DiffRisk: Diffusion Representation Learning with Informative Missingness for Health Risk Prediction

Shailesh Dahal Ratri Mukherjee Nicholas Mathews Kishlay Jha
Department of Electrical and Computer Engineering
University of Iowa
Iowa City, IA 52242
{shailesh-dahal, ratri-mukherjee, nicholas-mathews, kishlay-jha}@uiowa.edu

Abstract

Health risk prediction aims to predict a patient's potential health risks (e.g., mortality) using the longitudinal information present in their electronic health record (EHR). While deep learning-based risk prediction models have shown great promise, they are susceptible to the issue of missing data prevalent in real-world EHR. Most prior works mitigate this challenge by treating missing data as an uninformative perturbation to be ignored or imputed. However, inaccurate imputation may lead to the generation of sub-optimal patient representations that adversely impact risk prediction performance. Moreover, recent studies show that missing data in EHR does not necessarily imply error or noisy observation and instead may communicate informative clinical signals useful for health risk prediction. To address this, we propose a novel diffusion-based representation learning approach (namely **DiffRisk**) that explicitly captures *informative missingness* by modeling conditional dependencies induced by partial observations in EHR. Specifically, DiffRisk learns latent representations with coherent conditional structure, where any subset of observed features defines a distribution over the unobserved components. To achieve this, we introduce a conditional score-based objective and approximate it via denoising diffusion that enables efficient learning without explicit density estimation. Empirical results on five risk prediction tasks show that the proposed approach consistently outperforms state-of-the-art baselines. Our source code is available at `https://github.com/shaileshdahal/DiffRisk`.

1 Introduction

Deep learning-based risk prediction models [1, 2, 3] have shown great promise in leveraging the longitudinal information present in electronic health record (EHR) data to predict health outcomes such as mortality, Length of Stay, and adverse events. For example, models such as RETAIN [1] and Dipole [2] leveraged recurrent neural networks [4] to capture the temporality of EHR data valuable for risk prediction. Similarly, approaches such as BEHRT [5] and HiTANet [3] have leveraged the transformer architecture [6] for encoding the longitudinal information at a granular level, leading to improved risk prediction. Despite significant advances, existing risk prediction approaches are still susceptible to the issue of *missing data* prevalent in real-world EHR [7]. In healthcare settings, missing data can occur due to various reasons, including recording errors, irregular sampling, and inconsistent medical visits. Failing to properly account for missing data can degrade the quality of learned patient representations and result in inaccurate health risk predictions.

Prior works address the challenge of missing data in EHR by developing statistical or machine learning-based imputation strategies [8, 9, 10, 11] that aim to recover missing information potentially useful for health risk prediction. However, they still suffer from two key limitations: **(i)** Inaccurate imputations can introduce additional noise into the data, causing unintended shifts in the data

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

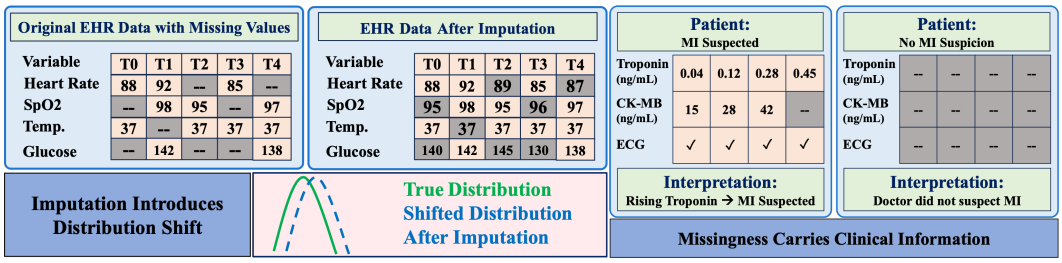


Figure 1: An illustration of EHR data imputation. (Left) Imputation introduces a distribution shift by filling missing values with potentially noisy estimates. (Right) Missingness carries clinical information: serial Troponin tests and Creatine Kinase-Myocardial Band (CK-MB) indicate suspected Myocardial Infarction (MI), while absent Troponin tests reflect no MI suspicion.

distribution (left block of Figure 1), which adversely impacts downstream risk prediction accuracy. **(ii)** Existing imputation methods mainly assume *missingness is at random*, which limits their potential in real-world healthcare applications, where missingness can be informative and actually reflect the underlying patient state. For example, multiple measurements for cardiac troponins - sensitive biomarkers of myocardial injury - can indicate that doctors suspect a cardiac-related diagnosis or complication. By the same intuition, missing measurements for cardiac troponins can indicate that the patient is not suspected of developing such complications, as the clinical guidelines only recommend this measurement for diagnosing myocardial injury [12] (see Figure 1, right). Leveraging this implicit information encoded in missingness patterns can enhance patient representation learning and improve health risk prediction; however, existing approaches have largely overlooked it.

To address the above challenges, we propose a novel representation learning approach (namely **DiffRisk**) that exploits the *informativeness of missing data* to learn patient representations that are both robust and offer superior predictive performance. Specifically, the proposed approach learns patient representations that characterize conditional relationships induced by partial observations in longitudinal EHR data. Explicitly modeling these conditional dependencies encourages the model to learn patient representations with well-defined, learnable distributions that remain robust under variations in observation patterns, thereby improving generalization at inference time. To achieve this, we design a novel patient representation learning objective that enforces conditional structure in the latent space such that conditioning on any subset of observed components yields a coherent distribution over the remaining (unobserved) components in longitudinal EHR data. Specifically, this objective is formalized through a conditional score function that characterizes conditional dependencies between subsets of representations in the latent space. However, directly enforcing such score-based constraints during patient representation learning is intractable due to the unknown and complex data distributions underlying high-dimensional EHR data. To mitigate this, we leverage recent advances in diffusion-based learning [8, 13, 14, 15] and reformulate score estimation as a denoising problem. This enables efficient approximation of score functions without explicit density estimation. Unlike existing diffusion-based approaches [8, 10, 9] that operate directly in input space, DiffRisk performs conditional diffusion in latent space, where dimensionality is reduced, feature geometry is homogeneous, and gradients from the downstream prediction task directly shape the learned conditional structure. This design choice allows diffusion to serve as a principled mechanism for enforcing conditional structure in learned representations, while preserving uncertainty induced by missingness and avoiding explicit value reconstruction. We further provide theoretical analysis demonstrating how minimizing the denoising score loss induces learnable conditional structure in latent representations and why such representations remain stable under missingness shifts. Altogether, our contributions are summarized as follows:

- We address the challenge of missing data prevalent in real-world EHR data from a new perspective, i.e., rather than treating missing values as noise to be ignored or imputed, we exploit their informativeness to learn high-quality patient representations that greatly improve performance on various downstream health risk prediction tasks.
- The proposed approach models conditional relationships implicit in longitudinal EHR data to learn patient representations that remain robust under arbitrary shifts in missingness

patterns, thus improving generalization at inference time. Existing approaches designed for health risk prediction have largely overlooked capturing these conditional dependencies.

- Experiments conducted on five clinical tasks across benchmark EHR datasets show that the proposed approach consistently and significantly outperforms strong baseline algorithms. We provide theoretical analysis to rigorously support our claims and ablation studies to evaluate the contribution of model design choices.

2 Related Work

Risk prediction from longitudinal EHR data has been widely studied using recurrent, attention-based, and transformer architectures. Early models such as RETAIN [1], Dipole [2], and HiTANet [3] leverage temporal clinical patterns for outcome prediction. However, real-world EHR data are often sparse and irregularly observed, making these models sensitive to missing measurements and observation shifts. To address this, several statistical and machine learning-based approaches have been proposed. Existing methods for partially observed EHR data can be broadly grouped into discriminative and generative approaches. Discriminative methods directly optimize downstream prediction while incorporating missingness information. GRU-D [16] uses masks and time gaps, SAITS [17] models missingness with self-attention, Warpformer [18] handles irregular clinical time series across multiple temporal scales, and SMART [11] improves prediction through missing-aware pretraining. STraTS [19] further addresses sparse and irregularly sampled clinical time series through self-supervised representation learning over observed (time, variable, value) triplets. Although effective, these methods do not explicitly enforce conditional consistency among latent representations under different observed subsets. Generative models such as GP-VAE [20], CSDI [8], and MTSCI [14] instead model input distributions and reconstruct missing values, but their uncertainty is not necessarily aligned with downstream prediction and may introduce spurious correlations. PISA/SupNotMIWAE [21] considers supervised probabilistic imputation under informative missingness, while MPP-VAE [22] employs latent-variable modeling for structured missingness. Diffusion models, grounded in DDPM [23] and score-based generative modeling [24], have recently been explored for EHR generation, augmentation, imputation, multi-view learning, and representation learning. MedDiffusion [25], MedDiff [26], TabDDPM [27], ScoEHR [28], and Diffmv [29] use diffusion for generation, augmentation, or multi-view healthcare prediction, while RLEG [30] and D2R2 [31] study diffusion-based representation learning. Most of these models operate in input space for reconstruction or generation. DiffRisk instead uses diffusion as a latent-space denoising objective to learn task-aligned conditional score structure induced by partial observations, rather than imputing or generating missing values. See Appendix A for details.

3 Methodology

In this section, we introduce **DiffRisk**, which learns task-aligned representations under informative missingness by combining supervised prediction with latent conditional score learning.

3.1 Problem Setup

Let $\mathcal{D} = \{(\mathbf{X}^{(i)}, \mathbf{M}^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote N patient episodes, where $\mathbf{X}^{(i)} \in \mathbb{R}^{K\times L}$ contains K variables over L timesteps and $\mathbf{M}^{(i)} \in \{0,1\}^{K\times L}$ is the observation mask, with $M_{k,l}^{(i)} = 1$ if variable k is observed at time l and 0 otherwise. Each episode has an outcome $y^{(i)} \in \mathcal{Y}$, which may be binary, multiclass, or multilabel. Unlike standard datasets where missingness is uninformative noise, EHR missingness often deviates from the Missing Completely At Random (MCAR) assumption and exhibits structured dependencies induced by clinical measurement processes. We summarize this structure using the global observation rate $\rho^{(i)} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} M_{k,l}^{(i)}$ and the per-variable observation-rate vector $\boldsymbol{\delta}_i \in [0,1]^K$, where $\delta_i^k = \frac{1}{L}\sum_{l=1}^{L} M_{k,l}^{(i)}$. While $\rho^{(i)}$ measures overall observation, $\boldsymbol{\delta}_i$ summarizes variable-specific monitoring frequency, whereas the full observation mask $\mathbf{M}^{(i)}$ preserves temporal observation patterns. Let $\Omega = [K] \times [L]$ denote all variable-time positions and $\mathcal{S}_{\text{obs}}(\mathbf{M}) = \{(k,l) \in \Omega : M_{k,l} = 1\}$ the observed set. For any $\mathcal{S} \subseteq \Omega$, $\mathbf{X}^{\mathcal{S}} = \{X_{k,l} : (k,l) \in \mathcal{S}\}$ denotes the restricted entries. We partition $\mathcal{S}_{\text{obs}}(\mathbf{M})$ into conditioning and target subsets $(\mathcal{C}, \mathcal{T})$, where $\mathcal{C} \cup \mathcal{T} = \mathcal{S}_{\text{obs}}(\mathbf{M})$ and $\mathcal{C} \cap \mathcal{T} = \emptyset$. This partitioning is structured to emphasize informative

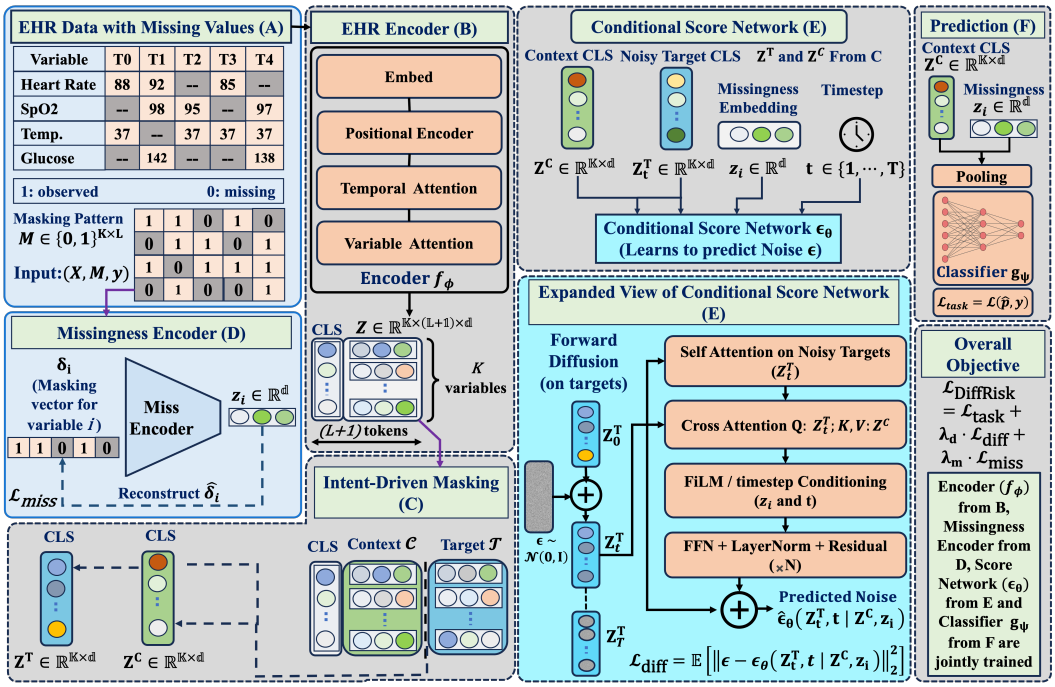


Figure 2: Overview of *DiffRisk*. Partially observed EHR data $(\mathbf{X}, \mathbf{M}, y)$ are encoded into latent representations, from which intent-driven masking forms conditioning and target CLS summaries $\mathbf{Z}^C$ and $\mathbf{Z}^T$. The missingness encoder maps $\boldsymbol{\delta}_i$ to $\mathbf{z}_i$. The score network denoises $\mathbf{Z}_t^T$ conditioned on $\mathbf{Z}^C$, $\mathbf{z}_i$, and t, while the prediction head uses $\mathbf{Z}^C$ and $\mathbf{z}_i$, and all components are trained jointly.

missingness and learn representations aligned with observation patterns. The encoder maps partially observed EHR data to latent representations as $f_\phi : \mathbb{R}^{K\times L} \times \{0,1\}^{K\times L} \to \mathbb{R}^{K\times(L+1)\times d}$, with $\mathbf{Z} = f_\phi(\mathbf{X}, \mathbf{M})$. Each variable has L temporal tokens and one learnable CLS token $\mathbf{Z}_{k,0} \in \mathbb{R}^d$ summarizing variable k's temporal dynamics. For a partition $(\mathcal{C}, \mathcal{T})$, we use $\mathbf{Z}^{\mathcal{C}} = \{\mathbf{z}_{k,l} : (k,l) \in \mathcal{C}\}$ and $\mathbf{Z}^{\mathcal{T}} = \{\mathbf{z}_{k,l} : (k,l) \in \mathcal{T}\}$ for token-level conditioning and target representations. We use compact per-variable CLS summaries $\mathbf{Z}^C, \mathbf{Z}^T \in \mathbb{R}^{K\times d}$ for score learning in Section 3.4; distinct from token-level restrictions $\mathbf{Z}^{\mathcal{C}}$ and $\mathbf{Z}^{\mathcal{T}}$.

3.2 Overview and Motivation

Given a partially observed EHR data $(\mathbf{X}, \mathbf{M})$, we encode the clinical measurements and observation mask into a latent tensor $\mathbf{Z} = f_\phi(\mathbf{X}, \mathbf{M}) \in \mathbb{R}^{K\times(L+1)\times d}$. The per-variable CLS representations, $\mathbf{Z}^C = \mathbf{Z}_{:,0,:} \in \mathbb{R}^{K\times d}$ summarizes temporal dynamics for downstream prediction. In parallel, the per-variable observation-rate vector $\boldsymbol{\delta}_i$ is encoded into an informative missingness context embedding $\mathbf{z}_i \in \mathbb{R}^d$. The prediction pipeline is $(\mathbf{X}, \mathbf{M}) \xrightarrow{f_\phi} (\mathbf{Z}^C, \mathbf{z}_i) \xrightarrow{g_\psi} \hat{\mathbf{p}}$, where $g_\psi : \mathbb{R}^{K\times d} \times \mathbb{R}^d \to \Delta^{|\mathcal{Y}|-1}$ produces a predictive distribution over outcomes. The standard supervised objective is $\mathcal{L}_{\text{task}} = \mathbb{E}_{(\mathbf{X},\mathbf{M},y)\sim\mathcal{D}}\big[\ell(g_\psi(\mathbf{Z}^C, \mathbf{z}_i), y)\big]$, where $\ell(\hat{\mathbf{p}}, y) = -\log \hat{p}_y$ is the cross-entropy loss. The overall architecture is shown in Figure 2. Predictive supervision alone does not constrain the internal geometry of $\mathcal{Z}_\phi$, the representation space induced by f_ϕ, or the conditional relationships among latent components, making representations sensitive to missingness shifts. DiffRisk therefore enforces conditional relationships in latent space rather than raw input space. Direct input-space modeling requires capturing interactions among KL heterogeneous (continuous, binary, categorical) variable-time entries, whereas latent-space modeling operates in a shared continuous space with $d \ll KL$, where Gaussian diffusion and score matching are tractable. Since f_ϕ is trained jointly with the downstream objective, the learned conditional structure is task-aligned rather than merely reconstructive. See Appendix B for further discussion. We next formalize this idea via conditional score-based learning.

3.3 Conditional Score-Based Representation Learning

Robust prediction under partial observability requires latent representations to support coherent conditional distributions across arbitrary partitions of observed positions. Directly enforcing such properties through density estimation is impractical in high-dimensional spaces due to intractable normalizing constants. DiffRisk therefore characterizes conditional structure through *score functions*, which describe distributions using gradients of log-densities without explicit density evaluation.

Score and Conditional Score Functions. Let $p(\mathbf{z})$ be a probability density over $\mathbf{z} \in \mathbb{R}^n$. Its *score function* is defined as $\mathbf{s}(\mathbf{z}) := \nabla_{\mathbf{z}} \log p(\mathbf{z}) = \frac{\nabla_{\mathbf{z}} p(\mathbf{z})}{p(\mathbf{z})} \in \mathbb{R}^n$, which describes the local geometry of the distribution. In partially observed EHR data, we are interested in conditional relationships among latent subsets. For a partition $(\mathcal{C}, \mathcal{T})$ of $\mathcal{S}_{\text{obs}}(\mathbf{M})$, the token-level *conditional score* is: $\mathbf{s}(\mathbf{Z}^{\mathcal{T}} \mid \mathbf{Z}^{\mathcal{C}}) := \nabla_{\mathbf{Z}^{\mathcal{T}}} \log p(\mathbf{Z}^{\mathcal{T}} \mid \mathbf{Z}^{\mathcal{C}}) \in \mathbb{R}^{|\mathcal{T}| \times d}$. This score captures how the log-probability of target representations changes with respect to $\mathbf{Z}^{\mathcal{T}}$, while holding the conditioning context $\mathbf{Z}^{\mathcal{C}}$ fixed.

Learnable Conditional Structure. Our goal is to learn latent representations whose conditional dependencies can be estimated across observed-target partitions. Let $\mathbf{Z} = f_\phi(\mathbf{X}, \mathbf{M}) \in \mathbb{R}^{K \times (L+1) \times d}$ denote the encoder-induced representation. Since $\mathbf{Z}$ is deterministic given $(\mathbf{X}, \mathbf{M})$, we consider the expectation over patient episodes and conditioning–target partitions sampled from $\pi(\cdot \mid \mathbf{M})$, where π denotes the distribution over partitions of $\mathcal{S}_{\text{obs}}(\mathbf{M})$. Specifically, the representation space admits learnable conditional structure if there exists a neural score network $\mathbf{s}_\theta : \mathbb{R}^{|\mathcal{T}| \times d} \times \mathbb{R}^{|\mathcal{C}| \times d} \to \mathbb{R}^{|\mathcal{T}| \times d}$ and a tolerance $\eta > 0$ such that

$$\mathbb{E}_{\substack{(\mathbf{X},\mathbf{M}) \sim \mathcal{D} \\ (\mathcal{C},\mathcal{T}) \sim \pi(\cdot \mid \mathbf{M})}} \left[\left\| \mathbf{s}_\theta(\mathbf{Z}^{\mathcal{T}}, \mathbf{Z}^{\mathcal{C}}) - \nabla_{\mathbf{Z}^{\mathcal{T}}} \log p_\phi(\mathbf{Z}^{\mathcal{T}} \mid \mathbf{Z}^{\mathcal{C}}) \right\|_2^2 \right] \leq \eta. \tag{1}$$

The tolerance η measures the expected conditional score approximation error over sampled episodes and partitions. Smaller η indicates more accurate estimation of conditional score structure in the encoder-induced representation space.

Conditional Score-Constrained Learning. Given a dataset $\mathcal{D}$, DiffRisk aims to learn representations that are both predictive of downstream outcomes and consistent with the conditional score property above. This gives the idealized constrained objective:

$$\min_{\phi, \psi} \mathcal{L}_{\text{task}}(\phi, \psi) \quad \text{subject to} \quad \mathcal{Z}_\phi \text{ satisfies Equation 1.} \tag{2}$$

Here, ϕ and ψ parameterize the encoder f_ϕ and classifier g_ψ, respectively, and $\mathcal{Z}_\phi$ denotes the representation space induced by f_ϕ. Since Equation 1 depends on the unknown conditional density $p(\mathbf{Z}^{\mathcal{T}} \mid \mathbf{Z}^{\mathcal{C}})$, directly enforcing this constraint is intractable. We therefore approximate it through latent denoising diffusion, which serves as a training mechanism for conditional score learning rather than as a generative model for imputing missing values.

3.4 Diffusion-Based Representation Learning Objective

Diffusion-Based Score Estimation. The objective in Equation 2 can be relaxed by introducing a *Lagrange multiplier* $\lambda \geq 0$ that penalizes deviations from the conditional score constraint:

$$\min_{\phi, \psi} \max_{\lambda \geq 0} \; \mathcal{L}_{\text{task}}(\phi, \psi) + \lambda \, d(\mathcal{Z}_\phi), \tag{3}$$

where $d(\mathcal{Z}_\phi)$ measures how far the encoder-induced representation space is from satisfying Equation 1. Since $d(\mathcal{Z}_\phi)$ depends on the unknown conditional score, we approximate it using *Denoising Score Matching* in latent space. Denoising score matching shows that learning to remove Gaussian noise from corrupted samples estimates the score of the underlying distribution without evaluating the density or its normalizing constant [32]. We adopt this principle within the multi-scale DDPM framework [23], using diffusion as a training mechanism.

Per-Variable CLS Realization. The score-structure bound in Equation 1 is stated at the token level for generality. In practice, we instantiate it at the variable-summary level using per-variable `CLS` representations: the conditioning summary $\mathbf{Z}^C \in \mathbb{R}^{K \times d}$ is obtained by applying the online encoder f_ϕ to the intent-masked input, while the target summary $\mathbf{Z}^T \in \mathbb{R}^{K \times d}$ is obtained by applying a separate exponential moving average (EMA) target encoder f_ξ to the full observed input. This yields a tractable, homogeneous space in which Gaussian diffusion is well-posed and the score network

can operate jointly across all K variables. The intent-masked input is constructed using the masking strategy described next.

Intent-Driven Masking. To construct $(\mathcal{C}, \mathcal{T})$, we partition $\mathcal{S}_{\text{obs}}(\mathbf{M})$ using an intent-driven strategy. Let $\mathcal{D}_c = \{i \in \mathcal{D}_{train} : y_i = c\}$ and define class-level prototypes $\delta_{\text{proto}}[c,k] = |\mathcal{D}_c|^{-1} \sum_{i \in \mathcal{D}_c} \delta_i^k$. For each patient and variable, we compute $\text{int}_i^k = |\delta_i^k - \delta_{\text{proto}}[y_i, k]|$, measuring how much the patient's observation pattern deviates from typical class behavior. Each observed cell receives score $s_{k,l}^{(i)} = u_{k,l} + \alpha \cdot \widetilde{\text{int}}_i^k$, where $u_{k,l} \sim \text{Uniform}(0,1)$, $\widetilde{\text{int}}_i^k = \text{int}_i^k / (\max_k \text{int}_i^k + \epsilon)$, and $\alpha \in [0,1]$ controls the blend between uniform and intent-driven masking. The highest-scoring observed cells form the target subset $\mathcal{T}$, while the remaining observed cells form $\mathcal{C} = \mathcal{S}_{\text{obs}}(\mathbf{M}) \setminus \mathcal{T}$. Naturally missing cells are never selected as denoising targets. Class labels are used only to construct training-time masking partitions and are never provided as model inputs. Neither labels nor class-level prototypes are required at inference. This procedure biases the learning objective toward informative missingness rather than uniformly sampling observed cells.

Informative Missingness Context. DiffRisk introduces an informative missingness context embedding $\mathbf{z}_i \in \mathbb{R}^d$ to summarize the patient's observation pattern. Given the per-variable observation-rate vector $\boldsymbol{\delta}_i$, a MissEncoder produces $\mathbf{z}_i = \text{MLP}_{\text{miss}}(\boldsymbol{\delta}_i)$. A reconstruction head $\hat{\boldsymbol{\delta}}_i = \sigma(\mathbf{W}_{\text{recon}} \mathbf{z}_i)$ preserves this information through a self-supervised loss: $\mathcal{L}_{\text{miss}} = \|\hat{\boldsymbol{\delta}}_i - \boldsymbol{\delta}_i\|_2^2$. Importantly, $\boldsymbol{\delta}_i$ is computed from the original observation mask $\mathbf{M}^{(i)}$, not from the further masked conditioning input, ensuring that $\mathbf{z}_i$ reflects the patient's true informative missingness pattern independent of how cells are assigned to the target set $\mathcal{T}$ during training.

Conditional Score Network. We apply a cosine variance schedule $\{\bar{\alpha}_t\}_{t=1}^T$ to the target representation $\mathbf{Z}^T$ as $\mathbf{Z}_t^T = \sqrt{\bar{\alpha}_t}\mathbf{Z}^T + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \text{Uniform}\{1, \ldots, T\}$, and $\bar{\alpha}_t = \prod_{s=1}^t (1-\beta_s)$ with $0 < \beta_1 < \cdots < \beta_T < 1$. A score network $\boldsymbol{\epsilon}_\theta$ predicts the injected noise given the noised target $(\mathbf{Z}_t^T)$, conditioning representation$(\mathbf{Z}^C)$, missingness context $(\mathbf{z}_i)$, and timestep (t). Diffusion loss is the inverse-frequency weighted denoising mean squared error (MSE):

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\frac{1}{K} \sum_{k=1}^{K} w_k \left\| \boldsymbol{\epsilon}_k - \boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i)_k \right\|_2^2 \right], \tag{4}$$

where $w_k = 1/(\bar{\delta}_k + 0.1)$ are normalized so $\frac{1}{K}\sum_k w_k = 1$, and $\bar{\delta}_k$ is the dataset-level mean observation rate for variable k. This weighting prevents frequently measured variables such as heart rate ($\bar{\delta}_k \approx 0.95$) from dominating the loss, while preserving gradient signal for rarely measured but clinically informative variables such as troponin ($\bar{\delta}_k \approx 0.05$). The score network conditions on two granularities: $\mathbf{z}_i$ provides patient-level FiLM [33] modulation, while $\mathbf{Z}^C$ provides per-variable cross-attention so each noisy target attends to its corresponding conditioning summary. To support the *Lipschitz assumption* in Section 3.6, every linear layer of $\boldsymbol{\epsilon}_\theta$ uses spectral normalization, enforcing $\text{Lip}(\boldsymbol{\epsilon}_\theta) \leq 1$ by composition. Minimizing $\mathcal{L}_{\text{diff}}$ estimates $\nabla_{\mathbf{Z}_t^T} \log p_t(\mathbf{Z}_t^T \mid \mathbf{Z}^C, \mathbf{z}_i)$ across noise scales, encouraging the representation space to satisfy Equation 1.

3.5 Training Objective

The final end-to-end unified training objective combines three complementary terms:

$$\mathcal{L}_{\text{DiffRisk}} = \mathcal{L}_{\text{task}} + \lambda_d \cdot \mathcal{L}_{\text{diff}} + \lambda_m \cdot \mathcal{L}_{\text{miss}}. \tag{5}$$

$\mathcal{L}_{\text{task}}$ from Section 3.2 drives clinically discriminative prediction; $\mathcal{L}_{\text{diff}}$ (Equation 4) enforces conditional score structure between $\mathbf{Z}^C$ and $\mathbf{Z}^T$ under informative missingness conditioning; and $\mathcal{L}_{\text{miss}}$ (Section 3.4) ensures the context embedding $\mathbf{z}_i$ faithfully summarizes $\boldsymbol{\delta}_i$. The score network parameters θ are optimized jointly with (ϕ, ψ) through the diffusion loss $\mathcal{L}_{\text{diff}}$. The encoder f_ϕ, score network $\boldsymbol{\epsilon}_\theta$, classifier g_ψ, and MissEncoder are optimized jointly. At inference, only the online encoder f_ϕ, MissEncoder, and classifier g_ψ are used; the target encoder, score network, and diffusion sampling are not required. See Appendix C for implementation details.

3.6 Theoretical Support for DiffRisk

We analyze why denoising score loss induces learnable conditional structure in latent space and why this structure improves stability under missingness shifts. Results are stated under mild regularity assumptions. The detailed proofs are provided in Appendix D.

Table 1: Statistics of clinical time-series datasets used in evaluation. Vars., Time, Obs., and Label Stat. denote the number of variables, horizon in hours, observation rate, and label statistics, respectively.

Dataset	Task	#Samples	Vars.	Time (hrs)	Obs. (%)	Label Stat.	Train/Val/Test
MIMIC-III	In-Hospital-Mortality	21,139	17	48	43.18	13.23	16,911 / 2,114 / 2,114
MIMIC-III	Decompensation	41,726	17	24	43.04	3.53	33,381 / 4,173 / 4,172
MIMIC-III	Phenotyping	41,902	17	60	41.72	16.51[†]	33,522 / 4,190 / 4,190
MIMIC-III	Length of Stay	33,352	17	24	43.65	10-class[‡]	26,681 / 3,335 / 3,336
PhysioNet	Cardiology	11,988	37	48	20.42	14.24	9,590 / 1,198 / 1,200

[†] Label density for multilabel Phenotyping. [‡] Length of Stay discretized into 10 bins: 1–8 days, 8–14 days, and >14 days.

Proposition 1 (**Score Loss Enables Conditional Score Approximation**). *Let* $(\mathbf{X}, \mathbf{M}) \sim \mathcal{D}$, $\mathbf{Z} = f_\phi(\mathbf{X}, \mathbf{M})$, *and let* $\mathbf{Z}^C, \mathbf{Z}^T \in \mathbb{R}^{K\times d}$ *denote the per-variable* `CLS` *conditioning and target representations. For the noised target* $\mathbf{Z}_t^T = \sqrt{\bar{\alpha}_t}\mathbf{Z}^T + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, *with* $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, *define*

$$\mathcal{L}_{\text{score},t}(\theta) = \mathbb{E}\Big[\big\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i)\big\|_2^2\Big]. \tag{6}$$

Let $\theta^\star$ *be a Bayes-optimal minimizer. If* $\mathcal{L}_{\text{score},t}(\theta) - \mathcal{L}_{\text{score},t}(\theta^\star) \le \eta$, *then*

$$\mathbb{E}\left[\left\|\boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i) + \sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{Z}_t^T}\log p_t(\mathbf{Z}_t^T \mid \mathbf{Z}^C, \mathbf{z}_i)\right\|_2^2\right] \le \eta. \tag{7}$$

Thus, minimizing the denoising objective yields a bounded-error estimator of the conditional score field $\nabla_{\mathbf{Z}_t^T}\log p_t(\mathbf{Z}_t^T \mid \mathbf{Z}^C, \mathbf{z}_i)$ *across noise scales.*

Interpretation. Accurate denoising directly translates to accurate conditional score estimation. Jointly training the encoder and score network encourages latent representations to admit reliable conditional structure under observed-target partitions.

Proposition 2 (**Robustness Under Missingness Shift**). *Let* $\mathbf{M}_1, \mathbf{M}_2$ *be two missingness patterns with observed sets* $\mathcal{S}_{\text{obs}}(\mathbf{M}_1)$ *and* $\mathcal{S}_{\text{obs}}(\mathbf{M}_2)$. *For* $r \in \{1, 2\}$, *define* $\bar{\mathbf{z}}_r = g_\psi(\mathbf{Z}_r^C, \mathbf{z}_r)$, *where* $\mathbf{Z}_r^C = f_\phi(\mathbf{X}, \mathbf{M}_r)_{:,0,:}$ *is the per-variable* `CLS` *representation under mask* $\mathbf{M}_r$ *and* $\mathbf{z}_r$ *is the corresponding missingness embedding. Under smoothness, bounded-variance, and Lipschitz assumptions, and assuming that the expected perturbation of the* `CLS` *representations caused by each changed observation position is bounded by* $C_1\sqrt{\eta}$, *if the conditional score approximation error in Proposition 1 is bounded by* η, *then*

$$\|\mathbb{E}[\bar{\mathbf{z}}_1] - \mathbb{E}[\bar{\mathbf{z}}_2]\|_2 \le C(\sqrt{\eta} + 1)\,|\mathcal{S}_{\text{obs}}(\mathbf{M}_1) \triangle \mathcal{S}_{\text{obs}}(\mathbf{M}_2)|\,, \tag{8}$$

where $\triangle$ *denotes symmetric difference and* $C > 0$ *depends on the Lipschitz constants of* g_ψ *and the missingness encoder, and the conditional latent distribution. The additive term accounts for the direct effect of changes in the missingness embedding.*

Interpretation. The bound shows that representation drift depends on the score approximation error η and the number of changed observation positions. Thus, accurate conditional score learning controls the `CLS`-related contribution to representation drift, while the missingness embedding introduces a separate bounded contribution.

4 Experiments

We evaluate DiffRisk on partially observed EHR data across five clinical prediction tasks. We address three research questions: **RQ1.** Does DiffRisk improve prediction under missing data? **RQ2.** Is DiffRisk robust to missingness changes? **RQ3.** Does the conditional diffusion score contribute to performance? Next, we concisely describe the datasets, evaluation metrics, and baseline algorithms. More detailed descriptions are provided in Appendices E, F, and G, respectively.

4.1 Experimental Settings

Datasets. Table 1 summarizes the benchmark EHR datasets. **PhysioNet Cardiology** [34] contains $11,988$ ICU episodes with 37 vital signs and biomarkers over a 48-hour window for In-Hospital

Mortality prediction. **MIMIC-III** [35] is a large-scale critical care database containing multivariate time series of 17 physiological variables from ICU admissions, on which we evaluate four tasks: In-Hospital Mortality (IHM), Decompensation, Phenotyping, and Length of Stay (LOS). Following prior works [36, 37, 11], we use an 80/10/10 train/validation/test split at the episode level. For MIMIC-III, we additionally evaluate a patient-disjoint split to assess sensitivity to patient overlap across partitions. To assess applicability beyond ICU time series, we add a non-ICU inpatient evaluation using the Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS) from 2000–2022 [38]. The cohort comprises 97,600 admissions involving Takotsubo syndrome, with 134 features and approximately 67% missing entries. We evaluate both cross-sectional representations and procedure-day longitudinal sequences without changing the DiffRisk architecture.

Evaluation Metrics. We evaluate the performance of DiffRisk using standard task-appropriate metrics that capture both discriminative ability and robustness under class imbalance. For binary tasks such as Cardiology, IHM, and Decompensation, we report F1 score, Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and the minimum of precision and recall min(Se,P+). For multilabel Phenotyping and multiclass LOS prediction, we report micro- and macro-averaged AUROC (MI-AUROC and MA-AUROC).

Implementation. We implement DiffRisk in PyTorch [39] on a single NVIDIA A30 GPU (24 GiB). The encoder is a two-layer transformer with hidden dimension 32, four heads, and dropout 0.1; the missingness encoder is a two-layer MLP; and the score network uses 500 diffusion steps with a cosine schedule. Input sequences are padded to task-specific lengths of 24, 48, or 60 timesteps. Models are trained for 50 epochs using Adam [40] with learning rate 5×10^{-4}, batch size 128, and loss weights $\lambda_d = 0.1$, $\lambda_m = 0.01$. Intent-driven masking samples ratios from $[0, 0.75]$ with $\alpha = 0.5$, and the target encoder uses EMA [41] momentum annealed from 0.996 to 1.0. Results are averaged over three seeds as mean $\pm$ std; details are in Appendix C.

Baselines. We compare DiffRisk with three groups of baselines. Architectural and missingness-aware models include AdaCare [42], StageNet [43], RainDrop [44], Warpformer [18], PrimeNet [45], and SMART [11]. Diffusion-based baselines include MedDiffusion [25], CSDI [8], and MTSCI [14]. Diffusion-based baselines are evaluated under a unified *Impute+Predict* protocol with a shared downstream predictor. All baselines are evaluated under consistent settings using publicly available implementations where possible to ensure a fair comparison. We additionally compare DiffRisk with STraTS [19], PISA/SupNotMIWAE [21], and MPP-VAE [22] on Cardiology and Decompensation. Details are provided in Appendix G.

4.2 Results

Main Results. Table 2 reports performance across five clinical tasks. DiffRisk achieves the best overall results across binary, multilabel, and multiclass settings, with the clearest gains on imbalance-sensitive metrics: AUPRC, F1, and min(Se,P+). On binary tasks, DiffRisk improves AUPRC over the strongest discriminative baseline SMART by $+5.90$ points on Cardiology, $+3.70$ on IHM, and $+4.98$ on Decompensation. A paired comparison over ten matched seeds on Cardiology further confirms the improvement over SMART ($t(9) = 3.80$, $p = 0.0042$; Appendix M). Several baselines are competitive in specific settings: SMART is consistently strong among discriminative models, AdaCare and StageNet perform well on Decompensation AUROC, and RainDrop/PrimeNet are competitive on LOS. However, their performance is less consistent across tasks and metrics, suggesting sensitivity to task type, class imbalance, and observation patterns. Diffusion-based baselines also show mixed behavior: CSDI and MTSCI perform reasonably on Cardiology AUPRC, confirming that diffusion captures useful EHR structure, but degrade on harder tasks such as Decompensation and Phenotyping, where task-aligned conditional dependencies are more critical. MedDiffusion performs weaker overall, suggesting that synthetic augmentation alone does not sufficiently shape predictive representations. DiffRisk outperforms the strongest diffusion baseline MTSCI by $+4.97$, $+5.37$, and $+9.69$ AUPRC points on Cardiology, IHM, and Decompensation, respectively. DiffRisk also shows smaller or comparable seed variance, indicating stable optimization despite joint encoder, score-network, and classifier training. Overall, these results support **RQ1**: DiffRisk improves prediction under missing data by learning task-aligned conditional structure rather than imputing missing values or relying only on discriminative supervision.

Non-ICU Inpatient Evaluation. On the HCUP NIS Takotsubo cohort, DiffRisk achieves 90.34 ± 0.54 AUROC and 42.92 ± 2.70 AUPRC using cross-sectional representations. With procedure-day

Table 2: Main results across five clinical tasks. Mean ± std over three seeds; best scores are bold and second-best scores are underlined. Scores are percentages.

Method	PhysioNet Cardiology				MIMIC-III In-Hospital Mortality			
	AUROC	AUPRC	F1	min(Se,P+)	AUROC	AUPRC	F1	min(Se,P+)
AdaCare	75.10 ± 0.50	39.35 ± 1.38	32.23 ± 5.46	39.74 ± 0.91	<u>85.52</u> ± 0.41	48.52 ± 2.02	36.35 ± 4.79	48.64 ± 0.96
StageNet	77.44 ± 0.95	43.89 ± 1.76	30.53 ± 2.17	44.51 ± 1.96	85.18 ± 0.32	49.07 ± 1.04	36.83 ± 1.62	47.33 ± 0.65
RainDrop	82.13 ± 0.41	44.05 ± 1.70	44.73 ± 1.37	40.31 ± 1.73	77.76 ± 0.26	37.91 ± 1.13	38.24 ± 0.44	33.52 ± 1.66
Warpformer	81.98 ± 1.90	45.00 ± 3.04	36.73 ± 3.62	44.30 ± 0.45	83.69 ± 0.54	47.91 ± 0.95	40.44 ± 2.73	48.69 ± 0.48
PrimeNet	84.98 ± 0.38	50.54 ± 0.66	40.54 ± 0.89	28.45 ± 1.49	81.93 ± 0.04	42.55 ± 0.36	36.42 ± 0.90	27.68 ± 1.06
MedDiffusion	81.82 ± 0.15	48.78 ± 0.17	35.78 ± 0.75	45.71 ± 0.27	83.37 ± 0.22	48.74 ± 0.25	<u>45.63</u> ± 0.52	47.89 ± 0.29
CSDI	84.12 ± 0.03	53.91 ± 0.33	41.49 ± 0.89	51.02 ± 0.26	83.73 ± 0.23	48.41 ± 0.29	41.88 ± 0.34	47.85 ± 0.20
MTSCI	84.26 ± 0.08	<u>54.35</u> ± 0.14	44.94 ± 0.94	51.68 ± 0.54	83.92 ± 0.13	48.68 ± 0.15	41.16 ± 0.55	47.35 ± 0.11
SMART	<u>85.67</u> ± 0.62	53.42 ± 2.52	<u>45.52</u> ± 2.06	<u>52.77</u> ± 1.55	84.88 ± 1.53	<u>50.35</u> ± 3.12	43.65 ± 3.01	<u>49.03</u> ± 2.42
DiffRisk (Ours)	**86.73** ± 0.44	**59.32** ± 0.43	**47.73** ± 3.23	**54.61** ± 1.11	**87.20** ± 0.81	**54.05** ± 2.45	**45.70** ± 3.24	**51.31** ± 1.98

Method	MIMIC-III Decompensation				MIMIC-III Phenotyping		MIMIC-III Length of Stay	
	AUROC	AUPRC	F1	min(Se,P+)	MI-AUROC	MA-AUROC	MI-AUROC	MA-AUROC
AdaCare	95.69 ± 0.19	72.03 ± 1.23	61.72 ± 3.00	67.02 ± 1.25	<u>81.55</u> ± 0.16	76.42 ± 0.13	80.34 ± 0.30	61.53 ± 0.33
StageNet	<u>95.97</u> ± 0.19	<u>72.76</u> ± 0.41	58.61 ± 1.04	<u>67.41</u> ± 1.70	79.59 ± 0.27	73.83 ± 0.44	80.27 ± 0.12	61.29 ± 0.36
RainDrop	89.92 ± 0.24	51.33 ± 0.75	40.54 ± 3.27	30.12 ± 3.95	77.78 ± 0.19	70.58 ± 0.28	81.65 ± 0.22	66.82 ± 0.30
Warpformer	82.93 ± 0.99	29.91 ± 2.83	15.21 ± 5.11	34.41 ± 2.57	76.35 ± 0.07	68.41 ± 0.13	80.16 ± 0.13	59.33 ± 0.31
PrimeNet	93.03 ± 0.14	59.15 ± 0.45	60.51 ± 0.45	44.08 ± 2.75	72.81 ± 0.18	61.92 ± 0.43	81.47 ± 0.03	65.58 ± 0.11
MedDiffusion	91.84 ± 0.22	60.86 ± 0.79	54.38 ± 0.60	55.01 ± 0.45	76.76 ± 0.62	68.29 ± 1.12	81.99 ± 0.47	67.53 ± 0.55
CSDI	92.15 ± 0.18	65.73 ± 0.35	55.37 ± 0.11	57.67 ± 0.38	78.87 ± 0.03	72.76 ± 0.62	81.68 ± 0.11	66.67 ± 0.84
MTSCI	91.71 ± 0.26	67.30 ± 0.14	56.78 ± 0.30	60.79 ± 0.58	79.37 ± 0.19	73.54 ± 0.28	81.82 ± 0.21	67.11 ± 0.45
SMART	93.62 ± 1.81	72.01 ± 2.23	<u>67.12</u> ± 0.64	65.01 ± 1.54	81.49 ± 0.08	<u>76.52</u> ± 0.10	<u>82.55</u> ± 0.14	<u>69.35</u> ± 0.64
DiffRisk (Ours)	**96.33** ± 0.21	**76.99** ± 0.32	**70.97** ± 0.49	**69.82** ± 0.62	**82.48** ± 0.59	**77.78** ± 0.83	**83.18** ± 0.12	**70.15** ± 0.30

Table 3: Ablation studies. Top: core design ablations on two benchmarks. Bottom: component and masking ablations using AUPRC. Scores are percentages.

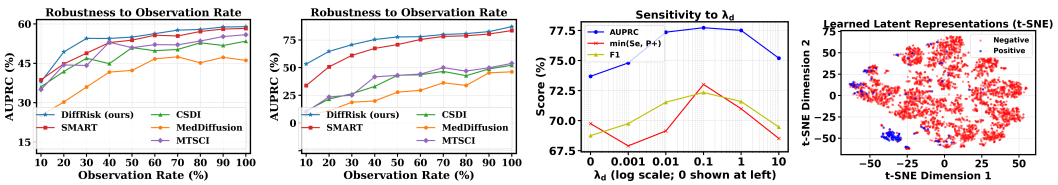
Variant	Cardiology			Decompensation		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1
DiffRisk (Full)	**86.73**	**59.32**	**47.73**	**96.33**	**76.99**	**70.97**
Unconditional (no value conditioning)	68.90	27.39	1.04	86.04	36.76	14.57
Random Partition (random $\mathcal{T}/\mathcal{C}$ split)	67.05	30.14	28.21	46.74	3.65	0.00
Partial Context (50% observed retained)	83.32	51.52	44.44	94.43	63.65	60.77
Partial Context (25% observed retained)	79.61	45.15	35.54	89.88	37.13	37.23
No Context (0% observed retained)	50.11	15.95	0.00	49.83	3.31	6.40

Dataset	Metric	Full	NoMissEnc	Δ	RandomObs Δ
MIMIC-III Mortality	AUPRC	52.75	49.50	-3.25	51.56 (-1.19)
PhysioNet Cardiology	AUPRC	56.34	54.55	-1.79	53.46 (-2.88)
MIMIC-III Decompensation	AUPRC	78.72	77.77	-0.95	76.66 (-2.06)

longitudinal sequences, it achieves 92.06 ± 0.56 AUROC and 53.50 ± 1.68 AUPRC. These results demonstrate the feasibility of applying DiffRisk beyond ICU physiological time series without architectural modification. Further cohort and preprocessing details are provided in Appendix E.

Additional Baseline Comparisons. Table 4 compares DiffRisk with STraTS, PISA/SupNotMIWAE, and MPP-VAE on Cardiology and Decompensation. DiffRisk achieves the highest AUPRC on both tasks, outperforming the strongest additional baseline by 4.95 and 5.18 percentage points, respectively. Implementation details are provided in Appendix G.

Core Design Ablation Study. The top panel of Table 3 shows that each major DiffRisk component contributes to performance. Removing conditional score learning sharply reduces results, confirming the importance of modeling dependencies between conditioning and target representations. Randomizing the $\mathcal{T}/\mathcal{C}$ split causes severe degradation, showing that missingness structure is informative rather than noise. Performance also drops as retained context decreases, indicating that DiffRisk effectively uses observed measurements for conditional inference. These results support **RQ2** and **RQ3**: robustness comes from missingness-aware conditional structure.



(a) Card. Robustness (b) Decomp. Robustness (c) λ_d Decomp. (d) t-SNE Decomp.

Figure 3: Robustness to observation-rate shifts, λ_d sensitivity, and representation visualization.

Table 4: Additional baseline comparisons. AUPRC (%), mean $\pm$ standard deviation over three seeds. Best scores are in bold.

Method	Cardiology	Decompensation
STraTS	51.11 ± 3.66	71.81 ± 3.39
PISA/SupNotMIWAE	54.37 ± 4.00	59.39 ± 1.66
MPP-VAE	44.24 ± 3.62	61.90 ± 2.56
DiffRisk (Ours)	$\mathbf{59.32 \pm 0.43}$	$\mathbf{76.99 \pm 0.32}$

Component and Masking Ablation Study. The bottom panel of Table 3 evaluates informative missingness modeling. Removing the missingness encoder consistently reduces AUPRC, with the largest drop on MIMIC-III Mortality (-3.25), showing that per-variable observation patterns improve prediction. Replacing intent-driven masking with RandomObs also degrades performance across tasks, confirming that uniform masking induces an MCAR-style bias and misses clinically informative missingness. These results support **RQ2** and **RQ3** by showing that DiffRisk benefits from both missingness encoding and structured masking.

Robustness, Sensitivity, and Qualitative Analysis. Figure 3 summarizes additional analyses. Figures 3a–3b show that DiffRisk consistently maintains higher AUPRC than SMART and diffusion-based imputation baselines across observation rates, with the largest margins in low-observation regimes where reconstructing missing values becomes unreliable. This supports **RQ2**, showing that latent conditional score learning improves robustness under missingness shifts. Figure 3c shows that performance improves as λ_d increases from near zero and peaks around $\lambda_d = 0.1$, indicating that moderate score regularization provides useful structural bias, while overly large weights can weaken the task objective. Finally, Figure 3d shows that DiffRisk learns a label-consistent latent geometry for Decompensation, with positive cases forming relatively compact regions despite severe class imbalance. Together, these results support **RQ3**: conditional score learning shapes the latent space beyond purely discriminative training. Further analyses are provided in Appendices H, J, and L.

5 Conclusion

In this research, we proposed **DiffRisk**, a diffusion-based representation learning approach that leverages the informativeness of missing data to generate high-quality representations useful for predicting the health status of patients. Specifically, DiffRisk encodes the conditional dependencies implicit in partial observations of longitudinal EHR data to learn patient representations that remain robust under missingness pattern shifts and offer better generalizability at test time. Experiments on five risk prediction tasks across benchmark EHR datasets validate the effectiveness of DiffRisk. Moreover, the ablation studies show that all components in DiffRisk contribute to the performance improvement. In future work, we will focus on the interpretability of DiffRisk to facilitate its integration into real-world clinical workflows.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers for their constructive comments. This work is supported in part by the U.S. National Science Foundation (NSF) and the National Institutes of Health (NIH) under grants SCH-2500344 and R01LM014012-01A1. Any recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NIH or NSF.

The authors have no competing interests to declare.

References

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

[2] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

[3] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 647–656, 2020.

[4] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.

[5] Yikuan Li, Shishir Rao, José 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.

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

[7] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3):1035, 2013.

[8] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

[9] Zongyu Dai, Emily Getzen, and Qi Long. Sadi: Similarity-aware diffusion model-based imputation for incomplete temporal ehr data. In *International Conference on Artificial Intelligence and Statistics*, pages 4195–4203. PMLR, 2024.

[10] Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. Pristi: A conditional diffusion framework for spatiotemporal imputation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 1927–1939. IEEE, 2023.

[11] Zhihao Yu, Chu Xu, Yujie Jin, Yasha Wang, and Junfeng Zhao. Smart: Towards pre-trained missing-aware model for patient health status prediction. *Advances in Neural Information Processing Systems*, 37:63986–64009, 2024.

[12] Andrew R Chapman, Philip D Adamson, and Nicholas L Mills. Assessment and classification of patients with myocardial injury and infarction in clinical practice. *Heart*, 103(1):10–18, 2017.

[13] Xinyu Yang, Yu Sun, Xiaojie Yuan, and Xinyang Chen. Frequency-aware generative models for multivariate time series imputation. *Advances in Neural Information Processing Systems*, 37:52595–52623, 2024.

[14] Jianping Zhou, Junhao Li, Guanjie Zheng, Xinbing Wang, and Chenghu Zhou. Mtsci: A conditional diffusion model for multivariate time series consistent imputation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3474–3483, 2024.

[15] Shailesh Dahal, Ratri Mukherjee, Nicholas Mathews, and Kishlay Jha. Taxodiff: Improving taxonomy completion with diffusion guided dynamic negative sampling. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM '26, page 122–131, New York, NY, USA, 2026. Association for Computing Machinery.

[16] 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.

[17] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

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

[19] Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

[20] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pages 1651–1661. PMLR, 2020.

[21] SeungHyun Kim, Hyunsu Kim, Eunggu Yun, Hwangrae Lee, Jaehun Lee, and Juho Lee. Probabilistic imputation for time-series classification with missing data. In *International Conference on Machine Learning*, pages 16654–16667. PMLR, 2023.

[22] Maksim Sinelnikov, Manuel Haussmann, and Harri Lähdesmäki. Latent variable model for high-dimensional point process with structured missingness. *arXiv preprint arXiv:2402.05758*, 2024.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[25] Yuan Zhong, Suhan Cui, Jiaqi Wang, Xiaochen Wang, Ziyi Yin, Yaqing Wang, Houping Xiao, Mengdi Huai, Ting Wang, and Fenglong Ma. Meddiffusion: Boosting health risk prediction via diffusion-based data augmentation. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 499–507. SIAM, 2024.

[26] H He, S Zhao, Y Xi, and JC Ho. Meddiff: generating electronic health records using accelerated denoising diffusion model. arxiv. *arXiv preprint arXiv:2302.04355*, 2023.

[27] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pages 17564–17579. PMLR, 2023.

[28] Ahmed Ammar Naseer, Benjamin Walker, Christopher Landon, Andrew Ambrosy, Marat Fudim, Nicholas Wysham, Botros Toro, Sumanth Swaminathan, and Terry Lyons. Scoehr: generating synthetic electronic health records using continuous-time diffusion models. In *Machine Learning for Healthcare Conference*, pages 489–508. PMLR, 2023.

[29] Chuang Zhao, Hui Tang, Hongke Zhao, and Xiaomeng Li. Diffmv: A unified diffusion framework for healthcare predictions with random missing views and view laziness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3933–3944, 2025.

[30] Liming Zhao, Kecheng Zheng, Yun Zheng, Deli Zhao, and Jingren Zhou. Rleg: Vision-language representation learning with diffusion-based embedding generation. In *International Conference on Machine Learning*, pages 42247–42258. PMLR, 2023.

[31] Ruoxue Liu, Linjiajie Fang, Wenjia Wang, and Bingyi Jing. D2r2: Diffusion-based representation with random distance matching for tabular few-shot learning. *Advances in Neural Information Processing Systems*, 37:36890–36913, 2024.

[32] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[33] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

[35] 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.

[36] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 833–840, 2020.

[37] 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.

[38] Healthcare Cost and Utilization Project (HCUP). HCUP National Inpatient Sample (NIS). Agency for Healthcare Research and Quality, Rockville, MD, 2000–2022. `https://hcup-us.ahrq.gov/nisoverview.jsp`.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[40] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[42] Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 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, pages 825–832, 2020.

[43] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of the web conference 2020*, pages 530–540, 2020.

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

[45] Ranak Roy Chowdhury, Jiacheng Li, Xiyuan Zhang, Dezhi Hong, Rajesh K Gupta, and Jingbo Shang. Primenet: Pre-training for irregular multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7184–7192, 2023.

[46] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 432–440. SIAM, 2016.

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

[48] Shuai Liu, Xiucheng Li, Gao Cong, Yile Chen, and Yue Jiang. Multivariate time-series imputation with disentangled temporal representations. In *The Eleventh international conference on learning representations*, 2023.

A Related Works

Risk prediction from longitudinal EHR data is fundamental to healthcare applications such as disease diagnosis, treatment planning, and medical resource allocation [46]. Early approaches, including RETAIN [1], Dipole [2], and HiTANet [3], primarily leveraged temporal clinical trajectories through recurrent, attention-based, or transformer architectures for downstream prediction. However, real-world EHR data are often sparse, irregularly sampled, and partially observed, making these models highly sensitive to missing measurements and observation shifts. To address this challenge, existing methods can generally be categorized into discriminative and generative approaches. Discriminative methods directly optimize predictive performance while incorporating missingness information into model design. For example, GRU-D [16] integrates masks and temporal decay, Phased-LSTM [47] models irregular event timing, Warpformer [18] captures multi-scale temporal irregularity, and SMART [11] improves robustness through missing-aware pretraining. While these methods enhance downstream prediction, they do not explicitly enforce conditional consistency among latent representations under varying observed subsets.

Generative approaches instead model the underlying distribution of partially observed data and reconstruct missing values probabilistically. GP-VAE [20], CSDI [8], and MTSCI [14] exemplify this paradigm, while PriSTI [10], FGTI [13], SADI [9], and TIDER [48] introduce additional structural inductive biases for improved imputation fidelity. More broadly, diffusion models grounded in DDPM [23] and score-based generative modeling [24] have recently been explored for EHR generation, augmentation, imputation, multi-view learning, and representation learning through methods such as MedDiffusion [25], MedDiff [26], TabDDPM [27], ScoEHR [28], DiffMV [29], RLEG [30], and D2R2 [31]. Despite their strong reconstruction or generative capabilities, most prior diffusion-based approaches operate primarily in input space and treat missingness as a nuisance variable to be imputed. In contrast, DiffRisk uses diffusion as a latent-space denoising objective to learn task-aligned conditional score structures induced by partial observations, explicitly modeling the semantic informativeness of missingness patterns rather than reconstructing missing values. This distinction enables improved robustness and generalizability for downstream risk prediction under distributional shifts in missingness.

B Conditional Modeling in Latent Space

For completeness, we restate and expand the argument from Section 3.2, providing additional intuition underlying representation-space conditional score learning. Following Section 3.1, let $\mathbf{Z} = f_\phi(\mathbf{X}, \mathbf{M}) \in \mathbb{R}^{K\times(L+1)\times d}$ denote the latent representation tensor produced by the encoder, with per-variable `CLS` summaries $\mathbf{Z}_{k,0} \in \mathbb{R}^d$. Modeling conditional structure directly in input space is problematic for three reasons. **First, statistical efficiency.** The input space is high-dimensional. Capturing conditional dependencies among raw measurements requires modeling interactions on the order of $O((KL)^2)$, which is statistically inefficient under the extreme missingness rates typical of EHR data (e.g., 20–40% in our benchmarks reported in Table 1). In contrast, the conditional score-based bound in Equation 1 operates in latent space and requires only $O(d^2)$ parameters with $d \ll KL$, allowing the conditional structure to be learnable from realistic sample sizes. **Second, homogeneity.** Clinical measurements are heterogeneous, consisting of continuous vital signs, binary indicators, and categorical labels with different scales and non-Gaussian noise characteristics. This heterogeneity induces complex conditional distributions in input space for which Gaussian denoising score matching in Equation 4 is ill-posed. The encoder f_ϕ maps all variables into a shared continuous space $\mathbb{R}^d$, where Gaussian diffusion is well-defined and the score network can operate jointly across all K variables. **Third, task alignment.** Modeling conditional structure in representation space allows it to be shaped by the downstream prediction objective. Because f_ϕ is trained jointly with the classifier g_ψ under the unified objective in Equation 5, the conditional dependencies captured by $\mathcal{L}_{\text{diff}}$ are directly informed by what is predictive of clinical outcomes, rather than reflecting generic correlations in the raw input. For these reasons, DiffRisk enforces conditional modeling in the representation space, where structure is statistically tractable, geometrically homogeneous, and task-aligned with the downstream clinical risk prediction objective.

B.1 Limitation of Prior Works

Existing approaches for learning from partially observed EHR data fall into two categories: discriminative and generative. Both exhibit structural limitations under partial observability that DiffRisk explicitly addresses through the conditional score-based objective in Equation 5.

B.1.1 Discriminative Approaches.

Discriminative models (e.g., GRU-D [16], SAITS [17], SMART [11]) learn a deterministic encoder $f_\phi : (\mathbf{X}, \mathbf{M}) \mapsto \mathbf{z} \in \mathbb{R}^d$ together with a downstream predictor h, trained by minimizing the supervised risk:

$$\min_{\phi} \ \mathbb{E}_{(\mathbf{X},\mathbf{M},y)\sim\mathcal{D}}\big[\ell\big(h(f_\phi(\mathbf{X},\mathbf{M})), y\big)\big]. \tag{9}$$

Limitation. This objective constrains representations only through predictive accuracy and imposes no explicit requirements on how latent components relate conditionally across variable-time positions. The learned representation $\mathbf{z}$ collapses all uncertainty induced by missing observations into a point estimate, leaving the representation space without any probabilistic or geometric constraints tied to the missingness pattern. This is precisely the gap formalized by Equation 1: discriminative encoders do not enforce any constraint analogous to the conditional score bound, so small changes in $\mathbf{M}$ at test time can produce uncontrolled latent space distortions. DiffRisk closes this gap by augmenting the supervised loss with $\mathcal{L}_{\text{diff}}$ (Equation 4), which enforces the score-structure constraint between the per-variable `CLS` representations $\mathbf{Z}^C$ and $\mathbf{Z}^T$. Furthermore, prior pretraining strategies typically mask positions uniformly at random, inducing a MCAR train-test mismatch with informative missingness in EHR data. DiffRisk avoids this via intent-driven masking (Section 3.4), where the partition $(\mathcal{C}, \mathcal{T})$ is biased toward observation patterns that are most informative about the clinical outcome.

B.1.2 Generative Approaches.

Generative methods (e.g., GP-VAE [20], CSDI [8]) instead model uncertainty directly in the input space by learning a conditional distribution $p_\theta(\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \mathbf{M})$ over missing values, optimized via a likelihood objective:

$$\min_{\theta} \ \mathbb{E}_{(\mathbf{X},\mathbf{M})\sim\mathcal{D}}\big[-\log p_\theta(\mathbf{X}^{\text{miss}} \mid \mathbf{X}^{\text{obs}}, \mathbf{M})\big]. \tag{10}$$

Limitation. While these approaches capture input-space variability, the learned uncertainty is not aligned with task-relevant decision boundaries. Accurate imputation does not necessarily preserve the information required for downstream prediction. Sampling plausible but incorrect values can introduce spurious correlations that degrade performance. Two structural issues compound this. First, generative methods operate in the input space, where heterogeneity across continuous, binary, and categorical variables makes Gaussian-based density modeling not well-suited (see Appendix B). Second, their objective is disconnected from the downstream classifier: the encoder is shaped by reconstruction fidelity rather than predictive utility. DiffRisk addresses both limitations. By enforcing conditional structure in latent space at the variable-summary level (Section 3.4), it operates in a homogeneous continuous domain. By training f_ϕ jointly with g_ψ under $\mathcal{L}_{\text{DiffRisk}}$ (Equation 5), the learned conditional dependencies are directly aligned with clinical prediction. Crucially, $\boldsymbol{\epsilon_\theta}$ is used as a training mechanism for learning conditional scores rather than as an imputation model, therefore no missing values are synthesized.

C Implementation Details

Encoder and missingness module. The encoder f_ϕ is a two-layer transformer with hidden dimension 32, four attention heads, and dropout 0.1. It maps $(\mathbf{X}, \mathbf{M})$ to $\mathbf{Z} \in \mathbb{R}^{K\times(L+1)\times d}$ using observed values, observation indicators, and temporal encodings. Each variable sequence is prepended with a learnable `CLS` token, producing fixed-dimensional per-variable summaries $\mathbf{Z}^C \in \mathbb{R}^{K\times d}$. Alternating temporal and cross-variable attention blocks capture dependencies across time and variables while remaining agnostic to missing entries. The MissEncoder is a two-layer MLP that maps the per-variable observation-rate vector $\boldsymbol{\delta}_i$ to a 32-dimensional missingness context embedding $\mathbf{z}_i$.

Conditional score network. The score network $\boldsymbol{\epsilon}_\theta$ is a transformer operating on noisy CLS targets $\mathbf{Z}_t^T \in \mathbb{R}^{K\times d}$ and conditioned on $\mathbf{Z}^C$, $\mathbf{z}_i$, and timestep t. Timestep embeddings are broadcast across variables, $\mathbf{z}_i$ modulates representations via FiLM, and $\mathbf{Z}^C$ provides per-variable cross-attention context, while self-attention captures cross-variable interactions. Spectral normalization is applied to all linear layers to enforce Lipschitz continuity, supporting the robustness analysis in Section 3.6. This architecture approximates the conditional score without explicit density modeling.

Classifier and training. The classifier g_ψ takes $\mathbf{Z}^C$ and $\mathbf{z}_i$, injects the missingness context through a learned projection, and produces predictions through a linear head. Cross-entropy is used for binary and multiclass classification, and binary cross-entropy is used for multilabel phenotyping. The encoder f_ϕ, MissEncoder, score network $\boldsymbol{\epsilon}_\theta$, and classifier g_ψ are trained end-to-end under the objective $\mathcal{L}_{\text{DiffRisk}} = \mathcal{L}_{\text{task}} + \lambda_d \mathcal{L}_{\text{diff}} + \lambda_m \mathcal{L}_{\text{miss}}$ from Equation 5. The target encoder f_ξ is initialized from f_ϕ and updated using an exponential moving average of its parameters, $\xi \leftarrow \mu\xi + (1-\mu)\phi$, without direct gradient updates. The momentum μ is annealed from 0.996 to 1.0 during training.

D Proof of Propositions

D.1 Proposition 1

Let $(\mathbf{X}, \mathbf{M}) \sim \mathcal{D}$ and let $\mathbf{Z} = f_\phi(\mathbf{X}, \mathbf{M}) \in \mathbb{R}^{K\times(L+1)\times d}$ denote the latent tensor produced by the encoder. The per-variable CLS representations $\mathbf{Z}^C, \mathbf{Z}^T \in \mathbb{R}^{K\times d}$ are constructed using the online encoder f_ϕ on the intent-masked input and the EMA target encoder f_ξ on the full observed input, respectively (Section 3.4). The forward noising process is $\mathbf{Z}_t^T = \sqrt{\bar\alpha_t}\,\mathbf{Z}^T + \sqrt{1-\bar\alpha_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K\times d})$. $\mathbf{z}_i \in \mathbb{R}^d$ is the missingness context produced by the MissEncoder. The single-timestep conditional denoising objective is

$$\mathcal{L}_{\text{score},t}(\theta) = \mathbb{E}\Big[\big\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i)\big\|_2^2\Big].$$

Proof. We prove the stated mean-squared learnability of the conditional score in four steps.

Step 1: Bayes-optimal predictor under squared loss. Fix t and define the random variables

$$Y := \boldsymbol{\epsilon} \in \mathbb{R}^{K\times d}, \qquad U := \big(\mathbf{Z}_t^T,\, \mathbf{Z}^C,\, \mathbf{z}_i\big).$$

The objective takes the form of a conditional regression problem,

$$\mathcal{L}_{\text{score},t}(\theta) = \mathbb{E}\big[\|Y - g_\theta(U)\|_2^2\big], \quad g_\theta(U) := \boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i). \tag{11}$$

The Bayes-optimal regressor over all measurable functions is the conditional expectation,

$$\boldsymbol{\epsilon}_{\theta^\star}(\mathbf{z}_t, t \mid \mathbf{c}, \mathbf{m}) = \mathbb{E}\Big[\boldsymbol{\epsilon} \;\Big|\; \mathbf{Z}_t^T = \mathbf{z}_t,\, \mathbf{Z}^C = \mathbf{c},\, \mathbf{z}_i = \mathbf{m}\Big]. \tag{12}$$

Step 2: Conditional Tweedie identity. Let $\sigma_t := \sqrt{1-\bar\alpha_t}$. Conditioning on $\mathbf{Z}^C = \mathbf{c}$ and $\mathbf{z}_i = \mathbf{m}$, the conditional density of $\mathbf{Z}_t^T$ is

$$p_t(\mathbf{z}_t \mid \mathbf{c}, \mathbf{m}) = \int p(\mathbf{z}^T \mid \mathbf{c}, \mathbf{m})\, \varphi_{\sigma_t}\big(\mathbf{z}_t - \sqrt{\bar\alpha_t}\mathbf{z}^T\big)\, d\mathbf{z}^T, \tag{13}$$

where φ_{σ_t} is the Gaussian density with covariance $\sigma_t^2\mathbf{I}$. Differentiating under the integral (justified by dominated convergence) and dividing by $p_t(\mathbf{z}_t \mid \mathbf{c}, \mathbf{m})$,

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{c}, \mathbf{m}) = -\frac{1}{\sigma_t^2}\,\mathbb{E}\Big[\mathbf{Z}_t^T - \sqrt{\bar\alpha_t}\mathbf{Z}^T \;\Big|\; \mathbf{Z}_t^T = \mathbf{z}_t,\, \mathbf{Z}^C = \mathbf{c},\, \mathbf{z}_i = \mathbf{m}\Big]. \tag{14}$$

Since $\mathbf{Z}_t^T - \sqrt{\bar\alpha_t}\mathbf{Z}^T = \sigma_t\boldsymbol{\epsilon}$ by the forward model,

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t \mid \mathbf{c}, \mathbf{m}) = -\frac{1}{\sqrt{1-\bar\alpha_t}}\,\mathbb{E}\big[\boldsymbol{\epsilon} \mid \mathbf{Z}_t^T = \mathbf{z}_t,\, \mathbf{Z}^C = \mathbf{c},\, \mathbf{z}_i = \mathbf{m}\big]. \tag{15}$$

Step 3: Identify the Bayes-optimal denoiser with the score. Combining (12) and (15),

$$\boldsymbol{\epsilon}_{\theta^\star}(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i) = -\sqrt{1-\bar\alpha_t}\,\nabla_{\mathbf{Z}_t^T} \log p_t(\mathbf{Z}_t^T \mid \mathbf{Z}^C, \mathbf{z}_i). \tag{16}$$

Step 4: Convert excess risk to score approximation bound. For squared loss, the excess risk equals the L^2 distance to the Bayes predictor,

$$\mathcal{L}_{\mathrm{score},t}(\theta) - \mathcal{L}_{\mathrm{score},t}(\theta^\star) = \mathbb{E}\Big[\big\|\boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i) - \boldsymbol{\epsilon}_{\theta^\star}(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i)\big\|_2^2\Big]. \tag{17}$$

The hypothesis $\mathcal{L}_{\mathrm{score},t}(\theta) - \mathcal{L}_{\mathrm{score},t}(\theta^\star) \le \eta$ together with (16) yields

$$\mathbb{E}\Big[\big\|\boldsymbol{\epsilon}_\theta(\mathbf{Z}_t^T, t \mid \mathbf{Z}^C, \mathbf{z}_i) + \sqrt{1-\bar{\alpha}_t}\,\nabla_{\mathbf{Z}_t^T} \log p_t(\mathbf{Z}_t^T \mid \mathbf{Z}^C, \mathbf{z}_i)\big\|_2^2\Big] \le \eta, \tag{18}$$

which is the stated claim. □

Inverse-Frequency Weighted Training Loss. The training objective in Equation 4 is the inverse-frequency weighted multi-timestep loss $\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}[(1/K)\sum_k w_k\| \cdot \|_2^2]$ with $w_k = 1/(\bar{\delta}_k + 0.1)$ normalized so $\frac{1}{K}\sum_k w_k = 1$. Each w_k is bounded between $w_{\min} = 1/(\max_k \bar{\delta}_k + 0.1)$ and $w_{\max} = 1/(\min_k \bar{\delta}_k + 0.1)$. The conclusion in Equation 18 therefore extends to the trained model with η replaced by $\eta/w_{\min}$, preserving its qualitative content. The unweighted form is used in Proposition 2 for clarity.

D.2 Proposition 2

Proof. Let $S_\cap = \mathcal{S}_{\mathrm{obs}}(\mathbf{M}_1) \cap \mathcal{S}_{\mathrm{obs}}(\mathbf{M}_2)$ denote shared observed positions and let $S_\Delta = \mathcal{S}_{\mathrm{obs}}(\mathbf{M}_1) \triangle \mathcal{S}_{\mathrm{obs}}(\mathbf{M}_2)$ denote positions whose observation status differs between the two missingness patterns. For $r \in \{1, 2\}$, the encoder produces the per-variable `CLS` representation $\mathbf{Z}_r^C = f_\phi(\mathbf{X}, \mathbf{M}_r)_{:,0,:} \in \mathbb{R}^{K\times d}$, and the missingness encoder produces $\mathbf{z}_r = \mathrm{MLP}_{\mathrm{miss}}(\boldsymbol{\delta}_r) \in \mathbb{R}^d$, where $\boldsymbol{\delta}_r$ is computed from $\mathbf{M}_r$. The aggregated patient representation is: $\bar{\mathbf{z}}_r = g_\psi(\mathbf{Z}_r^C, \mathbf{z}_r)$. By Proposition 1, minimizing the denoising score loss gives a bounded approximation to the conditional score, $\nabla_{\mathbf{Z}_t^T} \log p_t(\mathbf{Z}_t^T \mid \mathbf{Z}^C, \mathbf{z})$. Under the representation-stability assumption stated in Proposition 2, each changed observation position $j \in S_\Delta$ satisfies

$$\mathbb{E}\Big[\big\|\Delta\mathbf{Z}_j^C\big\|_2\Big] \le C_1\sqrt{\eta}, \tag{19}$$

where $\Delta\mathbf{Z}_j^C$ denotes the perturbation in the corresponding `CLS`-level representation induced by changing the observation status of position j. Intuitively, each mismatched observed position contributes at most $O(\sqrt{\eta})$ expected drift. Applying the per-position stability assumption successively over the changed positions and using the triangle inequality gives

$$\mathbb{E}\big[\big\|\mathbf{Z}_1^C - \mathbf{Z}_2^C\big\|_F\big] \le C_1\sqrt{\eta}\,|S_\Delta|. \tag{20}$$

Since g_ψ is Lipschitz, there exists $L_g > 0$ such that

$$\|g_\psi(\mathbf{U}, \mathbf{a}) - g_\psi(\mathbf{V}, \mathbf{b})\|_2 \le L_g\left(\|\mathbf{U} - \mathbf{V}\|_F + \|\mathbf{a} - \mathbf{b}\|_2\right). \tag{21}$$

Since the missingness encoder is Lipschitz with constant L_m, $\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \le L_m\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\|_2 \le (L_m/L)|S_\Delta|$, where L is the number of timesteps. Combining this with Equation 21 yields

$$\|\mathbb{E}[\bar{\mathbf{z}}_1] - \mathbb{E}[\bar{\mathbf{z}}_2]\|_2 \le L_g\left(C_1\sqrt{\eta} + \frac{L_m}{L}\right)|S_\Delta|. \tag{22}$$

Substituting the definition of S_Δ yields

$$\|\mathbb{E}[\bar{\mathbf{z}}_1] - \mathbb{E}[\bar{\mathbf{z}}_2]\|_2 \le C(\sqrt{\eta} + 1)\,|\mathcal{S}_{\mathrm{obs}}(\mathbf{M}_1) \triangle \mathcal{S}_{\mathrm{obs}}(\mathbf{M}_2)|\,. \tag{23}$$

Here, $C = L_g \max\{C_1, L_m/L\}$. This completes the proof. □

E Datasets

We evaluate DiffRisk on five clinical prediction tasks across two publicly available EHR benchmark datasets. Table 1 reports the statistics. **PhysioNet Cardiology** [34] dataset comprises $11,988$ ICU patient episodes with 37 vital signs and biomarkers recorded over a 48-hour window, with an observation rate of 20.42% and a positive event rate of 14.24%. The task is to predict In-Hospital

Mortality. **MIMIC-III** [35] is a large-scale critical care database containing multivariate time series of 17 physiological variables from ICU admissions. We conduct four tasks on MIMIC-III: In-Hospital Mortality (IHM), Decompensation, Phenotyping, and Length of Stay (LOS). These tasks share a common cohort but differ in label type and temporal horizon. IHM and Decompensation are binary classification tasks with positive rates of 13.23% and 3.53%, respectively, while Phenotyping is a multilabel task with 25 diagnostic categories and an average label density of 16.51%. LOS prediction is formulated as a 10-class classification task using 24-hour windows, comprising 33,352 samples. The remaining stay is discretized into 10 bins where bins 1–8 correspond to 1–8 days respectively, bin 9 represents 8–14 days, and bin 10 represents more than 14 days. We follow standard preprocessing protocols where observations are aligned to hourly intervals, continuous features are z-score normalized, and categorical features are one-hot encoded. Following prior works [36, 37, 11], we randomly split each dataset into training, validation, and test sets with an 80/10/10 ratio.

Non-ICU Inpatient Cohort. We additionally evaluate DiffRisk on a cohort constructed from the Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS) [38], comprising 97,600 Takotsubo syndrome admissions from general inpatient care between 2000 and 2022. After excluding outcome-related variables, identifiers, and diagnosis and procedure code strings, each admission contains 134 structured clinical features with approximately 67% missing entries. Missingness is preserved through the observation mask rather than imputed. We evaluate both a cross-sectional representation and a longitudinal within-admission representation constructed using procedure-day information. DiffRisk is applied in both settings without architectural modification.

Patient-Disjoint Evaluation. The main MIMIC-III experiments use episode-level splits, allowing admissions from the same patient to appear across partitions. To assess the effect of patient overlap, we additionally evaluate DiffRisk on Decompensation using a patient-grouped split, assigning all episodes from each patient to a single partition. Under this protocol, DiffRisk achieves 78.10 ± 0.78 AUPRC, compared with 76.99 ± 0.32 under the original episode-level split. The two results use different data partitions and are not directly comparable as a controlled performance improvement.

F Evaluation Metrics

This section provides the detailed formulations of the evaluation metrics used in our experiments. Let $y_i \in \{0, 1\}$ denote the ground-truth label for sample i, $\hat{y}_i \in \{0, 1\}$ denote the predicted label, and $s_i \in [0, 1]$ denote the predicted confidence score. Let TP, FP, TN, and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively.

F.1 Binary Classification Metrics

Accuracy. Accuracy measures the overall proportion of correctly classified samples defined as: $\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$. **Precision and Recall.** Precision measures the fraction of predicted positives that are correct, while recall measures the fraction of true positives that are correctly identified. Mathematically, $\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$, and $\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$. **F1 Score.** The F1 score is the harmonic mean of precision and recall, calculated as: $\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$. **Minimum of Precision and Recall.** To capture worst-case performance between false positives and false negatives, we report: $min(Se, P+) = \min(\text{Precision}, \text{Recall})$. **Area Under the Receiver Operating Characteristic Curve (AUROC).** AUROC measures the probability that a randomly selected positive sample receives a higher predicted score than a randomly selected negative sample, defined as: $\text{AUROC} = \mathbb{P}(s^+ > s^-)$, where s^+ and s^- denote the predicted scores for positive and negative samples, respectively. **Area Under the Precision-Recall Curve (AUPRC).** AUPRC is defined as the area under curve formed by plotting precision as a function of recall across all possible thresholds. Mathematically, $\text{AUPRC} = \int_0^1 \text{Precision}(r)\, dr$, where r denotes recall.

F.2 Multilabel and Multiclass Metrics

Let C denote the number of labels or classes. We define:

Micro-Averaged AUROC (MI-AUROC). MI-AUROC aggregates predictions across all labels and samples before computing the ROC curve: $\text{MI-AUROC} = \text{AUROC}\left(\bigcup_{c=1}^{C} \{(y_{i,c}, s_{i,c})\}_{i=1}^{N}\right)$, where $y_{i,c}$ and $s_{i,c}$ denote the ground-truth label and predicted score for class c. **Macro-Averaged**

AUROC (MA-AUROC). MA-AUROC computes the AUROC independently for each label and then averages across labels: $\text{MA-AUROC} = \frac{1}{C}\sum_{c=1}^{C} \text{AUROC}_c$.

G Baseline Algorithms

We compare our DiffRisk model against a diverse set of strong baselines that reflect the main directions in recent EHR modeling research for downstream clinical prediction under partial observability. These baselines are grouped into three categories. **Architectural and sequence modeling methods. AdaCare** [42] captures temporal patterns at multiple resolutions using dilated convolutions and adaptive feature recalibration. **StageNet** [43] is an LSTM-based model that modifies recurrent dynamics to explicitly model disease stage progression over time. **Missingness-aware and irregular time-series methods. RainDrop** [44] models dependencies between clinical variables through message passing on observation graphs, explicitly handling missingness patterns. **Warpformer** [18] addresses irregular sampling via sampling-frequency estimation and multi-scale attention. **PrimeNet** [45] introduces self-supervised pre-training for irregular time series by combining reconstruction and contrastive objectives in the input space. **SMART** [11] integrates missing-aware temporal and variable attention with latent-space self-supervised pre-training, and represents a strong state-of-the-art baseline across multiple clinical benchmarks. **STraTS** [19] learns representations of sparse and irregularly sampled clinical time series through self-supervised transformer modeling of observed measurement triplets.

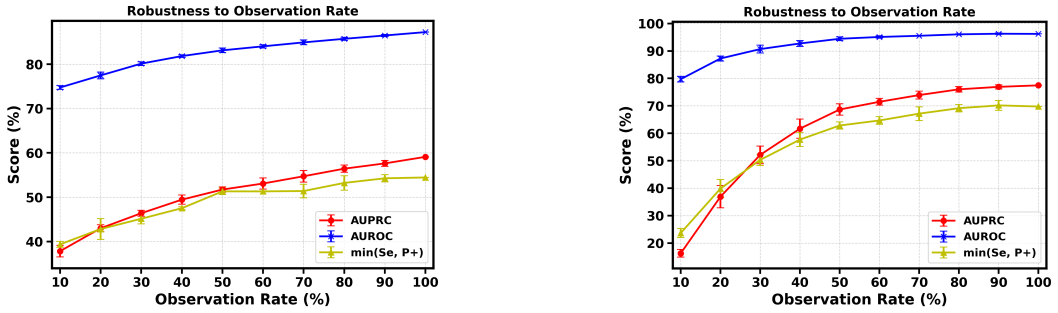
Generative imputation and augmentation methods. We additionally include diffusion-based approaches that operate via reconstruction or data generation. **CSDI** [8] and **MTSCI** [14] are state-of-the-art imputation models designed to reconstruct missing values in time series. Since their primary objective is value reconstruction rather than downstream prediction, we evaluate them under a unified *Impute+Predict* pipeline: each method first imputes missing values, and the completed sequences are then fed into a shared 2D transformer encoder and prediction head identical to that used in DiffRisk. We further include **MedDiffusion** [25], originally proposed for discrete ICD code augmentation. As it is not directly applicable to continuous EHR time series, we adapt its step-wise LSTM conditioning mechanism to continuous imputation using a CSDI-style training objective. The adapted model is evaluated under the same Impute+Predict pipeline as CSDI and MTSCI. **PISA/SupNotMIWAE** [21] uses supervised probabilistic imputation for time-series classification with missing observations. **MPP-VAE** [22] employs latent-variable modeling to capture structured missingness and is adapted to continuous time-series imputation under the shared Impute+Predict pipeline. Both methods are evaluated on Cardiology and Decompensation, alongside STraTS, using the same episode-level splits and three-seed protocol. Unlike these methods, which rely on architectural inductive biases or reconstruction-based objectives, DiffRisk directly shapes the geometry of the representation space by learning score-based conditional dependencies via diffusion-based training, leading to improved robustness under diverse missingness patterns. For fair comparison, all the experiments are conducted under consistent experimental settings, and baselines are reproduced using their publicly available implementations.

H Robustness to Missingness Rate

We evaluate robustness to varying levels of missingness by artificially masking additional observations at inference time. Figure 4 shows the performance of DiffRisk across observation rates on Cardiology and Decompensation. Across both tasks, DiffRisk exhibits smooth and monotonic improvements as observation rate increases, without abrupt degradation at low observation levels. Even under severe missingness (10–20% observed), the model maintains strong discriminative performance, indicating that it effectively leverages partial observations. As more observations become available, improvements are more pronounced in AUPRC and min(Se,P+), suggesting better precision-recall trade-offs rather than overfitting to dominant classes. The consistent trends across tasks demonstrate that DiffRisk learns stable representations under varying missingness patterns, addressing **RQ2**.

I Computational Efficiency

We report computational efficiency on MIMIC-III Mortality in Table 5. DiffRisk is trained jointly in a single stage and requires 296.7K parameters, with a total training time of 9.3 minutes. This



(a) PhysioNet Cardiology

(b) MIMIC-III Decompensation

Figure 4: Robustness of DiffRisk to variation in observation rate.

Table 5: Computational efficiency on MIMIC-III Mortality task. Params denotes the total parameters in the full model/pipeline. Train Time denotes total wall-clock training time across all stages, including pretraining/imputation and classifier training when applicable.

Method	Stages	Params	Train Time	Infer ms/sample	Peak GPU
DiffRisk	1	296.7K	9.3 min	0.342	1.08 GB
SMART	2	80.0K	8.6 min	0.306	1.23 GB
CSDI+Classify	2	618.5K	202.5 min	126.37	3.92 GB
MTSCI+Classify	2	618.5K	220.0 min	125.98	7.09 GB
MedDiffusion+Classify	2	742.7K	15.3 min	1.207	2.23 GB

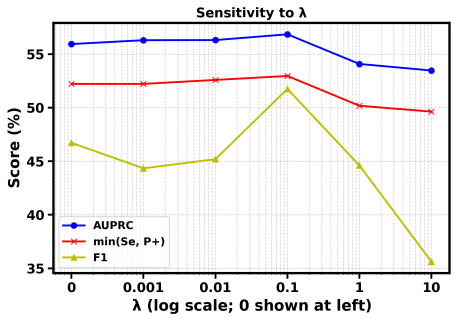
is comparable to SMART, which requires 8.6 minutes across pretraining and finetuning, while DiffRisk achieves stronger predictive performance with similar inference latency. The main efficiency advantage appears against diffusion-based *impute+classify* pipelines. CSDI and MTSCI require more than 200 minutes of training and around 126 ms per sample at inference because they must reconstruct missing values before prediction. MedDiffusion is lighter than CSDI/MTSCI due to its augmentation-style design and recurrent conditioning, but still has higher parameter count, memory usage, and inference latency than DiffRisk. In contrast, DiffRisk uses diffusion only as a training-time representation regularizer; no reverse diffusion or input-space imputation is performed at test time. Thus, DiffRisk obtains the benefit of diffusion-based conditional learning without the deployment cost of input-space diffusion.

J Sensitivity to λ_d for IHM

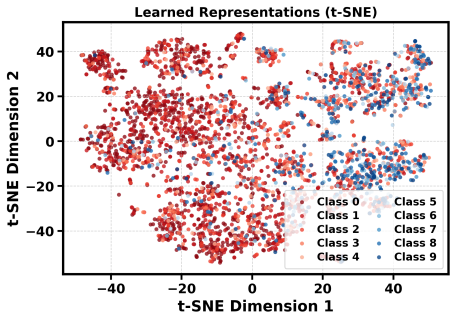
We investigate the impact of the trade-off parameter λ_d on the In-Hospital Mortality task. The performance on the IHM task follows a similar trend to that in Decompensation. Figure 5a shows that when λ_d is set close to zero, the model behaves similarly to a purely discriminative baseline and achieves noticeably lower AUPRC, F1, and min(Se,P+), indicating that ignoring conditional score constraints limits representation quality. As λ_d increases, performance consistently improves and peaks around $\lambda_d = 0.1$, suggesting that moderate score enforcement provides the most meaningful structural bias. Further increasing λ_d leads to gradual degradation, implying that excessive emphasis on score matching may overshadow the task objective. Overall, the stable peak around $\lambda_d = 0.1$ demonstrates that DiffRisk benefits from conditional score learning while remaining robust to moderate hyperparameter variation, supporting **RQ3**.

K Sensitivity Analysis for Missingness Loss Weight λ_m

Table 6 evaluates representative values of λ_m. Performance remains stable when moving from $\lambda_m = 0$ to the default setting $\lambda_m = 0.01$ and a larger value $\lambda_m = 0.1$. This indicates that $\mathcal{L}_{\text{miss}}$



(a) Sensitivity to λ_d for In-Hospital Mortality.



(b) t-SNE visualization for Length of Stay task.

Figure 5: Sensitivity analysis and representation visualization across tasks.

Table 6: Sensitivity to missingness reconstruction weight λ_m. The default value used in main experiments is $\lambda_m = 0.01$. Scores are percentages.

Dataset	Metric	$\lambda_m = 0$	$\lambda_m = 0.01$	$\lambda_m = 0.1$
PhysioNet Cardiology	AUPRC	56.45	**56.34**	56.33
MIMIC-III Mortality	AUPRC	52.81	**52.75**	54.80
MIMIC-III Phenotyping	Macro-AUC	75.03	**75.02**	74.95
MIMIC-III Decompensation	AUPRC	78.47	**78.72**	78.48
MIMIC-III LOS	Macro-AUC	69.55	**69.55**	69.61

provides useful regularization for the missingness embedding, while the overall framework remains robust across reasonable choices of λ_m.

L Qualitative Analysis

For the Length of Stay task (Figure 5b), which involves 10 ordinal classes representing different stay durations, the embedding reveals a clear global organization aligned with LOS duration. Lower LOS bins are predominantly concentrated toward one side of the embedding space, while higher LOS bins occupy the opposite side, with intermediate classes forming smooth transition regions.

M Paired Statistical Significance Analysis

We compare DiffRisk and SMART on Cardiology using ten matched random seeds. DiffRisk achieves higher AUPRC in all ten runs. A two-sided paired t-test yields $t(9) = 3.80$ and $p = 0.0042$, with a 95% confidence interval of $[0.70,\ 2.75]$ AUPRC percentage points for the paired improvement. This analysis uses a separate ten-seed evaluation from the three-seed summary in Table 2.

N Limitations

While DiffRisk shows promising results on benchmark EHR datasets, its robustness may suffer in EHR datasets where the observations are sparser, less regular, and influenced by different clinical workflows. Moreover, our experiments are conducted under simulated observation-rate shifts; however, these experiments do not exhaustively represent all forms of measurement bias, workflow variation, or institutional-specific data collection patterns. Further, while the multi-component architecture of DiffRisk is modular, it introduces computational overhead that may hinder deployment in low-resource clinical settings. Finally, similar to many existing deep-learning based patient representation learning methods, the learned latent representations may not always provide transparent or easily interpretable reasoning for individual predictions. This can reduce clinician trust and make it more difficult to understand whether predictions are driven by clinically meaningful evidence or mere correlations in the EHR data.

O Broader Impact

Deep learning-based risk prediction models are being increasingly used to support clinical decision-making and improve patient care. This work contributes to building robust risk prediction models that remain effective in the presence of missing data that are prevalent in real-world EHR. By explicitly modeling the informativeness of missing data, the proposed approach learns patient representations that are not only robust but also possess higher predictive power useful for clinical applications such as early risk assessment, patient deterioration monitoring, In-Hospital Mortality prediction, and early warning systems for critical conditions such as sepsis. Earlier identification of high-risk patients may help improve the timeliness of medical interventions and potentially contribute to better patient outcomes. Beyond its clinical use-case, the proposed approach may also be relevant to other domains involving partially observed data, such as scientific literature modeling, recommendation systems, longitudinal behavioral modeling, and sensor-based monitoring systems.

Despite numerous benefits, the proposed model should be deployed with careful consideration in the real-world clinical settings. Similar to existing risk prediction models, the proposed model may reflect existing biases related to patient demographics, healthcare access, documentation practices, or institutional workflows. As a result, prediction errors or biased risk estimates could unintentionally affect the quality or fairness of clinical decision-making for certain patient populations. Thus, the proposed model is intended to function as a clinical decision-support tool rather than an autonomous decision-making system. Its predictions should complement, not replace, physician expertise, clinical judgment, and patient-centered care. Any practical deployment should involve rigorous validation across diverse healthcare settings and careful human oversight to ensure that the system is used safely and responsibly in support of patient care.

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 main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The main claims are supported by the proposed method and experiments in Sections 3 and 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discussed the limitations of proposed work in Appendix N.

   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: [Yes]

Justification: The main paper, specifically Section 3.6 under Methodology, states the theoretical supports and key assumptions, while Appendix D provides the detailed assumptions and proofs for the propositions.

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: The paper reports the datasets, preprocessing and split strategy, evaluation metrics, baseline details, hyperparameters, training settings, and hardware information needed to reproduce the main experimental results. The supplementary material and public GitHub repository contain the code and dataset access instructions to support reproducibility.

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: We provide a public GitHub repository containing the code, configuration files, and instructions for reproducing the main experiments. For the restricted-access MIMIC-III dataset, the supplementary material provides access procedures and preprocessing instructions.

   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: The paper reports the data splits, hyperparameters and their tuning, optimizer, and task-specific settings needed to understand and reproduce the results. Specifically, Section 4 and Appendix E describe the dataset and experimental setup.

   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: The paper reports the main experimental results as mean $\pm$ standard deviation over three random seeds, as stated in Section 4.1. These error bars quantify variability across independent runs. Additionally, we report the paired statistical significance test in Appendix M.

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: The paper provides the compute setup and resource required to reproduce the experiments. Section 4.1 reports the hardware setup, and Appendix I reports computational efficiency details, including model parameters, total wall-clock training time, inference time per sample, and peak GPU memory required for the proposed model.

   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 research conducted in this paper conforms with the NeurIPS Code of Ethics.

   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 discussed the broader impacts of proposed work 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 pose such risks, as it does not release high-risk data or models such as pretrained language models, image generators, or scraped datasets.

    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: The paper credits the original datasets, baseline methods, and models through citations, and their licenses, access conditions, and terms of use are respected.

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: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: The research does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The use of LLM was limited to writing, editing, or formatting assistance and did not affect the scientific methodology, experimental design, or reported results.

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.