Causal Representation Learning from Multimodal EHRs under Non-Random Modality Missingness

Zihan Liang* **Ziwen Pan*** **Ruoxuan Xiong**

Emory University
{zihan.liang,ziwen.pan,ruoxuan.xiong}@emory.edu

Abstract

Clinical notes contain rich patient information, such as diagnoses or medications, making them valuable for *patient representation learning*. Recent advances in large language models have further improved the ability to extract meaningful representations from clinical texts. However, clinical notes are often missing. For example, in our analysis of the MIMIC-IV dataset, 24.5% of patients have no available discharge summaries. In such cases, representations can be learned from other modalities such as structured data, chest X-rays, or radiology reports. Yet the availability of these modalities is influenced by clinical decision-making and varies across patients, resulting in modality missing-not-at-random (*MMNAR*) patterns. We propose a *causal representation learning* framework that leverages observed data and informative missingness in multimodal clinical records. It consists of: (1) an MMNAR-aware modality fusion component that integrates structured data, imaging, and text while conditioning on missingness patterns to capture patient health and clinician-driven assignment; (2) a modality reconstruction component with contrastive learning to ensure semantic sufficiency in representation learning; and (3) a multitask outcome prediction model with a rectifier that corrects for residual bias from specific modality observation patterns. Comprehensive evaluations on MIMIC-IV show consistent gains over the strongest baselines, achieving up to 13.8% AUC improvement for hospital readmission and 13.1% for ICU admission.

1 Introduction

Language plays a central role in clinical communication. Learning patient representations from clinical notes has become an important focus in clinical NLP. These unstructured texts–written by clinicians to document observations, diagnoses, and decisions–encode rich contextual information that complements structured patient data. Since the introduction of contextualized language models like BERT [Devlin et al., 2019], the field has advanced rapidly with medical-domain adaptations, such as ClinicalBERT [Alsentzer et al., 2019, Huang et al., 2019]. More recently, large language models (LLMs) fine-tuned or adapted to clinical tasks have shown promise in medical reasoning, outcome prediction, and clinical decision support [Yang et al., 2022, Singhal et al., 2023].

Yet clinical text is often missing in real-world settings. In our analysis of the MIMIC-IV dataset, a large publicly available collection of de-identified electronic health records (EHR) [Johnson et al., 2024], 24.5% of patients lack discharge summaries. In such cases, other EHR modalities such as structured data, chest X-rays (CXR), and radiology reports may still be available and can be leveraged to learn patient representations.

Crucially, the availability of these modalities is **not random**. It is often determined by physician decision-making and patient conditions. For example, clinical notes and radiology reports are more

NeurIPS 2025 Workshop on Learning from Time Series for Health

likely to be recorded for patients with more severe conditions or complex diagnostic needs. As shown in Figure 1, patients with more complete modality combinations (i.e., structured data, CXR, clinical notes and radiology reports) have significantly higher post-discharge ICU admission and 30-day readmission rates. This reflects modality missing-not-at-random (MMNAR) patterns, where the absence of data itself encodes latent clinical state and correlates with outcomes.

In this paper, we propose CRL-MMNAR (*Causal Representation Learning under MMNAR*), a novel framework that explicitly uses both observed data and informative missingness in multimodal clinical records. CRL-MMNAR is structured in two stages. Together, they improve patient representation when clinical notes are available, and enable learning patient representation when text is missing.

The first stage consists of two complementary components for patient representation learning. The first component is **MMNAR-aware modality fusion**, which integrates structured data, imaging, and text using large language models and modality-specific encoders, while explicitly conditioning on modality missingness patterns. It serves three objectives: (i) increasing the estimation precision of latent patient representation by combining signals shared across modalities; (ii) preserving modality-specific information; and (iii) uncovering latent factors that influence clinical decision-making, such as a physician's judgment in ordering labs or imaging. By combining multimodal content with the clinician-assigned observation pattern, the fused representation reflects both the underlying health state and the reasons why specific modalities are observed or missing.

The second component is a **modality reconstruction with contrastive learning**. Its purpose is to ensure that the fused representation captures the essential content of each modality and can recover missing inputs. We achieve this using two complementary loss functions: a reconstruction loss that encourages recovery of masked modalities and a contrastive loss that aligns reconstructions with their originals while distinguishing them from other patients. Together, these objectives improve generalization across missingness patterns and yield robust, clinically meaningful representations.

The second stage is **multitask outcome prediction**, where the learned patient representations are applied to downstream tasks such as 30-day readmission, post-discharge ICU admission, and in-hospital mortality. The patient representation from Stage 1 serves as a shared backbone across all tasks, improving statistical efficiency by pooling information from common features of patient health. On top of this backbone, task-specific heads capture heterogeneity unique to each clinical outcome.

Crucially, modality observation patterns may themselves act as treatment variables, influencing outcomes through clinician decisions such as ordering additional tests or prescribing medications. To capture these observation-pattern-specific effects, we introduce a **rectifier mechanism** that applies post-training corrections inspired by semiparametric debiasing methods [Robins et al., 1994, Robins and Rotnitzky, 1995]. This adjustment ensures that predictions remain robust, even when modality assignment patterns encode systematic biases not captured by the base model.

We validate our approach with extensive experiments on MIMIC-IV. Our method consistently outperforms 13 state-of-the-art baselines. On MIMIC-IV, it achieves AUC gains of $+\mathbf{8.4}\%$ for readmission (from 0.7989 to 0.8657), $+\mathbf{13.1}\%$ for ICU admission (from 0.8687 to 0.9824), and $+\mathbf{4.7}\%$ for in-hospital mortality (from 0.9045 to 0.9472). Subgroup analyses underscore the robustness of MMNAR modeling, with pronounced benefits in underrepresented modality configurations. Ablation studies confirm that each component contributes meaningfully, with MMNAR-aware fusion and the rectifier mechanism driving the largest improvements.

2 Problem Formulation

Let $\mathcal{M}$ be the set of all available modalities. For each patient i, let $\mathcal{M}_i \subseteq \mathcal{M}$ denote the subset of modalities actually observed. For any modality $m \in \mathcal{M}$, let $x_i^{(m)}$ denote the corresponding raw input for patient i. We define $\boldsymbol{x}_i = \{x_i^{(m)}\}_{m \in \mathcal{M}}$ as the collection of all modality inputs (whether observed or not), and $\boldsymbol{x}_i^{\text{obs}} = \{x_i^{(m)}\}_{m \in \mathcal{M}_i}$ as the subset of observed modalities for patient i.

To encode the modality observation pattern, we define a binary vector $\boldsymbol{\delta}_i = [\delta_i^{(m)}]_{m \in \mathcal{M}}$. This pattern is typically determined by clinician decisions, such as whether to order imaging or write detailed notes. For each modality $m \in \mathcal{M}$, we set $\delta_i^{(m)} = 1$ if it is observed for patient i, and 0 otherwise. For example, in MIMIC-IV, we have $\mathcal{M} = \{\mathrm{S}, \mathrm{I}, \mathrm{T}, \mathrm{R}\}$, where "S" represents structured EHR data, "I" chest X-ray images, "T" discharge summaries, and "R" radiology reports.

For each patient i, we consider multiple clinical outcomes of interest. Let $\mathcal{T}$ denote the set of all outcome tasks. For each $t \in \mathcal{T}$, let $y_{i,t}$ denote the outcome for patient i corresponding to task t and let $\boldsymbol{y}_i = [y_{i,t}]_{t\in\mathcal{T}}$ collect all outcomes.

Our goal is to build a predictive model that uses both the observed modalities and the clinician-assigned observation pattern to estimate outcomes as accurately as possible. Formally, we define the outcome model as

$$y_{i,t} = f_{\boldsymbol{\theta}_t}\left(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{\delta}_i\right) + \varepsilon_{i,t}\,, \tag{1}$$

where $f_{\boldsymbol{\theta}_t}$ is the prediction function parameterized by $\boldsymbol{\theta}_t$, and $\varepsilon_{i,t}$ denotes random noise.

3 Method

Our CRL-MMNAR method learns the outcome model (1) under the causal diagram in Figure 2. The diagram posits a latent patient health state $\boldsymbol{h}_i$ that drives both the observed modalities $\boldsymbol{x}_i = \{x_i^{(m)}\}_{m\in\mathcal{M}}$ and clinician-assigned observation pattern $\boldsymbol{\delta}_i$. Both $\boldsymbol{h}_i$ and $\boldsymbol{\delta}_i$ in turn influence outcomes $\boldsymbol{y}_i$. CRL-MMNAR proceeds in two stages.

In the first stage, we learn a patient representation $\boldsymbol{h}_i$ that captures both observed modalities and the observation pattern:

$$\boldsymbol{h}_i = r_{\boldsymbol{\eta}}\left(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{\delta}_i\right), \tag{2}$$

where $r_{\boldsymbol{\eta}}$ is parameterized by shared weights $\boldsymbol{\eta}$ across all prediction tasks. This stage corresponds to Section 3.1 covering two core components: modality fusion and modality reconstruction, with preprocessing of raw data detailed in Appendix B.1.

In the second stage, we adopt a multitask outcome prediction framework. For each outcome task $t \in \mathcal{T}$, we model

$$y_{i,t} = g_{\boldsymbol{\psi}_t}\left(\boldsymbol{h}_i, \boldsymbol{\delta}_i\right) + \varepsilon_{i,t}\,, \tag{3}$$

where $g_{\boldsymbol{\psi}_t}$ is a task-specific predictor with parameters $\boldsymbol{\psi}_t$. Here, $\boldsymbol{\delta}_i$ is explicitly included to account for the observation pattern, which may itself act as a treatment variable. For example, certain patterns may reflect patients' health awareness (e.g., adherence to follow-up visits) or clinical decision-making (e.g., physicians ordering additional tests or prescribing medications). To account for these observation-pattern-specific effects, we introduce a rectifier mechanism (detailed in Section 3.2) that corrects residual biases from such patterns, thereby improving robustness in outcome prediction.

Finally, the overall end-to-end training integrates modality fusion, modality reconstruction, and outcome prediction, as summarized in Appendix B.2. For each outcome task t, the parameter set is $\boldsymbol{\theta}_t = (\boldsymbol{\eta}, \boldsymbol{\psi}_t)$. The outcome model (1) can thus be expressed as $f_{\boldsymbol{\theta}_t} = g_{\boldsymbol{\psi}_t} \circ r_{\boldsymbol{\eta}}$.

3.1 Patient Representation Learning

3.1.1 MMNAR-Aware Modality Fusion

The *MMNAR-aware modality fusion* component integrates modality embeddings $\{e_i^{(m)}\}_{m\in\mathcal{M}_i}$ into a unified representation $\boldsymbol{h}_i$ (Equation (2)). Unlike standard fusion strategies, this module explicitly treats missingness patterns $\boldsymbol{\delta}_i$ as structured contextual signals rather than random noise. The design ensures that the low-dimensional information in $\boldsymbol{\delta}_i$ is preserved and not overwhelmed by the high-dimensional embeddings $e_i^{(m)}$. The component proceeds in two main steps.

Step 1: Missingness-aware transformation. We first compute a missingness embedding $\boldsymbol{z}_i = \text{MLP}(\boldsymbol{\delta}_i)$, which transforms the binary observation pattern into a dense vector. To ensure that $\boldsymbol{z}_i$ retains structural information about clinician-assigned modality assignment, it is trained in a self-supervised encoder–decoder fashion: $\mathcal{L}_{\text{miss}} = \lambda_{\text{miss}}\text{CrossEntropy}\big(\hat{\boldsymbol{\delta}}_i, \boldsymbol{\delta}_i\big)$, where $\hat{\boldsymbol{\delta}}_i$ is decoded from $\boldsymbol{z}_i$ and λ_{miss} is the hyperparameter controlling the loss weight.

Each modality-specific embedding $e_i^{(m)}$ is then reweighted according to the missingness embedding $\boldsymbol{z}_i$: $e_{i,\text{gate}}^{(m)} = \delta_i^{(m)} \cdot \sigma\left(W^{(m)}\boldsymbol{z}_i + b^{(m)}\right) \cdot e_i^{(m)}$, where $\sigma(\cdot)$ is a sigmoid gating function and $(W^{(m)}, b^{(m)})$ are modality-specific parameters. Missing modalities ($\delta_i^{(m)} = 0$) are imputed with zero, while observed ones are adaptively emphasized or attenuated based on $\boldsymbol{z}_i$.

Step 2: Attention-based fusion. The gated embeddings $\{e_{i,\text{gate}}^{(m)}\}_{m\in\mathcal{M}}$ are then aggregated with a multi-head self-attention mechanism: $\boldsymbol{h}_i = r_{\text{fuse}}\left(\{e_{i,\text{gate}}^{(m)}\}_{m\in\mathcal{M}}\right)$. This produces the final patient representation $\boldsymbol{h}_i$. The attention mechanism contextualizes each modality relative to others and adapts dynamically to incomplete inputs–for example, placing greater weight on imaging data when clinical text is unavailable. The loss functions for learning $\boldsymbol{h}_i$ are described in the next subsection.

3.1.2 Modality Reconstruction with Contrastive Learning

The *modality reconstruction with contrastive learning* component defines the loss for representation learning. Its goal is to ensure that the fused patient representation $\boldsymbol{h}_i$ retains sufficient semantic information to recover missing inputs and generalize across observation patterns. It consists of two complementary objectives.

Objective 1: Cross-modality reconstruction. For each observed modality $m \in \mathcal{M}_i$, we randomly mask it to simulate a missing-at-random scenario, ensuring the masking itself does not encode clinical decisions. Using only the remaining modalities, we compute a partial representation $\boldsymbol{h}_i^{\setminus m}$ via the fusion process in Section 3.1.1. A modality-specific decoder $\phi^{(m)}$ then attempts to reconstruct the excluded embedding: $e_{i,\text{rec}}^{(m)} = \phi^{(m)}\left(\boldsymbol{h}_i^{\setminus m}\right)$. If $e_{i,\text{rec}}^{(m)}$ is close to the true embedding $e_i^{(m)}$, this indicates that the fusion mechanism has captured the necessary information from the other modalities. The reconstruction loss for modality m is $\mathcal{L}_{\text{rec}}^{(m)} = \sum_i \delta_i^{(m)} \cdot \left\| e_i^{(m)} - e_{i,\text{rec}}^{(m)} \right\|_2^2$, computed only when $e_i^{(m)}$ is available ($\delta_i^{(m)} = 1$).

Objective 2: Contrastive alignment. To prevent trivial reconstructions, we introduce a contrastive term that aligns each embedding with its own reconstruction while distinguishing it from reconstructions of other patients. For patient i and modality m, $(e_i^{(m)}, e_{i,\text{rec}}^{(m)})$ forms a positive pair, while $(e_i^{(m)}, e_{j,\text{rec}}^{(m)})$ for other patients $j \neq i$ are negative pairs. The contrastive loss is then defined as $\mathcal{L}_{\text{cont}}^{(m)} = -\log \frac{\exp(\text{sim}(e_i^{(m)}, e_{i,\text{rec}}^{(m)})/\tau_{\text{cont}})}{\sum_j \exp(\text{sim}(e_i^{(m)}, e_{j,\text{rec}}^{(m)})/\tau_{\text{cont}})}$, where $\text{sim}(\cdot,\cdot)$ cosine similarity and τ_{cont} is the InfoNCE temperature.

3.2 Multitask Outcome Prediction with Rectifier

The final part of CRL-MMNAR is *multitask outcome prediction with rectifier*. We consider a simplified form of Equation (3), where for each outcome t, the predictor decomposes into two parts:

$$y_{i,t} = g_{\boldsymbol{\psi}_t'}\left(\boldsymbol{h}_i\right) + \tau_{\boldsymbol{\delta}_i,t} + \varepsilon_{i,t}\,. \tag{4}$$

The first term, $g_{\boldsymbol{\psi}_t'}$, captures variation explained by the representation $\boldsymbol{h}_i$ and is invariant across all modality observation patterns, enabling parameter sharing across outcome tasks. The second term, $\tau_{\boldsymbol{\delta}_i,t}$, is a scalar parameter specific to each modality observation pattern $\boldsymbol{\delta}_i$, representing the treatment effect of the observation pattern $\boldsymbol{\delta}_i$ on the outcome t. This model is estimated in two steps.

Step 1: Multitask prediction loss. We jointly train outcome-specific predictors using the shared representation $\boldsymbol{h}_i$.

Step 2: Observation pattern specific rectifier. After training the base model (yielding $\boldsymbol{h}_i$ and $\hat{y}_{i,t}$), we estimate $\tau_{\boldsymbol{\delta},t}$ separately for each task t and modality pattern $\boldsymbol{\delta}$. To mitigate overfitting, we adopt cross-fitting. Specifically, for a fixed $\boldsymbol{\delta} \in \{0,1\}^{|\mathcal{M}|}$, we define the index set as $\mathcal{V}_{\boldsymbol{\delta}} = \{\, i : \boldsymbol{\delta}_i = \boldsymbol{\delta} \,\}$, which contains all patients with modality configuration $\boldsymbol{\delta}$. We partition this set into two disjoint folds, $\mathcal{V}_{\boldsymbol{\delta}}^{(1)}$ and $\mathcal{V}_{\boldsymbol{\delta}}^{(2)}$, satisfying $\mathcal{V}_{\boldsymbol{\delta}} = \mathcal{V}_{\boldsymbol{\delta}}^{(1)} \cup \mathcal{V}_{\boldsymbol{\delta}}^{(2)}$ and $\mathcal{V}_{\boldsymbol{\delta}}^{(1)} \cap \mathcal{V}_{\boldsymbol{\delta}}^{(2)} = \varnothing$.

On each fold $k \in \{1,2\}$ and for each task t, we estimate the average residual using predictions obtained without using that fold: $\hat{\tau}_{\boldsymbol{\delta},t}^{(k)} = |\mathcal{V}_{\boldsymbol{\delta}}^{(k)}|^{-1} \sum_{i\in\mathcal{V}_{\boldsymbol{\delta}}^{(k)}} \left(y_{i,t} - \hat{y}_{i,t}\right)$. This provides a fold-specific estimate of the systematic bias associated with observation pattern $\boldsymbol{\delta}$ for outcome t.

We then apply the correction estimated from one fold to the other to obtain rectified predictions: $\hat{y}_{i,t}^{\text{rect}} = \hat{y}_{i,t} + \mathbb{1}\left\{\left|\hat{\tau}_{\boldsymbol{\delta}_i,t}^{(k)}\right| > \kappa\right\} \cdot \hat{\tau}_{\boldsymbol{\delta}_i,t}^{(k)}$, for $i \in \mathcal{V}_{\boldsymbol{\delta}_i}^{(\bar{k})}$ and $\bar{k} \neq k$, where $\kappa \geq 0$ is a threshold that enables *selective correction*–that is, the rectifier is applied only when the estimated effect is non-

negligible. The threshold κ is chosen via cross-validation to balance correction effectiveness and stability, typically ranging from 0.01 to 0.05 depending on the scale of prediction errors.

4 Experiments

We use the MIMIC-IV v3.1 database [Johnson et al., 2024], a large-scale, de-identified clinical dataset containing structured EHRs, clinical notes, and chest radiographs. We construct a cohort of 20,000 adult patients ($\geq$18 years) with a single ICU stay, excluding those with multiple admissions or missing key records.

Among all patients, 15,098 (75.5%) have discharge summaries, 17,010 (85.0%) have radiology reports, and all patients (100%) have structured data. Overall, 17,903 (89.5%) have at least one text modality and 5,228 patients (26.1%) have CXRs. For downstream modeling, we encode each patient's modality availability as a binary vector.

We evaluate on three prediction tasks: (1) **30-day Hospital Readmission**: Predicting readmission within 30 days post-discharge; (2) **Post-discharge ICU Admission**: Predicting ICU admission within 90 days post-discharge; (3) **In-hospital Mortality**: Predicting death during hospital stay.

Table 1: Performance comparison across datasets and clinical tasks. Best results in **bold**. Results reported as mean $\pm$ standard deviation, computed over five independent runs with different random seeds for initialization.

Model	MIMIC-IV Dataset								
	30-day Readmission			Post-discharge ICU			In-hospital Mortality		
	AUC	AUPRC	Brier	AUC	AUPRC	Brier	AUC	AUPRC	Brier
CM-AE	$0.6892_{\pm 0.041}$	$0.4012_{\pm 0.052}$	$0.1798_{\pm 0.026}$	$0.6978_{\pm 0.043}$	$0.2687_{\pm 0.048}$	$0.1287_{\pm 0.021}$	$0.8423_{\pm 0.029}$	$0.4187_{\pm 0.039}$	$0.1043_{\pm 0.018}$
MT	$0.7134_{\pm 0.036}$	$0.4298_{\pm 0.047}$	$0.1745_{\pm 0.024}$	$0.7382_{\pm 0.038}$	$0.2998_{\pm 0.042}$	$0.1243_{\pm 0.018}$	$0.8612_{\pm 0.027}$	$0.4342_{\pm 0.037}$	$0.0998_{\pm 0.016}$
SMIL	$0.7087_{\pm 0.034}$	$0.4456_{\pm 0.046}$	$0.1732_{\pm 0.022}$	$0.6845_{\pm 0.044}$	$0.2623_{\pm 0.051}$	$0.1312_{\pm 0.023}$	$0.8489_{\pm 0.031}$	$0.4276_{\pm 0.041}$	$0.1018_{\pm 0.018}$
GRAPE	$0.7045_{\pm 0.038}$	$0.4267_{\pm 0.050}$	$0.1756_{\pm 0.025}$	$0.7234_{\pm 0.040}$	$0.2934_{\pm 0.044}$	$0.1267_{\pm 0.020}$	$0.8698_{\pm 0.025}$	$0.4428_{\pm 0.035}$	$0.0978_{\pm 0.016}$
HGMF	$0.7289_{\pm 0.032}$	$0.4823_{\pm 0.043}$	$0.1698_{\pm 0.023}$	$0.7123_{\pm 0.042}$	$0.2798_{\pm 0.048}$	$0.1289_{\pm 0.020}$	$0.8567_{\pm 0.027}$	$0.4234_{\pm 0.039}$	$0.1012_{\pm 0.018}$
M3Care	$0.7256_{\pm 0.034}$	$0.4612_{\pm 0.046}$	$0.1712_{\pm 0.023}$	$0.7267_{\pm 0.038}$	$0.2956_{\pm 0.044}$	$0.1276_{\pm 0.018}$	$0.8776_{\pm 0.025}$	$0.4456_{\pm 0.037}$	$0.0967_{\pm 0.016}$
COM	$0.7634_{\pm 0.029}$	$0.4178_{\pm 0.050}$	$0.1678_{\pm 0.020}$	$0.8298_{\pm 0.032}$	$0.3678_{\pm 0.039}$	$0.1198_{\pm 0.016}$	$0.8789_{\pm 0.023}$	$0.3823_{\pm 0.042}$	$0.0978_{\pm 0.016}$
DrFuse	$0.7687_{\pm 0.027}$	$0.4098_{\pm 0.052}$	$0.1656_{\pm 0.020}$	$0.8687_{\pm 0.029}$	$0.3634_{\pm 0.037}$	$0.1134_{\pm 0.014}$	$0.8923_{\pm 0.020}$	$0.3812_{\pm 0.039}$	$0.0934_{\pm 0.014}$
MissModal	$0.7478_{\pm 0.032}$	$0.4034_{\pm 0.054}$	$0.1689_{\pm 0.023}$	$0.8145_{\pm 0.034}$	$0.3312_{\pm 0.042}$	$0.1212_{\pm 0.018}$	$0.8667_{\pm 0.025}$	$0.3945_{\pm 0.044}$	$0.0998_{\pm 0.016}$
FLEXGEN-EHR	$0.7845_{\pm 0.025}$	$0.4323_{\pm 0.048}$	$0.1634_{\pm 0.018}$	$0.8476_{\pm 0.027}$	$0.3798_{\pm 0.035}$	$0.1123_{\pm 0.014}$	$0.8956_{\pm 0.018}$	$0.4078_{\pm 0.037}$	$0.0912_{\pm 0.014}$
MUSE+	$0.7989_{\pm 0.030}$	$0.4812_{\pm 0.042}$	$0.1543_{\pm 0.020}$	$0.8678_{\pm 0.036}$	$0.4067_{\pm 0.046}$	$0.1078_{\pm 0.018}$	$0.9045_{\pm 0.016}$	$0.4678_{\pm 0.033}$	$0.0889_{\pm 0.012}$
GRU-D	$0.7212_{\pm 0.042}$	$0.4134_{\pm 0.057}$	$0.1723_{\pm 0.027}$	$0.7645_{\pm 0.040}$	$0.2834_{\pm 0.053}$	$0.1278_{\pm 0.023}$	$0.8267_{\pm 0.034}$	$0.4189_{\pm 0.046}$	$0.1034_{\pm 0.020}$
Raindrop	$0.7434_{\pm 0.034}$	$0.4298_{\pm 0.050}$	$0.1678_{\pm 0.023}$	$0.7889_{\pm 0.036}$	$0.3112_{\pm 0.044}$	$0.1234_{\pm 0.018}$	$0.8623_{\pm 0.027}$	$0.4423_{\pm 0.039}$	$0.0967_{\pm 0.016}$
CRL-MMNAR	$\mathbf{0.8657}_{\pm 0.018}$	$\mathbf{0.5627}_{\pm 0.028}$	$\mathbf{0.1167}_{\pm 0.012}$	$\mathbf{0.9824}_{\pm 0.016}$	$\mathbf{0.5821}_{\pm 0.024}$	$\mathbf{0.0589}_{\pm 0.007}$	$\mathbf{0.9472}_{\pm 0.014}$	$\mathbf{0.4767}_{\pm 0.026}$	$\mathbf{0.0759}_{\pm 0.009}$

We compare CRL-MMNAR with 13 state-of-the-art methods, spanning three major paradigms in multimodal learning. The details about the benchmark methods are provided in Appendix D.2. Table 1 reports results across both datasets and all clinical tasks, using Area Under the ROC Curve (AUC), Area Under the Precision-Recall Curve (AUPRC), and Brier score as evaluation metrics.

CRL-MMNAR achieves substantial improvements across all tasks and metrics. On MIMIC-IV, the largest gains occur in ICU admission prediction, where AUC rises from 0.8687 with DrFuse to 0.9824 ($+\mathbf{13.1}\%$), and in 30-day readmission, from 0.7989 with MUSE+ to 0.8657 ($+\mathbf{8.4}\%$). For in-hospital mortality, our model improves from 0.9045 with MUSE+ to 0.9472 ($+\mathbf{4.7}\%$). On eICU, improvements are similarly consistent: readmission AUC increases from 0.8167 with MUSE+ to 0.9294 ($+\mathbf{13.8}\%$), while mortality prediction rises from 0.9334 with MUSE+ to 0.9380 ($+\mathbf{0.5}\%$).

Importantly, these gains are accompanied by consistent reductions in Brier scores, confirming that improvements in discrimination are matched by better-calibrated predictions. We further provide component ablation studies in Appendix D.3.

5 Conclusion

We introduce CRL-MMNAR, a causal multimodal framework that explicitly models MMNAR in clinical data. The framework combines missingness-aware fusion, cross-modal reconstruction, and multitask prediction with rectification to learn robust patient representations. Evaluations on MIMIC-IV show consistent improvements over 13 state-of-the-art baselines, with notable gains of $\mathbf{8.4}\%$ AUC for readmission and $\mathbf{13.1}\%$ for ICU admission. This work highlights the importance of treating missingness as structured signal and offers a principled approach for robust patient representation learning under realistic data constraints.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019. URL https://arxiv.org/abs/1904.03323.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. URL https://ieeexplore.ieee.org/abstract/document/6472238.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David A. Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *CoRR*, abs/1606.01865, 2016. URL http://arxiv.org/abs/1606.01865.

Jiayi Chen and Aidong Zhang. Hgmf: Heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*, pages 1295–1304, 2020. doi: 10.1145/3394486.3403182.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Junting Duan, Markus Pelger, and Ruoxuan Xiong. Factor analysis for causal inference on large non-stationary panels with endogenous treatment. *Available at SSRN 4823360*, 2024a. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4823360.

Junting Duan, Markus Pelger, and Ruoxuan Xiong. Target pca: Transfer learning large dimensional panel data. *Journal of Econometrics*, 244(2):105521, 2024b. URL https://www.sciencedirect.com/science/article/pii/S0304407623002373.

Huan He, William hao, Yuanzhe Xi, Yong Chen, Bradley Malin, and Joyce Ho. A flexible generative model for heterogeneous tabular EHR with missing modality. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=W2tCmRrj7H.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019. URL https://arxiv.org/abs/1904.05342.

Alistair Johnson, Luigi Bulgarelli, Tom Pollard, Benjamin Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV (version 3.1). https://doi.org/10.13026/kpb9-mt58, 2024. PhysioNet.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1617–1626, 2018. URL http://arxiv.org/abs/1806.06270.

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 04, pages 4485–4492, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/5876.

Dongyue Li, Haotian Ju, Aneesh Sharma, and Hongyang R Zhang. Boosting multitask learning on graphs through higher-order task affinities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1213–1222, 2023a. URL https://dl.acm.org/doi/abs/10.1145/3580305.3599265.

Dongyue Li, Huy L Nguyen, and Hongyang R Zhang. Identification of negative transfers in multitask learning using surrogate models. *Transactions on Machine Learning Research*, 2023b. URL https://arxiv.org/abs/2303.14582.

Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. Scalable multitask learning using gradient-based estimation of task affinity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1542–1553, 2024a. URL https://dl.acm.org/doi/abs/10.1145/3637528.3671835.

Dongyue Li, Ziniu Zhang, Lu Wang, and Hongyang R Zhang. Scalable fine-tuning from multiple data sources: A first-order approximation approach. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024b. URL https://arxiv.org/abs/2409.19458.

Dongyue Li, Ziniu Zhang, Lu Wang, and Hongyang R Zhang. Efficient ensemble for fine-tuning language models on multiple datasets. *The 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2505.21930.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023. doi: 10.1162/tacl_a_00628.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3069–3077, 2021. URL https://arxiv.org/abs/2103.05677.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18177–18186, 2022. URL https://arxiv.org/abs/2204.05454.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 689–696, 2011. URL http://www.icml-2011.org/papers/399_icmlpaper.pdf.

Shuwei Qian and Chongjun Wang. COM: Contrastive Masked-attention model for incomplete multimodal learning. *Neural Networks*, 162:443–455, May 2023. ISSN 08936080. doi: 10.1016/j.neunet.2023.03.003. URL https://linkinghub.elsevier.com/retrieve/pii/S089360802300120X.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476494.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1994.10476818.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, and Stephen Pfohl. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. URL https://www.nature.com/articles/s41586-023-06291-2.

Sen Wu, Hongyang R Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2005.00944.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Je5SHCKpPa.

Ruoxuan Xiong and Markus Pelger. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301, 2023. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2022.04.005. URL https://www.sciencedirect.com/science/article/pii/S0304407622000914.

Fan Yang, Hongyang R Zhang, Sen Wu, Christopher Re, and Weijie J Su. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26 (113):1–88, 2025. URL https://www.jmlr.org/papers/v26/24-0454.html.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022. doi: 10.1038/s41746-022-00742-2. URL https://www.nature.com/articles/s41746-022-00742-2.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, number 15, pages 16416–16424, 2024. URL http://arxiv.org/abs/2403.06197.

Jiaxuan You, Xiaobai Ma, Daisy Yi Ding, Mykel Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. In *NeurIPS 2020: Advances in Neural Information Processing Systems 33*, 2020. URL https://arxiv.org/abs/2010.16418.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, pages 2545–2554, 2022a. doi: 10.1145/3534678.3539388.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022b. URL https://arxiv.org/abs/2110.05357.

A Figures

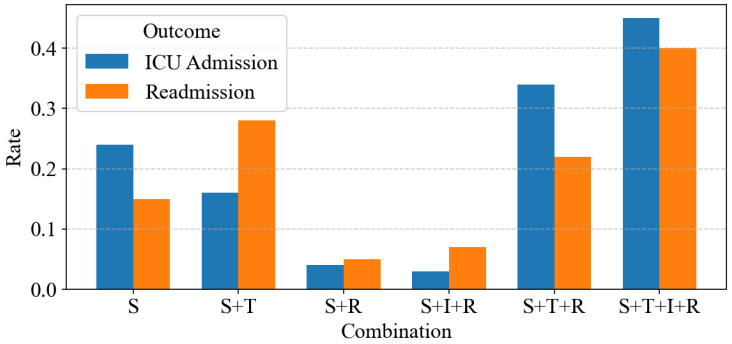

Figure 1: **Modality availability patterns are predictive of clinical outcomes (ICU admission and 30-day readmission).** S = structured EHR data; I = chest X-ray images; T = discharge summaries; R = radiology report.

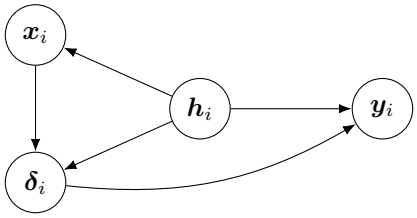

Figure 2: Causal diagram: The latent patient health state $\boldsymbol{h}_i$ influences the modality contents $\boldsymbol{x}_i = \{x_i^{(m)}\}_{m \in \mathcal{M}}$ and drives the clinician-assigned observation pattern $\boldsymbol{\delta}_i = [\delta_i^{(m)}]_{m \in \mathcal{M}}$. Outcomes $\boldsymbol{y}_i = [y_{i,t}]_{t \in \mathcal{T}}$ are caused by $\boldsymbol{h}_i$ and may also be directly affected by $\boldsymbol{\delta}_i$ (e.g., ordering additional tests or prescribing medications).

B Additional Method Details

B.1 Preprocessing Multimodal Data

For each observed modality $x_i^{(m)}$, we obtain a semantic embedding $e_i^{(m)} \in \mathbb{R}^d$ using a modality-specific encoder, by applying distinct preprocessing and encoding procedures to text, imaging, and structured data.

For text data such as discharge summaries and radiology reports, we apply standard preprocessing steps, including artifact removal, abbreviation normalization, and segmentation of long sequences. We then obtain embeddings using large language models pre-trained on biomedical corpora (e.g., domain-adapted BERT variants), which can be further fine-tuned for downstream prediction tasks.

For imaging data, such as chest X-rays, we first apply standard preprocessing (e.g., normalization and resizing) and then extract embeddings using models pre-trained on large-scale image datasets.

For structured data such as demographics, laboratory results, and vital signs, we apply standard preprocessing to normalize continuous features and embed categorical variables. For temporal signals like labs and vitals, we align measurements across time and then encode them with lightweight feed-forward or recurrent models designed for tabular and longitudinal data.

B.2 End-to-End Training Procedure

We now describe the complete training procedure for the outcome model $y_{i,t} = f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{\delta}_i) + \varepsilon_{i,t}$. The total training objective combines three losses:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{miss}}}_{\text{Section } 3.1.1} + \underbrace{\mathcal{L}_{\text{rep}}}_{\text{Section } 3.1.2} + \underbrace{\mathcal{L}_{\text{pred}}}_{\text{Section } 3.2},$$

where $\mathcal{L}_{\text{miss}}$ learns missingness context, $\mathcal{L}_{\text{rep}}$ ensures semantic sufficiency in representation learning, and $\mathcal{L}_{\text{pred}}$ optimizes outcome prediction. Here $\mathcal{L}_{\text{miss}}$ is defined as

$$\mathcal{L}_{\text{miss}} = \lambda_{\text{miss}} \text{CrossEntropy}\big(\hat{\boldsymbol{\delta}}_i, \boldsymbol{\delta}_i\big).$$

$\mathcal{L}_{\text{rep}}$ is defined as

$$\mathcal{L}_{\text{rep}} = \sum_{m \in \mathcal{M}} \Big(\lambda_{\text{rec}} \mathcal{L}_{\text{rec}}^{(m)} + \lambda_{\text{cont}} \mathcal{L}_{\text{cont}}^{(m)}\Big),$$

where λ_{rec} and λ_{cont} are hyperparameters controlling the relative importance of reconstruction and contrastive objectives. $\mathcal{L}_{\text{pred}}$ is defined as

$$\mathcal{L}_{\text{pred}} = \sum_t \lambda_{\text{pred},t} \cdot \text{CrossEntropy}\big(\hat{y}_{i,t}, y_{i,t}\big),$$

where $\hat{y}_{i,t} = g_{\boldsymbol{\psi}_t'}(\boldsymbol{h}_i)$ is the predicted outcome t using $\boldsymbol{h}_i$ and $\lambda_{\text{pred},t}$ balances prevalence and clinical importance across tasks.

The first two terms, $\mathcal{L}_{\text{miss}} + \mathcal{L}_{\text{rep}}$, define the loss used to train representation encoder parameters $\boldsymbol{\eta}$ in Equation (2). The final term, $\mathcal{L}_{\text{pred}}$ defines the loss for the outcome-specific parameters $\boldsymbol{\psi}_t'$ for all tasks t in Equation (4). The end-to-end training jointly learns both the patient representation $\boldsymbol{h}_i$ and the outcome predictors $\boldsymbol{\psi}_t'$ for all t.

After training, we estimate modality-pattern-specific corrections $\{\tau_{\boldsymbol{\delta},t}\}_{\boldsymbol{\delta} \in \{0,1\}^{|\mathcal{M}|}}$ via cross-fitting. These parameters account for the direct effect of modality assignment patterns $\boldsymbol{\delta}_i$ on outcomes, which may not be fully explained by $\boldsymbol{h}_i$.

Putting everything together, the parameter set is $\boldsymbol{\theta}_t = (\boldsymbol{\eta},\ \boldsymbol{\psi}_t',\ \{\tau_{\boldsymbol{\delta},t}\}_{\boldsymbol{\delta} \in \{0,1\}^{|\mathcal{M}|}})$ for each outcome task t.

C Related Work

Our work is most closely related to the growing literature on multimodal representation learning, and in particular to methods that address missing modalities. Early approaches, such as CM-AE [Ngiam et al., 2011], introduced autoencoder-based frameworks for cross-modal imputation. Subsequent methods developed more sophisticated modality-specific encoders with fusion mechanisms, including multimodal Transformers [Ma et al., 2022], graph-based aggregation (GRAPE [You et al., 2020]), and heterogeneous graph-based factorization (HGMF [Chen and Zhang, 2020]). While these models can operate under partial inputs, they treat missingness as a nuisance to be mitigated, rather than a source of signal. A complementary line of work treats missing modalities as a primary modeling objective rather than a nuisance. Examples include SMIL [Ma et al., 2021], COM [Qian and Wang, 2023], and MissModal [Lin and Hu, 2023], which design Bayesian, contrastive, or dropout-based strategies to directly handle sparsity. These approaches directly relate to our setting by treating missingness as a key modeling consideration.

Within this stream, several methods have been developed specifically for healthcare applications, including M3Care [Zhang et al., 2022a], MUSE+ [Wu et al., 2024], FLEXGEN-EHR [He et al., 2024], and DrFuse [Yao et al., 2024], which achieve strong performance under real-world modality sparsity. In parallel, temporal models such as GRU-D [Che et al., 2016] and Raindrop [Zhang et al., 2022b] are developed to capture implicit temporal missingness patterns. However, these approaches do not explicitly account for the causes of missingness. In contrast, our CRL-MMNAR explicitly models observation patterns as informative signals to recover clinically meaningful structure and improve predictive accuracy.

Our work also connects to the growing literature at the intersection of causal inference and machine learning. In spirit, we share the idea of leveraging causal principles (e.g., balancing and weighting)

to improve predictive accuracy in observational settings [Kuang et al., 2018, 2020]. More closely, our work relates to research on uncovering latent structures when data are missing not at random, where missingness is endogenously driven by unobserved factors [Xiong and Pelger, 2023, Duan et al., 2024a,b]. Finally, our rectifier mechanism parallels semiparametric approaches for handling MNAR data [Robins et al., 1994, Robins and Rotnitzky, 1995], as it corrects for the effect arises from the observation patterns to reduce bias.

Finally, our work also relates to the literature on multitask learning, which leverages commonalities across related prediction tasks to share statistical strength [Bengio et al., 2013]. In our setting, multitask outcome prediction illustrates positive transfer, but as the number of outcomes grows, negative transfer may occur, where shared representations harm accuracy for certain tasks [Wu et al., 2020, Yang et al., 2025], making its mitigation an important future direction. Recent advances in task modeling [Li et al., 2023a,b], adaptive and scalable fine-tuning for individual tasks [Li et al., 2024a,b, 2025] offer promising approaches to mitigate negative transfer by dynamically controlling when and how knowledge is shared across tasks.

D Additional Experiments on MIMIC Dataset

D.1 Dataset Preprocessing

For the MIMIC-IV data, we construct a cohort of 20,000 adult patients ($\geq$18 years) with a single ICU stay, excluding those with multiple admissions or missing key records. we retrieve structured data tables (`admissions`, `patients`, `chartevents`, `labevents`, etc.) via BigQuery, filtered to a fixed subject list. Extracted data are cached for consistency across experiments. For each patient, we aggregate diagnostic, laboratory, and medication events into summary features over predefined windows before ICU admission. Missing values are retained and augmented with indicator flags to preserve potential MMNAR signals.

D.2 Baselines and Implementation Details

We compare CRL-MMNAR with 13 state-of-the-art methods, spanning three major paradigms in multimodal learning. The second paradigm is explicitly designed to handle missing data.

Traditional Multimodal Methods.
1. **CM-AE** [Ngiam et al., 2011]: Cross-modality autoencoder for imputation and prediction.

2. **MT** [Ma et al., 2022]: Multimodal Transformer with late fusion.

3. **GRAPE** [You et al., 2020]: Bipartite graph neural network capturing patient-modality relations.

4. **HGMF** [Chen and Zhang, 2020]: Heterogeneous graph-based matrix factorization.

Missing Modality Specialists.
5. **SMIL** [Ma et al., 2021]: Bayesian meta-learning with modality-wise priors.

6. **M3Care** [Zhang et al., 2022a]: Modality-wise similarity graph with Transformer aggregation.

7. **COM** [Qian and Wang, 2023]: Contrastive multimodal framework.

8. **DrFuse** [Yao et al., 2024]: Disentangled clinical fusion network.

9. **MissModal** [Lin and Hu, 2023]: Robust modality dropout framework.

10. **FLEXGEN-EHR** [He et al., 2024]: Generative framework for heterogeneous EHR data.

11. **MUSE+** [Wu et al., 2024]: Bipartite patient-modality graph with contrastive objectives.

Irregular Time Series Methods.
12. **GRU-D** [Che et al., 2016]: Gated recurrent unit with decay mechanisms.

13. **Raindrop** [Zhang et al., 2022b]: Graph-guided network for irregularly sampled time series.

Implementation Details All models are trained with AdamW (learning rate 2×10^{-4}, weight decay 1×10^{-6}, batch size 32) using early stopping (patience 30) and automatic mixed precision. To address class imbalance, we adopt focal loss with task-specific parameters tuned on validation sets.

We use standardized 5-fold stratified cross-validation with grid search for hyperparameter tuning. For each model, results are reported as mean $\pm$ standard deviation, computed over five independent runs with different random seeds for initialization.

Training uses NVIDIA RTX A6000 GPUs (48GB VRAM) with 256GB RAM. The architecture is optimized for this hardware while remaining compatible with clinical environments, and adopts end-to-end training (Appendix B.2).

D.3 Component Ablation Studies

Table 2 presents systematic ablation results demonstrating each component's contribution. Starting from a multimodal baseline with standard feature concatenation, we progressively add: (1) MMNAR-aware fusion with $\mathcal{L}_{\text{miss}}$; (2) modality reconstruction with $\mathcal{L}_{\text{rep}}$; and (3) the rectifier mechanism.

MMNAR-aware fusion provides the largest improvements across both datasets, confirming that explicitly modeling clinician-driven missingness patterns captures meaningful clinical signals beyond standard fusion approaches. Modality reconstruction delivers consistent but modest gains, validating that cross-modal semantic sufficiency enhances representation quality. The rectifier shows task-dependent benefits, particularly for ICU admission prediction, suggesting bias correction is most valuable for outcomes closely tied to clinical decision-making patterns.

The complete framework achieves substantial cumulative improvements, with each component contributing meaningfully to the final performance across all clinical tasks and datasets.

Table 2: Component ablation analysis showing incremental performance gains on the MIMIC-IV dataset. MMNAR-Aware Fusion is abbreviated as MMNAR; Modality Reconstruction as MR; Multitask with Rectifier as Rectifier.

Component	**MIMIC-IV Dataset**								
	30-day Readmission			**Post-discharge ICU**			**In-hosp. Mortality**		
	AUC	**APR**	**ΔAUC**	**AUC**	**APR**	**ΔAUC**	**AUC**	**APR**	**ΔAUC**
Basic Baseline	.717	.347	–	.801	.307	–	.814	.369	–
+ MMNAR	.793	.470	+.076	.853	.372	+.052	.894	.381	+.080
+ MR	.811	.519	+.018	.889	.404	+.036	.929	.456	+.035
+ Rectifier	**.866**	**.563**	**+.055**	**.982**	**.582**	**+.093**	**.947**	**.477**	**+.018**

Table 3: Performance Across Modality Combinations on MIMIC-IV

Modality Configuration	**Readmission AUC**	**ICU AUC**
All modalities	0.8657	0.9824
Discharge summaries + Radiology reports	0.7808	0.8843
Structured-Only	0.7234	0.8156