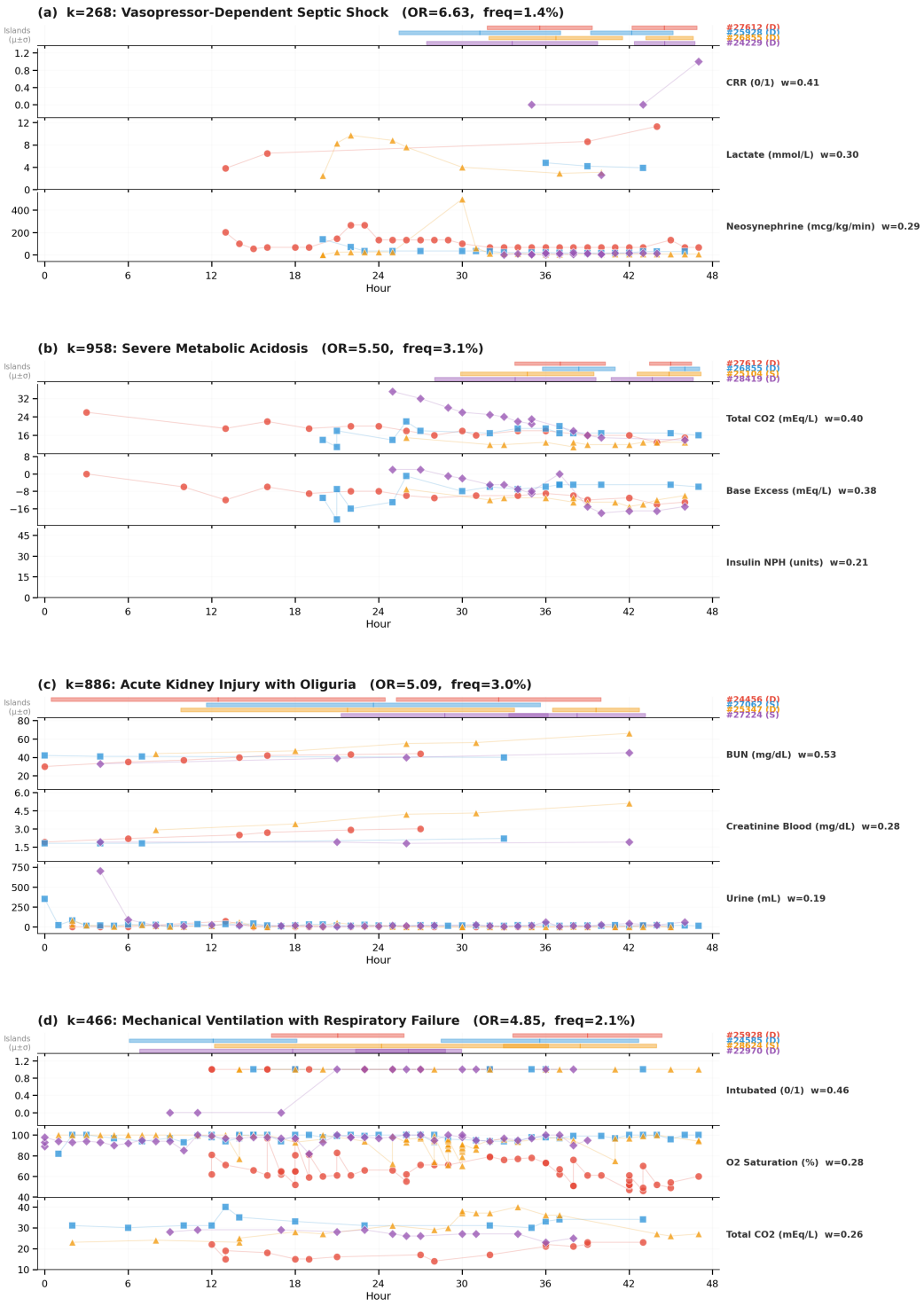


Figure 5: Each row: island temporal window (top, colored bars = $\mu \pm \sigma$ per episode) and raw variable traces (bottom) on clinical y-axes. (D) = died, (S) = survived.

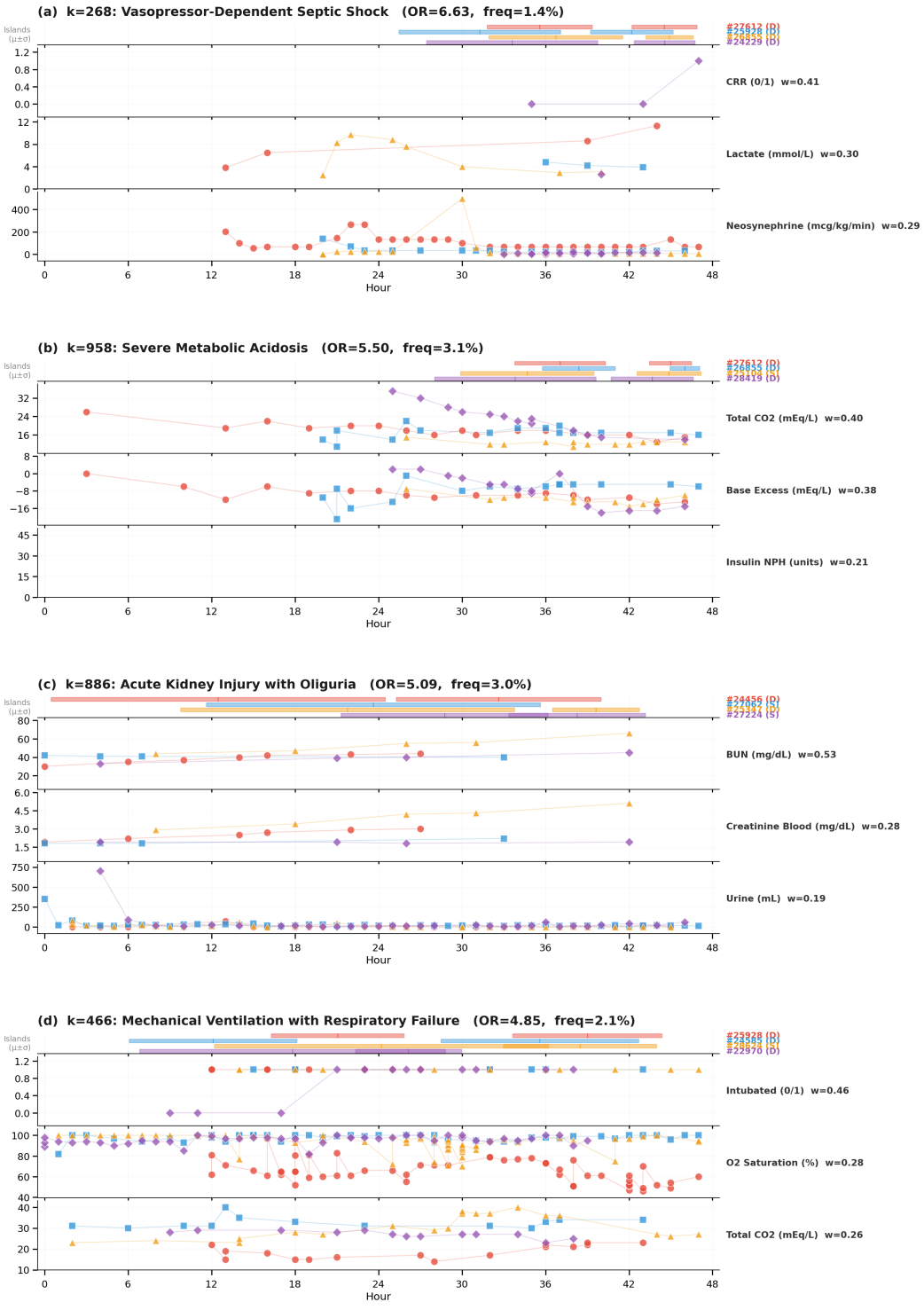


Figure 6: Protective exemplar latents. Same format as Figure 5.

G.5 Clinician evaluation of dictionary labels

To validate that LLM-generated dictionary labels are clinically meaningful, four independent tertiary-hospital clinicians (all $\geq$10 years of experience) rated a stratified random sample of N=30 labels (Discrete backbone, MIMIC-III mortality) on a 5-point Likert scale (1: completely wrong; 2: partially related but misleading; 3: vaguely correct but imprecise; 4: mostly accurate, minor issues; 5: accurate and clinically meaningful). Each rater saw the latent's top-3 decoder variables, mortality odds ratio, and up to 10 exemplar episodes with raw variable traces — the same information shown to the LLM labeler.

Results. Mean ratings were 4.20 (rater 1), 4.07 (rater 2), 4.17 (rater 3), and 4.40 (rater 4); 16 of 30 labels (53%) were rated ≥ 4 by all three raters, 4 of 30 (20%) were rated 5 unanimously, and no label received a rating of 1. Inter-rater agreement on the ordinal scale was fair (quadratic-weighted Cohen's $\kappa = 0.11$–0.47, Pearson $r = 0.12$–0.51, mean absolute rating difference 0.77 points).

Table 10: Per-rater Likert score distributions (N=30 labels).

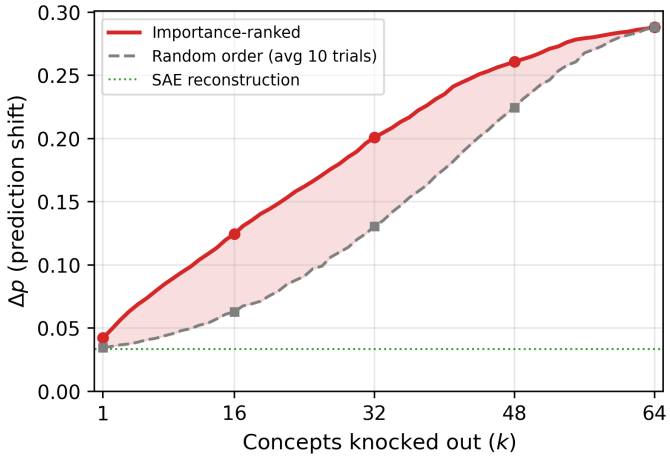
Rater	1	2	3	4	5	Mean
Rater 1	0	5	2	6	17	4.17
Rater 2	0	1	8	9	12	4.07
Rater 3	0	2	5	8	15	4.20
Rater 4	0	0	3	12	15	4.40

Of the six labels rated ≤ 2 by at least one rater, the most common disagreements were: (i) labels naming a specific etiology (e.g., *Cardiogenic Shock*) where decoder variables alone could not distinguish among shock subtypes; (ii) labels for protective concepts where exemplar episodes showed mixed trajectories.

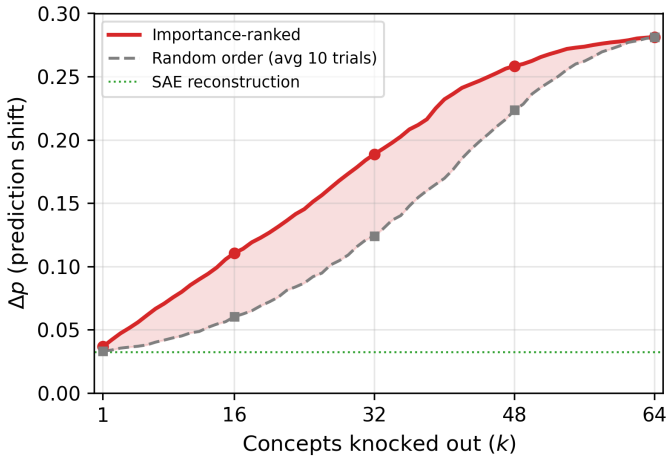
G.6 Concept ablation

We zero-ablate the top-k importance-ranked concepts (by prediction selector score) and measure the prediction shift $\Delta p = |\sigma(\ell) - \sigma(\ell_{\text{ko}})|$, where ℓ is the full backbone prediction and ℓ_{ko} is the prediction after reconstructing from the remaining concepts through the SAE decoder and backbone prediction head. As a control, we repeat the experiment with k concepts removed in random order (averaged over 10 trials).

Figure 7a shows monotonically increasing Δp as more concepts are removed. Importance-ranked removal consistently exceeds random-order removal across all k, with the largest gap at k=26 ($\Delta p = 0.170$ vs. 0.099). The two curves converge at k=64 (all selected concepts removed), showing that the gap reflects ranking quality.



(a) In-distribution (MIMIC-III)



(b) Out-of-distribution (MIMIC-IV)

Figure 7: Concept ablation. Prediction shift Δp when zero-ablating the top-k concepts from the selected set. Importance-ranked removal (red) degrades predictions more than random-order removal (grey) at all k in both settings. Dashed green: SAE reconstruction baseline (no concepts removed). Discrete backbone, mortality task, N=500.

G.7 Concept seed reproducibility

We assess whether TI-SAE's learned concepts are reproducible across random seeds by training three models (seeds 42, 43, 44) on the same backbone (Discrete, MIMIC-III mortality) with identical hyperparameters.

In an overcomplete dictionary (K=1280), multiple latents may share similar decoder weights but differ in temporal behavior. Matching by decoder similarity can pair temporally unrelated latents, therefore, we use two-stage alignment: optimal (Hungarian) assignment on decoder cosine similarity, refined by selecting the best temporal trace correlation among structurally similar candidates ($\cos > 0.7$). The random baseline applies the same temporal search with the same per-latent candidate counts, but draws candidates at random instead of by decoder similarity.

We evaluate five metrics on matched pairs: *Decoder cosine* measures the cosine similarity of matched decoder weight vectors $\mathbf{w}_k$. *Variable overlap* reports the fraction of pairs sharing at least two of three top decoder variables (by $|\mathbf{w}_k|$). *Latent energy correlation* is the Pearson correlation of per-latent activation energy across test samples, measuring whether matched concepts activate on the same patients. *Latent temporal cosine* is the mean cosine similarity of per-sample activation traces $\hat{A}[k,:]$ (T=48 time steps), measuring whether matched concepts fire at the same times on the same patients. *Prediction selector NDCG* is binary NDCG@64: treating one seed's top-64 prediction selector-ranked latents as relevant, it measures how well the other seed's ranking recovers them.

Table 11: Concept seed reproducibility (K=1280, Discrete backbone, MIMIC-III mortality). Mean $\pm$ std over 3 seed pairs.

Metric	Actual	Random
Decoder cosine	.724±.003	.034
Variable overlap ($\geq 2/3$)	77±1%	1.2%
Latent energy correlation	.408±.006	.093
Latent temporal cosine	.724±.003	.382
Prediction selector NDCG	.401±.016	.074

Structural reproducibility is strong: 77% of matched latent pairs share at least two of three decoder variables (vs. 1.2% random), with mean decoder cosine similarity of .724. Temporal firing patterns are moderately reproducible, indicating that matched concepts activate at similar time steps on the same patients. The prediction selector preserves importance ranking across seeds, though competitive top-k selection introduces variability.

Per-stage reproducibility. The analysis above retrains both stages per seed, so dictionary and selector variabilities are mixed. Here we separate the stages: (i) retraining only the dictionary across five seeds (42–46) on the fixed backbone; and (ii) fixing the seed-42 dictionary and retraining only the prediction selector across five seeds, measuring per-patient consistency of the selected concepts against a random-subset null.

Table 12: Per-stage seed reproducibility (K=1280, Discrete backbone, MIMIC-III mortality). Stage A: dictionary retrained, selector excluded. Stage B: selector retrained on the fixed seed-42 dictionary. Stage B metrics are computed per patient and averaged over patients and seed pairs. Mean $\pm$ std over 5 seed pairs.

Stage	Metric	Actual	Random null
A — dictionary	Decoder cosine	$.735 \pm .002$	.029
	Variable overlap ($\geq 2/3$)	$82.4\% \pm 1.1\%$	0.8%
	Latent energy correlation	$.361 \pm .010$	.083
	Latent temporal cosine	$.731 \pm .013$	.398
B — selector (dict. fixed)	Top-64 concept Jaccard	$.708 \pm .017$	.026
	PS-score rank corr. (Spearman)	$.870 \pm .029$	$\sim$0

Every metric is well above its matched null. With the dictionary fixed, the selector's importance ranking is highly consistent across seeds (Spearman .87) and its faithfulness is stable (neighborhood-based AOPC $.151 \pm .008$).

G.8 Cross-dataset generalization

We evaluate whether the dictionary learned on MIMIC-III transfers to out-of-distribution data without retraining. Because MIMIC-III and MIMIC-IV share patients admitted in 2008–2012, we construct a disjoint cohort using MIMIC-IV's `anchor_year_group`, restricting it to 2014–2016 and 2017–2019 bands. The MIMIC-III–trained Discrete backbone, TI-SAE, and prediction selector are applied directly to MIMIC-IV mortality data (132 of 135 variables shared; 3 MIMIC-III–only variables receive no observations). Raw values are re-quantized using MIMIC-III quantile boundaries to match the backbone's learned value embeddings.

Table 13 reports in-distribution and out-of-distribution performance. Backbone prediction is maintained (AUROC .861→.864). Faithfulness drops modestly (AOPC .149→.130), while clinical plausibility is nearly unchanged (AUROC .756→.756). Of 1280 latents, 500 are selected in $\geq$1% of MIMIC-IV episodes (vs. 479 on MIMIC-III), while 440 are never selected—indicating that some MIMIC-III–specific concepts do not appear in the MIMIC-IV population. The concept knockout curve (Figure 7b) retains the same monotonic shape and importance-ranked vs. random-order gap on OOD data (Appendix G.6), showing that individual concepts remain causally meaningful in a new patient population. While the disjoint cohort eliminates patient and temporal overlap, both datasets originate from the same institution, so this evaluation captures temporal and population shift rather than fully external, cross-institutional generalization.

Table 13: Cross-dataset generalization. MIMIC-III–trained Discrete backbone evaluated on disjoint MIMIC-IV mortality (N=500 for faithfulness and plausibility; prediction AUROC on the full sets). Neighborhood-based perturbation strategy. Baselines recomputed on MIMIC-IV.

	In-distribution (MIMIC-III)	Out-of-distribution (MIMIC-IV)
Backbone		
Prediction AUROC	.861	.864
Faithfulness (AOPC ↑, neighborhood-based)		
TI-SAE (soft)	.149	.130
TI-SAE (hard)	.134	.121
Attention	.100	.097
IG	.086	.045
TSR	.088	.046
DynaMask	.028	.021
TimeSHAP	.009	.012
WinIT	.035	.028
Random	.001	.005
Clinical Plausibility (AUROC / AUPRC ↑)		
TI-SAE (soft)	.756 / .386	.756 / .370
TI-SAE (hard)	.639 / .315	.652 / .306
Attention	.616 / .259	.631 / .261
IG	.567 / .255	.552 / .224
TSR	.593 / .280	.587 / .245
DynaMask	.507 / .213	.505 / .201
TimeSHAP	.521 / .212	.516 / .203
WinIT	.601 / .265	.601 / .253
Random	.502 / .193	.503 / .186
Dictionary		
$\geq$1%	479	500
Dead	26	440

G.9 Cross-backbone concept reuse

We investigate whether TI-SAE discovers similar latent concepts across different backbones by comparing the Discrete backbone's dictionary against each of the four remaining backbones (Disc.+PT, HyMaTE, STraTS, BAT), all trained on the MIMIC-III mortality task. We apply the same two-stage matching used for seed reproducibility

Table 14 reports five metrics for each backbone pair (Discrete vs. target). TI-SAE consistently discovers similar variable-level groupings across backbones. However, latent energy correlation, temporal trace cosine, and prediction selector agreement are substantially lower, indicating that the temporal concepts identified by models and predictive importance of matched concepts diverge across backbones.

Table 14: Cross-backbone concept reuse (reference: Discrete, MIMIC-III mortality)

Backbone	Dec. cos	Var $\geq$2/3	Energy corr	Temp. cos	PS NDCG
Disc.+PT	.801	88.2%	.230	.674	.163
HyMaTE	.731	77.5%	.246	.592	.155
BAT	.716	82.0%	.228	.583	.089
STraTS	.712	82.0%	.108	.596	.099
Random	.029	0.7%	.051	.390	.055

Disc.+PT shows the strongest agreement (which shares same architecture). The results suggest that the decoder's variable groupings are largely data-driven, but the temporal pattern grouping and functional importance over those groupings are shaped by each backbone's learned representation.

H Extended evaluation results

H.1 Bootstrap standard deviations and alternative perturbation strategies

Table 15 reproduces the faithfulness results from Table 2 with bootstrap standard deviations (2000 resamples over tasks × datasets × samples). Table 16 reports the same for plausibility AUROC and AUPRC. Tables 17 and 18 present AOPC under removal and marginal perturbation, averaged across tasks and datasets per backbone.

Table 15: AOPC under **neighborhood-based** perturbation (all methods), averaged across tasks and datasets. Best per column in **bold**, second-best underlined. ± bootstrap standard deviation (2000 within-condition resamples).

	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	**.086**±.003	**.076**±.003	**.097**±.004	**.043**±.002	.016±.001
TI-SAE (hard)	.079±.003	.068±.002	.090±.003	.041±.002	.016±.001
Attention	.053±.002	.019±.002	.051±.003	–	.016±.001
IG	.060±.003	.051±.002	.049±.003	.031±.003	.067±.003
TSR	.063±.003	.054±.003	.062±.003	.037±.003	.071±.003
DynaMask	.019±.002	.032±.002	.012±.001	.028±.002	.034±.002
TimeSHAP	.009±.001	.011±.001	.012±.001	.016±.002	.029±.002
FIT	.011±.002	.011±.001	.022±.002	.012±.001	.021±.002
WinIT	.020±.002	.030±.002	.032±.002	.021±.002	.032±.002
Cluster-TCAV	.046±.002	.026±.001	.031±.002	.023±.001	.022±.001
Random	.002±.001	.003±.000	.002±.001	.006±.001	.007±.001

Table 16: Plausibility with bootstrap standard deviations, averaged across tasks and datasets. Best per column in **bold**, second-best underlined. ± bootstrap standard deviation (2000 within-condition resamples).

	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
AUROC ↑					
TI-SAE (soft)	**.593**±.002	.589±.002	.613±.002	.588±.002	.536±.002
TI-SAE (hard)	.569±.002	.568±.002	.580±.002	.567±.002	.535±.002
Attention	.570±.002	.539±.002	.554±.002	–	.552±.002
IG	.545±.002	.548±.002	.596±.002	.583±.002	.576±.002
TSR	.566±.002	.573±.002	**.625**±.002	**.608**±.002	.594±.002
DynaMask	.510±.001	.504±.001	.494±.002	.493±.001	.497±.002
TimeSHAP	.512±.001	.517±.001	.516±.001	.519±.001	.513±.001
FIT	.511±.001	.537±.002	.528±.002	.494±.001	.551±.002
WinIT	.579±.002	**.611**±.002	.611±.002	.565±.002	**.621**±.002
Cluster-TCAV	.524±.002	.462±.002	.490±.002	.512±.002	.476±.002
Random	.502±.001	.502±.001	.502±.001	.502±.001	.502±.001
AUPRC ↑					
TI-SAE (soft)	**.286**±.003	.255±.003	**.276**±.002	.251±.002	.227±.003
TI-SAE (hard)	.267±.002	.243±.002	.257±.002	.243±.002	.228±.003
Attention	.241±.002	.219±.002	.218±.002	–	.217±.002
IG	.225±.002	.228±.002	.244±.002	.239±.002	.222±.002
TSR	.243±.002	.241±.002	.271±.002	**.260**±.002	.233±.002
DynaMask	.201±.002	.200±.002	.187±.002	.189±.002	.194±.002
TimeSHAP	.202±.002	.206±.002	.205±.002	.208±.002	.201±.002
FIT	.196±.002	.209±.002	.197±.002	.190±.002	.213±.002
WinIT	.251±.002	**.280**±.002	.273±.003	.246±.002	**.279**±.002
Cluster-TCAV	.220±.002	.206±.002	.237±.002	.218±.002	.202±.002
Random	.190±.002	.190±.002	.190±.002	.190±.002	.190±.002

Table 17: AOPC under **removal** perturbation, averaged across tasks and datasets. Best per column in **bold**, second-best underlined. ± bootstrap standard deviation (2000 within-condition resamples).

	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	**.197**±.005	**.071**±.003	.118±.003	.074±.004	.067±.003
TI-SAE (hard)	.172±.004	.066±.003	.102±.003	.081±.003	.053±.002
Attention	.191±.006	.068±.004	**.143**±.004	–	.038±.002
IG	.112±.005	.061±.003	.066±.004	.091±.005	.095±.003
TSR	.112±.005	.060±.004	.068±.003	**.100**±.005	.100±.003
DynaMask	.070±.003	.068±.003	.098±.003	.043±.003	**.252**±.004
TimeSHAP	.038±.002	.019±.001	.033±.002	.037±.002	.036±.002
FIT	.036±.004	.028±.003	.041±.003	.046±.003	.025±.002
WinIT	.036±.004	.027±.003	.053±.002	.064±.004	.044±.002
Cluster-TCAV	.143±.003	.069±.002	.042±.002	.092±.003	.050±.002
Random	.006±.001	.005±.001	.004±.001	.012±.002	.009±.001

Table 18: AOPC under **marginal** perturbation, averaged across tasks and datasets. Best per column in **bold**, second-best underlined. ± bootstrap standard deviation (2000 within-condition resamples).

	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	.092±.003	.071±.003	**.099**±.004	.038±.003	.015±.003
TI-SAE (hard)	.084±.003	.066±.003	.090±.004	.047±.003	.013±.001
Attention	.085±.004	.033±.003	.082±.004	–	.019±.001
IG	**.100**±.004	**.086**±.003	.098±.005	.069±.004	**.106**±.004
TSR	.098±.004	.085±.003	.098±.005	**.071**±.004	.105±.004
DynaMask	.080±.003	.043±.002	.041±.003	.049±.003	.076±.003
TimeSHAP	.058±.003	.047±.003	.063±.004	.053±.004	.087±.004
FIT	.048±.003	.033±.002	.055±.003	.024±.002	.038±.002
WinIT	.052±.003	.040±.003	.058±.004	.031±.003	.041±.002
Cluster-TCAV	.064±.003	.025±.002	.027±.002	.023±.002	.027±.002
Random	.022±.001	.014±.001	.023±.002	.014±.002	.025±.002

H.2 Per-fraction detail

Tables 19–33 present per-fraction comprehensiveness and sufficiency under neighborhood-based perturbation for each backbone (MIMIC-III and MIMIC-IV). Best AOPC per sub-table in **bold**, second-best underlined.

Table 19: Per-fraction faithfulness: **Discrete** backbone (neighborhood, mortality).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.082	.166	.200	.176	.146	**.149**	−.004	.009	.019	.017	.011
TI-SAE (hard)	.074	.149	.168	.160	.150	.134	.007	.021	.028	.014	.009
Attention	.045	.091	.131	.132	.128	.100	.091	.059	.045	.033	.021
IG	.057	.082	.101	.108	.110	.086	.164	.123	.072	.055	.050
TSR	.052	.079	.099	.111	.121	.088	.093	.072	.064	.054	.050
GradSHAP	.058	.080	.099	.109	.108	.086	.168	.124	.073	.053	.049
DynaMask	.015	.021	.029	.035	.049	.028	.108	.079	.061	.038	.034
TimeSHAP	.005	.007	.010	.012	.016	.009	.029	.047	.037	.021	.020
FIT	.003	.008	.015	.028	.037	.015	.099	.062	.035	.017	.011
WinIT	.010	.028	.038	.046	.064	.035	.095	.071	.041	.029	.016
Cluster-TCAV	.042	.074	.082	.097	.111	.074	.065	.078	.068	.044	.024
Random	.000	.000	.001	.001	.005	.001	.056	.040	.027	.019	.012
MIMIC-IV											
TI-SAE (soft)	.086	.116	.112	.110	.107	.102	.048	.043	.042	.038	.029
TI-SAE (hard)	.079	.117	.127	.123	.113	**.107**	.049	.046	.026	.017	.023
Attention	.031	.057	.077	.086	.085	.063	.046	.051	.044	.023	.008
IG	.040	.059	.078	.093	.101	.070	.063	.067	.060	.051	.046
TSR	.044	.068	.085	.103	.114	.077	.033	.033	.037	.036	.036
GradSHAP	.040	.061	.079	.093	.103	.070	.064	.066	.061	.052	.046
DynaMask	.025	.030	.039	.043	.066	.038	.059	.048	.032	.020	.015
TimeSHAP	.010	.012	.013	.012	.016	.012	.025	.046	.028	.024	.019
FIT	.005	.009	.020	.027	.042	.018	.061	.035	.019	.009	.005
WinIT	.018	.030	.035	.047	.063	.036	.067	.038	.025	.018	.006
Cluster-TCAV	.028	.048	.106	.163	.208	.099	.056	.048	.019	.010	.001
Random	.002	.003	.004	.005	.009	.004	.056	.036	.025	.017	.013

Table 20: Per-fraction faithfulness: **Discrete+PT** backbone (neighborhood, mortality).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.033	.080	.109	.129	.137	**.091**	.022	.002	.003	−.001	−.003
TI-SAE (hard)	.032	.071	.093	.110	.123	.081	.022	.008	.004	.003	.015
Attention	.012	.030	.042	.052	.060	.036	.250	.244	.220	.171	.117
IG	.035	.062	.086	.103	.113	.076	.065	.052	.041	.036	.037
TSR	.038	.068	.092	.108	.119	.079	.032	.035	.031	.033	.031
GradSHAP	.035	.063	.088	.103	.112	.075	.069	.054	.039	.034	.037
DynaMask	.019	.023	.041	.045	.061	.035	.069	.053	.031	.015	.008
TimeSHAP	.009	.010	.012	.015	.019	.012	.018	.023	.021	.025	.018
FIT	.005	.009	.012	.014	.020	.011	.077	.049	.030	.019	.012
WinIT	.015	.033	.051	.058	.063	.041	.068	.043	.030	.021	.013
Cluster-TCAV	.014	.019	.021	.022	.038	.020	.121	.115	.097	.093	.084
Random	.002	.003	.002	.004	.009	.003	.051	.026	.018	.008	.006
MIMIC-IV											
TI-SAE (soft)	.042	.102	.144	.153	.138	**.109**	.040	.036	.040	.027	.025
TI-SAE (hard)	.037	.085	.130	.133	.118	.094	.034	.033	.039	.024	.022
Attention	.009	.028	.054	.067	.076	.043	.118	.129	.132	.128	.128
IG	.032	.056	.076	.095	.101	.066	.061	.056	.054	.051	.048
TSR	.032	.058	.080	.092	.103	.068	.038	.036	.039	.040	.039
GradSHAP	.031	.056	.074	.093	.102	.066	.061	.054	.052	.049	.048
DynaMask	.031	.045	.061	.071	.082	.054	.070	.041	.032	.022	.015
TimeSHAP	.004	.005	.007	.005	.008	.006	.037	.063	.043	.035	.027
FIT	.001	.004	.008	.022	.030	.012	.079	.046	.026	.011	.005
WinIT	.011	.021	.031	.042	.051	.028	.081	.059	.038	.025	.020
Cluster-TCAV	.016	.028	.036	.028	.027	.026	.101	.083	.067	.063	.060
Random	.004	.001	.002	.003	.010	.004	.051	.032	.025	.017	.010

Table 21: Per-fraction faithfulness: **HyMaTE** backbone (neighborhood, mortality).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.082	.141	.172	.180	.182	**.143**	.071	.053	.041	.026	.009
TI-SAE (hard)	.065	.124	.156	.170	.165	.127	.048	.040	.057	.051	.036
Attention	.041	.086	.115	.136	.140	.096	.035	.042	.027	.017	.015
IG	.049	.079	.094	.108	.116	.085	.118	.088	.050	.030	.018
TSR	.051	.086	.110	.120	.133	.094	.070	.054	.031	.016	.011
GradSHAP	.051	.076	.095	.110	.124	.087	.123	.087	.051	.030	.018
DynaMask	.005	.011	.013	.033	.027	.016	.048	.044	.026	.021	.010
TimeSHAP	.011	.013	.018	.019	.025	.017	.029	.028	.029	.023	.016
FIT	.006	.023	.037	.052	.065	.033	.076	.055	.041	.030	.018
WinIT	.028	.047	.065	.071	.077	.056	.052	.043	.039	.029	.020
Cluster-TCAV	.024	.055	.075	.086	.090	.062	.146	.140	.107	.076	.064
Random	.001	.002	.001	.005	.010	.003	.064	.027	.013	.006	.005
MIMIC-IV											
TI-SAE (soft)	.048	.123	.150	.156	.155	**.120**	−.007	−.005	.001	.001	.006
TI-SAE (hard)	.041	.081	.124	.141	.151	.100	−.009	−.003	−.002	.002	−.001
Attention	.014	.031	.052	.077	.092	.047	.030	.023	.015	.013	.010
IG	.016	.033	.051	.065	.074	.045	.084	.073	.051	.032	.018
TSR	.027	.051	.065	.081	.084	.058	.041	.037	.021	.018	.013
GradSHAP	.016	.035	.047	.064	.069	.044	.081	.073	.052	.032	.018
DynaMask	.005	.015	.022	.036	.040	.020	.050	.023	.019	.007	.002
TimeSHAP	.003	.004	.002	.004	.010	.004	.026	.042	.038	.032	.025
FIT	−.006	.004	.015	.026	.034	.012	.049	.029	.015	.011	.005
WinIT	.009	.009	.024	.040	.044	.022	.046	.039	.030	.020	.019
Cluster-TCAV	.013	.018	.049	.056	.055	.035	.074	.072	.057	.041	.027
Random	.001	.000	.001	.000	.007	.001	.045	.029	.016	.010	.008

Table 22: Per-fraction faithfulness: **BAT** backbone (neighborhood, mortality).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.022	.052	.076	.091	.109	.065	.065	.067	.064	.067	.052
TI-SAE (hard)	.021	.055	.080	.099	.121	**.070**	.055	.039	.050	.048	.055
Attention	–	–	–	–	–	–	–	–	–	–	–
IG	.023	.055	.064	.075	.089	.057	.095	.079	.056	.031	.024
TSR	.024	.050	.066	.081	.095	.059	.059	.045	.030	.017	.012
GradSHAP	.024	.053	.069	.077	.093	.058	.097	.080	.057	.033	.022
DynaMask	.017	.042	.049	.066	.079	.046	.043	.030	.024	.011	.003
TimeSHAP	.020	.019	.022	.024	.029	.022	.047	.062	.051	.036	.031
FIT	.010	.013	.018	.029	.038	.020	.101	.048	.033	.024	.013
WinIT	.023	.033	.043	.040	.048	.035	.071	.065	.047	.032	.023
Cluster-TCAV	.016	.047	.075	.073	.090	.056	.110	.098	.089	.071	.051
Random	.002	.004	.007	.010	.020	.008	.075	.047	.033	.018	.016
MIMIC-IV											
TI-SAE (soft)	.021	.058	.091	.132	.162	**.085**	.041	.046	.040	.035	.032
TI-SAE (hard)	.031	.055	.081	.110	.143	.076	.039	.043	.040	.032	.020
Attention	–	–	–	–	–	–	–	–	–	–	–
IG	.017	.028	.037	.051	.064	.037	.095	.047	.024	.014	.009
TSR	.019	.031	.043	.062	.076	.042	.065	.031	.017	.011	.010
GradSHAP	.017	.029	.039	.049	.059	.036	.105	.049	.023	.016	.009
DynaMask	.003	.007	.009	.014	.027	.011	.081	.064	.054	.041	.035
TimeSHAP	.019	.020	.023	.024	.028	.022	.070	.105	.071	.055	.045
FIT	.006	.014	.021	.023	.030	.018	.111	.063	.036	.027	.018
WinIT	.017	.024	.031	.033	.044	.028	.100	.074	.057	.043	.032
Cluster-TCAV	.007	.014	.031	.037	.073	.028	.157	.162	.150	.136	.083
Random	.005	.006	.008	.010	.018	.008	.095	.055	.047	.029	.022

Table 23: Per-fraction faithfulness: **STraTS** backbone (neighborhood, mortality).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.004	.007	.023	.024	.025	.015	.089	.080	.070	.060	.044
TI-SAE (hard)	.003	.008	.017	.021	.032	.013	.080	.072	.059	.051	.045
Attention	.004	.008	.008	.010	.020	.010	.089	.098	.090	.083	.073
IG	.037	.058	.073	.084	.084	**.064**	.051	.037	.021	.014	.008
TSR	.034	.058	.071	.079	.082	.061	.023	.021	.012	.007	.004
GradSHAP	.038	.060	.071	.085	.083	.064	.049	.036	.021	.014	.009
DynaMask	.015	.030	.036	.048	.054	.033	.016	.014	.013	.004	.001
TimeSHAP	.014	.014	.017	.017	.019	.016	.034	.052	.042	.033	.022
FIT	.006	.010	.010	.012	.017	.010	.071	.037	.021	.013	.008
WinIT	.011	.024	.030	.031	.043	.025	.066	.042	.032	.024	.016
Cluster-TCAV	−.002	.003	.012	.024	.029	.012	.077	.065	.057	.048	.047
Random	.001	.001	.007	.009	.012	.005	.055	.032	.023	.015	.012
MIMIC-IV											
TI-SAE (soft)	.003	.012	.015	.028	.033	.016	.119	.120	.113	.101	.093
TI-SAE (hard)	.005	.009	.020	.030	.048	.021	.115	.118	.116	.105	.088
Attention	.003	.006	.014	.029	.039	.016	.141	.120	.115	.102	.096
IG	.044	.074	.103	.130	.139	.093	.014	.007	.000	−.001	−.001
TSR	.049	.080	.117	.116	.125	.092	.010	.004	.001	.001	.000
GradSHAP	.045	.075	.104	.130	.138	.093	.015	.007	−.001	−.001	−.001
DynaMask	.098	.172	.229	.274	.319	**.204**	−.047	−.056	−.057	−.057	−.055
TimeSHAP	.027	.032	.032	.034	.039	.031	.032	.058	.052	.035	.031
FIT	.017	.024	.034	.047	.051	.033	.050	.033	.021	.013	.005
WinIT	.031	.044	.056	.066	.080	.052	.058	.047	.035	.026	.013
Cluster-TCAV	.005	.040	.065	.081	.097	.053	.115	.086	.077	.050	.037
Random	.003	.004	.008	.014	.018	.009	.076	.056	.045	.032	.027

Table 24: Per-fraction faithfulness: **Discrete** backbone (neighborhood, length-of-stay).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.034	.040	.036	.046	.062	.041	.064	.046	.037	.030	.020
TI-SAE (hard)	.032	.043	.037	.042	.051	.039	.060	.050	.043	.034	.015
Attention	.011	.036	.050	.051	.055	.037	.047	.025	.016	.015	.013
IG	.021	.038	.057	.071	.086	.050	.016	−.012	−.024	−.023	−.017
TSR	.021	.035	.050	.071	.089	.049	.011	−.007	−.013	−.012	−.011
GradSHAP	.021	.038	.056	.073	.087	**.051**	.016	−.012	−.023	−.023	−.017
DynaMask	.013	.020	.026	.026	.030	.022	.055	.035	.017	.010	.008
TimeSHAP	.012	.013	.013	.012	.014	.012	.004	.009	.007	.005	.004
FIT	.004	.004	.009	.013	.021	.009	.033	.020	.011	.005	.002
WinIT	.005	.007	.013	.013	.022	.011	.058	.039	.023	.014	.011
Cluster-TCAV	.004	.008	.011	.016	.019	.011	.053	.032	.018	.016	.021
Random	.001	.002	.002	.002	.003	.002	.034	.020	.012	.011	.006
MIMIC-IV											
TI-SAE (soft)	.013	.023	.040	.052	.049	.032	.054	.044	.036	.032	.028
TI-SAE (hard)	.011	.021	.040	.059	.058	**.035**	.053	.043	.040	.027	.018
Attention	.003	.011	.015	.019	.034	.015	.034	.018	.010	.007	.004
IG	.010	.018	.031	.042	.054	.029	.048	.033	.019	.011	.008
TSR	.012	.023	.027	.049	.067	.033	.026	.015	.009	.007	.007
GradSHAP	.010	.018	.033	.042	.053	.029	.048	.033	.018	.010	.007
DynaMask	.003	.003	.007	.014	.017	.008	.055	.039	.025	.016	.011
TimeSHAP	.003	.003	.003	.004	.003	.003	.010	.017	.016	.013	.015
FIT	.003	.000	.003	.011	.016	.006	.038	.027	.017	.011	.007
WinIT	.005	.007	.013	.017	.023	.012	.050	.033	.023	.016	.010
Cluster-TCAV	.001	.005	.013	.020	.027	.012	.068	.052	.036	.037	.032
Random	.001	.000	.002	.004	.005	.002	.036	.021	.013	.006	.003

Table 25: Per-fraction faithfulness: **Disc.+PT** backbone (neighborhood, length-of-stay).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.010	.025	.043	.057	.068	**.037**	.031	.019	.013	.006	.001
TI-SAE (hard)	.007	.022	.038	.055	.061	.034	.033	.023	.016	.007	.003
Attention	−.012	−.006	.000	.010	.025	.002	.076	.050	.047	.041	.029
IG	.010	.022	.034	.049	.058	.032	.035	.017	.007	.003	.003
TSR	.011	.028	.036	.049	.061	.034	.024	.012	.007	.005	.004
GradSHAP	.011	.022	.033	.047	.058	.032	.035	.017	.007	.003	.002
DynaMask	.017	.026	.036	.045	.053	.033	.039	.024	.010	.003	−.002
TimeSHAP	.008	.009	.010	.011	.015	.010	.015	.021	.013	.016	.011
FIT	.007	.012	.016	.018	.021	.013	.046	.029	.018	.012	.007
WinIT	.030	.033	.039	.040	.051	<u>.036</u>	.050	.034	.026	.016	.010
Cluster-TCAV	.006	.018	.024	.035	.047	.024	.075	.072	.067	.055	.048
Random	.000	−.001	.001	.001	.005	.001	.039	.029	.021	.016	.009
MIMIC-IV											
TI-SAE (soft)	.039	.060	.069	.060	.062	<u>.056</u>	.018	.010	.004	.007	.005
TI-SAE (hard)	.034	.059	.073	.063	.069	**.057**	.025	.013	.022	.030	.034
Attention	.005	.014	.025	.032	.034	.020	.051	.038	.033	.030	.021
IG	.015	.025	.038	.055	.064	.037	.048	.047	.040	.031	.025
TSR	.018	.035	.051	.062	.070	.045	.029	.028	.023	.021	.018
GradSHAP	.015	.025	.038	.054	.064	.037	.047	.049	.039	.031	.025
DynaMask	.021	.029	.038	.043	.051	.035	.034	.020	.005	−.001	−.001
TimeSHAP	.013	.012	.011	.011	.011	.012	.010	.024	.021	.015	.014
FIT	.006	.008	.013	.019	.025	.013	.035	.019	.010	.005	.004
WinIT	.015	.030	.031	.038	.042	.030	.029	.021	.014	.007	.002
Cluster-TCAV	.004	.011	.019	.034	.018	.017	.064	.064	.064	.047	.054
Random	.001	.002	.003	.004	.006	.003	.040	.024	.015	.009	.006

Table 26: Per-fraction faithfulness: **HyMaTE** backbone (neighborhood, length-of-stay).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.112	.129	.142	.138	.130	<u>.124</u>	−.039	−.029	−.023	−.018	−.008
TI-SAE (hard)	.117	.131	.147	.149	.144	**.130**	−.028	−.030	−.017	.004	.005
Attention	.053	.064	.074	.074	.085	.067	.016	.024	.017	.011	.008
IG	.033	.045	.058	.069	.082	.054	.055	.045	.027	.019	.012
TSR	.065	.089	.096	.107	.112	.089	.034	.024	.013	.010	.007
GradSHAP	.034	.043	.058	.058	.082	.052	.059	.046	.027	.019	.011
DynaMask	.003	.010	.019	.037	.046	.021	.034	.011	.009	.007	.001
TimeSHAP	.007	.006	.008	.010	.011	.008	.020	.022	.021	.016	.015
FIT	.010	.019	.037	.046	.056	.031	.048	.028	.019	.016	.008
WinIT	.055	.063	.076	.073	.069	.066	.043	.020	.014	.007	.006
Cluster-TCAV	.002	.005	.011	.014	.012	.008	.130	.114	.128	.129	.132
Random	.000	.001	.000	.001	.000	.001	.032	.014	.012	.007	.003
MIMIC-IV											
TI-SAE (soft)	.022	.047	.056	.051	.049	**.044**	.007	.006	.005	.005	.005
TI-SAE (hard)	.024	.046	.044	.048	.057	<u>.043</u>	.010	.013	.014	.022	.031
Attention	.012	.019	.026	.034	.044	.026	.014	.015	.009	.004	.002
IG	.011	.013	.019	.024	.035	.018	.075	.055	.036	.024	.017
TSR	.011	.021	.031	.044	.052	.030	.051	.037	.023	.015	.010
GradSHAP	.010	.012	.019	.022	.034	.017	.076	.056	.036	.024	.018
DynaMask	−.002	.006	.017	.022	.042	.015	.029	.011	.003	−.001	.001
TimeSHAP	.007	.007	.008	.008	.010	.008	.016	.034	.026	.019	.012
FIT	.006	.010	.013	.018	.027	.014	.028	.011	.005	.003	.004
WinIT	.001	.007	.014	.022	.022	.013	.034	.023	.011	.004	.002
Cluster-TCAV	.000	.002	.009	.017	.029	.009	.060	.056	.050	.054	.050
Random	.001	.001	.000	.001	.001	.001	.045	.026	.019	.009	.008

Table 27: Per-fraction faithfulness: **BAT** backbone (neighborhood, length-of-stay).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.011	.017	.024	.028	.038	.022	.040	.026	.019	.010	.002
TI-SAE (hard)	.010	.017	.025	.025	.034	.020	.036	.028	.018	.011	.003
Attention	–	–	–	–	–	–	–	–	–	–	–
IG	.016	.023	.029	.032	.036	.025	.037	.023	.016	.011	.008
TSR	.017	.034	.037	.043	.047	**.034**	.019	.013	.009	.005	.004
GradSHAP	.014	.025	.034	.037	.042	.028	.034	.024	.016	.011	.009
DynaMask	.008	.011	.019	.034	.042	.021	.014	.010	.008	.004	.002
TimeSHAP	.009	.008	.011	.009	.014	.010	.015	.024	.024	.019	.019
FIT	.007	.009	.012	.012	.020	.011	.026	.009	.008	.007	.006
WinIT	.002	.005	.011	.018	.021	.010	.021	.013	.010	.010	.007
Cluster-TCAV	.001	.006	.009	.015	.019	.009	.046	.031	.023	.015	.014
Random	.001	.003	.002	.005	.011	.004	.022	.019	.014	.007	.007
MIMIC-IV											
TI-SAE (soft)	.014	.019	.037	.057	.070	**.035**	.036	.015	.010	.014	.014
TI-SAE (hard)	.011	.019	.032	.044	.058	.030	.030	.020	.016	.006	.014
Attention	–	–	–	–	–	–	–	–	–	–	–
IG	.013	.023	.033	.041	.043	.028	.062	.059	.045	.030	.018
TSR	.018	.030	.039	.046	.057	.035	.039	.031	.022	.016	.012
GradSHAP	.014	.022	.033	.036	.042	.027	.064	.060	.047	.030	.019
DynaMask	.003	.009	.013	.015	.020	.011	.029	.025	.020	.014	.012
TimeSHAP	.017	.017	.015	.017	.019	.017	.025	.038	.037	.029	.022
FIT	.002	.008	.011	.016	.020	.010	.041	.028	.013	.008	.005
WinIT	.011	.013	.019	.026	.030	.018	.029	.029	.016	.009	.003
Cluster-TCAV	.005	.010	.015	.021	.022	.014	.072	.080	.062	.037	.026
Random	−.001	.002	.005	.007	.008	.004	.037	.022	.016	.013	.009

Table 28: Per-fraction faithfulness: **STraTS** backbone (neighborhood, length-of-stay).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.014	.025	.026	.021	.032	.023	.135	.143	.131	.127	.108
TI-SAE (hard)	.010	.021	.032	.028	.027	.023	.131	.134	.121	.122	.110
Attention	.012	.012	.015	.029	.040	.020	.120	.106	.094	.080	.061
IG	.044	.052	.058	.073	.085	.060	.022	.018	.011	.006	.002
TSR	.048	.064	.084	.106	.123	**.080**	.013	.007	.005	.003	.002
GradSHAP	.043	.051	.058	.077	.090	.061	.023	.018	.011	.006	.002
DynaMask	.019	.040	.051	.074	.083	.048	.022	.009	.007	.008	.005
TimeSHAP	.030	.031	.032	.031	.033	.031	.024	.038	.027	.021	.017
FIT	.009	.015	.018	.024	.033	.018	.058	.026	.019	.010	.006
WinIT	.025	.033	.039	.047	.058	.037	.037	.027	.020	.015	.011
Cluster-TCAV	.002	.005	.012	.031	.049	.017	.122	.097	.067	.042	.039
Random	.001	.003	.008	.010	.012	.006	.068	.039	.030	.018	.009
MIMIC-IV											
TI-SAE (soft)	.010	.016	.024	.031	.040	.022	.117	.098	.080	.081	.076
TI-SAE (hard)	.003	.017	.021	.032	.038	.022	.104	.102	.080	.076	.065
Attention	.006	.026	.037	.044	.052	.029	.124	.114	.101	.095	.074
IG	.037	.049	.057	.069	.082	.056	.028	.020	.011	.003	.002
TSR	.042	.055	.063	.071	.083	**.059**	.022	.005	.004	.002	.001
GradSHAP	.037	.051	.059	.071	.081	.057	.027	.022	.010	.003	.002
DynaMask	.027	.050	.054	.071	.080	.051	.038	.020	.009	.007	.005
TimeSHAP	.031	.033	.032	.034	.035	.033	.026	.063	.051	.043	.031
FIT	.020	.028	.035	.038	.044	.032	.068	.039	.021	.012	.008
WinIT	.026	.035	.045	.059	.069	.044	.061	.044	.027	.020	.011
Cluster-TCAV	.007	.012	.022	.036	.050	.023	.125	.117	.109	.121	.100
Random	.003	.008	.012	.015	.018	.010	.082	.055	.046	.028	.025

Table 29: Per-fraction faithfulness: **Discrete** backbone (neighborhood, readmission).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.066	.112	.126	.127	.128	**.106**	−.063	−.049	−.041	−.029	−.022
TI-SAE (hard)	.054	.090	.102	.110	.110	.088	−.050	−.037	−.019	−.006	−.006
Attention	.025	.051	.064	.078	.088	.056	.003	.010	.010	.008	.004
IG	.037	.054	.058	.059	.060	.051	.007	.017	.018	.015	.010
TSR	.035	.052	.058	.060	.064	.051	.008	.009	.005	.005	.004
GradSHAP	.037	.054	.058	.059	.060	.051	.007	.017	.018	.015	.010
DynaMask	.010	.014	.015	.018	.023	.016	.018	.011	.010	.009	.007
TimeSHAP	.008	.010	.010	.012	.016	.011	−.010	−.008	−.008	−.009	−.008
FIT	−.005	−.001	.002	.006	.011	.002	−.001	.002	.008	.008	.005
WinIT	.006	.010	.014	.017	.023	.013	.011	.018	.019	.016	.013
Cluster-TCAV	.029	.035	.046	.055	.071	.043	−.004	.014	.012	.005	.008
Random	.000	.000	.001	.001	.001	.001	.014	.008	.003	.004	.003
MIMIC-IV											
TI-SAE (soft)	.039	.074	.100	.120	.122	**.085**	−.001	−.008	−.003	.001	.007
TI-SAE (hard)	.035	.058	.080	.093	.100	.068	.009	.003	.010	.006	.003
Attention	.025	.041	.050	.056	.063	.044	.057	.044	.033	.023	.016
IG	.050	.071	.082	.091	.100	.074	.020	.004	−.005	−.006	−.004
TSR	.050	.074	.086	.095	.101	.077	.017	.003	−.002	−.003	−.004
GradSHAP	.050	.071	.082	.091	.099	.074	.020	.003	−.005	−.006	−.004
DynaMask	.001	.004	.004	.008	.010	.005	.047	.036	.032	.021	.010
TimeSHAP	.006	.008	.009	.011	.014	.009	.008	.008	.006	.001	.001
FIT	.003	.012	.016	.025	.029	.015	.020	.012	.006	.004	.005
WinIT	.003	.011	.013	.024	.028	.014	.040	.021	.017	.013	.008
Cluster-TCAV	.014	.030	.040	.055	.065	.038	.028	.003	.005	.002	−.002
Random	.001	.001	.001	.003	.004	.002	.034	.017	.009	.006	.002

Table 30: Per-fraction faithfulness: **Disc.+PT** backbone (neighborhood, readmission).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.050	.095	.117	.123	.129	**.097**	.008	−.011	−.015	−.009	−.006
TI-SAE (hard)	.044	.083	.099	.111	.116	.086	.008	.001	.001	.004	.020
Attention	.003	.005	.003	.007	.010	.005	.119	.103	.096	.090	.083
IG	.049	.062	.071	.081	.090	.068	.039	.034	.031	.026	.023
TSR	.050	.065	.079	.093	.101	.074	.025	.021	.014	.015	.014
GradSHAP	.049	.062	.070	.080	.090	.067	.040	.035	.032	.026	.023
DynaMask	.020	.031	.031	.035	.040	.030	.044	.024	.021	.017	.012
TimeSHAP	.018	.021	.028	.035	.039	.027	.003	.010	.007	.006	.005
FIT	.006	.014	.018	.025	.027	.017	.035	.033	.019	.011	.008
WinIT	.020	.035	.044	.053	.060	.040	.040	.024	.019	.010	.006
Cluster-TCAV	.017	.047	.049	.062	.068	.045	.065	.045	.036	.031	.025
Random	.000	.001	.003	.004	.007	.003	.048	.027	.015	.009	.008
MIMIC-IV											
TI-SAE (soft)	.020	.047	.078	.096	.096	**.063**	.022	.016	.014	.011	.009
TI-SAE (hard)	.015	.043	.071	.093	.094	.058	.014	.011	.010	.009	.008
Attention	.003	.003	.009	.018	.028	.010	.061	.055	.047	.045	.037
IG	.011	.020	.027	.038	.047	.026	.036	.041	.038	.035	.030
TSR	.008	.015	.027	.035	.051	.024	.029	.028	.028	.027	.026
GradSHAP	.010	.020	.028	.039	.047	.026	.035	.042	.038	.035	.030
DynaMask	.000	.004	.003	.004	.005	.003	.052	.035	.027	.022	.013
TimeSHAP	−.002	−.002	−.002	.000	.002	−.001	.016	.023	.021	.017	.014
FIT	.000	.001	.002	.004	.004	.002	.031	.023	.020	.012	.008
WinIT	.002	.002	.008	.013	.015	.007	.043	.028	.022	.015	.008
Cluster-TCAV	.010	.017	.023	.031	.040	.022	.046	.037	.024	.013	.009
Random	.001	.003	.004	.004	.005	.003	.025	.010	.006	.004	.002

Table 31: Per-fraction faithfulness: **HyMaTE** backbone (neighborhood, readmission).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.025	.057	.077	.069	.058	**.055**	−.024	−.010	.007	.014	.010
TI-SAE (hard)	.025	.051	.066	.069	.072	.054	−.028	−.025	−.017	.003	−.003
Attention	.033	.035	.039	.043	.048	.038	.048	.044	.038	.026	.016
IG	.030	.033	.036	.039	.041	.035	.055	.049	.035	.024	.018
TSR	.033	.038	.041	.042	.047	.039	.042	.035	.026	.018	.013
GradSHAP	.030	.034	.038	.041	.043	.036	.053	.047	.034	.024	.017
DynaMask	.001	−.001	−.002	−.001	.000	.000	.031	.028	.015	.005	.001
TimeSHAP	.009	.011	.012	.011	.016	.011	.021	.021	.017	.012	.009
FIT	.005	.013	.018	.027	.034	.018	.031	.036	.026	.014	.008
WinIT	.019	.019	.021	.023	.027	.022	.026	.032	.029	.026	.020
Cluster-TCAV	.034	.036	.038	.049	.058	.040	.030	.033	.031	.017	.015
Random	.000	.001	.002	.004	.005	.002	.033	.020	.015	.009	.006
MIMIC-IV											
TI-SAE (soft)	.060	.095	.110	.118	.118	**.096**	−.052	−.043	−.031	−.023	−.014
TI-SAE (hard)	.061	.090	.101	.103	.105	.088	−.051	−.037	−.025	−.018	−.008
Attention	.021	.030	.036	.043	.047	.033	.034	.028	.028	.026	.022
IG	.050	.056	.058	.062	.066	.057	.026	.027	.023	.017	.011
TSR	.053	.059	.062	.066	.069	.060	.023	.019	.013	.011	.008
GradSHAP	.050	.056	.059	.062	.066	.057	.025	.026	.022	.016	.010
DynaMask	−.001	−.001	−.003	−.004	−.001	−.002	.046	.036	.014	.005	.002
TimeSHAP	.022	.026	.028	.030	.033	.026	.003	.008	.004	.000	.000
FIT	.008	.019	.027	.031	.034	.022	.035	.029	.024	.016	.010
WinIT	.005	.010	.018	.019	.024	.015	.041	.039	.033	.031	.028
Cluster-TCAV	.003	.017	.030	.055	.076	.032	.073	.057	.047	.023	.008
Random	.000	.001	.002	.005	.009	.003	.048	.024	.013	.008	.002

Table 32: Per-fraction faithfulness: **BAT** backbone (neighborhood, readmission).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.005	.019	.028	.038	.051	.025	.084	.067	.059	.047	.040
TI-SAE (hard)	.006	.018	.030	.041	.057	.028	.084	.074	.061	.046	.035
Attention	–	–	–	–	–	–	–	–	–	–	–
IG	−.002	.011	.020	.033	.042	.018	.143	.111	.088	.063	.045
TSR	−.003	.015	.034	.038	.048	.023	.135	.103	.074	.051	.037
GradSHAP	−.002	.010	.026	.033	.046	.020	.148	.111	.087	.064	.046
DynaMask	.005	.013	.023	.039	.044	.024	.069	.059	.041	.031	.022
TimeSHAP	.015	.019	.021	.021	.030	.019	.077	.069	.056	.054	.042
FIT	.002	.007	.012	.017	.027	.012	.132	.105	.075	.057	.044
WinIT	.015	.020	.029	.036	.040	.026	.102	.082	.062	.047	.037
Cluster-TCAV	.005	.021	.034	.041	.052	**.028**	.091	.077	.061	.042	.031
Random	.002	.006	.010	.015	.019	.010	.078	.046	.029	.022	.019
MIMIC-IV											
TI-SAE (soft)	.007	.019	.027	.034	.040	.024	.025	.013	−.002	−.004	.000
TI-SAE (hard)	.010	.019	.020	.028	.045	.022	.020	.016	.011	.013	.000
Attention	–	–	–	–	–	–	–	–	–	–	–
IG	.011	.018	.024	.032	.037	.022	.007	.005	.000	−.002	−.002
TSR	.015	.026	.035	.043	.048	**.031**	.009	.001	−.002	−.001	−.001
GradSHAP	.011	.017	.023	.029	.035	.021	.005	.002	−.002	−.002	−.002
DynaMask	.002	.002	−.001	.002	.006	.002	.020	.009	.004	.001	.001
TimeSHAP	.006	.009	.008	.007	.009	.007	−.001	.003	.000	−.001	.000
FIT	.001	.003	.003	.005	.008	.003	.027	.010	.005	.006	.004
WinIT	.004	.011	.009	.013	.016	.010	.003	.003	.004	.006	.003
Cluster-TCAV	.002	.004	.007	.007	.008	.005	.019	.005	.011	.010	.007
Random	.000	.001	.002	.001	.002	.001	.016	.008	.006	.005	.002

Table 33: Per-fraction faithfulness: **STraTS** backbone (neighborhood, readmission).

	Comprehensiveness ↑						Sufficiency ↓				
Method	@.10	@.20	@.30	@.40	@.50	AOPC	@.10	@.20	@.30	@.40	@.50
MIMIC-III											
TI-SAE (soft)	.003	.008	.013	.014	.017	.010	.075	.074	.075	.071	.061
TI-SAE (hard)	.003	.008	.010	.014	.016	.010	.076	.074	.074	.074	.066
Attention	.002	.004	.009	.013	.020	.008	.073	.064	.057	.047	.035
IG	.055	.059	.060	.058	.059	.057	.013	.013	.010	.007	.005
TSR	.055	.058	.059	.060	.062	<u>.058</u>	.008	.009	.006	.004	.003
GradSHAP	.055	.058	.060	.058	.060	.057	.012	.014	.010	.006	.005
DynaMask	.066	.083	.091	.097	.105	**.085**	−.032	−.031	−.031	−.031	−.031
TimeSHAP	.031	.032	.034	.036	.039	.034	.010	.014	.009	.007	.005
FIT	.005	.012	.018	.023	.029	.016	.023	.010	.008	.006	.005
WinIT	.006	.010	.012	.020	.029	.014	.044	.039	.030	.023	.015
Cluster-TCAV	.000	.007	.009	.020	.026	.011	.066	.050	.037	.030	.025
Random	.001	.005	.006	.007	.013	.006	.052	.035	.023	.019	.012
MIMIC-IV											
TI-SAE (soft)	.003	.007	.008	.013	.022	.010	.092	.091	.086	.082	.071
TI-SAE (hard)	.004	.007	.007	.011	.022	.009	.092	.092	.083	.078	.069
Attention	.003	.012	.014	.020	.033	.014	.093	.077	.065	.058	.045
IG	.048	.064	.080	.088	.094	.071	.009	.004	.003	.002	.001
TSR	.052	.069	.081	.093	.100	<u>.074</u>	.009	.003	.001	.001	.000
GradSHAP	.048	.066	.081	.088	.095	.072	.010	.004	.002	.001	.001
DynaMask	.080	.107	.125	.144	.158	**.116**	−.061	−.067	−.073	−.073	−.068
TimeSHAP	.030	.031	.033	.036	.039	.032	.004	.012	.013	.008	.007
FIT	.005	.011	.016	.027	.037	.017	.036	.017	.011	.007	.004
WinIT	.016	.020	.023	.022	.030	.021	.044	.032	.027	.021	.016
Cluster-TCAV	.003	.009	.016	.020	.035	.015	.085	.073	.062	.045	.034
Random	.002	.003	.005	.009	.014	.006	.060	.035	.024	.015	.007

H.3 Per-task and per-dataset plausibility

Tables 34–39 report per-task per-dataset plausibility. Best AUROC in **bold**, second-best underlined.

Table 34: Per-task plausibility: **Mortality, MIMIC-III** (AUROC/AUPRC).

Method	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	**.756**/.384	**.631**/.303	**.707**/.356	.598/.249	.516/.216
TI-SAE (hard)	.639/.313	.625/.289	.621/.301	.575/.240	.506/.218
Attention	.615/.258	.468/.189	.568/.231	–	.610/.249
IG	.567/.254	.555/.245	.641/.278	.620/.265	.621/.251
TSR	.594/.279	.586/.262	.673/.320	**.649**/.296	.636/.264
DynaMask	.507/.212	.507/.205	.492/.189	.491/.189	.492/.189
TimeSHAP	.521/.212	.521/.214	.525/.214	.526/.218	.523/.214
FIT	.523/.205	.547/.219	.558/.218	.489/.191	.564/.227
WinIT	.601/.264	.622/.283	.657/.342	.584/.261	**.644**/.298
Cluster-TCAV	.541/.223	.520/.217	.598/.256	.569/.264	.494/.207
Random	.502/.193	.502/.193	.502/.193	.502/.193	.502/.193

Table 35: Per-task plausibility: **Mortality, MIMIC-IV** (AUROC/AUPRC).

Method	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	**.705**/.362	.569/.256	.611/.255	**.638**/.296	.545/.248
TI-SAE (hard)	.632/.314	.533/.245	.556/.232	.621/.276	.561/.252
Attention	.668/.312	**.624**/.268	.523/.203	–	.497/.190
IG	.562/.231	.530/.218	.595/.246	.572/.230	.541/.208
TSR	.578/.244	.552/.232	**.619**/.269	.592/.248	.564/.218
DynaMask	.506/.199	.491/.193	.495/.188	.515/.202	.484/.203
TimeSHAP	.514/.203	.512/.203	.520/.208	.519/.207	.513/.203
FIT	.523/.204	.532/.207	.505/.189	.495/.191	.562/.221
WinIT	.601/.270	.607/.276	.576/.247	.570/.256	**.621**/.279
Cluster-TCAV	.599/.257	.471/.234	.513/.222	.559/.243	.410/.176
Random	.501/.191	.501/.191	.501/.191	.501/.191	.501/.191

Table 36: Per-task plausibility: **LOS>7d, MIMIC-III** (AUROC/AUPRC).

Method	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	**.699**/.360	.514/.197	.623/.295	**.633**/.279	.559/.242
TI-SAE (hard)	.659/.335	.508/.193	.614/.275	.612/.266	.555/.241
Attention	.617/.282	.453/.184	.560/.243	–	.637/.260
IG	.543/.222	.543/.234	.587/.257	.591/.243	.593/.239
TSR	.569/.242	.581/.248	**.626**/.290	.611/.264	.629/.261
DynaMask	.510/.203	.505/.204	.492/.189	.493/.189	.492/.189
TimeSHAP	.513/.203	.521/.216	.517/.210	.517/.209	.517/.210
FIT	.510/.199	.538/.214	.528/.209	.504/.195	.556/.221
WinIT	.575/.242	**.612**/.279	.602/.265	.562/.253	**.640**/.310
Cluster-TCAV	.463/.196	.453/.195	.423/.198	.556/.225	.421/.199
Random	.502/.193	.502/.193	.502/.193	.502/.193	.502/.193

Table 37: Per-task plausibility: **LOS>7d, MIMIC-IV** (AUROC/AUPRC).

Method	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	.312/.181	**.629**/.263	.588/.255	.525/.219	.601/.244
TI-SAE (hard)	.417/.217	.572/.237	.541/.244	.499/.217	.606/.249
Attention	.523/.220	.619/.260	.577/.236	–	.477/.181
IG	.516/.214	.535/.232	.606/.263	.564/.239	.552/.220
TSR	.544/.230	.565/.249	**.644**/.298	**.587**/.259	.580/.237
DynaMask	.500/.198	.500/.199	.495/.188	.474/.186	.494/.188
TimeSHAP	.512/.203	.521/.212	.525/.217	.517/.212	.514/.204
FIT	.519/.210	.542/.220	.521/.193	.466/.180	.551/.222
WinIT	**.567**/.241	.618/.296	.595/.252	.563/.247	**.610**/.272
Cluster-TCAV	.420/.171	.341/.153	.348/.162	.382/.165	.465/.188
Random	.501/.191	.501/.191	.501/.191	.501/.191	.501/.191

Table 38: Per-task plausibility: **Readmission, MIMIC-III** (AUROC/AUPRC).

Method	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	.533/.213	**.608**/.261	.593/.261	.577/.241	.522/.195
TI-SAE (hard)	.534/.211	.596/.251	.616/.262	.558/.232	.515/.194
Attention	.464/.175	.463/.175	.573/.203	–	**.627**/.249
IG	.553/.220	.574/.223	.593/.227	.590/.240	.605/.219
TSR	**.573**/.245	.584/.228	.618/.249	**.615**/.258	.603/.218
DynaMask	.506/.200	.515/.208	.494/.183	.495/.184	.497/.192
TimeSHAP	.513/.212	.512/.194	.510/.197	.515/.197	.505/.189
FIT	.493/.180	.532/.201	.553/.204	.525/.209	.535/.197
WinIT	.565/.261	.602/.267	**.645**/.306	.557/.228	.599/.259
Cluster-TCAV	.567/.232	.498/.213	.617/.391	.534/.215	.536/.240
Random	.503/.188	.503/.188	.503/.188	.503/.188	.503/.188

Table 39: Per-task plausibility: **Readmission, MIMIC-IV** (AUROC/AUPRC).

Method	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
TI-SAE (soft)	.556/.215	.581/.247	.558/.234	.559/.220	.471/.215
TI-SAE (hard)	.534/.210	.572/.240	.534/.228	.537/.224	.464/.214
Attention	.532/.200	**.606**/.239	.524/.189	–	.462/.171
IG	.527/.210	.553/.219	.552/.191	.564/.215	.544/.193
TSR	.542/.219	.571/.229	.569/.199	**.592**/.235	.554/.197
DynaMask	.530/.195	.503/.191	.495/.182	.488/.182	.521/.201
TimeSHAP	.498/.180	.512/.196	.501/.184	.500/.183	.505/.187
FIT	.499/.179	.530/.193	.503/.172	.488/.176	.536/.192
WinIT	**.563**/.230	.603/.279	**.589**/.225	.552/.233	**.610**/.258
Cluster-TCAV	.553/.244	.490/.226	.439/.191	.470/.195	.530/.202
Random	.502/.183	.502/.183	.502/.183	.502/.183	.502/.183

I Clinical plausibility ground truth generation

We construct soft ground-truth severity maps from established ICU scoring systems to evaluate whether attributions align with clinically meaningful observations. The ground truth is a per-cell score $\mathbf{S} \in [0,1]^{B\times T\times V}$ defined only at observed cells, representing clinical severity.

SOFA scoring. We implement the Sequential Organ Failure Assessment (SOFA) [Vincent et al., 1996] scoring for six organ systems:

- **Respiratory**: PaO_2/FiO_2 ratio. Thresholds $< 400, 300, 200, 100$ yield SOFA 1–4. Requires co-observation of PO_2 and FiO_2 within 6 time bins.
- **CNS**: Glasgow Coma Scale total (eye + verbal + motor). GCS $\leq 5, 9, 12, 14$ yield SOFA 4, 3, 2, 1.
- **Cardiovascular**: MAP < 70 yields SOFA 1; vasopressor presence and dose (Dopamine, Norepinephrine, Epinephrine) yield SOFA 2–4.
- **Coagulation**: Platelet count $< 150, 100, 50, 20 \times 10^3$ yields SOFA 1–4.
- **Liver**: Bilirubin $> 1.2, 2.0, 6.0, 12.0$ mg/dL yields SOFA 1–4.
- **Renal**: Creatinine $> 1.2, 2.0, 3.5, 5.0$ mg/dL yields SOFA 1–4; urine output $< 500, 200$ mL/day yields SOFA 3–4.

Each cell's score is the organ sub-score divided by 4 (normalizing to $[0,1]$). For multi-variable organs (respiratory, CNS, cardiovascular), all contributing variables at a given time bin receive the same organ score.

Metrics. Given an attribution map $\mathbf{A} \in \mathbb{R}^{B\times T\times V}$ and severity map $\mathbf{S}$, both restricted to observed cells, we compute:

- **AUROC**: ROC-AUC for classifying $S > 0$ (any severity) vs. $S = 0$.
- **AUPRC**: Precision-recall AUC for the same binary classification.

Higher scores indicate that the attribution method assigns higher importance to clinically severe cells.

This ground truth is a minimal surrogate for clinical relevance. Actual clinical concepts are more complex, involving entangled multi-variable interactions and temporal dynamics that static threshold-based scoring cannot capture. The SOFA score evaluates organ systems independently and does not reflect inter-system dependencies or context-dependent severity. We use this metric as a necessary but not sufficient condition for clinical plausibility.

J Ablation study

J.1 Layer selection

TI-SAE operates on a single backbone layer's hidden states. The choice of layer involves a tradeoff: early layers preserve per-variable locality but contain less contextual information; deeper layers form richer concepts but may lose variable distinctiveness.

Table 40 reports AOPC (neighborhood) across layers for each backbone. For the discrete transformer, layer -1 (last) outperforms other layers across all tasks. For Disc.+PT, earlier layers outperform, suggesting that pretraining shifts local representations to earlier depths. We select layer -3, however, as layer -4 plausibility degrades to near-random on several tasks. For HyMaTE, layer -2 is optimal. For STraTS, all layers yield low TI-SAE AOPC. We attribute this to STraTS's residual-averaging architecture ($\mathbf{h} \leftarrow (\mathbf{x} + \mathbf{h})/2$) they used instead of layernorm for simplification, which causes representations converge toward uniformity at all depths, eliminating the per-variable per-time distinctiveness that TI-SAE's routing requires. For BAT, layer -2 is optimal, suggesting that round-1 representations (before cross-axis mixing) better preserve separability.

Table 40: Layer selection: AOPC (neighborhood) by backbone layer, MIMIC-III mortality. Best per backbone in **bold**.

Backbone	Layer -1	Layer -2	Layer -3	Layer -4
Discrete	**.149**	.121	.114	.105
Discrete+PT	.029	.037	.091	**.216**
HyMaTE	.109	**.143**	.121	–
STraTS	.002	.003	.003	**.015**
BAT	.061	**.065**	–	–

J.2 Selection method

We compare island selection strategies at evaluation time. "SAE (energy)" uses the energy-gated score from Stage A (Eq. 5). "Grad×Act" uses gradient-times-activation scoring as a non-learned heuristic prior. "prediction selector (soft)" uses the learned prediction-aware selector from Stage B.

Table 41: Selection method: AOPC (neighborhood), averaged across tasks.

	Discrete	Disc.+PT	HyMaTE	BAT	STraTS
Activation (energy)	.037	.008	.028	.006	.008
Gradient×Activation	.035	.027	.032	.022	.014
Prediction selector (hard)	.079	.068	.090	.035	.020
Prediction selector (soft)	.086	.076	.097	.043	.016

J.3 Structural mechanism ablation

We ablate the structural mechanisms of TI-SAE by progressively removing components: temporal Gaussian islands, variable routing (W_k=3 $\rightarrow$ 135), channel purity regularization (λ_{ch}=0), and grid structure entirely (standard event-level top-k SAE on $\mathbf{H}[\![B, L, D]\!]$). Additionally, to isolate the dictionary's contribution from the prediction-aware selector, we train the selector directly on raw hidden states with no dictionary. All conditions use the Discrete backbone on MIMIC-III mortality.

Quantitative results. Table 42 reports faithfulness (neighborhood-based AOPC and sufficiency at 10%) and plausibility (AUROC and AUPRC against SOFA-derived ground truth). Removing temporal structure or channel purity has minimal impact on faithfulness. Removing channel purity slightly improves AOPC, likely because the model gains more freedom in variable grouping. In contrast, removing variable routing causes a 58–75% AOPC drop, showing that decoder sparsity is the critical mechanism for faithful attribution. The standard event-level top-k SAE produces near-random attributions (AOPC .023) based on activation.

Table 42: Structural mechanism ablation (Discrete backbone, MIMIC-III mortality). AOPC = comprehensiveness (fractions 0.05–0.5); Suff = sufficiency at 10%; AUROC/AUPRC = plausibility against SOFA prior. All metrics use PS-soft selection. For Standard top-k SAE, activation(dictionary)-based AOPC is shown.

	Faithfulness		Plausibility	
Condition	AOPC↑	Suff@.1↓	AUROC↑	AUPRC↑
Full TI-SAE	.149	$-$.004	.756	.386
− Temporal	.145	.006	.777	.418
− Channel org	.184	.003	.734	.358
− Var routing	.061	.052	.629	.280
Standard SAE	.023 (.155)	.126 (.041)	.487 (.568)	.189 (.224)

However, temporal structure, channel purity, and variable routings are all important components for generating a meaningful clinical dictionary. We examine the interpretability of the dictionary by *channel coherence* and *temporal localization*.

Channel coherence. Without channel purity (λ_{ch}=0), latents freely mix variables across organ systems. In the full dictionary, only 36.9% of latents have at least two decoder variables within a single clinical channel (compared to 91.9% in the default model), directly degrading the clinical interpretability of individual concepts.

Table 43 illustrates this on a representative mortality episode. TI-SAE discovers organ-specific concepts (e.g., a pure Renal latent: BUN + Creatinine + Creatinine Urine, labeled "Chronic Kidney Disease"). Without channel purity, the corresponding latents mix variables from unrelated systems, making them difficult to label or act upon clinically.

Table 43: Top-5 prediction-relevant latents for episode 23020 (mortality): default vs. no channel purity. Channel annotations in brackets.

Default		
k=113	Temperature, CRP, Eosinophils [all Infection]	*Hypothermia w/ Inflammatory Markers*
k=1246	TPN [Other], RDW [Heme], Gastric [Other]	*TPN-Dependent Feeding Intolerance*
k=1119	GCS_verbal, GCS_eye [Neuro], PO2 [Resp]	*Coma with Mechanical Ventilation*
k=931	BUN, Creatinine, Creatinine Urine [all Renal]	*Chronic Kidney Disease with Azotemia*
k=886	BUN, Creatinine, Urine [all Renal]	*Acute Kidney Injury with Oliguria*
No channel organization		
k=136	Platelets, Fibrinogen [Coag], **Bilirubin [Liver]**	—
k=789	Albumin [**Liver**], Saline [**Other**], MgSO4 [**Metabolic**]	—
k=1270	PCO2 [**Resp**], MCH [**Heme**], Platelets [**Coag**]	—
k=1119	GCS_verbal, GCS_eye [Neuro], **Cr Urine [Renal]**	—
k=817	BUN [**Renal**], GCS_eye [**Neuro**], Insulin [**Metabolic**]	—

Temporal localization. Without Gaussian islands, the encoder still produces per-time-step activations $\hat{A}[k, t]$, but it does not learn temporally concentrated patterns. Latents activate nearly uniformly across all 48 time bins (temporal entropy 0.957 vs. default version's localized islands). While this does not substantially hurt faithfulness metrics (the model still identifies which variables matter), it eliminates the ability to say *when* a clinical concept is active. In the same episode, the default model distinguishes early renal decline (k=931 at hour 20) from late-stage injury (k=886 at hour 33) and late coma onset (k=1119 at hour 46). The no-temporal model collapses this temporal narrative.

Variable routing. Removing decoder sparsity degrades both faithfulness and interpretability. Each latent activates 22–37 variables, making individual concepts unidentifiable.

Standard SAE. We attempted to label latents by showing an LLM the full clinical records of each latent's top-activating episodes. For the same mortality episode (23020) where TI-SAE identifies five distinct organ-system concepts (Table 43), the standard SAE's top 10 latents all fire on 100% of patients and receive contradictory labels—the same patient is simultaneously described as having "stable vital signs," "tachycardia with hemodynamic instability," and "hypotension requiring vasopressors." These labels seem to describe the common phenotypes of different top-activating patients, not the latent's mechanism. Unlike TI-SAE, where the decoder structurally binds each latent to specific variables and time, the standard SAE's latents have no such anchor, making labels unfalsifiable. We also trained a modified prediction selector that scores latents globally, recovering neighborhood-based AOPC. However, this required injecting variable identity information that the SAE itself does not encode (using per-variable activation profiles computed from batch data), and weights each latent uniformly across all temporal positions. The low plausibility metrics imply that the attribution does not reflect genuine clinical relevance. Still, the small change in AOPC suggests that the specific TI-SAE architecture is not required for the faithfulness gain; the islands instead primarily improve concept plausibility and interpretability, which is the structural information needed to read the dictionary.

Isolating the dictionary from the selector. The dictionary plays an independent role in turning a sufficient set of cells into a decomposable, spatiotemporally grounded concept shared across patients, which the prediction selector alone cannot produce. Still, it is important to decompose where the performance gain comes from. We tried to isolate the prediction selector from the dictionary, using the

selector trained directly on raw hidden states with no dictionary, under the same selection objective and budget (cell-level selection). TI-SAE substantially improves AOPC over raw hidden states (.149 vs. .038), consistent with a representation-level contribution beyond prediction-aware selection alone.

J.4 Island window shape

A Gaussian window in this study is a temporal support gate that indicates when a concept is active, while the actual temporal trajectory (the activation strength) is carried by the encoder output. Nevertheless, here we test whether it restricts the temporal patterns a concept can express, by replacing the symmetric Gaussian with a split-normal (skewed) window ($\sigma_L = \sigma(1-a)$, $\sigma_R = \sigma(1+a)$), with the skew a either learned per island or fixed, and re-running the full pipeline over seeds 42–44 (Table 44).

Table 44: Island window shape ablation (Discrete backbone, MIMIC-III mortality; mean $\pm$ std over three seeds). AOPC = neighborhood-based comprehensiveness; plausibility = SOFA AUROC.

Island shape	AOPC↑	Plausibility↑
Symmetric Gaussian	$.140 \pm .006$	$.739 \pm .013$
Learned asymmetric	$.142 \pm .014$	$.725 \pm .013$
Fixed asymmetric (a=0.5)	$.141 \pm .016$	$.715 \pm .031$

When the skew is free the model uses it heavily (mean $|a| \approx 0.34$, with 85–96% of islands skewed), so asymmetry is available and actively used, yet performance is unchanged. This supports that the symmetric window is not a restrictive choice for these tasks.

Also, the windows are already diverse at the concept level. A concept's mask is the clamped sum of its M windows (Eq. 6), and on the canonical model most concepts use more than one: 30/21/23/26% of selected concepts use 1/2/3/4 windows (70% use two or more). Therefore, most masks are composed from several windows rather than a single symmetric bump.

J.5 Dictionary capacity and sparsity budget

We sweep the dictionary size K with fixed selection budget n_{global}=128, sweep n_{global} with fixed K=1280, and sweep the number of islands per latent M with proportional budget n_{global}=$32M$. All conditions use the Discrete backbone on MIMIC-III mortality.

Table 45: Dictionary capacity sweep (n_{global}=128 fixed). Dead% = fraction of latents never selected across 28,626 episodes; >1% = latents active in >1% of episodes.

	Utilization		Faithfulness		Plausibility	
K	Dead%	>1%	AOPC↑	Suff@.1↓	AUROC↑	AUPRC↑
320	0.3%	281	.137	.030	.711	.353
640	0.0%	417	.165	−.005	.734	.375
1280	2.0%	469	.149	−.004	.756	.386
2560	23.5%	524	.143	−.005	.700	.333

Table 46: Sparsity budget sweep (K=1280 fixed). n/KM = fraction of island slots selected per episode.

		Utilization		Faithfulness		Plausibility	
n_{global}	n/KM	Dead%	>1%	AOPC↑	Suff@.1↓	AUROC↑	AUPRC↑
32	0.6%	26.2%	136	.134	.015	.713	.333
64	1.2%	8.2%	259	.151	−.007	.716	.353
128	2.5%	2.0%	469	.149	−.004	.756	.386
256	5.0%	0.1%	767	.141	.000	.714	.364
512	10.0%	0.2%	1022	.128	.019	.763	.395

Table 47: Island count sweep (K=1280, n_{global}=32M proportional).

	Faithfulness		Plausibility	
M	AOPC↑	Suff@.1↓	AUROC↑	AUPRC↑
1	.131	−.021	.723	.346
2	.152	−.004	.731	.341
4	.149	−.004	.756	.386
8	.144	.005	.703	.357

Faithfulness peaks at moderate capacity: AOPC is highest at K=640 and n_{global}=64, then declines as larger dictionaries spread attribution across more latents. Plausibility improves with capacity up to K=1280 but drops at K=2560, and increases monotonically with n_{global} as more active latents cover more clinically relevant cells. Increasing M from 1 to 4 improves faithfulness (AOPC .131→.149) by enabling temporal decomposition within latents, but performance declines at M=8, suggesting overly fine fragmentation is counterproductive. We select K=1280, n_{global}=128, and M=4 as the operating point where faithfulness and plausibility remain competitive and latent utilization is efficient.

J.6 Loss component analysis

We ablate each regularization loss by setting its weight to zero (Discrete backbone, MIMIC-III mortality). Without auxK revival (λ_{auxk}=0), 80.0% of latents are never selected across the full cohort (vs. 2.0% with auxK), and temporal spread within latents collapses to 1.80h (vs. 5.46h). Without μ-repulsion (λ_{repul}=0), islands within each latent overlap temporally, and the model compensates by widening them (σ_{alive}: 7.80 vs. 4.92). Removing mask coverage (λ_{mcov}=0) has minimal effect at the default operating point (σ_{sel}: 4.97 vs. 5.25); residualization and μ-repulsion already constrain island width, and we retain mask coverage at a small λ=0.005 as a safeguard for higher-budget regimes.

J.7 STraTS residual architecture

STraTS replaces the standard additive residual with an averaging operation $\mathbf{h} \leftarrow (\mathbf{x} + \mathbf{h})/2$. Each averaging step blends the cell representation toward a shared mean, reducing per-(t, v) distinctiveness. We hypothesize that this undermines TI-SAE's structure, which requires distinct per-cell features to assign observations to latent channels. To isolate this effect, we replace only the residual with $\mathbf{h} \leftarrow \text{LayerNorm}(\mathbf{x} + \mathbf{h})$, keeping all other hyperparameters identical.

Table 48 shows the comparison before and after the residual change. On MIMIC-III mortality (layer −4, neighborhood-based masking), this change recovers TI-SAE to competitive faithfulness (PS-soft AOPC .015 → .044) and best plausibility (AUROC .516 → .652), showing that the averaging residual is the primary bottleneck for TI-SAE on this backbone.

Table 48: STraTS residual architecture ablation (MIMIC-III mortality, layer −4, neighborhood-based masking).

	AOPC ↑		AUROC ↑		AUPRC ↑	
Method	Orig	LN	Orig	LN	Orig	LN
TI-SAE (soft)	.015	.044	.516	.652	.216	.296
TI-SAE (hard)	.013	.042	.506	.637	.218	.285
Attention	.010	.046	.611	.542	.250	.220
IG	.064	.048	.622	.608	.252	.244
TSR	.061	.050	.637	.625	.265	.260
DynaMask	.033	.033	.492	.492	.189	.189
TimeSHAP	.016	.016	.524	.522	.214	.214
FIT	.010	.011	.564	.547	.227	.218
WinIT	.025	.026	.644	.619	.299	.278

K Backbone prediction performance

Table 49 reports test AUROC for each backbone across prediction tasks and datasets. All backbones achieve comparable performance on mortality (.855–.870) and length-of-stay (.819–.843). Readmission is consistently harder (.618–.652), which is typical for this task.

Table 49: Backbone test AUROC per task and dataset.

	Mortality		LOS >7d		Readmission	
Backbone	III	IV	III	IV	III	IV
Discrete	.861	.867	.832	.824	.632	.652
Discrete+PT	.861	.865	.834	.827	.618	.651
HyMaTE	.861	.872	.830	.829	.638	.649
STraTS	.873	.867	.843	.823	.644	.649
BAT	.858	.868	.826	.822	.612	.648

L Latent dictionary labeling

Each active latent k with activation frequency $>1\%$ in the learned dictionary is labeled by an LLM (Claude Sonnet 4.6) to produce a clinical concept name and description. The system message instructs the model to act as an expert clinical informaticist interpreting TI-SAE latents from MIMIC-III, responding only with a JSON object.

The per-latent prompt contains five sections:

1. **Context.** Describes the TI-SAE setup (48-hour irregularly-sampled ICU time-series, mortality prediction, latent concepts with Gaussian islands and channel assignments) and a note that zero observations within a window reflect clinical ordering patterns, not missing data.
2. **Variable composition.** Top-3 decoder variables with weights from $\mathbf{W}[k, :]$.
3. **Activation statistics.** Population activation frequency and mortality odds ratio.
4. **Exemplar episodes.** Up to 10 episodes ranked by SAE activation strength, plus up to 10 ranked by prediction selector weight. Each shows island windows (μ, σ), raw variable traces, and unobserved variables. Example:

```
Episode 4821 (mortality: YES, strength: 0.847)
Island: hours 18-30 (mu=24.2, sigma=3.1)
- Creatinine: t20h=1.82, t24h=2.14, t28h=2.61
- BUN: t18h=34.0, t24h=42.0
- Urine Output: NOT MEASURED in this window
```

5. **Task.** Requests JSON with `clinical_label` (2–6 word name) and `clinical_description` (1–3 sentence interpretation).

LLM implementation details. Dictionary labeling uses Claude Sonnet 4.6 (Anthropic API, temperature 0.0). Verbalization uses the same model. Pairwise judging uses GPT-4o (OpenAI API, temperature 0.0) as the judge model, to avoid self-preference bias.

M Verbalized explanation evaluation

M.1 Verbalization and explanation quality evaluation protocol

The verbalization evaluation (Section 4.3.3) uses two LLM stages: a verbalizer (Claude Sonnet 4.6) that generates explanations, and a pairwise judge (GPT-4o) that compares them per dimension.

Verbalizer prompt. Both conditions share the same system instruction:

> *You are an ICU physician writing a clinical explanation of a mortality prediction. In under 150 words, explain why this patient is at this risk level. Explain exclusively based on the model-provided analysis below. Plain text, no markdown.*

Explanation conditions. Both conditions receive the prediction (risk level and probability) and the full 48-hour raw value table. They differ in what model-derived information is provided.

Attribution-augmented. Top-K cells ranked by attribution score:

```
Patient prediction:  HIGH mortality risk (probability:  0.82)
The model's attribution analysis identified these observations
as most important for its prediction:
Hour 18:  Creatinine (importance:  0.1842, value:  2.14)
Hour 24:  Heart Rate (importance:  0.1203, value:  112)
...
Clinical values over the 48-hour ICU stay:
[full raw value table]
```

TI-SAE-augmented. Top-5 prediction selector-ranked concepts with dictionary labels (Appendix L):

```
Patient prediction:  HIGH mortality risk (probability:  0.82)
The model identified these clinical concepts as most important
for its prediction:
Concept 1 [Renal] "Progressive AKI" (total weight=3.5):
Creatinine (wt=0.482), BUN (wt=0.318)
Rising creatinine and BUN suggesting acute kidney injury.
Island h18-30 (wt=2.10):
Creatinine:  h20=1.82, h24=2.14, h28=2.61
BUN: h18=34.0, h24=42.0
...
Clinical values over the 48-hour ICU stay:
[full raw value table]
```

Pairwise judge protocol. For each episode and baseline, the judge compares the TI-SAE explanation against the baseline explanation. A/B order is randomized per pair. Each of the three dimensions is evaluated in an independent API call to avoid holistic bias [Zheng et al., 2023].

Concept-level reasoning and Clinical actionability dimensions are *blinded*: the judge sees only the patient data table and the two explanations, with no indication of which method produced which. Faithfulness dimension is *unblinded*.

1. **Concept-level reasoning**: Which explanation better identifies clinical concepts (co-evolving variable groups) rather than listing individual abnormalities?
2. **Clinical actionability**: Which explanation would be more useful for a clinician deciding on next steps for this patient?
3. **Faithfulness**: The judge additionally sees each explanation's source model analysis (attribution cells or concept structure) and evaluates which explanation more faithfully represents the findings from its source, accurately conveying the key patterns, priorities, and values identified by the model.

Evaluation protocol. $N{=}50$ test episodes selected via stratified random sampling (25 high-risk with $p > 0.5$, 25 low-risk).

Budget matching. K for the attribution condition equals the mean number of distinct observed (t, v) cells covered by TI-SAE's top-5 concepts ($K{=}75$), ensuring comparable information budgets.

M.2 Clinician evaluation of verbalized explanations

To validate the LLM judge alignment, four independent tertiary-hospital clinicians (all board-certified specialists with $\geq$10 years of experience; the same panel as the dictionary label evaluation in Appendix G.5) re-ran the pairwise comparison on the same 50 episodes used in the LLM judge evaluation, restricted to the IG baseline (for completely blinded evaluation). Each clinician saw two explanations (A and B, randomized order) and scored three dimensions: concept-level reasoning, actionability, and faithfulness. Phase 1 (concept, actionability) was blinded to the source method; faithfulness was scored after locking Phase 1, with source analyses revealed in a separate window. Each rater could optionally view raw patient data and per-variable trend charts.

Aggregate marginal alignment. Aggregate clinician majority votes matched the LLM judge's ranking direction across all three dimensions (Table 50).

Table 50: Marginal preference distributions (N=50 episode pairs, IG baseline). Format: TI-SAE preferred / IG preferred / Tie.

	Concept	Actionability	Faithfulness
Clinician majority	39 / 4 / 7	23 / 9 / 18	4 / 4 / 42
LLM judge	33 / 15 / 2	32 / 18 / 0	15 / 11 / 24

Per-rater marginals. Individual raters varied substantially in tie usage (Table 51), particularly on faithfulness.

Table 51: Per-rater preference marginals (N=50 each).

Rater	Concept	Actionability	Faithfulness
Rater 1	38 / 8 / 4	24 / 5 / 21	3 / 2 / 45
Rater 2	30 / 20 / 0	22 / 22 / 6	21 / 29 / 0
Rater 3	30 / 3 / 17	23 / 15 / 12	14 / 13 / 23
Rater 4	36 / 9 / 5	25 / 10 / 15	6 / 3 / 41

Per-item agreement. Per-item inter-rater agreement was low across all rater pairs: clinician-clinician Fleiss κ was $+0.04$ for concept, $+0.18$ for actionability, and -0.06 for faithfulness; clinician-LLM agreement was in a simliar range (mean Cohen's $\kappa = -0.03$, -0.01, and $+0.08$ across raters for concept, actionability, and faithfulness, respectively). We attribute this to the inherent hardness in our setting, compounded by the verbalizer LLM's strong clinical prior, and the difficulty of consistently judging free-form text on multiple implicit quality criteria simultaneously [Saad-Falcon et al., 2025]. Small differences in explanation quality produce many borderline cases, so per-item agreement becomes dominated by rater-style differences such as individual tie-threshold preferences. In our data this manifested as substantial across-rater variation in tie usage (one rater tied $45/50$ faithfulness pairs while another never tied). When both raters committed to a direction on the concept dimension, they agreed in 68% of comparisons (153/226; 50–87% across the six rater pairs), indicating that the low κ reflects heterogeneous tie-calling together with genuine disagreement concentrated on borderline cases. We therefore report aggregate marginal preferences rather than per-item agreement as the primary clinician validation signal.

M.3 Verbalization advantage from concept labels

In the pairwise comparison of verbalized explanations, TI-SAE provides dictionary labels while baselines provide attributed cells (Section 4.3.3); here we isolate what the labels contribute. We ran two controls (same LLM verbalizer and judge, 50 episodes): *labels removed* strips the label and description from TI-SAE's concepts, keeping variables, temporal windows, weights, and per-window activation values; + *cross-episode summary, no label* adds back only the label-free cross-episode characterization of each concept (cohort frequency, outcome odds ratio, typical temporal positions). We focus on the concept dimension, where the label-inflation concern applies; without labels the concept format is also rated less actionable than per-cell attributions, and on faithfulness the two are mixed and tie-heavy.

Table 52: Label ablation on TI-SAE's input. Win counts: TI-SAE condition–baseline (ties), LLM judge, N=50 per comparison.

TI-SAE condition	vs. IG	vs. Attention	vs. TSR
Label removed	24–23 (3)	16–29 (5)	25–23 (2)
+ cross-episode summary, no label	31–19 (0)	26–20 (4)	27–21 (2)

With its label removed, TI-SAE no longer clearly beats raw attribution cells, so most of the verbalized advantage came from the label. Adding back only the label-free cross-episode summary recovers

a small concept lead. However, as noted in Section 5.3, the labels are themselves generated from cross-patient exemplar evidence that only the shared dictionary provides.

In addition, we compare TI-SAE against different baseline conditions: *LLM-summarized (per-episode)* lets the same LLM first group the baseline's attributed cells into per-episode concepts (channel, variables, time window, key values) before verbalization; *channel-grouped, no label* groups the same cells into organ-channel concepts, isolating structure without LLM wording.

Table 53: TI-SAE vs. different baseline conditions. Win counts: TI-SAE–baseline (ties), LLM judge, N=50 per comparison.

Baseline condition	vs. IG	vs. Attention	vs. TSR
LLM-summarized (per-episode)	11–39 (0)	11–37 (2)	18–30 (2)
Channel-grouped, no label	27–21 (2)	29–20 (1)	28–20 (2)

Given a fresh per-episode LLM summary, the baseline beats TI-SAE on concept. However, that summary is a per-patient description the LLM can word freely for each case. Against the matched label-free baseline that groups the same cells into organ-channel concepts (isolating structure, not LLM wording), TI-SAE wins across all three methods.

N Example explanations

Below are verbalized explanations for representative high-risk and low-risk episodes, generated by the same verbalizer.

High-risk example (episode 23415, p=0.646).

Attention:

> This patient carries a 64.6% predicted mortality risk driven by several converging physiological concerns over the 48-hour ICU stay. The most influential features include a low temperature at hour 47 (36.06°C), suggesting early hypothermia, and persistently impaired neurological status, with GCS eye opening of 3 and verbal score of 1 at multiple time points despite ongoing sedation with midazolam and fentanyl, making true neurological assessment difficult. The patient required prolonged mechanical ventilation and vasopressor support with neosynephrine throughout the stay, indicating refractory hemodynamic instability. Markedly elevated neutrophils (92.8%) with reciprocally low lymphocytes (4.9%) at hour 15 reflect a severe stress or septic response. Persistently low urine output, ongoing broad-spectrum antibiotics, amiodarone use suggesting arrhythmia burden, and elevated heart rates up to 160 bpm further compound the risk.

TI-SAE:

> This patient carries a 64.6% predicted mortality risk driven by five converging critical illness patterns. First, persistent hypothermia early in the admission (temperatures 35.3–35.5°C) signals impaired thermoregulation consistent with severe physiologic stress. Second, the patient demonstrates oliguria with urine outputs frequently 10–30 mL/hr in the latter half of the stay, accompanied by acidemic arterial pH values as low as 7.23, indicating acute kidney injury with metabolic acidosis. Third, the patient requires sustained mechanical ventilation throughout the 48-hour window with FiO2 ranging from 0.5 to 1.0, reflecting severe hypoxemic respiratory failure. Fourth, neurological assessment reveals a largely unresponsive state with GCS verbal scores of 1 and absent eye opening for much of the admission, indicating profound coma. Fifth, continuous amiodarone infusion signals serious cardiac arrhythmia requiring active suppression. The simultaneous presence of respiratory failure, renal failure with acidosis, coma, hemodynamic instability requiring vasopressors, and cardiac arrhythmia represents a high-acuity multi-organ dysfunction syndrome portending poor survival.

Low-risk example (episode 23134, p=0.016).

Attention:

This patient is predicted to have a low mortality risk (1.6%) over the 48-hour ICU stay. Despite markedly elevated transaminases (ALT peaking above 10,000 U/L) and persistently elevated INR consistent with acute hepatocellular injury, several reassuring features drove the model's favorable prediction. Temperature remained consistently normal throughout the stay, suggesting absence of significant infectious or inflammatory deterioration. Platelet counts were maintained in an acceptable range (164–218), indicating preserved hematopoietic reserve. Albumin levels, while mildly reduced (3.0–3.5 g/dL), remained relatively stable. Neurological status was intact, with GCS verbal and motor scores consistently at 5 and 6 respectively throughout the admission. Urine output was adequate, reflecting preserved renal perfusion. Hemodynamics, though occasionally borderline, largely stabilized over time. The patient was tolerating oral intake by the latter part of the admission. Collectively, these findings suggest a patient with acute liver injury who is nonetheless maintaining end-organ function and neurological integrity, supporting a favorable short-term prognosis.

TI-SAE:

This patient carries a low predicted mortality risk of 1.6%, driven primarily by sustained neurological and physiologic stability throughout the 48-hour ICU stay. The most protective finding is a consistently maximal GCS motor score of 6 and full verbal responsiveness (GCS verbal 5) maintained across the entire observation window, indicating intact consciousness without deterioration. Eye-opening was spontaneous (GCS eye 4) from approximately hour 9 onward. Platelet counts remained normal throughout, reflecting reserved hematologic function. Urine output was sustained and adequate across all time intervals, suggesting intact renal perfusion. Hemodynamics, while occasionally labile, generally remained acceptable without vasopressor requirement. Despite concerning laboratory findings — markedly elevated transaminases, elevated bilirubin, significantly prolonged INR, and initially very high LDH consistent with acute hepatocellular injury — these hepatic derangements showed a clear downward trend over 48 hours. The combination of preserved and stable neurological function, adequate renal output, and improving hepatic markers collectively drove the model's low mortality prediction.

O Broader impacts

TI-SAE provides interpretable explanations of clinical prediction models that may help clinicians audit individual predictions and surface unintended model dependencies. Concept-level explanations are easier to validate against domain knowledge, lowering the technical barrier for clinical users to inspect a model.

Potential risks include increasing automation bias (clinically-coherent explanations can produce false confidence in incorrect predictions) and inheritance of model bias (any disparate performance in the predictor is reflected in its concepts). We do not advocate clinical deployment without independent validation.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 state the core contributions (TI-SAE architecture with Gaussian islands and sparse variable composition, two-stage decoupled selector, evaluation across five backbones on MIMIC-III/IV) and these match the experimental scope reported in Section5 and G–N.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 explicitly discusses six limitations: fixed-length grid, backbone compatibility (per-variable locality requirement), post-hoc validation of concepts as clinical, coarseness of SOFA-based plausibility ground truth, single-token-per-event setting, and LLM-mediated verbalization.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper presents an empirical method with architectural and algorithmic contributions; it does not include theorems or formal proofs.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Appendix A provides complete hyperparameters, optimizer settings, learning-rate schedules, and architecture configurations for all five backbones and TI-SAE; Appendices C–F detail regularizer formulas, IslandParamNet specifics, the prediction selector feature vector, and the clinical ontology mapping.

   Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: MIMIC-III and MIMIC-IV are publicly available through PhysioNet under credentialed access (cited in Section 4.1). Code and trained-model checkpoints will be released as an anonymized repository linked in supplementary materials, with full instructions to reproduce results.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4 specifies dataset splits, prediction tasks, backbone configurations, and frozen-backbone training protocol; Appendix A reports the full hyperparameter table including optimizer, learning rates, batch sizes, schedule, and early-stopping criteria.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Appendix H reports bootstrap standard deviations (2000 resamples) for all faithfulness and plausibility metrics; Appendix G.7 reports mean ± std across three independent training seeds for concept reproducibility metrics.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix A specifies NVIDIA A100-SXM4-80GB GPUs as the compute platform and reports trainable parameter counts (TI-SAE $\approx$ 856K, prediction selector $\approx$ 38K).

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses publicly available critical care databases (MIMIC-III, MIMIC-IV) under PhysioNet credentialed access agreements, which include de-identification and HIPAA-compliant data handling.

   Guidelines:

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

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss potential positive impacts (clinical auditability, surfacing unintended model dependencies) and risks (automation bias, inheritance of model bias) in Appendix O.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release pre-trained generative models, scraped datasets, or assets with high misuse potential. Trained backbones are clinical mortality/LOS/readmission predictors on credentialed data and pose no dual-use risk beyond the underlying MIMIC dataset.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: MIMIC-III and MIMIC-IV are used under the PhysioNet Credentialed Health Data License. Backbone architectures (STraTS, HyMaTE, BAT, DuETT) and attribution baselines (Captum, Dynamask, TimeX, etc.) are cited to their original publications and used under their respective open-source licenses (MIT or Apache 2.0). Claude (Anthropic API) and GPT-4o (OpenAI API) are used under their respective commercial terms of service.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: TI-SAE code, training scripts, and configuration files will be released with documentation.

Guidelines:

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

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

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

Answer: [Yes]

Justification: This study did not investigate human participants as research subjects. Nevertheless, we provide here the evaluation instructions given to clinicians for their assessments. Four independent tertiary-hospital clinicians participated in two evaluations: a dictionary-label rating task (N=30, 5-point Likert; Appendix G.5) and a pairwise verbalization comparison (N=50, three-dimensional A/B with ties; Appendix M.2). Full instructions given to clinicians and the rating rubrics are described in those appendices. The HTML evaluation interface used by the clinicians is generated by code released with the repository. Clinicians were colleagues of the authors and did not receive monetary compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper uses de-identified, publicly available critical care datasets (MIMIC-III/IV) under PhysioNet's credentialed access agreement, which does not require additional IRB approval for secondary analysis. No new human subjects research was conducted.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: Claude Sonnet 4.6 is used for unsupervised post-hoc dictionary labeling (Appendix L) and for generating verbalized clinical explanations (Appendix M). GPT-4o serves as the pairwise judge in Section 5.3 / Appendix M to avoid self-preference bias.

    Guidelines:

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