Causal Representation Learning from Multimodal Clinical Records under Non-Random Modality Missingness

Zihan Liang[*]
Emory University
zihan.liang@emory.edu

Ziwen Pan[*]
Emory University
ziwen.pan@emory.edu

Ruoxuan Xiong
Emory University
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 across MIMIC-IV and eICU 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; Agrawal et al., 2022).

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, institutional protocols, and patient conditions. For example, clinical notes and radiology reports are more 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 leverages both observed data and informative missingness in multimodal clinical records. Our framework is structured in two stages. Together, they not only improve patient representation when clinical notes

[*]Equal contribution. Code and reproducibility materials are available at https://github.com/CausalMLResearch/CRL-MMNAR.

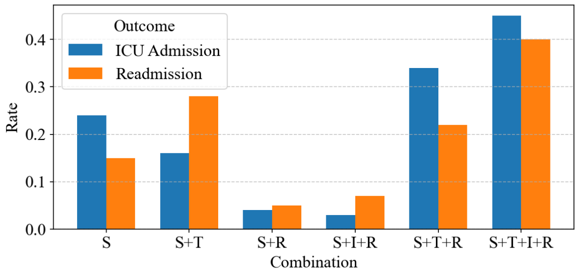

Figure 1: **Modality availability patterns are predictive of clinical outcomes (ICU admission and 30-day readmission).** S = structured EHR data; I = chest X-ray images; T = discharge summaries; R = radiology report.

are available, but also 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 two large-scale clinical datasets, MIMIC-IV and eICU (Pollard et al., 2018). 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). On eICU, our method achieves consistent improvements as well, including a $+\mathbf{13.8}\%$ relative gain for readmission (from 0.8167 to 0.9294) and $+\mathbf{0.5}\%$ for mortality (from 0.9334 to 0.9380). 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.

Although we focus on healthcare as the primary application, we note that the central idea extends more broadly. Modality observation patterns often contain meaningful signal, and in many domains these patterns themselves can influence outcomes.

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. In this case, $\boldsymbol{\delta}_i$ is a four-dimensional binary vector describing the clinician-determined modality configuration.

For each patient i, we consider multiple clinical outcomes of interest, such as 30-day hospital readmission, post-discharge ICU admission within 90 days, in-hospital mortality, and other relevant endpoints. 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 Sections 3.1-3.2, covering preprocessing of raw data together with the two core components: modality fusion and modality reconstruction.

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

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

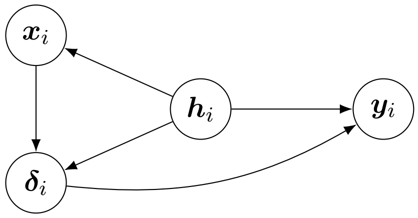

Figure 2: Causal diagram: The latent patient health state $\boldsymbol{h}_i$ influences the modality contents $\boldsymbol{x}_i = \{x_i^{(m)}\}_{m\in\mathcal{M}}$ and drives the clinician-assigned observation pattern $\boldsymbol{\delta}_i = [\delta_i^{(m)}]_{m\in\mathcal{M}}$. Outcomes $\boldsymbol{y}_i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

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 3.3) 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 Section 3.4. 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 the composition of the two stages, $f_{\boldsymbol{\theta}_t} = g_{\boldsymbol{\psi}_t} \circ r_{\boldsymbol{\eta}}$. A complete algorithmic summary of the framework is provided in Appendix C.8.

3.1 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

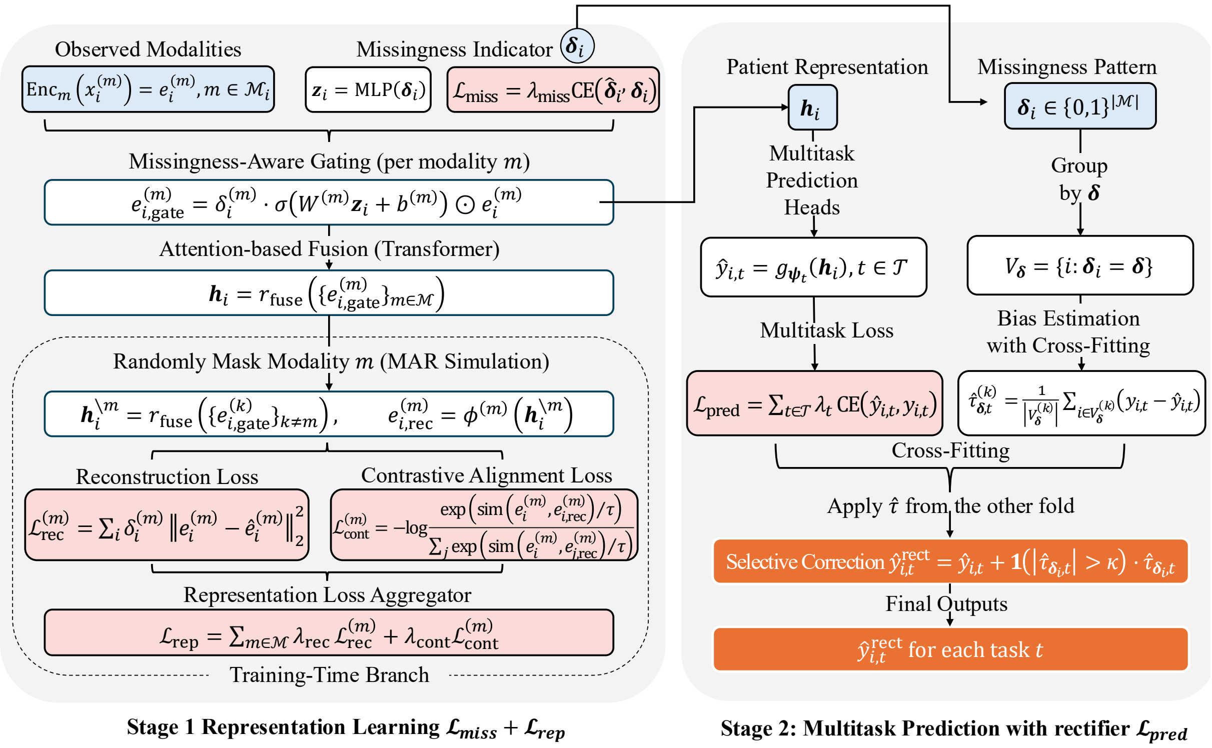
Observed Modalities
$\mathrm{Enc}_m\left(x_i^{(m)}\right) = e_i^{(m)}, m \in \mathcal{M}_i$
Missingness Indicator
$\boldsymbol{\delta}_i$
$\boldsymbol{z}_i = \mathrm{MLP}(\boldsymbol{\delta}_i)$
$\mathcal{L}_{\mathrm{miss}} = \lambda_{\mathrm{miss}}\mathrm{CE}(\hat{\boldsymbol{\delta}}_i, \boldsymbol{\delta}_i)$
Missingness-Aware Gating (per modality m)
$e_{i,\mathrm{gate}}^{(m)} = \delta_i^{(m)} \cdot \sigma(W^{(m)}\boldsymbol{z}_i + b^{(m)}) \odot e_i^{(m)}$
Attention-based Fusion (Transformer)
$\boldsymbol{h}_i = r_{\mathrm{fuse}}\left(\{e_{i,\mathrm{gate}}^{(m)}\}_{m\in\mathcal{M}}\right)$
Randomly Mask Modality m (MAR Simulation)
$\boldsymbol{h}_i^{\setminus m} = r_{\mathrm{fuse}}\left(\{e_{i,\mathrm{gate}}^{(k)}\}_{k\neq m}\right), \quad e_{i,\mathrm{rec}}^{(m)} = \phi^{(m)}\left(\boldsymbol{h}_i^{\setminus m}\right)$
Reconstruction Loss
Contrastive Alignment Loss
$\mathcal{L}_{\mathrm{rec}}^{(m)} = \sum_i \delta_i^{(m)} \left\| e_i^{(m)} - \hat{e}_i^{(m)} \right\|_2^2$
$\mathcal{L}_{\mathrm{cont}}^{(m)} = -\log \frac{\exp\left(\mathrm{sim}\left(e_i^{(m)}, e_{i,\mathrm{rec}}^{(m)}\right)/\tau\right)}{\sum_j \exp\left(\mathrm{sim}\left(e_i^{(m)}, e_{j,\mathrm{rec}}^{(m)}\right)/\tau\right)}$
Representation Loss Aggregator
$\mathcal{L}_{\mathrm{rep}} = \sum_{m\in\mathcal{M}} \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}^{(m)} + \lambda_{\mathrm{cont}}\mathcal{L}_{\mathrm{cont}}^{(m)}$
Training-Time Branch
Patient Representation
$\boldsymbol{h}_i$
Missingness Pattern
$\boldsymbol{\delta}_i \in \{0,1\}^{|\mathcal{M}|}$
Multitask Prediction Heads
Group by $\boldsymbol{\delta}$
$\hat{y}_{i,t} = g_{\psi_t}(\boldsymbol{h}_i), t \in \mathcal{T}$
$V_{\boldsymbol{\delta}} = \{i: \boldsymbol{\delta}_i = \boldsymbol{\delta}\}$
Multitask Loss
Bias Estimation with Cross-Fitting
$\mathcal{L}_{\mathrm{pred}} = \sum_{t\in\mathcal{T}} \lambda_t \,\mathrm{CE}(\hat{y}_{i,t}, y_{i,t})$
$\hat{\tau}_{\delta,t}^{(k)} = \frac{1}{|V_\delta^{(k)}|}\sum_{i\in V_\delta^{(k)}}(y_{i,t} - \hat{y}_{i,t})$
Cross-Fitting
Apply $\hat{\tau}$ from the other fold
Selective Correction $\hat{y}_{i,t}^{\mathrm{rect}} = \hat{y}_{i,t} + \mathbf{1}(|\hat{\tau}_{\delta_i,t}| > \kappa) \cdot \hat{\tau}_{\delta_i,t}$
Final Outputs
$\hat{y}_{i,t}^{\mathrm{rect}}$ for each task t
Stage 1 Representation Learning $\mathcal{L}_{miss} + \mathcal{L}_{rep}$
Stage 2: Multitask Prediction with rectifier $\mathcal{L}_{pred}$

Figure 3: Overview of our CRL-MMNAR method. Stage 1 learns patient representations $\boldsymbol{h}_i$ by fusing multimodal inputs with missingness embeddings and optimizing reconstruction and contrastive losses. Stage 2 predicts clinical outcomes with multitask heads and applies a rectifier to correct missingness-induced bias.

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.

3.2 Patient Representation Learning

3.2.1 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 $\boldsymbol{z}_i = \mathrm{MLP}(\boldsymbol{\delta}_i)$, 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:

$$\mathcal{L}_{\mathrm{miss}} = \lambda_{\mathrm{miss}}\mathrm{CrossEntropy}\left(\hat{\boldsymbol{\delta}}_i, \boldsymbol{\delta}_i\right),$$

where $\hat{\boldsymbol{\delta}}_i$ is decoded from $\boldsymbol{z}_i$ and λ_{miss} is the hyper-parameter controlling the loss weight.

Each modality-specific embedding $e_i^{(m)}$ is then reweighted according to the missingness embedding $\boldsymbol{z}_i$:

$$e_{i,\mathrm{gate}}^{(m)} = \delta_i^{(m)} \cdot \sigma\left(W^{(m)}\boldsymbol{z}_i + b^{(m)}\right) \cdot e_i^{(m)},$$

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 $\{e_{i,\mathrm{gate}}^{(m)}\}_{m\in\mathcal{M}}$ are then aggregated with a multi-head self-attention mechanism:

$$\boldsymbol{h}_i = r_{\mathrm{fuse}}\left(\{e_{i,\mathrm{gate}}^{(m)}\}_{m\in\mathcal{M}}\right),$$

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.

3.2.2 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 3.2.1. A modality-specific decoder $\phi^{(m)}$ then attempts to reconstruct the excluded embedding:

$$e_{i,\text{rec}}^{(m)} = \phi^{(m)}\big(\boldsymbol{h}_i^{\setminus m}\big)\,.$$

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.

Aggregating across modalities, the representation learning objective is

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

where λ_{rec} and λ_{cont} are hyperparameters controlling the relative importance of reconstruction and contrastive objectives.

3.3 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:

$$y_{i,t} = g_{\psi_t'}\big(\boldsymbol{h}_i\big) + \tau_{\boldsymbol{\delta}_i,t} + \varepsilon_{i,t}\,. \tag{4}$$

The first term, $g_{\psi_t'}$, captures variation explained by the shared 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. We estimate this model in two steps.

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

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

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

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:

$$\hat{\tau}_{\boldsymbol{\delta},t}^{(k)} = \frac{1}{|\mathcal{V}_{\boldsymbol{\delta}}^{(k)}|} \sum_{i \in \mathcal{V}_{\boldsymbol{\delta}}^{(k)}} \big(y_{i,t} - \hat{y}_{i,t}\big)\,.$$

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:

$$\hat{y}^{\text{rect}}_{i,t} = \hat{y}_{i,t} + \mathbb{1}\left\{\left|\hat{\tau}^{(k)}_{\boldsymbol{\delta}_i,t}\right| > \kappa\right\} \cdot \hat{\tau}^{(k)}_{\boldsymbol{\delta}_i,t}, \quad \text{for } i \in \mathcal{V}^{(\bar{k})}_{\boldsymbol{\delta}_i},$$

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 (see a numerical example in Appendix A). 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.

Our rectifier resembles debiasing strategies from the semiparametric inference literature on missing data and multiple imputation (Robins et al., 1994; Robins and Rotnitzky, 1995). In particular, it parallels augmented inverse probability weighting estimators, which achieve doubly robust consistency by adding augmentation terms to correct for model misspecification. In the same spirit, our rectifier introduces validation-based corrections that account for modality-assignment effects not captured by the base model. This shares conceptual similarity with predictive powered inference (Angelopoulos et al., 2023), where post-hoc adjustments improve predictive reliability under misspecification.

The key distinction is that here the "treatment effect" arises from the modality observation pattern itself, rather than an externally defined treatment as in traditional causal inference. To implement this robustly, we adopt cross-fitting, which mirrors the sample-splitting principle in semiparametric methods: residual-based corrections are estimated on held-out folds and then applied to complementary folds, thereby reducing overfitting and improving generalization (Chernozhukov et al., 2018). While two folds are sufficient in practice, the approach naturally extends to multiple folds.

3.4 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:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{miss}}}_{\text{Section 3.2.1}} + \underbrace{\mathcal{L}_{\text{rep}}}_{\text{Section 3.2.2}} + \underbrace{\mathcal{L}_{\text{pred}}}_{\text{Section 3.3}},$$

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.

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. Hyperparameter values and implementation details are provided in Appendix C.2.

4 Experiments

4.1 Datasets and Preprocessing

To demonstrate the robustness and generalizability of CRL-MMNAR across diverse healthcare settings, we conduct comprehensive experiments on two large-scale clinical datasets representing fundamentally different healthcare delivery models.

MIMIC-IV Dataset 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.

eICU Dataset We further evaluate on the eICU Collaborative Research Database (Pollard et al., 2018), which spans 208 hospitals across the United States. This multi-center design allows us to test generalizability across institutions and care models.

Since eICU consists primarily of structured data, we define *virtual modalities*. Features are grouped into 10 clinical categories: Demographics, Admission, APACHE, Diagnosis, Medication, Laboratory Values, Vital Signs, Respiratory, Fluid Balance, and Comorbidities. For each virtual modality m, which consists of a group of related features (e.g., all laboratory measurements or all vitals), we treat it as missing for patient i if more than 80% of the features within that group are NaN. This virtual modality approach follows established practices for handling heterogeneous EHR data, where clinical features are systematically grouped into meaningful categories to enable effective multimodal learning (Wang et al., 2024; Li et al., 2022).

Detailed preprocessing procedures, feature extraction methods, and virtual modality specifications for eICU are provided in Appendix C.1.

Clinical Tasks We evaluate on three clinically important 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.

4.2 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 ± standard deviation, computed over five independent runs with different random seeds for initialization. Full dependency and package versions are provided in Appendix C.7.

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 (Section 3.4).

4.3 Main Results

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.

Performance Analysis 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. The consistency of results across two distinct healthcare contexts highlights the generalizability of our MMNAR-aware framework.

Robustness and Validation Studies We conduct supplementary analyses to evaluate robustness, stability, and efficiency.

Table 1: Performance comparison across datasets and clinical tasks. Best results in **bold**. Results reported as mean ± 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	**0.8657**$_{\pm 0.018}$	**0.5627**$_{\pm 0.028}$	**0.1167**$_{\pm 0.012}$	**0.9824**$_{\pm 0.016}$	**0.5821**$_{\pm 0.024}$	**0.0589**$_{\pm 0.007}$	**0.9472**$_{\pm 0.014}$	**0.4767**$_{\pm 0.026}$	**0.0759**$_{\pm 0.009}$

Model	eICU Dataset					
	30-day Readmission			In-hospital Mortality		
	AUC	AUPRC	Brier	AUC	AUPRC	Brier
CM-AE	0.7345$_{\pm 0.048}$	0.4189$_{\pm 0.061}$	0.1812$_{\pm 0.031}$	0.8478$_{\pm 0.035}$	0.3745$_{\pm 0.052}$	0.1067$_{\pm 0.020}$
MT	0.7456$_{\pm 0.044}$	0.4298$_{\pm 0.057}$	0.1756$_{\pm 0.029}$	0.8634$_{\pm 0.031}$	0.3945$_{\pm 0.048}$	0.0998$_{\pm 0.018}$
SMIL	0.7412$_{\pm 0.046}$	0.4323$_{\pm 0.059}$	0.1767$_{\pm 0.029}$	0.8567$_{\pm 0.033}$	0.3898$_{\pm 0.050}$	0.1023$_{\pm 0.018}$
GRAPE	0.7523$_{\pm 0.042}$	0.4367$_{\pm 0.054}$	0.1723$_{\pm 0.027}$	0.8756$_{\pm 0.029}$	0.3978$_{\pm 0.046}$	0.0934$_{\pm 0.016}$
HGMF	0.7489$_{\pm 0.044}$	0.4312$_{\pm 0.057}$	0.1734$_{\pm 0.027}$	0.8678$_{\pm 0.031}$	0.3956$_{\pm 0.048}$	0.0956$_{\pm 0.018}$
M3Care	0.7423$_{\pm 0.046}$	0.4278$_{\pm 0.059}$	0.1756$_{\pm 0.029}$	0.8823$_{\pm 0.027}$	0.4012$_{\pm 0.046}$	0.0889$_{\pm 0.016}$
COM	0.7512$_{\pm 0.042}$	0.4267$_{\pm 0.061}$	0.1734$_{\pm 0.027}$	0.8667$_{\pm 0.031}$	0.3567$_{\pm 0.054}$	0.0967$_{\pm 0.018}$
DrFuse	0.7698$_{\pm 0.037}$	0.4445$_{\pm 0.052}$	0.1645$_{\pm 0.025}$	0.8778$_{\pm 0.027}$	0.3967$_{\pm 0.046}$	0.0912$_{\pm 0.016}$
MissModal	0.7378$_{\pm 0.048}$	0.4134$_{\pm 0.063}$	0.1789$_{\pm 0.031}$	0.8589$_{\pm 0.033}$	0.3723$_{\pm 0.054}$	0.1012$_{\pm 0.018}$
FLEXGEN-EHR	0.7745$_{\pm 0.035}$	0.4434$_{\pm 0.050}$	0.1634$_{\pm 0.022}$	0.8778$_{\pm 0.025}$	0.3998$_{\pm 0.044}$	0.0912$_{\pm 0.014}$
MUSE+	0.8167$_{\pm 0.033}$	0.4756$_{\pm 0.046}$	0.1589$_{\pm 0.020}$	0.9334$_{\pm 0.020}$	0.4334$_{\pm 0.039}$	0.0798$_{\pm 0.012}$
GRU-D	0.7178$_{\pm 0.054}$	0.3967$_{\pm 0.068}$	0.1878$_{\pm 0.035}$	0.8312$_{\pm 0.039}$	0.3234$_{\pm 0.059}$	0.1145$_{\pm 0.022}$
Raindrop	0.7743$_{\pm 0.037}$	0.4334$_{\pm 0.054}$	0.1634$_{\pm 0.023}$	0.8634$_{\pm 0.031}$	0.3789$_{\pm 0.050}$	0.0978$_{\pm 0.016}$
CRL-MMNAR	**0.9294**$_{\pm 0.048}$	**0.6543**$_{\pm 0.031}$	**0.1142**$_{\pm 0.009}$	**0.9380**$_{\pm 0.016}$	**0.4973**$_{\pm 0.033}$	**0.0672**$_{\pm 0.007}$

First, performance across modality configurations (Table 3, Appendix C.3) shows steady gains as more modalities are added and graceful degradation under severe missingness. Even with limited inputs such as structured data and text, our method outperforms traditional imputation baselines.

Second, hyperparameter sensitivity analysis (Table 4, Appendix C.4) demonstrates strong parameter stability, with sensitivity scores below 0.31% across all key hyperparameters.

Third, efficiency evaluation (Appendix C.5) confirms computational feasibility: training requires 12-16 hours on 20,000 patients with an RTX A6000 GPU, and inference takes only 50–100ms per patient, with modest memory usage.

Finally, embedding analysis (Appendix C.6) supports our MMNAR assumptions. Learned embeddings predict missingness patterns with 92.3% accuracy and correlate strongly with clinical outcomes ($r = 0.67, p < 0.001$).

4.4 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.

5 Related Work

Our work is most closely related to the growing literature on multimodal representation learning, and

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

Component	MIMIC-IV Dataset									eICU Dataset					
	30-day Readmission			Post-discharge ICU			In-hosp. Mortality			30-day Readmission			In-hosp. Mortality		
	AUC	APR	ΔAUC	AUC	APR	ΔAUC	AUC	APR	ΔAUC	AUC	APR	ΔAUC	AUC	APR	ΔAUC
Basic Baseline	.717	.347	–	.801	.307	–	.814	.369	–	.802	.352	–	.741	.346	–
+ MMNAR	.793	.470	+.076	.853	.372	+.052	.894	.381	+.080	.858	.516	+.056	.873	.388	+.132
+ MR	.811	.519	+.018	.889	.404	+.036	.929	.456	+.035	.929	.651	+.070	.900	.470	+.027
+ Rectifier	**.866**	**.563**	**+.055**	**.982**	**.582**	**+.093**	**.947**	**.477**	**+.018**	**.929**	**.654**	**+.001**	**.938**	**.497**	**+.038**

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. More related work can be found in Appendix B.

6 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 and eICU show consistent improvements over 13 state-of-the-art baselines, with notable gains of 8.4% AUC for readmission and 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.

7 Limitations

Our framework, though effective under Modality Missing-Not-at-Random (MMNAR) conditions, has several limitations. Its generalizability beyond MIMIC-IV and eICU remains uncertain without broader international validation, and reliance on proxy assumptions may not fully capture underlying causal mechanisms. Performance can degrade when missingness is random or driven by non-clinical factors, and rare modality patterns limit embedding reliability. The multi-component architecture, while modular, introduces complexity, computational demands, and interpretability challenges that may hinder deployment in low-resource settings. Future work should focus on lightweight variants, improved interpretability, and wider cross-institutional evaluation.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

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

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. 2023. Prediction-powered inference. *Science*, 382(6671):669–674.

Chandrashekhar Azad, Bharat Bhushan, Rohit Sharma, Achyut Shankar, Krishna Kant Singh, and Aditya Khamparia. 2022. Prediction model using smote, genetic algorithm and decision tree (pmsgd) for classification of diabetes mellitus. *Multimedia Systems*, 28.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

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

Jiayi Chen and Aidong Zhang. 2020. 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.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, pages 1–68.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Junting Duan, Markus Pelger, and Ruoxuan Xiong. 2024a. Factor analysis for causal inference on large non-stationary panels with endogenous treatment. *Available at SSRN 4823360*.

Junting Duan, Markus Pelger, and Ruoxuan Xiong. 2024b. Target pca: Transfer learning large dimensional panel data. *Journal of Econometrics*, 244(2):105521.

Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, and Yinghuan Shi. 2023. Towards semi-supervised learning with non-random missing labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16121–16131.

Huan He, William hao, Yuanzhe Xi, Yong Chen, Bradley Malin, and Joyce Ho. 2024. A flexible generative model for heterogeneous tabular EHR with missing modality. In *The Twelfth International Conference on Learning Representations*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

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

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

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. 2020. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 04, pages 4485–4492.

Dongyue Li, Haotian Ju, Aneesh Sharma, and Hongyang R Zhang. 2023a. 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.

Dongyue Li, Huy L Nguyen, and Hongyang R Zhang. 2023b. Identification of negative transfers in multitask learning using surrogate models. *Transactions on Machine Learning Research*.

Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. 2024a. 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.

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

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

Rui Li, Fenglong Ma, and Jing Gao. 2022. Integrating Multimodal Electronic Health Records for Diagnosis Prediction. *AMIA Annual Symposium Proceedings*, 2021:726–735.

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

Zhen Lin, Shubhendu Trivedi, Cao Xiao, and Jimeng Sun. 2023. Fast online value-maximizing prediction sets with conformal cost control. In *Proceedings of the 40th International Conference on Machine Learning*.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, Yang Yang, Lei Clifton, and David A. Clifton. 2023. A medical multimodal large language model for future pandemics. *npj Digital Medicine*, 6(1):226.

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

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

Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C Wallace. 2023. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 689–696.

Yasunobu Nohara, Koutarou Matsumoto, Hidehisa Soejima, and Naoki Nakashima. 2022. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214:106584.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178.

Shuwei Qian and Chongjun Wang. 2023. COM: Contrastive Masked-attention model for incomplete multimodal learning. *Neural Networks*, 162:443–455.

James M Robins and Andrea Rotnitzky. 1995. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

A. Sellergren, A. Kiraly, T. Pollard, W. Weng, Y. Liu, A. Uddin, and C. Chen. 2023. Generalized image embeddings for the mimic chest x-ray dataset (version 1.0).

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

Eline Stenwig, Giampiero Salvi, Pierluigi Salvo Rossi, and Nils Kristian Skjærvold. 2022. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Medical Research Methodology*, 22(1):53.

Brandon Theodorou, Lucas Glass, Cao Xiao, and Jimeng Sun. 2024. FRAMM: Fair ranking with missing modalities for clinical trial site selection. *Patterns*, 5(3):100944.

Yuanlong Wang, Changchang Yin, and Ping Zhang. 2024. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5).

Zifeng Wang, Cao Xiao, and Jimeng Sun. 2023. Autotrial: Prompting language models for clinical trial design. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12461–12472.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. *International Conference on Learning Representations*.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. 2024. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

Ruoxuan Xiong and Markus Pelger. 2023. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301.

Fan Yang, Hongyang R Zhang, Sen Wu, Christopher Re, and Weijie J Su. 2025. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26(113):1–88.

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. 2022. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194.

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

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

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022a. 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.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022b. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*.

A Rectifier Mechanism: Numerical Example

To illustrate the rectifier mechanism described in Section 3.3, we provide a concrete example.

Setup

- Missingness pattern δ: patients with no text notes
- Task t: 30-day readmission
- Validation fold estimate: $\hat{\tau}_{\delta,t}^{(k)} = +0.08$
- Threshold: $\kappa = 0.05$
- New patient i has base prediction $\hat{y}_{i,t} = 0.30$

Rectification Since $|\hat{\tau}_{\delta,t}^{(k)}| = 0.08 > \kappa$, we apply the correction:

$$\hat{y}_{i,t}^{\text{rect}} = 0.30 + \hat{\tau}_{\delta,t}^{(k)} = 0.30 + 0.08 = 0.38$$

Interpretation The rectifier increases the readmission risk from 30% to 38%, compensating for systematic underestimation observed among patients without clinical notes. This demonstrates how the MMNAR framework leverages missingness patterns as clinically meaningful signals.

B Additional Related Work

For completeness, we provide more detailed discussion of related research areas that complement the overview in Section 5. While Section 5 covered the baselines and immediate methodological context of our study, here we highlight additional streams of work in medical foundation models, interpretability, trial design, and multimodal data synthesis.

Medical Foundation Models and Prompting. Recent innovations in medical foundation models extend multimodal learning to large-scale LLMs. Med-MLLM (Liu et al., 2023) exemplifies a multimodal LLM that generalizes across visual and textual clinical data, achieving robust performance in few-shot pandemic prediction tasks. CHiLL (McInerney et al., 2023) leverages zero-shot prompting of LLMs to generate interpretable features from free-text notes, showing that clinically meaningful representations can be constructed without manual engineering. These efforts highlight complementary strategies for scalability beyond task-specific architectures.

Interpretability and Transparency. Model transparency remains essential for clinician adoption. Stenwig et al. (Stenwig et al., 2022) and Nohara et al. (Nohara et al., 2022) introduce SHAP-based interpretability frameworks for ICU mortality prediction, offering granular insight into model behavior. Such frameworks align with our emphasis on validating robustness under uncertainty and suggest directions for future introspective analyses of MMNAR-aware models.

Handling Non-Random Missingness. In the context of non-random missingness, Duan et al. (2023) propose PRG, a graph-based semi-supervised framework that improves label quality in MNAR settings, which conceptually parallels our MMNAR formulation.

Efficient Fusion and Clinical Applications. Efficient multimodal fusion has also been studied through architectural innovations. Nagrani et al. (2021) propose attention bottlenecks to optimize cross-modal communication while reducing computational costs, highlighting the benefits of architectural constraints for scalability. Fusion has also been extended to clinical trial design: AutoTrial (Wang et al., 2023) employs discrete-neural prompting to control eligibility criteria generation, illustrating the utility of prompt-based language model interventions in biomedical contexts.

Data Synthesis and Fairness. Healthcare applications include M3Care's multitask prediction under modality incompleteness (Zhang et al., 2022a), and methods addressing imbalanced data, uncertainty, or fairness (PMSGD (Azad et al., 2022), FavMac (Lin et al., 2023), FRAMM (Theodorou et al., 2024)). In data synthesis, EMIXER (Bengio et al., 2013) provides an end-to-end multimodal framework for generating diagnostic image–report pairs, offering strong benefits in low-label regimes. These directions highlight how multimodal learning can be extended to fairness, robustness, and decision-support applications beyond clinical prediction.

C Extended Experimental Results

C.1 Detailed Data Preprocessing

C.1.1 MIMIC-IV Dataset - Technical Details

Data Extraction 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.

Structured Features 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.

Text and Image Embeddings Discharge summaries and radiology reports are preprocessed and embedded using Bio_ClinicalBERT (Alsentzer et al., 2019), which is pre-trained on MIMIC-III clinical notes. Our preprocessing pipeline includes cleaning de-identification artifacts, normalizing medical abbreviations, segmenting long documents, and intelligently truncating to preserve sentence boundaries within the 512-token limit. For imaging, we use pretrained embeddings from the *Generalized Image Embeddings for the MIMIC Chest X-Ray dataset* (Sellergren et al., 2023), derived from frontal-view CXRs via a CNN trained on large-scale radiology corpora.

C.1.2 eICU Dataset - Virtual Modality Specifications

For the eICU dataset, we extracted information from patient, diagnosis, treatment, medication, laboratory, and APACHE tables, focusing on adult ICU stays (18–89 years old) with a minimum length of 12 hours and a maximum of 10 days. Admissions lacking crucial identifiers or features were excluded. Variables with over 80% missingness were removed, and other missing values were imputed using median or zero as appropriate.

The 10 virtual modalities are defined as follows:

1. **Demographics**: Patient characteristics including age, gender, ethnicity, admission weight, height, BMI, and hospital characteristics (teaching status, bed size).

2. **Admission**: Admission source (emergency department, ICU transfer, operating room, etc.) and ICU unit type (MICU, SICU, cardiac ICU, etc.).

3. **APACHE**: APACHE II components including age score, Glasgow Coma Scale components, acute physiology score, predicted ICU and hospital mortality, and ventilation status.

4. **Diagnosis**: Diagnostic information including diagnosis counts, primary diagnostic categories (cardiovascular, respiratory, neurologic, etc.), and surgical patient status.

5. **Medication**: Medication usage patterns including total and unique medication counts, and specific drug categories (antibiotics, vasopressors, sedatives).

6. **Laboratory Values**: Laboratory test results including complete blood count, basic metabolic panel, liver function tests, and statistical aggregations (mean, min, max, standard deviation) over the ICU stay.

7. **Vital Signs**: Physiological measurements including heart rate, respiratory rate, oxygen saturation, blood pressure (invasive and non-invasive), and derived parameters (shock index, calculated MAP).

8. **Respiratory**: Respiratory support parameters including mechanical ventilation status, fraction of inspired oxygen (FiO_2), and positive end-expiratory pressure (PEEP).

9. **Fluid Balance**: Fluid management data including total intake, output, net fluid balance over 24 hours, and positive fluid balance indicators.

10. **Comorbidities**: Pre-existing conditions including AIDS, hepatic failure, malignancies, immunosuppression, and chronic diseases (diabetes, hypertension, heart failure, COPD, chronic kidney disease).

Missingness Determination. The missingness indicator $\delta_i^{(m)} = 1$ if modality m is available for patient i (i.e., $\leq 80\%$ NaN values), and 0 otherwise. This threshold accounts for the clinical reality that some features within a group may be selectively unavailable while maintaining the group's overall clinical utility.

This virtual modality approach enables our MMNAR framework to model systematic patterns in clinical data collection across different hospitals and care protocols, where certain categories of information may be consistently missing due to institutional practices, patient acuity, or resource constraints. The same preprocessing strategies were applied to both datasets for consistency.

C.2 Training Hyperparameters

C.2.1 Phase 1: Self-Supervised Pretraining Hyperparameters

The self-supervised loss weights are set as follows:

- $\alpha = 1.0$ (cross-modal reconstruction loss weight)
- $\beta = 0.5$ (missingness-pattern classification loss weight)
- $\gamma = 0.3$ (contrastive loss weight)

These values were determined through validation set performance, with α receiving the highest weight to emphasize cross-modal consistency, β providing moderate supervision for missingness pattern learning, and γ contributing contrastive regularization.

C.2.2 Phase 2: Downstream Task Fine-tuning Hyperparameters

The multitask learning weights are configured as:

- $w_{\text{readm}} = 1.2$ (30-day readmission weight)
- $w_{\text{ICU}} = 1.0$ (post-discharge ICU admission weight)
- $w_{\text{mort}} = 1.5$ (in-hospital mortality weight)

The readmission task receives moderate upweighting due to class imbalance considerations, while mortality prediction receives the highest weight reflecting its clinical criticality and difficulty. ICU admission serves as the baseline task with unit weight.

C.3 Performance Across Modality Configurations

To evaluate robustness under varying data availability, Table 3 analyzes performance across different modality combinations on MIMIC-IV.

Our approach demonstrates consistent performance improvements as more modalities become available, with graceful degradation under severe missingness. Notably, even with only text modalities (S+T), our method substantially outperforms traditional imputation approaches using complete data.

Table 3: Performance across different modality availability patterns on MIMIC-IV dataset. S = Structured data, I = Imaging, T = Text, R = Radiology reports.

Configuration	Readmission		ICU Admission		Sample
	AUC	AUPRC	AUC	AUPRC	Count
S only (Structured)	0.7234	0.4583	0.8156	0.3421	857
S+R (Struct.+Radio.)	0.7645	0.4821	0.8743	0.4156	433
S+T (Struct.+Text)	0.7808	0.5127	0.8843	0.4398	177
S+T+R (No Imaging)	0.8234	0.5423	0.9156	0.4872	1,968
S+I+T+R (Complete)	**0.8657**	**0.5627**	**0.9824**	**0.5821**	931
Traditional Imputation:					
Zero-filling + Masking	0.7234	0.4412	0.8156	0.3298	–
Mean Imputation	0.7156	0.4298	0.8091	0.3187	–

Table 4: Hyperparameter sensitivity analysis showing stability across parameter ranges.

Parameter	Default	Range Tested	Max Drop	Stability	AUC Range
Learning Rate	0.0002	[1e-5, 1e-3]	0.0048	99.69%	0.767-0.777
Weight Decay	1e-05	[1e-6, 1e-4]	0.0018	99.80%	0.773-0.778
Hidden Dimension	128	[64, 512]	0.0029	99.78%	0.772-0.777
Dropout Rate	0.2	[0.1, 0.5]	0.0071	99.74%	0.767-0.774
Contrastive Weight	0.15	[0.05, 0.3]	0.0017	99.80%	0.774-0.779

C.4 Hyperparameter Sensitivity Analysis

Table 4 demonstrates the robustness of our framework across key hyperparameter variations.

All hyperparameters exhibit exceptional stability with sensitivity scores below 0.31% and performance variations within narrow bounds, significantly reducing deployment risk in clinical settings.

C.5 Computational Efficiency Analysis

Our framework is computationally efficient, requiring 12–16 hours of training on 20,000 patients using an RTX A6000 GPU (48GB) and only 50–100 milliseconds for a single patient prediction. Memory demands are modest, with about 12GB during training and 3GB for inference, primarily driven by Transformer components that scale predictably with data size and modality complexity. The modular design further allows selective activation of components, enabling deployment in standard clinical computing environments without specialized hardware.

C.6 MMNAR Embedding Validation

To provide empirical validation of our MMNAR modeling approach, we conduct systematic analysis of the learned missingness embeddings z_i across all modality patterns. We examine embedding norms, clinical outcome associations, and predictive capabilities to demonstrate that the learned representations capture meaningful clinical decision-making factors rather than random noise.

Systematic Embedding Analysis. We analyze embedding characteristics across modality availability patterns, revealing clear, clinically interpretable gradients. Embedding norms systematically vary from 10.72 to 13.44 across different missingness patterns, demonstrating learned differentiation beyond simple pattern counting. Strong correlation exists between embedding characteristics and clinical outcomes (r=0.67, p<0.001 for readmission rates), confirming that learned embeddings capture meaningful signals related to clinical decision pathways and patient risk.

Missingness Pattern Prediction. The learned embeddings achieve 92.3% accuracy in predicting the original 16 missingness patterns from the latent representation alone, providing direct evidence that z_i encodes systematic information about modality availability rather than capturing random variation.

This validation approach, while necessarily indirect given the observational nature of EHR data, provides strong proxy evidence that our MMNAR assumptions are well-founded and that the learned embeddings capture clinically meaningful factors influencing data collection decisions.

C.7 Dependencies and Package Versions

Table 5: Key Python package dependencies and versions

Package	Version	Package	Version
Python	3.8.10	torch	1.13.1+cu116
numpy	1.23.5	matplotlib	3.6.2
pandas	1.5.2	seaborn	0.12.1
scipy	1.9.3	tqdm	4.64.1
scikit-learn	1.1.3	imbalanced-learn	0.10.1

C.8 Algorithm

Algorithm 1 MMNAR-Aware Modality Fusion (Stage 1A)

```
Input: Observed-modality inputs $\{x_i^{(m)}\}_{m \in \mathcal{M}_i}$ for patient $i$; missingness vector $\boldsymbol{\delta}_i \in \{0,1\}^{|\mathcal{M}|}$
Output: Fused patient representation $\boldsymbol{h}_i$
// Encode each observed modality (Sec. 3.2.1)
foreach modality $m \in \mathcal{M}_i$ do
    if $\delta_i^{(m)} = 1$ then
        $e_i^{(m)} \leftarrow \mathsf{Enc}^{(m)}(x_i^{(m)})$ ;                // modality-specific encoder
// Missingness embedding and self-supervised reconstruction of $\boldsymbol{\delta}_i$ (Sec. 3.2.1)
$\boldsymbol{z}_i \leftarrow \mathsf{MLP}_{\text{miss}}(\boldsymbol{\delta}_i);\ \hat{\boldsymbol{\delta}}_i \leftarrow \mathsf{Dec}_{\text{miss}}(\boldsymbol{z}_i)$ ;     // Missingness embedding decoder
$\mathcal{L}_{\text{miss}} = \lambda_{\text{miss}} \sum_i \mathsf{CrossEntropy}(\hat{\boldsymbol{\delta}}_i,\ \boldsymbol{\delta}_i)$;
// Missingness-aware gating per modality (Sec. 3.2.1)
foreach $m \in \mathcal{M}$ do
    $g_i^{(m)} \leftarrow \sigma\big(W^{(m)}\boldsymbol{z}_i + b^{(m)}\big);\ e_{i,\text{gate}}^{(m)} \leftarrow \boldsymbol{\delta}_i^{(m)} \cdot g_i^{(m)} \odot e_i^{(m)}$ ;     // Hadamard $\odot$
// Attention-based fusion with mask $\boldsymbol{\delta}_i$ (Sec. 3.2.1)
$\boldsymbol{E}_i \leftarrow \mathsf{stack}\big(\{e_{i,\text{gate}}^{(m)}\}_{m \in \mathcal{M}}\big);\ \boldsymbol{H}_i \leftarrow \mathsf{TransformerFuse}(\boldsymbol{E}_i,\ \text{mask} = \boldsymbol{\delta}_i);\ \boldsymbol{h}_i \leftarrow \mathsf{MeanPooling}(\boldsymbol{H}_i)$;
return $\boldsymbol{h}_i$
```

Algorithm 2 Modality Reconstruction with Contrastive Alignment (Stage 1B)

```
Input: Gated modality embeddings $\{e_{i,\text{gate}}^{(m)}\}_{m \in \mathcal{M}}$; original embeddings $\{e_i^{(m)}\}_{m \in \mathcal{M}_i}$; missingness $\boldsymbol{\delta}_i$
Output: Reconstruction losses $\{\mathcal{L}_{\text{rec}}^{(m)}\}_{m \in \mathcal{M}}$ and contrastive losses $\{\mathcal{L}_{\text{cont}}^{(m)}\}_{m \in \mathcal{M}}$
// Cross-modality reconstruction (Sec. 3.2.2)
foreach $m \in \mathcal{M}$ do
    $\boldsymbol{h}_i^{\setminus m} \leftarrow \mathsf{Fuse}\big(\{e_{i,\text{gate}}^{(k)} \mid k \neq m\}\big);\ \hat{e}_i^{(m)} \leftarrow \phi^{(m)}\big(\boldsymbol{h}_i^{\setminus m}\big);\ \mathcal{L}_{\text{rec}}^{(m)} \leftarrow \sum_i \delta_i^{(m)} \cdot \big\|e_i^{(m)} - \hat{e}_i^{(m)}\big\|_2^2$;
// InfoNCE contrastive alignment (Sec. 3.2.2)
foreach $m \in \mathcal{M}$ do
    $\mathcal{L}_{\text{cont}}^{(m)} \leftarrow -\log \dfrac{\exp\big(\text{sim}(e_i^{(m)}, \hat{e}_i^{(m)})/\tau_{\text{cont}}\big)}{\sum_{j=1}^{N} \exp\big(\text{sim}(e_i^{(m)}, \hat{e}_j^{(m)})/\tau_{\text{cont}}\big)}$;
// Aggregate Stage-1 representation loss
$\mathcal{L}_{\text{rep}} \leftarrow \sum_{m \in \mathcal{M}} \big(\lambda_{\text{rec}}\mathcal{L}_{\text{rec}}^{(m)} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}}^{(m)}\big)$;
return $\{\mathcal{L}_{\text{rec}}^{(m)}\}_{m \in \mathcal{M}}, \{\mathcal{L}_{\text{cont}}^{(m)}\}_{m \in \mathcal{M}}$
```

Algorithm 3 Multitask Outcome Prediction with Cross-Fitted Rectifier (Stage 2)

```
Input: Shared representation $\boldsymbol{h}_i$; missingness $\boldsymbol{\delta}_i$; task heads $\{g_{\boldsymbol{\psi}'_t}\}_{t\in\mathcal{T}}$; threshold $\kappa \geq 0$
Output: Rectified predictions $\{\hat{y}^{\text{rect}}_{i,t}\}_{t\in\mathcal{T}}$
// Base multitask prediction (Eq. (4) first term; Sec. 3.3)
foreach $t \in \mathcal{T}$ do
    $\hat{y}_{i,t} \leftarrow g_{\boldsymbol{\psi}'_t}(\boldsymbol{h}_i)$;
$\mathcal{L}_{\text{pred}} \leftarrow \sum_{t\in\mathcal{T}} \lambda_{\text{pred},t} \cdot \texttt{CrossEntropy}(\hat{y}_{i,t},\, y_{i,t})$;
// Cross-fitting rectifier per pattern $\boldsymbol{\delta}$ (Sec. 3.3)
foreach pattern $\boldsymbol{\delta} \in \{0,1\}^{|\mathcal{M}|}$ do
    $V_{\boldsymbol{\delta}} \leftarrow \{\, i : \boldsymbol{\delta}_i = \boldsymbol{\delta} \,\}$; split $V_{\boldsymbol{\delta}}$ into $V^{(1)}_{\boldsymbol{\delta}}$ and $V^{(2)}_{\boldsymbol{\delta}}$ (disjoint);
    foreach $t \in \mathcal{T}$ do
        for $k \in \{1,2\}$ do
            $\hat{\tau}^{(k)}_{\boldsymbol{\delta},t} \leftarrow \frac{1}{|V^{(k)}_{\boldsymbol{\delta}}|} \sum_{i\in V^{(k)}_{\boldsymbol{\delta}}} \big(y_{i,t} - \hat{y}^{(\bar{k})}_{i,t}\big)$ ; // predictions from model not fit on fold $k$
            foreach $i \in V^{(\bar{k})}_{\boldsymbol{\delta}}$ do
                $\hat{y}^{\text{rect}}_{i,t} \leftarrow \hat{y}_{i,t} + \mathbf{1}\big(|\hat{\tau}^{(k)}_{\boldsymbol{\delta},t}| > \kappa\big) \cdot \hat{\tau}^{(k)}_{\boldsymbol{\delta},t}$;
return $\{\hat{y}^{\text{rect}}_{i,t}\}_{t\in\mathcal{T}}$
```

Algorithm 4 End-to-End Training and Inference

```
Input: Training data $\{(x^{(m)}_i)_{m\in\mathcal{M}_i},\, \boldsymbol{\delta}_i,\, (y_{i,t})_{t\in\mathcal{T}}\}$; hyperparameters $(\lambda_{\text{rec}}, \lambda_{\text{cont}}, \{\lambda_{\text{pred},t}\}_{t\in\mathcal{T}})$
Output: Trained encoder parameters $\boldsymbol{\eta}$, task heads $\{\boldsymbol{\psi}'_t\}_{t\in\mathcal{T}}$, and rectified predictions at inference
foreach minibatch do
    // Stage 1A: fusion
    $\boldsymbol{h}_i \leftarrow$ Alg. 1$(\{x^{(m)}_i\}_{m\in\mathcal{M}_i}, \boldsymbol{\delta}_i)$ ;        // Stage 1B: reconstruction + contrastive
    $\{\mathcal{L}^{(m)}_{\text{rec}}\}, \{\mathcal{L}^{(m)}_{\text{cont}}\} \leftarrow$ Alg. 2$(\{e^{(m)}_{i,\text{gate}}\}_{m\in\mathcal{M}}, \{e^{(m)}_i\}_{m\in\mathcal{M}_i}, \boldsymbol{\delta}_i)$ ; // Stage 2: multitask base loss
    compute $\mathcal{L}_{\text{pred}}$ with current $\{\boldsymbol{\psi}'_t\}_{t\in\mathcal{T}}$ as in Alg. 3 ;        // Total loss (Sec. 3.4)
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{miss}} + \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{pred}}$;
    update $\eta$ and $\{\boldsymbol{\psi}'_t\}_{t\in\mathcal{T}}$ by backprop on $\mathcal{L}_{\text{total}}$;
// Post-training rectification (Sec. 3.3)
apply Alg. 3 on validation/test folds to obtain $\{\hat{y}^{\text{rect}}_{i,t}\}_{t\in\mathcal{T}}$;
return trained $(\boldsymbol{\eta}, \{\boldsymbol{\psi}'_t\}_{t\in\mathcal{T}})$ and rectified predictions
```