SMILE: Learning from Patient-Specific Observation Geometry in Sparse Clinical Time Series

Anonymous submission

Abstract

Clinical time series are sparse and irregular, and their observation patterns can be informative. Existing methods primarily encode missingness at individual variable–time points, making it difficult to jointly capture local monitoring intensity and within-patient cross-variable observation structure. We introduce SMILE (Structured Missingness informed Learning), a representation-learning framework that models these first- and second-order patterns as patient-specific observation geometry. SMILE augments variable tokens with time-local observation density to encode local monitoring intensity and conditions cross-variable attention on centered co-missingness after accounting for each patient's marginal missing rates. Record-level relation gating and time-indexed affine modulation further adapt geometry-conditioned interactions across patients and over time. Across six ICU prediction tasks on PhysioNet 2012, PhysioNet 2019, and MIMIC-III, SMILE achieves state-of-the-art average performance under matched evaluation settings, including a +6.63 AUPRC improvement over SMART on MIMIC-III decompensation. Extensive controls—including capacity, permutation, cohort-prior, component, and density-window analyses—show that these gains stem from patient-specific observation geometry rather than increased model capacity, generic attention bias, or a particular density window. Note that additional descriptions and experimental results, as well as source code and datasets, are also submitted as the Supplements.

Introduction

ICU electronic health records (EHRs) are sparse and irregular. Across the six processed task cohorts used here, 55.1–80.2% of valid dynamic entries are missing; related MIMIC-III and PhysioNet/CinC benchmarks likewise exhibit substantial incompleteness (Sun et al. 2026; Sharafoddini et al. 2019). Measurements are not absent uniformly: their presence reflects clinical orders, monitoring protocols, disease severity, and workflow. The arrangement of observations can therefore carry predictive information beyond the recorded values (Rubin 1976; Agniel, Kohane, and Weber 2018; Che et al. 2018; Singh, Sato, and Ohkuma 2021; Groenwold 2020).

Conventional approaches address irregularity through imputation, timestamp encoding, missingness-aware prediction, or masked self-supervision. Although these strategies improve value modeling, they usually expose missingness as a pointwise binary state. This view does not represent whether monitoring is locally dense or sparse, or whether variables are observed together beyond their marginal rates. We call these coupled structures within one record *patient-specific observation geometry*.

We test whether masks contain structure beyond marginal missing rates with a problem-driven, train-only audit. Observation fractions are associated with binary endpoints, vary across trajectory stages, and show system-level co-missingness. We also compare the exact $\mathbf{C}_b$ used by SMILE with a within-record null that independently permutes each variable over time while preserving its observation rate. Among 800 sampled training records per binary task, 99.0–100% retain significant excess off-diagonal structure after task-wise correction. Figure 1 shows that nearly matched marginal rates can still yield distinct matrices. These descriptive audits establish structured, informative observation patterns but do not identify MAR versus MNAR or a causal monitoring policy.

Each centered observation trajectory defines a variable embedding whose inner products form the positive-semidefinite Gram kernel $\mathbf{C}_b$; its induced pseudometric motivates the geometric terminology, although SMILE uses the similarities directly. Local density and $\mathbf{C}_b$ are deterministic mask functions, so they add no information for an unrestricted predictor. Instead, they expose degree-one and degree-two statistics as a finite-capacity inductive bias, constructed from each trajectory at inference time.

The audit exposes three representation failures: *local-intensity collapse*, where a mask value omits surrounding monitoring density; *marginal-rate dominance*, where raw co-missingness reflects measurement rates and cohort relations erase record specificity; and *temporally uniform conditioning*, where one relation is applied across changing context.

SMILE encodes this geometry through one observation-conditioned interaction operator: local density supplies first-order intensity, centered co-missingness supplies second-order Gram similarities, and head-specific calibration, elapsed-time gating, and affine modulation adapt cross-variable attention. We implement it in a missing-aware dual-

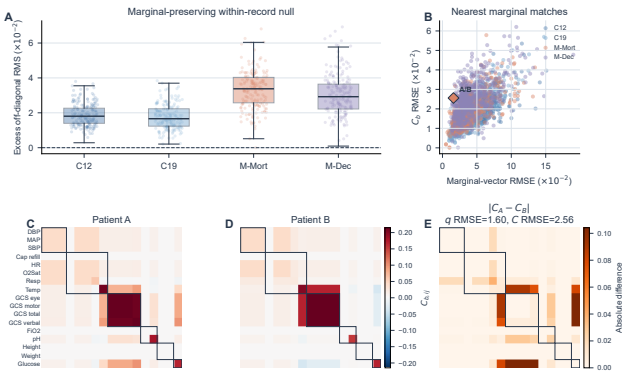

Figure 1: Patient-level observation geometry beyond marginal rates. **A**: excess off-diagonal RMS over the marginal-preserving null. **B**: nearest records in marginal-rate space can have different $\mathbf{C}_b$. **C–E**: two label-free MIMIC mortality training records with nearly matched rates, their matrices, and absolute difference. The audit does not identify a causal missingness mechanism.

channel masked autoencoder, but it applies to encoders with explicit variable interactions.

Our contributions are:

- A problem-driven train-only audit and patient-specific observation geometry that connect endpoint association, temporal variation, and cross-variable structure without causal missingness claims.
- A unified observation-conditioned interaction operator that converts patient-specific Gram similarities into calibrated attention energies and combines them with local density and temporal conditioning.
- A six-task ICU evaluation with capacity, permutation, cohort-prior, density-window, component, and sensor-removal controls that test the value and limits of patient-specific observation geometry.

Related Work

Irregular clinical time-series modeling. Early clinical systems commonly discretize records onto a fixed grid, which simplifies batching but discards the actual observation times (Choi et al. 2016; Harutyunyan et al. 2019). Missing-aware recurrent models instead expose the mask and the elapsed time to the predictor (Che et al. 2018); related imputation lines reconstruct the unobserved values with bidirectional recurrence or spatio-temporal message passing (Cao et al. 2018; Cini, Marisca, and Alippi 2022; Zhang et al. 2021b). Other methods represent the irregular process directly through attention or continuous-time dynamics (Shukla and Marlin 2021; Chen et al. 2023), set/graph interactions (Horn et al. 2020; Zhang et al. 2021b), and multi-scale or adaptive time encoding (Zhang et al. 2023; Lee et al. 2025; Chen et al. 2026; Hajisafi et al. 2025). These designs improve temporal interpolation and long-range dependency modeling, but their observation features are still predominantly pointwise, interval-based, or learned at the cohort level. They therefore leave open how to represent whether a particular patient is being monitored densely at a given time and which variables are observed together in that same record.

Missingness as a signal and self-supervised representation learning. Prediction with missing covariates differs from likelihood-based inference: even under a linear complete-data response model, the optimal predictor can depend nonlinearly on observed entries and their missingness pattern (Le Morvan et al. 2020). In routine health data, informative presence and observation distinguish whether a measurement occurs from its longitudinal timing or intensity (Sisk et al. 2021). These patterns can be statistically associated with clinical workflow and outcomes, so several studies append missingness features or use them for early sepsis prediction (Singh, Sato, and Ohkuma 2021; Singh et al. 2019; Groenwold 2020). SMART learns missing-aware temporal and variable attention and performs latent-space reconstruction (Yu et al. 2024a), while SPECTRA jointly encodes missingness patterns and class imbalance (Qin et al. 2026). In parallel, masked pretraining has become a general representation-learning strategy for time series, with models reconstructing points or patches (Tang and Zhang 2022; Nie et al. 2023), decoupled latent targets (Cheng et al. 2026), or reconstruction-error-guided pseudo-observations for irregular data (Liu, Cao, and Chen 2026). For irregular series, PrimeNet combines time-sensitive contrastive learning with reconstruction (Chowdhury et al. 2023). Image-based modeling provides another route to robustness under heavy missingness, through either pretrained vision transformers or sequence–image self-supervision (Li, Li, and Yan 2023; Chen et al. 2025). Taken together, existing approaches represent observation patterns mainly through pointwise masks and time gaps, reconstruction objectives, or sequence-level missingness encodings. They do not jointly distinguish a patient's marginal observation rates from excess cross-variable co-missingness while representing how monitoring intensity varies locally over time. SMILE addresses this gap by modeling time-local density and centered co-missingness as first- and second-order components of patient-specific observation geometry, which then conditions cross-variable interaction by converting patient-specific Gram similarities into calibrated attention energies. This formulation treats structured observation patterns as predictive signals without making a causal claim about the missingness mechanism. Clinical monitoring can vary within a stay.

Method

Method Overview

Figure 2 summarizes the pipeline. SMILE derives local density $\boldsymbol{\rho}_b$ and centered co-missingness $\mathbf{C}_b$ from each mask. Every encoder block fuses density at tokenization, uses gated $\mathbf{C}_b$ in variable attention, and applies time-indexed modulation while retaining temporal attention. Self-supervised pretraining is followed by task-specific fine-tuning.

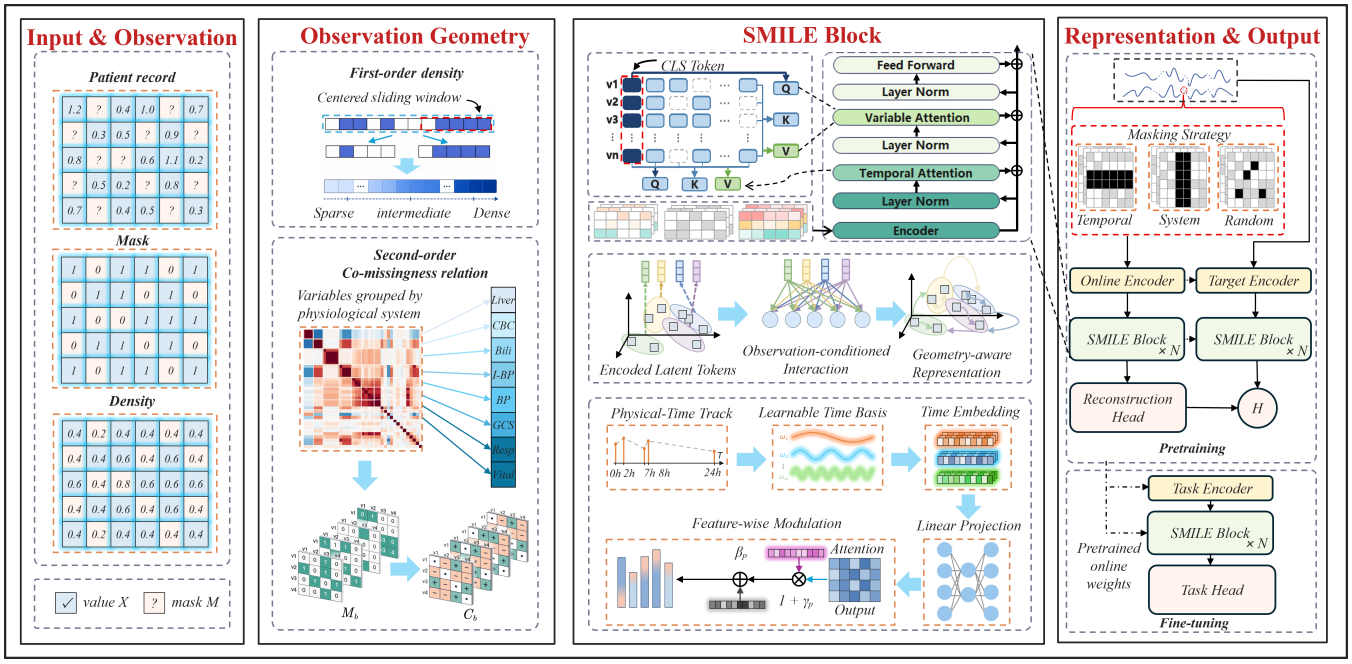

Figure 2: Overview of SMILE. Given values and their observation mask, SMILE constructs local densities and centered co-missingness. Each block uses them in density-augmented inputs, relation-conditioned variable attention, and elapsed-time modulation while retaining temporal attention. Pretraining reconstructs artificially hidden observations with online and target encoders; fine-tuning transfers the encoder to a task-specific head.

Problem Formulation and Geometric Properties

Let $\mathcal{D} = \{(\mathbf{X}_b, \mathbf{M}_b^{\text{orig}}, \mathbf{t}_b, y_b)\}_{b=1}^N$ denote the ICU records. Record b has T_b valid time steps and V variables, with $\mathbf{X}_b \in \mathbb{R}^{T_b \times V}$, $\mathbf{M}_b^{\text{orig}} \in \{0,1\}^{T_b \times V}$, $\mathbf{t}_b \in \mathbb{R}^{T_b}$, and label y_b. We use $M_{b,t,v} = 1$ when variable v is observed at time t. SMILE treats $\mathbf{M}_b^{\text{orig}}$, before imputation or training corruption, as a representation signal and constructs $\mathcal{G}_b = \{\boldsymbol{\rho}_b, \mathbf{C}_b\}$ from local first-order and record-level second-order statistics.

Structured feature lift. Because $\mathcal{G}_b$ is a deterministic function of the mask, augmenting $(\mathbf{X}_{b,\text{obs}}, \mathbf{M}_b^{\text{orig}}, \mathbf{t}_b)$ with $\mathcal{G}_b$ does not change the information available to an unrestricted predictor. SMILE instead supplies a finite-capacity inductive bias: $\boldsymbol{\rho}_b$ exposes local degree-one mask statistics and $\mathbf{C}_b$ exposes aggregated degree-two interactions (Le Morvan et al. 2020).

First-order monitoring intensity. Let $\widetilde{M}_{b,k,v}^{\text{orig}} = M_{b,k,v}^{\text{orig}}$ for $1 \le k \le T_b$ and zero otherwise. For $w = 2r + 1 = 5$, the local observation density is the length-preserving boxcar average

$$\rho_{b,t,v} = \frac{1}{w} \sum_{\ell=-r}^{r} \widetilde{M}_{b,t+\ell,v}^{\text{orig}}. \tag{1}$$

Unlike a pointwise mask value, $\rho_{b,t,v}$ distinguishes locally dense from sparse monitoring. Zero padding on both sides matches the implementation and prevents access beyond the available trajectory.

Second-order Gram geometry. Let $\mathbf{R}_b = \mathbf{1} - \mathbf{M}_b^{\text{orig}}$ and $\mathbf{Z}_b = \mathbf{R}_b - \mathbf{1}\mathbf{q}_b^\top$. SMILE defines

$$\begin{aligned} \mathbf{q}_b &= \frac{1}{T_b} \mathbf{R}_b^\top \mathbf{1}, \\ \boldsymbol{\phi}_b(v) &= T_b^{-1/2} \mathbf{Z}_{b,:,v}, \\ C_{b,uv} &= \langle \boldsymbol{\phi}_b(u), \boldsymbol{\phi}_b(v) \rangle, \\ \mathbf{C}_b &= \frac{1}{T_b} \mathbf{R}_b^\top \mathbf{R}_b - \mathbf{q}_b \mathbf{q}_b^\top = \frac{1}{T_b} \mathbf{Z}_b^\top \mathbf{Z}_b. \end{aligned} \tag{2}$$

The map $\boldsymbol{\phi}_b$ embeds each variable as its centered observation trajectory. Its Gram entry $C_{b,uv}$ measures whether variables u and v are synchronously missing more or less often than expected from their patient-specific marginal rates. The Gram form proves that $\mathbf{C}_b$ is symmetric positive semidefinite. Centering the observation indicators gives $-\mathbf{Z}_b$ and therefore the same Gram matrix as centering the missingness indicators. We accordingly refer to $\mathbf{C}_b$ as the centered co-missingness matrix throughout. It induces the patient-specific pseudometric

$$d_b^2(u,v) = C_{b,uu} + C_{b,vv} - 2C_{b,uv} = T_b^{-1} \|\mathbf{Z}_{b,:,u} - \mathbf{Z}_{b,:,v}\|_2^2. \tag{3}$$

The model uses the Gram similarities $C_{b,uv}$ directly rather than this induced distance; the pseudometric makes explicit the geometry underlying those similarities. Centering subtracts the product-of-marginals independence baseline but does not remove all marginal-rate dependence. Because $\mathbf{C}_b$ is invariant to time-bin permutations, it represents synchronous association rather than temporal order; proofs, bounds, and invariances are in the supplement.

Observation-Conditioned Interaction

SMILE preserves the backbone's temporal path and modifies its variable path. The input encoder first projects value, observation state, and local density:

$$\mathbf{e}_{b,t,v} = \mathrm{MLP}\big(\mathrm{stack}(x_{b,t,v}, m_{b,t,v}, \rho_{b,t,v})\big). \tag{4}$$

Here m is $\mathbf{M}^{\mathrm{recon}}$ for online pretraining and $\mathbf{M}^{\mathrm{input}}$ otherwise; geometry uses $\mathbf{M}^{\mathrm{orig}}$. For attention head h, a summary of the valid elapsed-hour grid produces a gate $g_b^{(h)}$, and the zero-initialized pairwise scale $S^{(h)}$ yields

$$\begin{aligned} B_{b,ij}^{(h)} &= g_b^{(h)} S_{ij}^{(h)} \langle \boldsymbol{\phi}_b(i), \boldsymbol{\phi}_b(j) \rangle = g_b^{(h)} S_{ij}^{(h)} C_{b,ij}, \\ \alpha_{b,ij}^{(h)} &= \frac{\pi_{b,ij}^{(h)} \exp(B_{b,ij}^{(h)})}{\sum_k \pi_{b,ik}^{(h)} \exp(B_{b,ik}^{(h)})}, \end{aligned} \tag{5}$$

where $\pi^{(h)}$ is content-only variable attention. Although $\mathbf{C}_b$ is a symmetric positive-semidefinite Gram kernel, $S^{(h)}$ is unconstrained and calibrates its entries into a head-specific, potentially directed interaction energy. Therefore $\mathbf{B}_b^{(h)}$ is geometry-derived but is not required to remain symmetric or positive semidefinite.

Geometry-calibrated interaction. For fixed (b, h, i) and $\boldsymbol{\pi}_{b,i}^{(h)}$ with positive entries,

$$\boldsymbol{\alpha}_{b,i}^{(h)} = \underset{\boldsymbol{a} \in \Delta^{V-1}}{\arg\max} \left\{ \mathbb{E}_{j \sim \boldsymbol{a}} \left[B_{b,ij}^{(h)} \right] - \mathrm{KL}\left(\boldsymbol{a} \| \boldsymbol{\pi}_{b,i}^{(h)} \right) \right\}. \tag{6}$$

The maximizer is unique, and the optimum equals $\log \sum_j \pi_{b,ij}^{(h)} \exp(B_{b,ij}^{(h)})$; the supplement gives the proof. Thus $\mathbf{C}_b$ supplies the patient-specific geometric substrate, while $S^{(h)}$ and $g_b^{(h)}$ map it to a decision energy. The content attention $\boldsymbol{\pi}_{b,i}^{(h)}$ provides the reference distribution: the optimum favors geometrically compatible targets when their calibrated energy justifies deviating from content evidence, and the KL term controls that deviation. This characterization explains the operator without claiming a prediction or generalization guarantee; those questions are evaluated empirically.

Position-wise affine parameters from the same elapsed-grid encoding then modulate the attention update. The public loaders provide a regularized hourly grid, so this conditioning does not model raw event times. Setting the new input weights, pairwise scales, and affine generators to zero recovers the backbone. With $P = T + 1$ positions (including the summary token), L blocks, width d, and density window w, the dual-attention backbone costs $O(Ld(VP^2 + PV^2) + LVPd^2)$ per record. Constructing the observation geometry adds $O(TVw + TV^2)$ time and $O(TV + V^2)$ memory, so it does not change the backbone's leading asymptotic order. Detailed cost accounting, parameter counts, and a controlled inference microbenchmark are in the supplement. Pooling, gating, and affine-update equations are also provided there.

Task	Train/Val/Test	V	$T_{\mathrm{med/max}}$	Miss. (%)	Test target
C12	9,590/1,198/1,200	37	47/48	79.6	15.9% pos.
C19	32,268/4,033/4,034	34	38/60	80.2	7.0% pos.
M-Mort	16,906/2,113/2,114	17	48/48	56.8	13.0% pos.
M-Dec	31,920/3,547/6,251	17	24/24	57.0	3.6% pos.
M-Pheno	33,515/4,189/4,190	17	48/48	58.3	25 labels
M-LOS	188,773/23,596/23,598	17	24/24	55.1	10 classes

Table 1: Dataset and task statistics under fixed split seed 42. T is median/maximum valid hourly record length, and missingness is pooled over valid dynamic entries across all splits. "Test target" gives positive prevalence for binary tasks and output dimensionality otherwise. M denotes MIMIC-III.

Two-Stage Training Strategy

We use three masks: $\mathbf{M}^{\mathrm{orig}}$ is the clean observation mask, $\mathbf{M}^{\mathrm{input}}$ is the mask channel after input dropout, and $\mathbf{M}^{\mathrm{recon}}$ is its visibility mask after pretraining corruption. Geometry always uses $\mathbf{M}^{\mathrm{orig}}$. The EMA target encoder receives original values with $\mathbf{M}^{\mathrm{input}}$; the online encoder receives hidden values with $\mathbf{M}^{\mathrm{recon}}$. Reconstruction uses $\Omega = \{(b, v, t) : M_{b,t,v}^{\mathrm{input}} = 1,\ M_{b,t,v}^{\mathrm{recon}} = 0\}$, and

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\Omega|} \sum_{(b,v,t) \in \Omega} \left\| \hat{\mathbf{h}}_{b,v,t} - \tilde{\mathbf{h}}_{b,v,t} \right\|_1, \tag{7}$$

where $\hat{\mathbf{h}}$ is the online reconstruction, $\tilde{\mathbf{h}}$ is the EMA target representation; this Ω excludes genuinely unobserved entries. Input dropout changes only the mask channel, ramping to 5% in pretraining and remaining at 5% in fine-tuning. Fine-tuning uses no reconstruction corruption and replaces the decoder with a prediction head; validation and test use $\mathbf{M}^{\mathrm{input}} = \mathbf{M}^{\mathrm{orig}}$. Binary tasks use binary cross-entropy, multi-class tasks use cross-entropy, and reconstruction loss is not combined with the supervised objective.

Auxiliary structured masking. As a secondary pretraining regularizer, the corruption process mixes random entry masking, contiguous temporal blocks, and physiological-system blocks, with the structured modes introduced gradually. This corruption produces $\mathbf{M}^{\mathrm{recon}}$ from $\mathbf{M}^{\mathrm{input}}$; only entries removed by this step contribute to Ω. The supplement gives the schedule, subsystems, and a random-only control, which retains the AUPRC gain over SMART on three of four binary tasks and identifies MIMIC mortality as the curriculum-sensitive exception.

Experiments

Experimental Protocol

The experiments address four questions: whether SMILE improves clinical prediction under a matched protocol, whether the learned representation remains useful when sensors are removed at test time, whether any gain depends on patient-specific observation geometry, and how the components of the interaction operator contribute.

Datasets and tasks. We evaluate PhysioNet 2012 (C12 mortality), PhysioNet 2019 (C19 sepsis), and four tasks

Model	C12		C19		MM		MD	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
AdaCare (Ma et al. 2020a)	40.54 ± 2.36	78.01 ± 1.16	41.13 ± 2.75	84.70 ± 0.43	41.26 ± 1.62	78.35 ± 1.12	52.41 ± 1.76	87.06 ± 1.23
StageNet (Gao et al. 2020)	47.01 ± 2.94	82.55 ± 1.75	59.11 ± 4.63	91.25 ± 1.50	51.77 ± 3.62	85.32 ± 0.62	64.97 ± 0.60	92.49 ± 0.47
ConCare (Ma et al. 2020b)	47.77 ± 3.62	82.74 ± 1.47	74.83 ± 3.47	94.39 ± 0.74	51.64 ± 3.34	84.97 ± 0.96	65.50 ± 1.30	92.43 ± 0.46
GRASP (Zhang et al. 2021a)	47.93 ± 5.73	82.98 ± 0.66	74.47 ± 3.79	94.43 ± 0.76	51.84 ± 3.42	84.79 ± 1.26	64.29 ± 0.64	92.25 ± 0.34
RainDrop (Zhang et al. 2021b)	37.23 ± 1.20	76.86 ± 1.67	72.82 ± 3.68	92.90 ± 0.79	46.21 ± 2.71	83.21 ± 0.99	61.16 ± 1.47	90.48 ± 1.60
SAFARI (Ma et al. 2022)	45.58 ± 3.59	81.84 ± 1.09	72.06 ± 1.74	94.64 ± 0.83	48.96 ± 3.01	84.48 ± 1.16	61.93 ± 1.57	90.69 ± 0.16
WarpFormer (Zhang et al. 2023)	48.10 ± 3.95	83.21 ± 2.09	73.03 ± 2.53	93.86 ± 0.77	51.49 ± 1.89	84.11 ± 1.10	68.21 ± 1.64	93.24 ± 0.96
PrimeNet (Chowdhury et al. 2023)	49.25 ± 2.36	84.19 ± 0.52	33.38 ± 2.29	83.60 ± 1.16	51.41 ± 0.89	84.82 ± 1.05	57.86 ± 1.32	91.15 ± 1.49
PPN (Yu et al. 2024b)	49.98 ± 3.97	83.42 ± 1.23	72.70 ± 4.99	94.25 ± 1.07	51.24 ± 2.96	84.90 ± 0.96	64.91 ± 0.33	91.50 ± 0.70
SMART (Yu et al. 2024a)	53.84 ± 2.24	85.80 ± 0.05	81.67 ± 0.84	95.92 ± 0.84	53.30 ± 0.12	85.61 ± 0.62	71.26 ± 0.81	94.20 ± 0.12
ATENet (Lee et al. 2025)	53.31 ± 2.02	85.54 ± 1.26	41.16 ± 3.02	84.02 ± 1.38	49.69 ± 1.15	85.29 ± 0.29	50.68 ± 0.83	90.96 ± 0.40
WaveGNN (Hajisafi et al. 2025)	49.40 ± 1.50	83.90 ± 1.20	57.10 ± 4.70	88.00 ± 0.90	49.22 ± 0.21	85.98 ± 0.38	72.19 ± 0.75	95.40 ± 0.29
ISTS-PLM (Zhang et al. 2025)	56.60 ± 3.30	85.60 ± 1.40	56.90 ± 5.00	89.40 ± 2.20	47.82 ± 1.66	85.11 ± 0.24	73.28 ± 1.38	95.16 ± 0.15
MissTSM (Neog et al. 2026)	43.80 ± 1.10	82.20 ± 0.50	56.50 ± 1.20	88.80 ± 1.30	43.34 ± 2.02	82.46 ± 0.81	63.05 ± 2.52	91.36 ± 1.40
SMILE(Ours)	**56.61 ± 0.97**	**86.02 ± 0.54**	**83.45 ± 1.42**	**96.30 ± 0.30**	**53.39 ± 4.79**	**86.00 ± 1.67**	**77.89 ± 1.03**	**96.31 ± 0.37**

Table 2: Binary-task threshold-free results (mean ± std, %). Rows are ordered by publication year, with SMILE(Ours) shown last. All methods use the same data splits and evaluation protocol; values for ATENet, WaveGNN, ISTS-PLM, and MissTSM are from our local adaptations with validation-AUPRC checkpoint selection and held-out test evaluation over seeds $\{1, 42, 3407\}$. Best values are in **bold**. MM: MIMIC mortality; MD: MIMIC decompensation.

Model	Phenotyping		Length of Stay	
	Micro	Macro	Micro	Macro
AdaCare (Ma et al. 2020a)	73.75 ± 0.26	63.28 ± 0.62	82.11 ± 0.40	66.06 ± 0.97
StageNet (Gao et al. 2020)	79.53 ± 0.31	73.33 ± 0.30	83.26 ± 0.15	69.54 ± 0.39
ConCare (Ma et al. 2020b)	78.03 ± 0.21	71.05 ± 0.36	83.19 ± 0.12	69.27 ± 0.50
GRASP (Zhang et al. 2021a)	76.97 ± 0.25	69.08 ± 0.47	83.20 ± 0.20	69.24 ± 0.47
RainDrop (Zhang et al. 2021b)	78.38 ± 0.42	71.59 ± 0.44	82.83 ± 0.18	68.23 ± 0.66
SAFARI (Ma et al. 2022)	75.69 ± 0.45	66.91 ± 0.84	82.93 ± 0.21	68.51 ± 0.48
WarpFormer (Zhang et al. 2023)	80.36 ± 0.23	74.65 ± 0.23	83.34 ± 0.28	69.53 ± 0.45
PrimeNet (Chowdhury et al. 2023)	76.66 ± 0.19	68.90 ± 0.17	82.89 ± 0.18	67.88 ± 0.40
PPN (Yu et al. 2024b)	76.55 ± 0.41	68.77 ± 0.54	83.13 ± 0.18	68.52 ± 0.48
SMART (Yu et al. 2024a)	81.41 ± 0.18	76.13 ± 0.14	83.92 ± 0.24	70.29 ± 0.44
ATENet (Lee et al. 2025)	75.87 ± 0.21	67.96 ± 0.27	77.60 ± 0.12	66.46 ± 0.29
WaveGNN (Hajisafi et al. 2025)	81.11 ± 0.04	75.66 ± 0.01	83.50 ± 0.20	70.00 ± 0.40
ISTS-PLM (Zhang et al. 2025)	80.21 ± 0.12	74.65 ± 0.10	83.00 ± 0.20	69.20 ± 0.40
MissTSM (Neog et al. 2026)	79.05 ± 0.11	72.95 ± 0.11	82.82 ± 0.04	72.25 ± 0.08
SMILE(Ours)	**82.51 ± 0.28**	**77.43 ± 0.19**	**84.46 ± 0.22**	**72.78 ± 0.39**

Table 3: Multi-class and multi-label AUROC results (mean ± std, %). Best values are in **bold**. Micro and macro denote micro- and macro-averaged AUROC. Values for ATENet, WaveGNN, ISTS-PLM, and MissTSM are from our local adaptations under the same data splits and evaluation protocol.

derived from the MIMIC-III Clinical Database v1.4: in-hospital mortality, decompensation, phenotyping, and length of stay (Johnson, Pollard, and Mark 2016; Johnson et al. 2016; Harutyunyan et al. 2019). MIMIC-III v1.4 contains deidentified critical-care records and is made available through PhysioNet under credentialed access, required training, and a Data Use Agreement (Johnson, Pollard, and Mark 2016; Goldberger et al. 2000). Its creation was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, with individual patient consent waived for the deidentified database (Johnson et al. 2016). Our study uses derived time-series task cohorts and reports aggregate results only. Results are mean ± standard deviation over seeds $\{1, 42, 3407\}$. The main binary-task table reports threshold-free AUROC and AUPRC. The supplement reports test-swept F1 and $\min(\mathrm{Se}, P^+)$ only as oracle operating-point diagnostics.

Baselines and metrics. We compare against recurrent, attention-based, graph-based, and self-supervised baselines. The matched comparisons use the same train/validation/test splits and evaluation code. For ATENet, WaveGNN, MissTSM, and ISTS-PLM, we locally adapt the released models to all four binary tasks and report complete three-seed results under the same splits, preprocessing, validation-AUPRC checkpoint selection, and held-out test evaluation; reproduction details are provided in the supplement. The missing-aware masked-autoencoder baseline is used both as the strongest matched baseline and as the implementation substrate for controlled comparison. AUROC measures global ranking and AUPRC is emphasized for imbalanced binary outcomes. For phenotyping and length of stay, we report micro- and macro-AUROC to distinguish sample-weighted from class-balanced behavior. All tabulated values are percentages.

Training and model selection. For MIMIC-III, mortality and phenotyping are generated through the MIMIC-III benchmark task pipeline (Harutyunyan et al. 2019); decompensation and length of stay use the documented local task-generation settings provided with our codebase. Within each

task cohort, all compared models use identical preprocessing and fixed train/validation/test splits. Values are normalized using training-set statistics. $\mathbf{M}^{\text{orig}}$ is one for measured entries before model-side corruption; training-only mask dropout and pretraining corruption follow the definitions above, while validation and test use the clean mask. For binary tasks, the classifier is selected by validation AUPRC; for phenotyping and length of stay, selection uses macro-AUROC and the configured length-of-stay AUROC protocol, respectively. The encoder dimension is 32, with two encoder layers, four attention heads, dropout 0.1, Adam optimization, learning rate 10^{-3}, batch size 64, and cosine learning-rate decay. Unless otherwise stated, pretraining uses 25 epochs and fine-tuning uses 25 epochs for the controlled SMILE/ablation runs, with five frozen fine-tuning warm-up epochs for the pretrained encoder. Experiments use seeds $\{1, 42, 3407\}$ and the same evaluation code across model variants. Code and generated scripts are organized under the accompanying SMART/SMILE workspace and will be released after de-anonymization.

The mechanism controls follow the same protocol. Architecture decisions use validation AUPRC only; after the design is locked, the selected checkpoints are evaluated once on the held-out test split. The permutation control preserves each patient-specific relation matrix but permutes variable identities, the cohort control uses a training-only running prior, and the capacity control widens the backbone without access to observation geometry.

Main Results

Tables 2 and 3 summarize the matched-protocol comparison across the six prediction tasks.

Binary event prediction. SMILE has the highest mean AUPRC and AUROC on all four binary endpoints in Table 2. The clearest gain over SMART occurs on MIMIC-III decompensation, where AUPRC increases by 6.63 points and AUROC by 2.11 points. C12 is numerically much tighter: SMILE exceeds ISTS-PLM by 0.01 AUPRC points and 0.42 AUROC points. On C19, SMILE improves over the strongest baseline, SMART, by 1.78 AUPRC points and 0.38 AUROC points. MIMIC-III mortality improves by only 0.09 AUPRC points over SMART and has substantial cross-seed variability, so we do not interpret this AUPRC difference as a stable gain. On the two matched MIMIC-III tasks, the strongest recent baselines reach 49.69 AUPRC and 85.98 AUROC for mortality, and 73.28 AUPRC and 95.40 AUROC for decompensation, compared with SMILE's 53.39/86.00 and 77.89/96.31.

Phenotyping and length of stay. Table 3 tests whether the gains extend beyond binary endpoints. SMILE improves phenotyping micro- and macro-AUROC from 81.41/76.13 to 82.51/77.43. For length of stay, the corresponding values increase from 83.92/70.29 to 84.46/72.78. The larger macro-AUROC gain on length of stay indicates a greater improvement in class-balanced than in sample-weighted ranking. Oracle operating-point diagnostics for the binary tasks are reported in the supplement.

Model	C12	C19	M-Mort	M-Dec
Capacity ctrl	55.49 ± 0.53	81.86 ± 0.28	51.14 ± 0.23	74.77 ± 0.89
Permutation	55.62 ± 0.44	82.28 ± 0.50	51.06 ± 0.75	77.59 ± 0.87
Cohort prior	55.02 ± 1.63	82.28 ± 0.97	49.21 ± 1.29	77.07 ± 0.32
Full SMILE	$\mathbf{56.61 \pm 0.97}$	$\mathbf{83.45 \pm 1.42}$	$\mathbf{53.39 \pm 4.79}$	$\mathbf{77.89 \pm 1.03}$

Table 4: Mechanism controls: test AUPRC (mean $\pm$ std over seeds 1/42/3407, %). Each control trails Full SMILE on all four tasks, with consistent paired-seed directions, so the ordering is uniform across endpoints. Best per column in **bold**.

Sensor-Removal Robustness

Protocol. We freeze each checkpoint and, without retraining, set the selected variables' values and mask columns to zero before evaluation. The perturbed mask becomes $\mathbf{M}^{\text{orig}}$ for that condition, so $\boldsymbol{\rho}_b$ and $\mathbf{C}_b$ are recomputed after removal. For each $k \in \{2, 3, 5, 7, 8\}$, five per-patient sensor subsets are drawn from a fixed manifest. We report absolute AUPRC and chance-adjusted retention relative to the intact input (k=0), averaged over seeds $\{1, 42, 3407\}$ with hierarchical-bootstrap intervals.

Results. At k=8, SMILE has the highest adjusted retention on MIMIC-III mortality (70.2%) and decompensation (78.9%). On C19, SMILE retains 91.6% and its paired AUPRC difference from SMART is inconclusive at the 95% level; ISTS-PLM and WaveGNN retain 65.3% and 51.3%, respectively. C12 is the boundary case: although SMILE leads on the intact input, ISTS-PLM has a small paired advantage at the most severe removal level, while SMILE remains competitive in absolute AUPRC. ATENet retains only 12–46% across the four endpoints. The stress test therefore supports robustness most clearly on the two MIMIC-III tasks and C19, rather than a universal advantage under every sensor-removal setting.

Ablation Study

Does patient-specific geometry matter? This experiment distinguishes patient-specific observation geometry from three simpler explanations. The capacity control enlarges the plain backbone without receiving $\boldsymbol{\rho}_b$ or $\mathbf{C}_b$. The permutation control applies the same symmetric row–column permutation to $\mathbf{C}_b$, preserving its entry distribution and magnitude while destroying the correspondence between clinical variables. The cohort control replaces $\mathbf{C}_b$ with a running prior estimated only from the training split. All controls use the same optimization budget, validation-AUPRC checkpoint selection, and seeds $\{1, 42, 3407\}$.

Full SMILE outperforms the capacity control, the permuted relation, and the cohort prior on all four binary tasks, with consistent paired-seed directions (Table 4). These comparisons show that added width and an arbitrary pre-softmax bias are not sufficient explanations. In particular, the gap over the cohort prior supports the use of record-level rather than fixed population relations, but does not by itself separate patient identity from all forms of sample-level variation. Density-window experiments over $w \in \{1, 3, 5, 7, 9\}$ and the no-density ablation likewise show that the first-order contribution is not an artifact of one receptive field. Together,

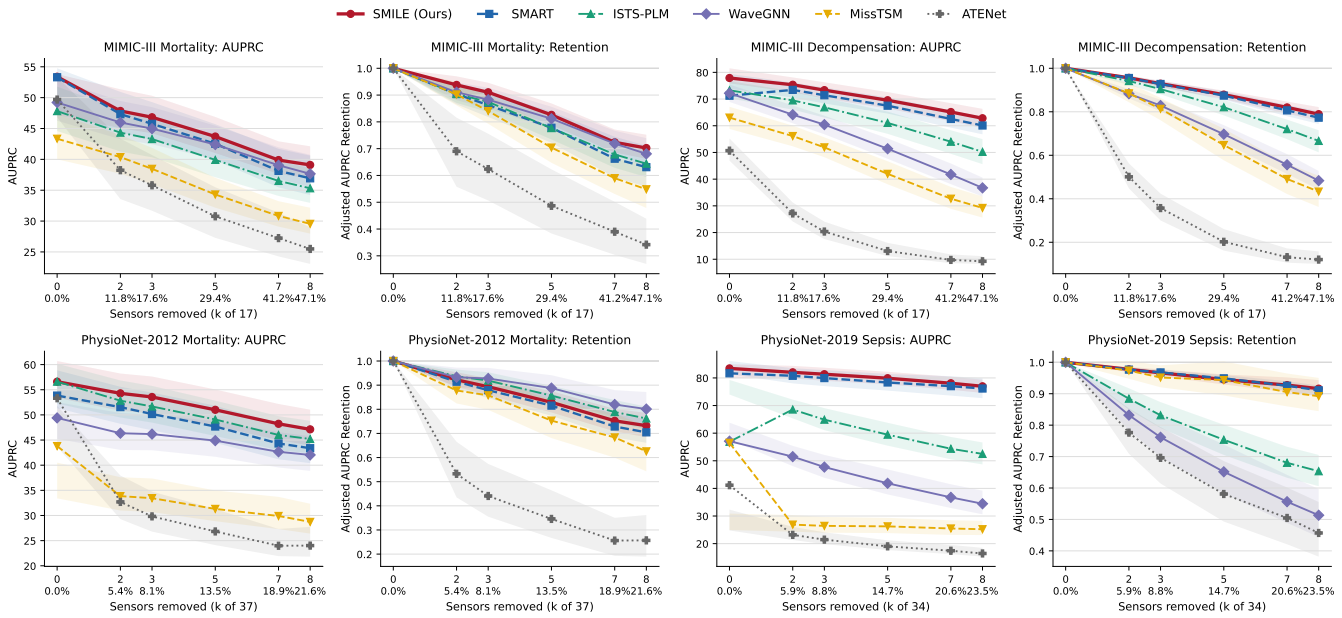

Figure 3: Sensor-removal robustness on all four binary endpoints. Top row: MIMIC-III mortality and decompensation (17 variables); bottom row: PhysioNet-2012 mortality (37 variables) and PhysioNet-2019 sepsis (34 variables). Each panel plots performance against the number of removed sensors k (the bottom axis also shows the removed fraction). "AUPRC" panels report absolute test AUPRC; "Retention" panels report AUPRC above prevalence retained relative to the intact input, $(\mathrm{AUPRC}_k - \pi)/(\mathrm{AUPRC}_0 - \pi)$, where π is test prevalence. Lines average seeds $\{1, 42, 3407\}$ and five sensor-subset replicates; shaded bands are 2,000-sample hierarchical-bootstrap 95% confidence intervals.

Configuration	C12	C19	M-Mort	M-Dec
SMILE-Full	**56.61 ± 0.97**	**83.45 ± 1.42**	**53.39 ± 4.79**	**77.89 ± 1.03**
SMILE w/o Density	54.89 ± 0.75	<u>82.39 ± 1.78</u>	48.78 ± 0.52	<u>77.28 ± 1.38</u>
SMILE w/o CoMiss Bias	<u>55.48 ± 0.30</u>	81.62 ± 1.14	<u>50.21 ± 0.24</u>	77.03 ± 1.25
SMILE w/o Time-Affine	54.89 ± 0.93	82.06 ± 0.84	49.14 ± 1.16	76.94 ± 1.89

Table 5: Component ablation: test AUPRC (mean ± std, %). Best values are in **bold**; second-best values are underlined. Full metrics and diagnostic ablations are in the supplement.

the controls provide converging evidence for the centered, patient-specific geometry as an effective signal within this architecture; they do not identify a causal missingness mechanism.

Component-wise comparison. The ablations decompose the unified operator rather than introducing three independent contributions. "w/o Density" removes the first-order statistic, "w/o CoMiss Bias" removes the second-order relation and its time gate, and "w/o Time-Affine" removes output conditioning while retaining the geometry-conditioned attention logits. Full SMILE improves over the backbone on all sixteen binary task–metric cells. The ordering among partial variants changes by task: removing the second-order co-missingness relation is least harmful on C12 and MIMIC mortality, whereas removing the first-order density is least harmful on C19 and MIMIC decompensation. These reversals show that the internal allocation of gain is task dependent, but they do not weaken the mechanism-control result: the matched permutation and cohort controls retain the same attention pathway yet remain below the patient-specific geometry. We therefore use the full operator as a fixed default rather than pruning it per endpoint. Full-metric component ablations, including operating-point diagnostics, are reported in the supplement.

Conclusion

SMILE models structured observation patterns in clinical time series as patient-specific observation geometry. It combines time-local observation density with a centered Gram kernel and converts its patient-specific similarities into calibrated cross-variable interaction energies. Across six ICU prediction tasks, SMILE obtains the best mean performance under the matched protocol, with the clearest gain on MIMIC-III decompensation. Sensor-removal tests support robustness most clearly on the two MIMIC-III tasks and C19, while C12 remains a boundary case. Capacity, permutation, cohort-prior, and component controls support record-level observation structure within the evaluated backbone, without establishing a causal missingness mechanism. Evidence is limited to fixed benchmark splits, three seeds, one dual-channel architecture, and regularized hourly grids; future work should test cross-cohort generalization, other encoders, raw event-time data, subgroup calibration, and robustness to shifts in measurement practice.

References

Agniel, D.; Kohane, I. S.; and Weber, G. M. 2018. Biases in Electronic Health Record Data Due to Processes within the Healthcare System: Retrospective Observational Study. *BMJ*, 361: k1479.

Cao, W.; Wang, D.; Li, J.; Zhou, H.; Li, L.; and Li, Y. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1): 6085.

Chen, H.; Zhang, J.; Yan, X.; Li, Z.; Lu, S.; and Tang, B. 2026. DynaMamba: Multi-Scale Dynamic Interacting Mamba Network for Irregular Clinical Time Series Classification. *Journal of Biomedical Informatics*, 178: 105027.

Chen, L.; Xiao, S.; Ding, S.; Hu, S.; and Sun, L. 2025. Integrating Sequence and Image Modeling in Irregular Medical Time Series Through Self-Supervised Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15821–15829.

Chen, Y.; Ren, K.; Wang, Y.; Fang, Y.; Sun, W.; and Li, D. 2023. ContiFormer: Continuous-Time Transformer for Irregular Time Series Modeling. In *Advances in Neural Information Processing Systems*, volume 36.

Cheng, M.; Tao, X.; Liu, Z.; Liu, Q.; Zhang, H.; Zhang, R.; and Chen, E. 2026. TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders. *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, 498–508. Conference Name: WSDM '26:The Nineteenth ACM International Conference on Web Search and Data Mining ISBN: 9798400722929.

Choi, E.; Bahadori, M. T.; Kulas, J. A.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3512–3520. Red Hook, NY, USA: Curran Associates Inc.

Chowdhury, R. R.; Li, J.; Zhang, X.; Hong, D.; Gupta, R. K.; and Shang, J. 2023. PrimeNet: Pre-training for Irregular Multivariate Time Series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7184–7192.

Cini, A.; Marisca, I.; and Alippi, C. 2022. FILLING THE G AP S: MULTIVARIATE TIME SERIES IMPUTATION BY GRAPH NEURAL NETWORKS. *International Conference on Learning Representations*.

Gao, J.; Xiao, C.; Wang, Y.; Tang, W.; Glass, L. M.; and Sun, J. 2020. StageNet: Stage-Aware Neural Networks for Health Risk Prediction. In *Proceedings of The Web Conference 2020*, 530–540.

Goldberger, A. L.; Amaral, L. A. N.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23): e215–e220.

Groenwold, R. H. H. 2020. Informative Missingness in Electronic Health Record Systems: The Curse of Knowing. *Diagnostic and Prognostic Research*, 4: 8.

Hajisafi, A.; Siampou, M. D.; Azarijoo, B.; Xiong, Z.; and Shahabi, C. 2025. WaveGNN: Integrating Graph Neural Networks and Transformers for Decay-Aware Classification of Irregular Clinical Time-Series. In *2025 IEEE International Conference on Big Data*.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask Learning and Benchmarking with Clinical Time Series Data. *Scientific Data*, 6(1): 96.

Horn, M.; Moor, M.; Bock, C.; Rieck, B.; and Borgwardt, K. 2020. Set Functions for Time Series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 4353–4363. PMLR.

Johnson, A.; Pollard, T.; and Mark, R. 2016. MIMIC-III Clinical Database. *PhysioNet*. Version 1.4.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3: 160035.

Le Morvan, M.; Prost, N.; Josse, J.; Scornet, E.; and Varoquaux, G. 2020. Linear Predictor on Linearly-Generated Data with Missing Values: Non Consistency and Solutions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 3165–3174. PMLR.

Lee, S.; Min, K.; Son, Y.; and Do, H. 2025. Adaptive Time Encoding for Irregular Multivariate Time-Series Classification. In *Advances in Neural Information Processing Systems*, volume 38.

Li, Z.; Li, S.; and Yan, X. 2023. Time Series as Images: Vision Transformer for Irregularly Sampled Time Series. In *Advances in Neural Information Processing Systems*, volume 36.

Liu, J.; Cao, M.; and Chen, S. 2026. Beyond Observations: Reconstruction Error-Guided Irregularly Sampled Time Series Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 23712–23720.

Ma, L.; Gao, J.; Wang, Y.; Zhang, C.; Wang, J.; Ruan, W.; Tang, W.; Gao, X.; and Ma, X. 2020a. AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 825–832.

Ma, L.; Zhang, C.; Wang, Y.; Ruan, W.; Wang, J.; Tang, W.; Ma, X.; Gao, X.; and Gao, J. 2020b. ConCare: Personalized Clinical Feature Embedding via Capturing the Healthcare Context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 833–840.

Ma, X.; Wang, Y.; Chu, X.; Ma, L.; Tang, W.; Zhao, J.; Yuan, Y.; and Wang, G. 2022. Patient Health Representation

Learning via Correlational Sparse Prior of Medical Features. *IEEE Transactions on Knowledge and Data Engineering*.

Neog, A.; Daw, A.; Fatemi, S.; Sawhney, M.; Pradhan, A.; Lofton, M. E.; McAfee, B. J.; Breef-Pilz, A.; Wander, H. L.; Howard, D. W.; Carey, C. C.; Hanson, P.; and Karpatne, A. 2026. Investigating a Model-Agnostic and Imputation-Free Approach for Irregularly-Sampled Multivariate Time-Series Modeling. *Transactions on Machine Learning Research*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*.

Qin, X.; Liu, M.; Wang, W.; Li, S.; Li, T.; Liu, X.; and Cheng, X. 2026. Beyond Missing Data Imputation: Information-Theoretic Coupling of Missingness and Class Imbalance for Optimal Irregular Time Series Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 24945–24953.

Rubin, D. B. 1976. Inference and Missing Data. *Biometrika*, 63(3): 581–592.

Sharafoddini, A.; Dubin, J. A.; Maslove, D. M.; and Lee, J. 2019. A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study. *JMIR Medical Informatics*, 7(1): e11605.

Shukla, S. N.; and Marlin, B. M. 2021. Multi-Time Attention Networks for Irregularly Sampled Time Series. ArXiv:2101.10318 [cs].

Singh, J.; Oshiro, K.; Krishnan, R.; Sato, M.; Ohkuma, T.; and Kato, N. 2019. Utilizing Informative Missingness for Early Prediction of Sepsis. In *Computing in Cardiology (CinC)*.

Singh, J.; Sato, M.; and Ohkuma, T. 2021. On Missingness Features in Machine Learning Models for Critical Care: Observational Study. *JMIR Medical Informatics*, 9(12): e25022.

Sisk, R.; Lin, L.; Sperrin, M.; Barrett, J. K.; Tom, B.; Diaz-Ordaz, K.; Peek, N.; and Martin, G. P. 2021. Informative Presence and Observation in Routine Health Data: A Review of Methodology for Clinical Risk Prediction. *Journal of the American Medical Informatics Association*, 28(1): 155–166.

Sun, C.; Song, M.; Cai, D.; Zhang, B.; Li, H.; and Hong, S. 2026. A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data. *Health Data Science*, 6: 0456.

Tang, P.; and Zhang, X. 2022. MTSMAE: Masked Autoencoders for Multivariate Time-Series Forecasting. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, 982–989.

Yu, Z.; Chu, X.; Jin, Y.; Wang, Y.; and Zhao, J. 2024a. SMART: Towards Pre-trained Missing-Aware Model for Patient Health Status Prediction. *Advances in Neural Information Processing Systems*, 37: 63986–64009.

Yu, Z.; Zhang, C.; Wang, Y.; Tang, W.; Wang, J.; and Ma, L. 2024b. Predict and Interpret Health Risk Using EHR Through Typical Patients. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1506–1510.

Zhang, C.; Gao, X.; Ma, L.; Wang, Y.; Wang, J.; and Tang, W. 2021a. GRASP: Generic Framework for Health Status Representation Learning Based on Incorporating Knowledge from Similar Patients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 715–723.

Zhang, J.; Zheng, S.; Cao, W.; Bian, J.; and Li, J. 2023. Warpformer: A Multi-scale Modeling Approach for Irregular Clinical Time Series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3273–3285. Long Beach CA USA: ACM.

Zhang, W.; Yin, C.; Liu, H.; and Xiong, H. 2025. Unleashing the Power of Pre-Trained Language Models for Irregularly Sampled Time Series. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 5499–5510.

Zhang, X.; Zeman, M.; Tsiligkaridis, T.; and Zitnik, M. 2021b. Graph-Guided Network for Irregularly Sampled Multivariate Time Series. In *International Conference on Learning Representations*.