CTPD: Cross-Modal Temporal Pattern Discovery for Enhanced Multimodal Electronic Health Records Analysis

Fuying Wang[1]*, Feng Wu[1]*, Yihan Tang[1], Lequan Yu[1]

[1]Department of Statistics and Actuarial Science,
School of Computing and Data Science,
The University of Hong Kong
{fuyingw,fengwu96,hank-tang}@connect.hku.hk, lqyu@hku.hk

Correspondence: lqyu@hku.hk

Abstract

Integrating multimodal Electronic Health Records (EHR) data—such as numerical time series and free-text clinical reports—has great potential in predicting clinical outcomes. However, prior work has primarily focused on capturing temporal interactions within individual samples and fusing multimodal information, overlooking critical temporal patterns across patients. These patterns, such as trends in vital signs like abnormal heart rate or blood pressure, can indicate deteriorating health or an impending critical event. Similarly, clinical notes often contain textual descriptions that reflect these patterns. Identifying corresponding temporal patterns across different modalities is crucial for improving the accuracy of clinical outcome predictions, yet it remains a challenging task. To address this gap, we introduce a **C**ross-Modal **T**emporal **P**attern **D**iscovery (CTPD) framework, designed to efficiently extract meaningful cross-modal temporal patterns from multimodal EHR data. Our approach introduces shared initial temporal pattern representations which are refined using slot attention to generate temporal semantic embeddings. To ensure rich cross-modal temporal semantics in the learned patterns, we introduce a Temporal Pattern Noise Contrastive Estimation (TP-NCE) loss for cross-modal alignment, along with two reconstruction losses to retain core information of each modality. Evaluations on two clinically critical tasks—48-hour in-hospital mortality and 24-hour phenotype classification—using the MIMIC-III database demonstrate the superiority of our method over existing approaches. The code is anonymously available at https://github.com/HKU-MedAI/CTPD.

1 Introduction

The increasing availability of Electronic Health Records (EHR) presents significant opportunities for advancing predictive modeling in healthcare (Acosta et al., 2022; Wang et al., 2024). EHR data is inherently multimodal and time-aware, encompassing structured data like vital signs, laboratory results, and medications, as well as unstructured data such as free-text clinical reports (Kim et al., 2023). Integrating these diverse data types is crucial for comprehensive patient monitoring and accurate prediction of clinical outcomes (Hayat et al., 2022; Wang et al., 2023; Zhang et al., 2023b). However, the irregularity and heterogeneity of multimodal data present significant challenges for precise outcome prediction.

Existing approaches primarily address either the irregularity of data (Shukla and Marlin, 2019; Horn et al., 2020; Zhang et al., 2021, 2023b) or the fusion of multiple modalities (Huang et al., 2020; Zhang et al., 2020b; Xu et al., 2021; Kline et al., 2022), but they often neglect broader temporal trends that span across patient cases. These cross-modal temporal patterns, present in both structured and unstructured data, can provide high-level semantic insights into a patient's health trajectory and potential risks (Conrad et al., 2018). As illustrated in Fig. 1, high-level semantic patterns related to medical conditions can emerge across multiple modalities. Capturing these patterns in a cross-modal, temporal manner is essential for improving predictive performance. Furthermore, critical temporal patterns in EHR data unfold at different time scales (Zhang et al., 2023a; Luo et al., 2020; Ma et al., 2020; Ye et al., 2020), yet existing methods struggle to capture variations across these granularities. For instance, sudden changes in vital signs—such as a sharp drop in oxygen saturation or a rapid heart rate increase—may indicate an acute health crisis. In contrast, longer-term trends, such as persistently high blood pressure or a gradual decline in respiratory function, may signal deteriorating health or an impending critical event. Effectively analyzing patient states across multiple time scales is crucial for comprehensive EHR modeling, yet remains an open challenge in current methodologies.

Findings of the Association for Computational Linguistics: ACL 2025, pages 6783–6799
July 27 - August 1, 2025

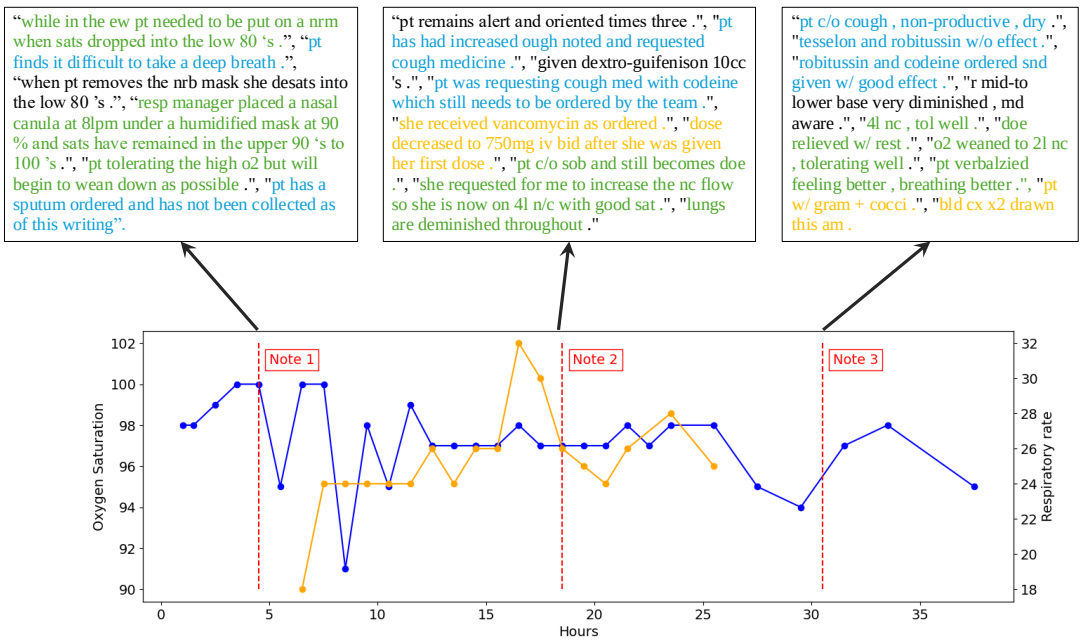


Figure 1: **Motivation of our proposed CTPD**: we visualized the time-series EHR with corresponding clinical notes in one ICU stay of the MIMIC-III dataset, and observed the temporal patterns across two modalities: **Blue text** highlights respiratory status. Oxygen requirements gradually decreased from 8L to 4L, and then to 2L nasal cannula, indicating steady respiratory improvement. Note that this pattern is also reflected from the time series. **Green text** captures cough progression and medication effects. Symptom relief was observed after administering Robitussin with codeine, demonstrating a delayed but positive response to treatment. **Yellow text** represents infection monitoring. The detection of Gram-positive cocci prompted blood culture collection (bld cx) for further evaluation, indicating active infection surveillance.

To address these limitations, we propose the **C**ross-modal **T**emporal **P**attern **D**iscovery (CTPD) framework, designed to extract meaningful temporal patterns from multimodal EHR data to improve the accuracy of clinical outcome predictions. The core innovation of our approach is a novel temporal pattern discovery module, which identifies corresponding temporal patterns (i.e., temporal prototypes) with meaningful semantics across both modalities throughout the dataset. This approach ensures that the model captures essential temporal semantics relevant to patient outcomes, providing a more comprehensive understanding of the data. To further enhance the quality of the learned temporal patterns, we introduce a Temporal Pattern Noise Contrastive Estimation (TP-NCE) loss for aligning pattern embeddings across modalities, along with auxiliary reconstruction losses to ensure that the patterns retain core information of the data. Moreover, our framework incorporates a transformer (Vaswani et al., 2017)-based fusion mechanism to effectively fuse the discovered temporal patterns with timestamp-level representations from both modalities, leading to more accurate predictions. We evaluate CTPD on two critical clinical prediction tasks: 48-hour in-hospital mortality prediction and 24-hour phenotype classification, using the MIMIC-III database. The results demonstrate the effectiveness of our approach, which significantly outperforms existing methods, and suggest a promising direction for improving multimodal EHR analysis for clinical prediction.

2 Related Works

2.1 EHR Time-Series Data Analysis

EHR data is critical for clinical tasks such as disease diagnosis, mortality prediction, and treatment planning (Harutyunyan et al., 2019; Zhang et al., 2023b). However, its high dimensionality and irregular nature pose challenges for traditional predictive models (Rani and Sikka, 2012; Lee et al., 2017). Deep learning models, such as RNNs and LSTMs, are often used to capture temporal dependencies in EHR data (Hayat et al., 2022; Deldari et al., 2023), but they struggle with irregular time intervals due to their reliance on fixed-length sequences (Xie et al., 2021). To address this, some methods update patient representations at each time step using graph neural networks (Zhang et al., 2021), while others employ time-aware embeddings to incorpo-

rate temporal information (Qian et al., 2023; Zhang et al., 2023b). Despite these advancements, existing approaches still struggle to model high-level temporal patterns essential for accurate clinical outcome prediction.

2.2 Prototype-based Pattern Learning

Prototype-based learning identifies representative instances (prototypes) and optimizes their distance from input data in latent space for tasks like classification (Li et al., 2023a; Ye et al., 2024). This approach has been widely applied in tasks such as anomaly detection, unsupervised learning, and few-shot learning (Tanwisuth et al., 2021; Li et al., 2023b). Recently, it has been extended to time-series data (Ghosal and Abbasi-Asl, 2021; Li et al., 2023a; Yu et al., 2024), demonstrating its potential for detecting complex temporal patterns. Additionally, prototype-based learning offers interpretable predictions, which is essential for healthcare applications (Zhang et al., 2024). However, learning efficient cross-modal temporal prototypes for multimodal EHR data remains an unexplored problem, as irregular time series and multi-scale patterns present significant challenges for existing methods.

2.3 Multi-modal Learning in Healthcare

In healthcare, patient data is typically collected in various forms—such as vital signs, laboratory results, medications, medical images, and clinical notes—to provide a comprehensive view of a patient's health. Integrating these diverse modalities significantly enhances the performance of clinical tasks (Hayat et al., 2022; Zhang et al., 2023b; Yao et al., 2024). However, fusing multimodal data remains challenging due to the heterogeneity and complexity of the sources. Earlier research on multimodal learning (Trong et al., 2020; Ding et al., 2022; Hayat et al., 2022) often rely on late fusion strategies, where unimodal representations are combined via concatenation or Kronecker products. While straightforward, these approaches often fail to capture complex inter-modal interactions, leading to suboptimal representations. Recent works have introduced transformer-based models that focus on cross-modal token interactions (Zhang et al., 2023b; Theodorou et al., 2024; Yao et al., 2024). While these models are effective at capturing inter-modal relationships, they often struggle to extract high-level temporal semantics from multimodal data, limiting the ability to achieve a comprehensive understanding of a patient's health conditions.

3 Methodology

3.1 Problem Formulation

In practice, multimodal EHR datasets contain multiple data types, specifically Multivariate Irregular Time Series (MITS) and free-text clinical notes. We represent the multimodal EHR data for the i-th admission as $\{(\mathbf{x}_{(i)}, \mathbf{t}^{\text{TS}}_{(i)}), (n_{(i)}, \mathbf{t}^{\text{Text}}_{(i)}), \mathbf{y}_{(i)}\}_{i=1}^{N}$. Here $\mathbf{x}_{(i)}$ represents the multivariate time series observations, with $\mathbf{t}^{\text{TS}}_{(i)}$ indicating their corresponding time points. The sequence of clinical notes is represented by $n_{(i)}$, and $\mathbf{t}^{\text{Text}}_{(i)}$ denotes the time points of these notes. The variable $\mathbf{y}_i$ denotes the clinical outcomes to predict. For simplicity, we omit the admission index i in subsequent sections. The MITS $\mathbf{x}$ comprises d_m variables, where each variable $j = 1, ..., d_m$ has $l^{TS}_{(j)}$ observations, with the rest missing. Similarly, each clinical note sequence n includes l^{Text} notes. Early-stage medical prediction tasks aim to forecast an outcome y for the admission i using their multimodal EHR data $\{(\mathbf{x}, \mathbf{t}^{\text{TS}}), (n, \mathbf{t}^{\text{Text}})\}$, specifically before a certain time point (e.g., 48 hours) after admission. Here, both $\mathbf{t}^{\text{TS}}$ and $\mathbf{t}^{\text{Text}}$ are constrained within a fixed temporal window following admission (e.g., 48 or 24 hours). Apart from this shared temporal boundary, the two sets of timestamps are recorded independently and are not inherently correlated.

3.2 Encoding MITS and Clinical Notes

Here, we introduce our time series encoder E_{TS} and text encoder E_{Text}, which separately encode MITS $(\mathbf{x}, \mathbf{t}^{\text{TS}})$ and clinical notes $(n, \mathbf{t}^{\text{Text}})$ into their respective embeddings $\mathbf{z}^{\text{TS}}$ and $\mathbf{z}^{\text{Text}} \in \mathbb{R}^{T \times D}$. Here T denotes the number of regular time points, and D denotes the embedding dimension.

For MITS, we utilize a gating mechanism that dynamically integrates both irregular time series embeddings $\mathbf{e}^{\text{TS}}_{\text{imp}}$ and imputed regular time series embeddings $\mathbf{e}^{\text{TS}}_{\text{mTAND}}$, following the approach in (Zhang et al., 2023b). Formally, the MITS embedding $\mathbf{z}^{\text{TS}}$ is computed as:

$$\mathbf{z}^{\text{TS}} = \mathbf{g} \odot \mathbf{e}^{\text{TS}}_{\text{imp}} + (1 - \mathbf{g}) \odot \mathbf{e}^{\text{TS}}_{\text{mTAND}} \quad (1)$$

where $\mathbf{g} = f(\mathbf{e}^{\text{TS}}_{\text{imp}} \oplus \mathbf{e}^{\text{TS}}_{\text{mTAND}}) \in \mathbb{R}^{T \times D}$, $f(\cdot)$ is a gating function implemented via an MLP, $\oplus$ denotes the concatenation, and $\odot$ denotes pointwise multiplication.

The regular time series $\mathbf{e}^{\text{TS}}_{\text{imp}}$ embedding is derived by applying a 1D convolution layer to the imputed time series. At each reference time point

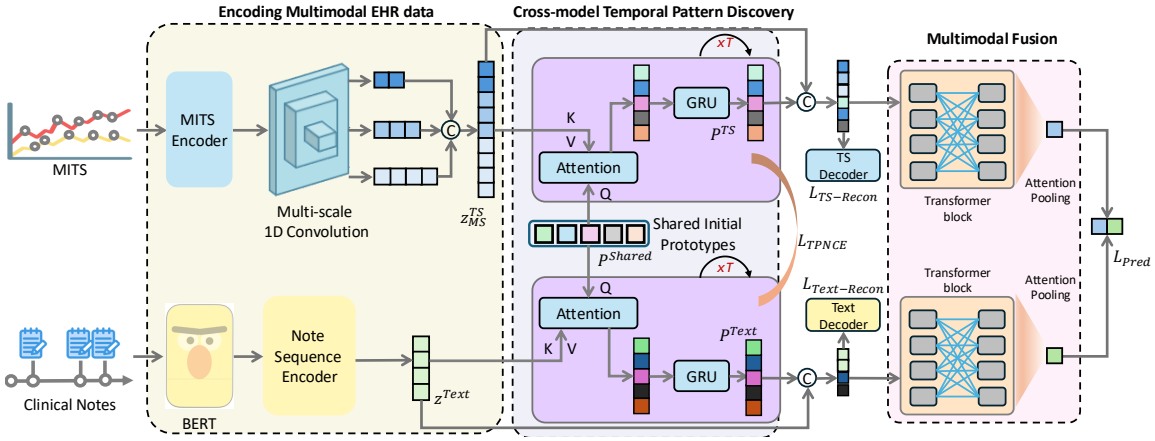


Figure 2: **CTPD overview**: the input Multivariate Irregular Time Series (MITS) and clinical note sequences are first encoded into regular embeddings. We then introduce the Cross-Modal Temporal Pattern Discovery (CTPD) module to extract meaningful temporal semantics. The extracted temporal patterns, along with the timestamp-level embeddings from both modalities, are fused to generate the final predictions.

$\alpha = 1, ..., T$, the imputed values are sourced from the nearest preceding values or replaced with a standard normal value if no prior data is available. Concurrently, mTAND (multi-time attention) (Shukla and Marlin, 2021) generates an alternative set of time series representations $\mathbf{e}^{\text{TS}}_{\text{mTAND}}$ with the same reference time points $\mathbf{r}$ with irregular time representations. Specifically, we leverage V different Time2Vec (Kazemi et al., 2019) functions $\{\theta_v(\cdot)\}_{v=1}^{V}$ to produce interpolation embeddings at each time point α, which are then concatenated and linearly projected to form $\mathbf{e}^{\text{TS}}_{\text{mTAND}}(\alpha) \in \mathbb{R}^D$.

For clinical notes, each of them is first encoded using Bert-tiny (a text encoder), where we extract the [CLS] token to represent this as a sequence of note-level embeddings $\mathbf{e}^{\text{Text}} \in \mathbb{R}^{T \times D}$. To tackle irregularity, we sort $\mathbf{e}^{\text{Text}}$ according to their timestamps $\mathbf{T}^{\text{Text}}$ and treat the pair $(\mathbf{e}^{\text{Text}}, \mathbf{T}^{\text{Text}})$ as a multimodal time series. Following the approach for time-series data, we leverage V different Time2Vec features $\{\theta_v\}_{v=1}^{V}$ to produce interpolation embeddings at each time point α:

$$\mathbf{z}^{\text{Text}} = \text{mTAND}(\alpha, \mathbf{T}^{\text{Text}}, \mathbf{e}^{\text{Text}}) \quad (2)$$

This yields temporally aligned text embeddings at each time point α, enabling effective interaction with time series data. Note that our choice of BERT-Tiny is based on prior work (Park et al., 2022) in multimodal EHR analysis, where its lightweight architecture has proven effective. Nonetheless, our framework can be seamlessly integrated with other text encoders.

3.3 Discover Cross-modal Temporal Patterns from Multimodal EHR

High-level temporal patterns in multimodal EHR data often encode rich medical condition-related semantics that are crucial for predicting clinical outcomes. However, previous works primarily focus on timestamp-level embeddings, frequently overlooking these important temporal patterns (Bahadori and Lipton, 2019; Xiao et al., 2023; Sun et al., 2024). Drawing inspiration from object-centric learning in the computer vision domain (Locatello et al., 2020; Li et al., 2021), we propose a novel temporal pattern discovery module to capture complex patterns within longitudinal data.

Considering the hierarchical nature (Yue et al., 2022; Cai et al., 2024) of time series data, the critical temporal patterns for EHR may manifest across multiple time scales. Consequently, our approach performs temporal pattern discovery on multi-scale time series embeddings.

Extracting Cross-modal Temporal Patterns. Owing to the correspondence within multi-modal data, our cross-modal temporal pattern discovery module focuses on extracting corresponding temporal patterns across both modalities for a better understanding of multimodal EHR. Starting with the time series embeddings $\mathbf{z}^{\text{TS}}$ in Eq. 1, we generate multi-scale embeddings (each scale can be divided by T) $\{\mathbf{z}^{\text{TS}}_{(1)}, \mathbf{z}^{\text{TS}}_{(2)}, \mathbf{z}^{\text{TS}}_{(3)}\}$ using three convolutional blocks followed by mean pooling with a stride of 2 along the time dimension. Each resulting embedding corresponds to a different temporal scale, where the sequence lengths are fractions of the original length T (e.g., T, $T/2$, and $T/4$). The concate-

nated multi-scale embedding $\mathbf{z}_{\text{MS}}^{\text{TS}} \in \mathbb{R}^{T' \times D}$ serves as the diverse temporal representation, where $T' = T + T/2 + T/4 = 1.75T$. Since T is set to either 48 or 24 in our experiments, T' remains an integer. We then enhance these embeddings by applying position encoding: $\hat{\mathbf{z}}_{\text{MS}}^{\text{TS}} = \mathbf{z}_{\text{MS}}^{\text{TS}} + \text{PE}(\mathbf{z}_{\text{MS}}^{\text{TS}})$, $\hat{\mathbf{z}}_{\text{MS}}^{\text{TS}} \in \mathbb{R}^{T' \times D}$, where $\text{PE}(\cdot)$ denotes the position embeddings in (Vaswani et al., 2017). Furthermore, to capture potential temporal patterns, we define a group of K learnable vectors as temporal prototypes, $\mathbf{P}^{\text{Shared}} \in \mathbb{R}^{K \times D}$, initially sampled from a normal distribution $\mathcal{N}(\mu, \text{diag}(\sigma)) \in \mathbb{R}^{K \times D}$ and refined during training. The shared prototype embeddings are designed to capture semantic-corresponding temporal patterns across modalities, respectively, with μ and σ randomly initialized and subsequently optimized.

To extract temporal patterns, we first calculate the assignment weights $\mathbf{W}$ between prototype embeddings and modality embeddings using a dot-product attention mechanism:

$$\begin{aligned} \mathbf{W}^{\text{TS}} &= \text{Attention}(\mathbf{P}^{\text{Shared}}, \hat{\mathbf{z}}_{\text{MS}}^{\text{TS}}), \\ \mathbf{W}^{\text{Text}} &= \text{Attention}(\mathbf{P}^{\text{Shared}}, \mathbf{z}^{\text{Text}}) \end{aligned} \tag{3}$$

The Attention mechanism is defined as:

$$\text{Attention}(\mathbf{q}, \mathbf{k})_{i,j} = \frac{e^{M_{i,j}}}{\sum_l e^{M_{i,l}}}, \tag{4}$$

where $\mathbf{M} = \frac{1}{\sqrt{D}} g_q(\mathbf{q}) \cdot g_k(\mathbf{k})^T$ and $g_q(\cdot)$ and $g_k(\cdot)$ are two learnable matrices. Next, we aggregate the input values to their assigned prototypes using a weighted mean to obtain updated embeddings $\mathbf{z}_{\text{updated}}^{\text{TS}}$ and $\mathbf{z}_{\text{updated}}^{\text{Text}}$:

$$\begin{aligned} \mathbf{z}_{\text{updated}}^{\text{TS}} &= \mathbf{W}^{\text{TS}} \cdot v(\hat{\mathbf{z}}_{\text{MS}}^{\text{TS}}) \in \mathbb{R}^{K \times D}, \\ \mathbf{z}_{\text{updated}}^{\text{Text}} &= \mathbf{W}^{\text{Text}} \cdot v(\hat{\mathbf{z}}^{\text{Text}}) \in \mathbb{R}^{K \times D}, \end{aligned} \tag{5}$$

where $v(\cdot)$ is a learnable matrix.

Finally, the prototype embeddings $\mathbf{P}^{\text{TS}}$ and $\mathbf{P}^{\text{Text}}$ are refined using the corresponding updated embeddings via a learned recurrent function:

$$\begin{aligned} \mathbf{P}^{\text{TS}} &= f(\text{GRU}(\mathbf{P}_{(0)}^{\text{TS}})) \in \mathbb{R}^{K \times D} \\ \mathbf{P}^{\text{Text}} &= f(\text{GRU}(\mathbf{P}_{(0)}^{\text{Text}})) \in \mathbb{R}^{K \times D} \end{aligned} \tag{6}$$

where $\mathbf{P}_{(0)}^{\text{TS}}$ and $\mathbf{P}_{(0)}^{\text{Text}}$ are the prototype embeddings from the previous step , $\text{GRU}(\cdot)$ is Gated Recurrent Unit (Cho et al., 2014), and $f(\cdot)$ denotes MLP. The above process is repeated for 3 iterations per step. Those refined embeddings denote the discovered temporal patterns for each modality. Note that our design of attention and GRU is inspired by Slot Attention (Locatello et al., 2020), which has been shown to be effective in learning object-centric representations from images.

TP-NCE Contraint. To ensure consistent semantics across modalities, we introduce a Temporal Pattern Noise Contrastive Estimation (TP-NCE) loss, inspired by InfoNCE (Oord et al., 2018), to enforce the similarity of multimodal prototype embeddings for the same ICU stay while increasing the distance between prototype embeddings from different ICU stays. For a minibatch of B samples, the TP-NCE loss from MITS to notes is defined as:

$$\mathcal{L}_{\text{TPNCE}}^{\text{TS}\to\text{Text}} = -\sum_{i=1}^{B} \left(\log \frac{\exp(\text{sim}(i,i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(i,j)/\tau)}\right) \tag{7}$$

where τ is a temperature parameter, and i, j ($1 \leq i, j \leq B$) denote sample indices within the minibatch. The similarity function $\text{sim}(i,j)$ measures the similarity between the i-th $\mathbf{P}^{\text{TS}}$ and j-th $\mathbf{P}^{\text{Text}}$, and is defined as (for convenience, we omit the indices i and j in the equation below):

$$\text{sim}(\cdot) = \sum_{k=1}^{K} (\beta_k < \mathbf{P}^{\text{TS}}(k), \mathbf{P}^{\text{Text}}(k) >) \tag{8}$$

where $\langle \cdot \rangle$ denotes cosine similarity, and k is the prototype index. The bidirectional TP-NCE loss is then given by: $\mathcal{L}_{\text{TPNCE}} = \frac{1}{2}(\mathcal{L}_{\text{TPNCE}}^{\text{TS}\to\text{Text}} + \mathcal{L}_{\text{TPNCE}}^{\text{Text}\to\text{TS}})$. To account for varying prototype importance, an attention mechanism is used to generate weights β for the slots, based on global MITS and text embeddings: $\beta = \text{MLP}(\text{concat}[\mathbf{g}_{\text{MS}}^{\text{TS}}, \mathbf{g}^{\text{Text}}])$, where $\mathbf{g}_{\text{MS}}^{\text{TS}}$ and $\mathbf{g}^{\text{Text}} \in \mathbb{R}^D$ are global embeddings obtained by averaging $\hat{\mathbf{z}}_{\text{MS}}^{\text{TS}}$ and $\mathbf{z}^{\text{Text}}$ along the time dimension.

Auxiliary Reconstruction. To ensure that the learned prototype representations capture core information from multimodal EHR data, we introduce two reconstruction objectives aimed at reconstructing imputed regular time series and text embeddings from the learned prototypes. Specifically, we implement a time series decoder to reconstruct the imputed regular time series from $\mathbf{P}^{\text{TS}}$, and a text embedding decoder to reconstruct text embeddings from $\mathbf{P}^{\text{Text}}$. Both decoders are based on a transformer decoder architecture (Vaswani et al., 2017), and two mean squared error (MSE) losses denoted by $\mathcal{L}_{\text{TS}-\text{Recon}}$ and $\mathcal{L}_{\text{Text}-\text{Recon}}$ are used

as the objective function. Here, we define the overall reconstruction loss $\mathcal{L}_{\text{Recon}} = \frac{1}{2}(\mathcal{L}_{\text{TS-Recon}} + \mathcal{L}_{\text{Text-Recon}})$.

3.4 Multimodal Fusion

Since information from both modalities is crucial for predicting medical conditions, we propose a multimodal fusion mechanism to integrate these inputs. First, we apply a 2-layer transformer encoder (Vaswani et al., 2017) to capture interactions between timestamp-level and prototype embeddings across both modalities for each sample. We continue to use $\mathbf{P}^{\text{TS}}$ and $\mathbf{P}^{\text{Text}}$ to represent the resulted prototype embeddings for time-series data and clinical notes, respectively. Similarly, $\hat{\mathbf{z}}^{\text{TS}}_{\text{MS}}$ and $\mathbf{z}^{\text{Text}}$ denote the corresponding timestamp-level embeddings. Then, we aggregate K prototype embeddings and T timestamp-level embeddings of each modality using an attention-based pooling mechanism:

$$\mathbf{F}^{\text{TS}} = \sum_{k=1}^{K} \gamma_k^{\text{TS}} \mathbf{P}^{\text{TS}}(k) + \sum_{t=1}^{T} \phi_t^{\text{TS}} \hat{\mathbf{z}}^{\text{TS}}_{\text{MS}}(t)$$
$$\mathbf{F}^{\text{Text}} = \sum_{k=1}^{K} \gamma_k^{\text{Text}} \mathbf{P}^{\text{Text}}(k) + \sum_{t=1}^{T} \phi_t^{\text{Text}} \mathbf{z}^{\text{Text}}(t) \tag{9}$$

Here k and t refer to the indices of prototype embeddings and timestamp-level embeddings, respectively. Here γ^{TS}, ϕ^{TS}, γ^{Text}, ϕ^{Text} are learned attention weights by passing the corresponding embeddings through a shared MLP. The resulting global embeddings from both modalities are concatenated along the feature dimension to form the final global representation.

3.5 Overall Learning Objectives

To optimize our framework, we employ four loss functions jointly. The overall objective is a weighted sum of these loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_1 * \mathcal{L}_{\text{TPNCE}} + \lambda_2 * \mathcal{L}_{\text{Recon}} \tag{10}$$

where λ_1, λ_2 are hyperparameters that control the weights of respective losses. Here $\mathcal{L}_{\text{pred}}$, is a cross-entropy loss used for classification.

4 Experiment

4.1 Experimental Setup

Dataset. We assess our model's efficacy using MIMIC-III v1.4[1], a comprehensive open-source multimodal clinical database (Johnson et al., 2016). We focus our evaluation on two critical tasks, *48-hour in-hospital mortality prediction* (48-IHM) and *24-hour phenotype classification* (24-PHE), as established in prior research (Zhang et al., 2023b; Hayat et al., 2022). We extracted raw, irregular time-series data (containing 17 clinical variables) from the MIMIC-III database and selected time series within the first 48-hour and 24-hour windows within each ICU stay for each respective task, as done in (Zhang et al., 2023b; Hayat et al., 2022; Harutyunyan et al., 2019). ICU stays shorter than 48 or 24 hours were excluded from our dataset, which explains the different sample size between these two tasks. Unlike (Harutyunyan et al., 2019), we did not apply imputation during preprocessing but instead retained the original irregular time-series structure, following (Zhang et al., 2023b). Following the dataset splitting by (Harutyunyan et al., 2019), we ensure that the model evaluation is robust by partitioning the data into 70% training, 10% validation, and 20% testing sets, based on unique subject IDs to prevent information leakage. For multimodal analysis, we paired numerical time-series data with corresponding clinical notes from each patient's ICU stay, consistent with (Zhang et al., 2023b). Note that the pipeline of processing clinical notes we used follows the practice in (Khadanga et al., 2019). The dataset statistics, including sample counts before and after multimodal paring, are presented in Table 1. The additional details of experimental setup can be found in Appendix A.

Table 1: Number of samples for EHR and fully paired EHR-clinical notes across the training, validation, and test sets.

	Training	Validation	Test
Number of EHR.			
48-IHM	16,093	1,810	3,236
24-PHE	26,891	2,955	5,282
Number of Paired EHR and Clinical Notes.			
48-IHM	15,425	1,727	3,107
24-PHE	25,435	2,807	5,013

Evaluation Metrics. The 48-hour In-Hospital Mortality (48-IHM) prediction is a binary classification with a marked label imbalance, indicated by a death-to-discharge ratio of approximately 1:6. Following previous work (Harutyunyan et al., 2019; Zhang et al., 2023b), we use AUROC, AUPR, and F1 score, for a comprehensive evaluation. The 24-

[1] https://physionet.org/content/mimiciii/1.4/

Table 2: Comparison of our method with baselines on 48-IHM and 24-PHE tasks using the MIMIC-III dataset. We report average performance on three random seeds, with standard deviation as the subscript. The **Best** and 2nd best methods under each setup are bold and underlined.

Model	48-IHM			24-PHE		
	AUROC (↑)	AUPR (↑)	F1 (↑)	AUROC (↑)	AUPR (↑)	F1 (↑)
Methods on MITS.						
CNN (LeCun et al., 1998)	$85.80_{0.32}$	$49.73_{0.65}$	$46.37_{3.01}$	$75.36_{0.18}$	$38.10_{0.26}$	$40.80_{0.37}$
RNN (Elman, 1990)	$84.75_{0.74}$	$46.57_{1.18}$	$45.60_{2.06}$	$73.78_{0.10}$	$36.76_{0.27}$	$33.99_{0.48}$
LSTM (Hochreiter, 1997)	$85.22_{0.67}$	$46.93_{1.28}$	$45.72_{1.41}$	$74.46_{0.23}$	$36.80_{0.28}$	$39.45_{0.49}$
Transformer (Vaswani et al., 2017)	$83.45_{0.97}$	$43.03_{1.65}$	$39.31_{3.83}$	$74.98_{0.14}$	$39.37_{0.26}$	$36.13_{1.42}$
IP-Net (Shukla and Marlin, 2019)	$81.76_{0.38}$	$39.50_{0.83}$	$43.89_{1.07}$	$73.98_{0.13}$	$35.31_{0.29}$	$39.38_{0.15}$
GRU-D (Che et al., 2018)	$49.21_{5.26}$	$12.85_{2.24}$	$19.63_{0.01}$	$52.11_{0.42}$	$17.99_{0.55}$	$26.17_{0.23}$
DGM-O (Wu et al., 2021)	$71.99_{7.30}$	$28.67_{7.34}$	$31.72_{11.02}$	$60.70_{0.82}$	$22.56_{0.77}$	$28.87_{0.64}$
mTAND (Shukla and Marlin, 2021)	$85.27_{0.20}$	$49.82_{0.97}$	$48.02_{1.93}$	$72.79_{0.09}$	$35.95_{0.14}$	$32.42_{0.72}$
SeFT (Horn et al., 2020)	$65.00_{0.84}$	$22.93_{1.27}$	$19.70_{15.59}$	$60.50_{0.09}$	$23.57_{0.07}$	$21.26_{0.21}$
UTDE (Zhang et al., 2023b)	$86.14_{0.45}$	$50.60_{0.28}$	$49.29_{0.62}$	$73.62_{0.57}$	$36.80_{0.87}$	$40.58_{0.64}$
Methods on Clinical Notes.						
Flat (Deznabi et al., 2021)	$85.71_{0.49}$	$50.96_{0.19}$	$45.80_{6.61}$	$81.77_{0.09}$	$53.79_{0.21}$	$\underline{52.13}_{0.79}$
HierTrans (Pappagari et al., 2019)	$84.32_{0.34}$	$47.92_{0.57}$	$42.01_{0.58}$	$80.79_{0.06}$	$51.97_{0.08}$	$50.85_{0.71}$
T-LSTM (Baytas et al., 2017)	$85.70_{0.21}$	$45.29_{0.86}$	$42.84_{5.53}$	$81.15_{0.04}$	$50.23_{0.07}$	$49.77_{0.24}$
FT-LSTM (Zhang et al., 2020a)	$84.28_{0.95}$	$43.93_{1.03}$	$36.87_{6.54}$	$81.66_{0.12}$	$51.71_{0.24}$	$50.21_{0.85}$
GRU-D (Che et al., 2018)	$72.58_{0.42}$	$26.38_{1.06}$	$31.64_{0.79}$	$49.52_{0.65}$	$16.90_{0.12}$	$27.80_{0.08}$
mTAND (Shukla and Marlin, 2021)	$85.40_{0.60}$	$49.68_{1.26}$	$35.76_{11.57}$	$\underline{82.14}_{0.07}$	$\underline{54.57}_{0.15}$	$52.01_{0.93}$
Methods on Multimodal EHR.						
MMTM (Joze et al., 2020)	$\underline{87.88}_{0.07}$	$\underline{53.58}_{0.24}$	$51.54_{1.58}$	$81.46_{0.25}$	$51.88_{0.12}$	$51.59_{0.19}$
DAFT (Pölsterl et al., 2021)	$87.53_{0.22}$	$52.40_{0.21}$	$\underline{51.95}_{0.64}$	$81.18_{0.08}$	$50.91_{0.31}$	$50.72_{0.39}$
MedFuse (Hayat et al., 2022)	$86.02_{0.29}$	$51.00_{0.22}$	$49.29_{0.75}$	$78.88_{0.14}$	$45.99_{0.21}$	$47.47_{0.23}$
DrFuse (Yao et al., 2024)	$85.97_{1.02}$	$49.94_{1.91}$	$49.75_{1.52}$	$80.88_{0.18}$	$49.62_{0.40}$	$50.18_{0.31}$
CTPD (Ours)	$\mathbf{88.15}_{0.28}$	$\mathbf{53.86}_{0.65}$	$\mathbf{53.85}_{0.16}$	$\mathbf{83.34}_{0.05}$	$\mathbf{56.39}_{0.17}$	$\mathbf{53.83}_{0.43}$

Table 3: Statistics analysis of CTPD on MIMIC-III dataset. p-values are computed from paired t-tests.

Model	48-IHM			24-PHE		
	AUROC (↑)	AUPR (↑)	F1 (↑)	AUROC (↑)	AUPR (↑)	F1 (↑)
SOTA	$87.88_{0.07}$	$53.58_{0.24}$	$51.95_{0.64}$	$82.14_{0.07}$	$54.57_{0.15}$	$52.13_{0.79}$
CTPD (Ours)	$88.15_{0.28}$	$53.86_{0.65}$	$53.85_{0.16}$	$83.34_{0.05}$	$56.39_{0.17}$	$53.83_{0.43}$
Gains	**+0.27**	**+0.28**	**+1.9**	**+1.2**	**+1.82**	**+1.7**
p values	0.016	0.233	7.74e−6	7.89e−12	1.10e−10	2.08e−4

hour Phenotype Classification (24-PHE) involves predicting the presence of 25 different medical conditions during an ICU stay, making it a multi-label classification task. For this task, we employ the AUROC, AUPR, and F1 score (Macro) for a thorough assessment of model efficacy. The F1 score threshold is determined by selecting the value that maximizes the F1 score on the validation set.

Implementation Details. We train the model with batch size of 128, learning rate of 4e-5, and Adam (Kingma and Ba, 2014) optimizer. We use a cosine annealing learning rate scheduler with a 0.2 warm-up proportion. To prevent overfitting, we implement early stopping when there is no increase in the AUROC on the validation set for 48-IHM or 24-PHENO over 5 consecutive epochs. All experiments are conducted on 1 RTX-3090 GPU card using about 1 hour per run. We clip the norm of gradient values with 0.5 for stable training. By default, we use Bert-tiny (Turc et al., 2019; Bhargava et al., 2021) as our text encoder.

Compared Methods. To ensure a comprehensive comparison, we compare our CTPD with three types baselines: MITS-only approaches, note-only approaches and multimodal approaches. For MITS-only setting, we compare CTPD with 4 baselines for imputed regular time series: RNN (Elman, 1990), LSTM (Hochreiter, 1997), CNN (LeCun et al., 1998) and Transformer (Vaswani et al., 2017), and 5 baselines for irregular time series, including IP-Net (Shukla and Marlin, 2019), GRU-D (Che et al., 2018), DGM-O (Wu et al., 2021), mTAND (Shukla and Marlin, 2021), SeFT (Horn et al., 2020), and UTDE (Zhang et al., 2023b). The imputation ap-

proach follows the MIMIC-III benchmark (Harutyunyan et al., 2019). For the note-only setting, we compare our model with 6 baselines: Flat (Deznabi et al., 2021), HierTrans (Pappagari et al., 2019), T-LSTM (Baytas et al., 2017), FT-LSTM (Zhang et al., 2020a), GRU-D (Che et al., 2018), and mTAND (Shukla and Marlin, 2021). In the multimodal setting, we compare our model with 4 baselines: MMTM (Joze et al., 2020), DAFT (Pölsterl et al., 2021), MedFuse (Hayat et al., 2022), and DrFuse (Yao et al., 2024). To ensure a fair comparison, we implement Bert-tiny (Bhargava et al., 2021; Turc et al., 2019) as the text encoder across all baselines. Details of baselines can be found in the Appendix B.

4.2 Comparison with SOTA Baselines

Results on MIMIC-III. Table 2 presents a comparison of our proposed CTPD against 3 types of baselines: MITS-based methods, clinical notes-based methods, and multimodal EHR-based methods. Our CTPD, which incorporates cross-modal temporal pattern embeddings, consistently achieves the best performance across all 6 metrics. Specifically, CTPD shows a 1.89% improvement in F1 score on the 48-IHM task, and a 1.2% improvement in AUROC and 1.92% in AUPR on the more challenging 24-PHE task, compared to the second-best results. Additionally, we have conducted statistical analysis in Table 3 to assess statistical significance. CTPD demonstarted significant gains over SOTAs (p value < 0.05) in 5 out of 6 settings, indicating its effectiveness in analyzing multimodal EHR.

Evaluation on More Tasks and Datasets. To further evaluate the generalizability of our CTPD framework, we have extended it to another important admission-level task: 30-day readmission prediction on MIMIC-III. Additionally, we also conducted experiments on the additional MIMIC-IV dataset to assess our framework's adaptability to different data sources. These results can be found in Appendix C.1 and Appendix C.2.

Discussion on Missing Modalities and Noisy Data Scenarios. Currently, CTPD framework is built on paired time-series and clinical notes. In practice, the dataset might have missing modalities, such as partial clinical notes or time-series data are missed. We acknowledge this is an interesting research direction. However, the study of missing modalities falls outside the scope of our work and will be explored in future research. Additionally, our evaluation is conducted on MIMIC-III (Johnson et al., 2016), a large-scale, de-identified real-world dataset containing patient records from critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012. Our approach is designed to be adaptable and can be seamlessly applied to other real-world databases.

Table 4: Ablation results show the impact of removing different types of input embeddings.

	48-IHM			24-PHE		
	AUROC	AUPR	F1	AUROC	AUPR	F1
Ours	**88.15**$_{0.28}$	**53.86**$_{0.65}$	**53.85**$_{0.16}$	**83.34**$_{0.05}$	**56.39**$_{0.17}$	**53.83**$_{0.43}$
:w/o prototype	86.89$_{0.97}$	53.67$_{0.65}$	48.47$_{6.13}$	82.24$_{0.07}$	54.06$_{0.07}$	52.88$_{0.11}$
:w/o timestamp embeddings	87.18$_{0.94}$	54.32$_{1.66}$	45.85$_{4.91}$	82.41$_{0.15}$	54.30$_{0.21}$	53.34$_{0.41}$
:w/o multi-scale embedding	87.59$_{0.49}$	53.38$_{1.79}$	49.74$_{3.50}$	83.11$_{0.09}$	55.95$_{0.05}$	53.81$_{0.46}$

Table 5: Ablation study of loss functions $\mathcal{L}_{\text{TPNCE}}$ and $\mathcal{L}_{\text{Recon}}$. The **Best** results are highlighted in bold.

		48-IHM			24-PHE		
Cont	Recon	AUROC	AUPR	F1	AUROC	AUPR	F1
		87.49$_{0.47}$	53.31$_{1.46}$	43.59$_{4.63}$	82.49$_{0.10}$	55.25$_{0.30}$	53.71$_{0.43}$
✓		87.15$_{0.38}$	53.45$_{1.09}$	44.76$_{4.48}$	82.94$_{0.05}$	55.62$_{0.22}$	53.99$_{0.19}$
	✓	86.93$_{0.69}$	52.70$_{1.97}$	41.52$_{0.90}$	82.86$_{0.03}$	55.43$_{0.05}$	**54.08**$_{0.29}$
✓	✓	**88.15**$_{0.28}$	**53.86**$_{0.65}$	**53.85**$_{0.16}$	**83.34**$_{0.05}$	**56.39**$_{0.17}$	53.83$_{0.43}$

4.3 Model Analysis

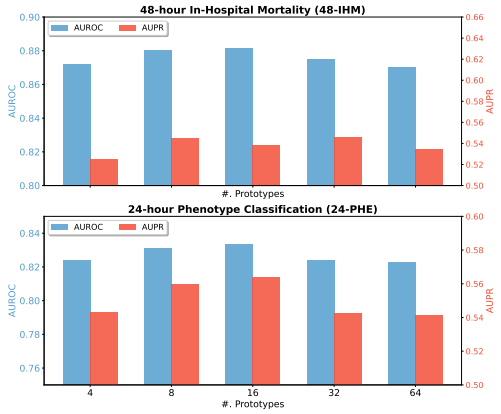
Ablation Results on Different Components. We conduct ablation studies by removing the prototype embeddings, timestamp-level embeddings, and multi-scale feature extractor respectively, and analyze their impacts on two clinical prediction tasks, as shown in Table 4. Notably, prototype embeddings play the most significant role among the three components, with their removal resulting in a 1.96% AUROC decrease in 48-IHM and a 1.1% decrease in 24-PHE. The results also show that all three embeddings are important for capturing effective information for prediction.

Ablation Results on Learning Objectives. Table 5 presents the ablation results of the learning objectives. Combining both $\mathcal{L}_{\text{TPNCE}}$ and $\mathcal{L}_{\text{Recon}}$ leads to the best performance across 5 out of 6 settings. Our model remains robust under different loss function configurations, with only a 0.66% AUROC drop in the model's performance in 48-IHM and a 0.85% drop in 24-PHE. Note that the first row of Table 5, i.e., CTPD without auxiliaries losses performs slightly worse than several SOTAs (Table 2) in 48-IHM task, while significantly outperforms compared approaches in 24-PHE task. We think The performance gap is mainly due to task complexity. The 48-IHM task is a simple binary classification problem with limited need for cross-modal reasoning, making our model more prone to

Table 6: Number of parameters of different modules in CTPD.

Module	Multi-scale Convolution	Shared Prototype	Dual Slot Attention	MITS Decoder	Text Decoder	Multimodal Fusion	Text Encoder	Others	**Total**
#. Trainable Parameters	312.2K	256	364K	1.3M	1.3M	1.23M	4.4M	0.294M	**9.4M**

Figure 3: Ablation study on the number of prototypes.



overfitting due to its added components. In contrast, the 24-PHE task is a more complex multi-label classification problem that benefits from stronger temporal and cross-modal understanding, leading to larger performance gains over SOTAs.

Complexity of the Proposed Method To provide a more detailed analysis of the computational complexity of our proposed framework, we report the number of trainable parameters for each module in the Table 6. Our overall model only has 9.2 M parametes, which is lightweight for multimodal EHR analysis. In addition, we evaluated the inference speed of our CTPD and found that it requires only **5.42 ms** per time series-text pair on average. This indicates the framework's potential for real-time deployment in clinical settings.

Ablation Results on Hyperparameters. We analyze the effects of the number of prototypes in Fig. 3. According to the results, we find that using 16 prototypes achieves the best results, though our model remains robust, with 8 prototypes yielding similar outcomes. Additional experimental results of hyperparameters and visualization are in Appendix C.3 and Appendix C.4.

5 Conclusion

In this paper, we present the Cross-Modal Temporal Pattern Discovery (CTPD) framework, which captures cross-modal temporal patterns and incorporates them with timestamp-level embeddings for more accurate clinical outcome predictions based on multimodal EHR data. To efficiently optimize the framework, we introduce a Temporal Pattern Noise Contrastive Estimation (TP-NCE) loss to enhance cross-modal alignment, along with two reconstruction objectives to retain core information from each modality. Our experiments on two clinical prediction tasks using the MIMIC-III dataset demonstrate the effectiveness of CTPD in multimodal EHR analysis.

6 Limitations

A limitation of our approach is its primary focus on extracting temporal semantics in the embedding space, which affects model interpretability. As a potential solution, we plan to explore retrieval-based methods (Patel et al., 2024) or pretrained generative models (Zhao et al., 2023) to enhance the interpretability of learned temporal pattern embeddings in future work. Besides, our work primarily focuses on exploring cross-modal temporal patterns to enhance representation learning, the scenario of missing modalities falls outside the current scope—since shared patterns across modalities cannot be fully captured in their absence. We plan to thoroughly investigate this direction in future work. Additionally, our framework is currently designed for specific clinical prediction tasks. In practice, there are various prediction tasks related to multimodal EHR analysis. Extending the proposed method into a more generalized or foundational model capable of handling multiple downstream tasks with minimal training annotations could be more practical and effective. Our future work will focus on resolving those aspects.

Potential Risks. The medical dataset used in our framework must be carefully reviewed to mitigate any potential identification risk. Additionally, our framework is developed solely for research purposes and is not intended for commercial use. Note that AI assistant was used only for polishing the writing of this paper.

7 Acknowledgement

This work was supported in part by the Research Grants Council of Hong Kong (27206123, C5055-24G, and T45-401/22-N), the Hong Kong Innovation and Technology Fund (ITS/273/22, ITS/274/22, and GHP/318/22GD), the National Natural Science Foundation of China (No. 62201483), and Guangdong Natural Science Fund (No. 2024A1515011875).

References

Julián N. Acosta, Guido J. Falcone1, Pranav Rajpurka, and Eric J. Topol. 2022. Multimodal biomedical AI. *Nature Medicine*, 28:1773–1784.

Rasha Assaf and Rashid Jayousi. 2020. 30-day hospital readmission prediction using mimic data. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6. IEEE.

Mohammad Taha Bahadori and Zachary Chase Lipton. 2019. Temporal-clustering invariance in irregular healthcare time series. *arXiv preprint arXiv:1904.12206*.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics. *Preprint*, arXiv:2110.01518.

Wanlin Cai, Yuxuan Liang, Xianggen Liu, Jianshuai Feng, and Yuankai Wu. 2024. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11141–11149.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Nathalie Conrad, Andrew Judge, Jenny Tran, Hamid Mohseni, Deborah Hedgecott, Abel Perez Crespillo, Moira Allison, Harry Hemingway, John G Cleland, John JV McMurray, et al. 2018. Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *The Lancet*, 391(10120):572–580.

Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora Salim, and Akhil Mathur. 2023. Latent masking for multimodal self-supervised learning in health timeseries. *arXiv preprint arXiv:2307.16847*.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 4026–4031.

Ning Ding, Sheng-wei Tian, and Long Yu. 2022. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Gaurav R Ghosal and Reza Abbasi-Asl. 2021. Multimodal prototype learning for interpretable multivariable time series classification. *arXiv preprint arXiv:2106.09636*.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. 2022. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. 2020. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR.

Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P Lungren. 2020. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):22147.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. 2020. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. 2019. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*.

Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. 2023. Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5141–5150.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. 2022. Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine*, 5(1):171.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Chonho Lee, Zhaojing Luo, Kee Yuan Ngiam, Meihui Zhang, Kaiping Zheng, Gang Chen, Beng Chin Ooi, and Wei Luen James Yip. 2017. Big healthcare data analytics: Challenges and applications. *Handbook of large-scale distributed computing in smart healthcare*, pages 11–41.

Bin Li, Carsten Jentsch, and Emmanuel Müller. 2023a. Prototypes as explanation for time series anomaly detection. *arXiv preprint arXiv:2307.01601*.

Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. 2021. Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1046–1055.

Yuxin Li, Wenchao Chen, Bo Chen, Dongsheng Wang, Long Tian, and Mingyuan Zhou. 2023b. Prototype-oriented unsupervised anomaly detection for multivariate time series.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538.

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

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 825–832.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 838–844. IEEE.

Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, and Edward Choi. 2022. Graph-text multi-modal pretraining for medical representation learning. In *Conference on Health, Inference, and Learning*, pages 261–281. PMLR.

Ravi Patel, Angus Brayne, Rogier Hintzen, Daniel Jaroslawicz, Georgiana Neculae, and Dane Corneil. 2024. Retrieve to explain: Evidence-driven predictions with language models. *arXiv preprint arXiv:2402.04068*.

Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. 2021. Combining 3d image and tabular data via the dynamic affine feature map transform. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 688–698. Springer.

Linglong Qian, Zina Ibrahim, Hugh Logan Ellis, Ao Zhang, Yuezhou Zhang, Tao Wang, and Richard Dobson. 2023. Knowledge enhanced conditional imputation for healthcare time-series. *arXiv preprint arXiv:2312.16713*.

Sangeeta Rani and Geeta Sikka. 2012. Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications*, 52(15).

Satya Narayan Shukla and Benjamin M Marlin. 2019. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*.

Satya Narayan Shukla and Benjamin M Marlin. 2021. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*.

Chenxi Sun, Hongyan Li, Moxian Song, Derun Cai, Baofeng Zhang, and Shenda Hong. 2024. Time pattern reconstruction for classification of irregularly sampled time series. *Pattern Recognition*, 147:110075.

Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. 2021. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:17194–17208.

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

Vo Hoang Trong, Yu Gwang-hyun, Dang Thanh Vu, and Kim Jin-young. 2020. Late fusion of multimodal deep neural networks for weeds classification. *Computers and Electronics in Agriculture*, 175:105506.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

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

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 2839. NIH Public Access.

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

Yinjun Wu, Jingchao Ni, Wei Cheng, Bo Zong, Dongjin Song, Zhengzhang Chen, Yanchi Liu, Xuchao Zhang, Haifeng Chen, and Susan B Davidson. 2021. Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 651–659.

Jingyun Xiao, Ran Liu, and Eva L Dyer. 2023. Gaformer: Enhancing timeseries transformers through group-aware embeddings. In *The Twelfth International Conference on Learning Representations*.

Feng Xie, Yuan Han, Ning Yilin, Marcus E. H. Ong, Feng Mengling, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. 2021. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *J Biomed Inform.*, 126:103980.

Zhen Xu, David R So, and Andrew M Dai. 2021. Mufasa: Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10532–10540.

Haiyang Yang, Li Kuang, and FengQiang Xia. 2021. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12:1–14.

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*, volume 38, pages 16416–16424.

Hangting Ye, Wei Fan, Xiaozhuang Song, Shun Zheng, He Zhao, Dandan Guo, and Yi Chang. 2024. Ptarl: Prototype-based tabular representation learning via space calibration. *arXiv preprint arXiv:2407.05364*.

Muchao Ye, Junyu Luo, Cao Xiao, and Fenglong Ma. 2020. Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1753–1762.

Zhihao Yu, Xu Chu, Liantao Ma, Yasha Wang, and Wenwu Zhu. 2024. Imputation with inter-series information from prototypes for irregular sampled time series. *arXiv preprint arXiv:2401.07249*.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8980–8987.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. 2020a. Time-aware transformer-based network for clinical notes series prediction. In *Machine learning for healthcare conference*, pages 566–588. PMLR.

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

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

Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. 2023b. Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pages 41300–41313. PMLR.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. 2024. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646*.

Yu-Dong Zhang, Zhengchao Dong, Shui-Hua Wang, Xiang Yu, Xujing Yao, Qinghua Zhou, Hua Hu, Min Li, Carmen Jiménez-Mesa, Javier Ramirez, et al. 2020b. Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation. *Information Fusion*, 64:149–187.

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

Appendix

Table 7: Dataset description of 48-IHM and 24-PHE.

	Training	Validation	Test
24-PHE	5046	573	1369
48-IHM	4301	466	1183
Positive	644	73	184
Negative	3657	393	999

A Experimental Setup Details

A.1 More details of Tasks.

We presented the class distributions for 48-IHM and 24-PHE in Table 7 and Table 8, respectively. 48-IHM is a binary classification task, whereas 24-PHE is a multi-label classification problem with 25 labels. Notably, our 24-PHE task differs from the phenotype classification problem in the MIMIC-III benchmark but follows the setup in (Zhang et al., 2023b). This setup focuses on acute care conditions that arise during ICU stays, where early prediction is critical for timely intervention. To enhance clinical relevance, we use only the first 24 hours of data for phenotype classification rather than the entire admission record. As highlighted in (Yang et al., 2021), early-stage diagnosis holds greater clinical significance.

The selection of 25 phenotype labels for the 24-PHE task follows established practices in the MIMIC-III benchmark (Harutyunyan et al., 2019), as also utilized in prior studies like UTDE (Zhang et al., 2023b) and MedFuse (Hayat et al., 2022). These labels cover conditions commonly observed in adult ICUs, including 12 critical and life-threatening conditions (e.g., respiratory failure, sepsis), 8 chronic conditions that are often considered comorbidities or risk factors (e.g., diabetes, metabolic disorders), and 5 'mixed' conditions that exhibit characteristics of both chronic and acute conditions. Phenotype labels were determined using the MIMIC-III ICD-9 diagnosis table. To facilitate the translation and conversion of the above-mentioned conditions, we use the Health Cost and Utilization (HCUP) Clinical Classification Software (CCS)[2]. We first mapped each ICD-9 code to its corresponding HCUP CCS category, retaining only 25 categories. Diagnoses were then linked to ICU stays using the hospital admission identifier, as ICD-9 codes in MIMIC-III are associated with hospital visits rather than specific ICU stays. To reduce label ambiguity, we excluded hospital admissions involving multiple ICU stays, ensuring each diagnosis could be associated with a single ICU stay. Please find the class distribution of 25 phenotypes in Table 8.

[2]https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

A.2 Additional Information on Datasets

The 17 variables from the MIMIC-III dataset that we use include 5 categorical variables (capillary refill rate, Glasgow coma scale eye opening, Glasgow coma scale motor response, Glasgow coma scale total, and Glasgow coma scale verbal response) and 12 continuous measures (diastolic blood pressure, fraction of inspired oxygen, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH).

B More Details on Baselines

B.1 Baselines only using MITS

- CNN (LeCun et al., 1998): CNN (Convolutional Neural Network) uses backpropagation to synthesize a complex decision surface that facilitates learning.
- RNN (Elman, 1990): RNN (Residual Neural Network) is trained to process data sequentially so as to model the time dimension of data.
- LSTM (Hochreiter, 1997): LSTM (Long short-term memory) is a variant of recurrent neural network. It excels at dealing with the vanishing gradient problem and is relatively insensitive to time gap length.
- Transformer (Vaswani et al., 2017): Transformer is a powerful deep learning architecture based on attention mechanism. It has great generalisability and has been adopted as foundation model in multiple research domains.
- IP-Net (Shukla and Marlin, 2019): IP-Net (Interpolation-Prediction Network) is a deep learning architecture for supervised learning focusing on processing sparse multivariate time series data that are sampled irregularly.
- GRU-D (Che et al., 2018): GRU-D is based on GRU (Gated Recurrent Unit). It incorporates features of missing data in EHR into

Table 8: Distribution of 25 phenotypes in 24-PHE task.

Phenotype	Type	Training		Validation		Test	
		Positive	Negative	Positive	Negative	Positive	Negative
Acute and unspecified renal failure	acute	6066	20825	2284	2284	1176	4106
Acute cerebrovascular disease	acute	2069	24822	2736	2736	363	4919
Acute myocardial infarction	acute	2870	24021	2632	2632	588	4694
Cardiac dysrhythmias	mixed	9011	17880	1975	1975	1785	3497
Chronic kidney disease	chronic	3663	23228	2564	2564	711	4571
Chronic obstructive pulmonary disease and bronchiectasis	chronic	3616	23275	2542	2542	685	4597
Complications of surgical procedures or medical care	acute	5880	21011	2293	2293	1202	4080
Conduction disorders	mixed	1971	24920	2744	2744	382	4900
Congestive heart failure; nonhypertensive	mixed	7623	19268	2150	2150	1486	3796
Coronary atherosclerosis and other heart disease	chronic	8840	18051	1964	1964	1787	3495
Diabetes mellitus with complications	mixed	2612	24279	2675	2675	503	4779
Diabetes mellitus without complication	chronic	5305	21586	2401	2401	1032	4250
Disorders of lipid metabolism	chronic	7841	19050	2081	2081	1541	3741
Essential hypertension	chronic	11340	15551	1689	1689	2238	3044
Fluid and electrolyte disorders	acute	7480	19411	2107	2107	1429	3853
Gastrointestinal hemorrhage	acute	2001	24890	2746	2746	427	4855
Hypertension with complications and secondary hypertension	chronic	3646	23245	2583	2583	706	4576
Other liver diseases	mixed	2485	24406	2688	2688	492	4790
Other lower respiratory disease	acute	1414	25477	2810	2810	305	4977
Other upper respiratory disease	acute	1130	25761	2814	2814	238	5044
Pleurisy; pneumothorax; pulmonary collapse	acute	2516	24375	2671	2671	518	4764
Pneumonia (except that caused by tuberculosis or sexually transmitted disease)	acute	4121	22770	2502	2502	769	4513
Respiratory failure; insufficiency; arrest (adult)	acute	5382	21509	2365	2365	1025	4257
Septicemia (except in labor)	acute	4136	22755	2494	2494	793	4489
Shock	acute	2263	24628	2700	2700	456	4826

the model architecture to improve prediction results.

- DGM-O (Wu et al., 2021): DGM-O (Dynamic Gaussian Mixture based Deep Generative Model) is a generative model derived from a dynamic Gaussian mixture distribution. It makes predictions based on incomplete inputs. DGM-O is instantiated with multilayer perceptron (MLP).

- mTAND (Shukla and Marlin, 2021): mTAND (Multi-Time Attention network) is a deep learning model that learns representations of continuous time values and uses an attention mechanism to generate a consistent representation of a time series based on a varying amount of observations.

- SeFT (Horn et al., 2020): SeFT (Set Functions for Time Series) addresses irregularly-sampled time series. It is based on differentiable set function learning.

- UTDE (Zhang et al., 2023b): UTDE (Unified TDE module) is built upon TDE (Temporal discretization-based embedding). It models asynchronous time series data by combining imputation embeddings and learned interpolation embeddings through a gating mechanism. It also uses a time attention mechanism.

B.2 Baselines only using clinical notes

- Flat (Deznabi et al., 2021): Flat encodes clinical notes using a fine-tuned BERT model. It also utilizes an LSTM model that takes in patients' vital signals so as to jointly model the two modalities. Furthermore, it addresses the temporal irregularity issue of modeling patients' vital signals.

- HierTrans (Pappagari et al., 2019): HierTrans (Hierarchical Transformers) is built upon BERT model. It achieves an enhanced ability to take in long inputs by first partitioning the inputs into shorter sequences and processing them separately. Then, it propagates each output via a recurrent layer.

- T-LSTM (Baytas et al., 2017): T-LSTM (Time-Aware LSTM) deals with irregular time intervals in EHRs by learning decomposed cell memory which models elapsed time. The final patient subtyping model uses T-LSTM in an auto-encoder module before doing patient subtyping.

- FT-LSTM (Zhang et al., 2020a): FT-LSTM (Flexible Time-aware LSTM Transformer) models the multi-level structure in clinical notes. At the base level, it uses a pre-trained ClinicalBERT model. Then, it merges sequential information and content embedding into a new position-enhanced representation. Then,

Table 9: Comparison between our method with other baselines on 30-day Readmission task on MIMIC-III. We report average performance on three random seeds, with standard deviation as the subscript.

Model	AUROC (↑)	AUPR (↑)	F1 (↑)
Methods on EHR.			
CNN (LeCun et al., 1998)	$0.757_{0.003}$	$0.447_{0.006}$	$0.449_{0.006}$
RNN (Elman, 1990)	$0.750_{0.001}$	$0.449_{0.008}$	$0.433_{0.003}$
LSTM (Hochreiter, 1997)	$0.757_{0.002}$	$0.443_{0.003}$	$0.430_{0.015}$
Transformer (Vaswani et al., 2017)	$0.749_{0.006}$	$0.409_{0.012}$	$0.433_{0.016}$
DGM-O (Wu et al., 2021)	$0.671_{0.017}$	$0.325_{0.032}$	$0.382_{0.020}$
mTAND (Shukla and Marlin, 2019)	$0.743_{0.002}$	$0.437_{0.001}$	$0.441_{0.007}$
UTDE (Zhang et al., 2023b)	$0.758_{0.002}$	$0.453_{0.004}$	$0.445_{0.008}$
Methods on Clinical Notes.			
Flat (Deznabi et al., 2021)	$0.755_{0.005}$	$0.447_{0.018}$	$0.437_{0.015}$
HierTrans (Pappagari et al., 2019)	$0.754_{0.001}$	$0.425_{0.005}$	$0.434_{0.005}$
mTAND (Shukla and Marlin, 2019)	$0.757_{0.001}$	$0.435_{0.001}$	$0.440_{0.014}$
Methods on Multiple modalities.			
MMTM (Joze et al., 2020)	$\underline{0.769_{0.005}}$	$\underline{0.469_{0.006}}$	$\underline{0.459_{0.007}}$
DAFT (Pölsterl et al., 2021)	$0.765_{0.006}$	$0.452_{0.018}$	$0.458_{0.005}$
MedFuse (Hayat et al., 2022)	$0.763_{0.009}$	$0.461_{0.012}$	$0.454_{0.013}$
DrFuse (Yao et al., 2024)	$0.748_{0.002}$	$0.443_{0.014}$	$0.442_{0.017}$
CTPD (Ours)	$\mathbf{0.777_{0.006}}$	$\mathbf{0.474_{0.009}}$	$\mathbf{0.461_{0.019}}$

it uses a time-aware layer that considers the irregularity of time intervals.

- GRU-D (Che et al., 2018): Please refer to the above subsection.
- mTAND (Shukla and Marlin, 2021): Please refer to the above subsection.

B.3 Baselines using multimodal EHR

- MMTM (Joze et al., 2020): MMTM (Multimodal Transfer Module) is a neural network module that leverages knowledge from various modalities in CNN. It can recalibrate features in each CNN stream via excitation and squeeze operations.
- DAFT (Pölsterl et al., 2021): DAFT (Dynamic Affine Feature Map Transform) is a general-purpose CNN module that alters the feature maps of a convolutional layer with respect to a patient's clinical data.
- MedFuse (Hayat et al., 2022): MedFuse is an LSTM-based fusion module capable of processing both uni-modal and multi-modal input. It treats multi-modal representations of data as a sequence of uni-modal representations. It handles inputs of various lengths via the recurrent inductive bias of LSTM.
- DrFuse (Yao et al., 2024): DrFuse is a fusion module that addresses the issue of missing modality by separating the unique features within each modality and the common ones across modalities. It also adds a disease-wise attention layer for each modality.

C More on Experimental Results

C.1 Results on 30-day Readmission

We further evaluated CTPD on the 30-day readmission prediction task (Assaf and Jayousi, 2020) using the MIMIC-III dataset. The results are presented in Table 9. This task involves predicting whether a patient will be readmitted based on data from their current ICU stay. Our results show that CTPD consistently outperforms baseline methods across all three evaluation metrics, further demonstrating its effectiveness.

C.2 Results on MIMIC-IV

To assess the adaptability of our framework to different data sources, we also conducted experiments on the MIMIC-IV dataset. Since MIMIC-IV lacks temporal clinical text, we evaluated our approach using tabular time-series data and chest radiographs as the two input modalities. The results are summarized in Table 10. CTPD achieved the best performance across all six evaluation settings. Given the relatively small standard deviations, our approach demonstrates statistically significant improvements over previous methods, further validating its generalizability.

Table 10: Comparison between our method with other baselines on 48-IHM and 24-PHE on MIMIC-IV. We report average performance on three random seeds, with standard deviation as the subscript. The **Best** and 2nd best methods under each setup are bold and underlined. "-" denotes the results are close to 0.

Model	48-IHM			24-PHE		
	AUROC (↑)	AUPR (↑)	F1 (↑)	AUROC (↑)	AUPR (↑)	F1 (↑)
Methods on MITS.						
CNN (LeCun et al., 1998)	$74.16_{0.79}$	$36.32_{0.15}$	$27.34_{8.89}$	$67.50_{0.14}$	$37.00_{0.02}$	$20.90_{0.81}$
RNN (Elman, 1990)	$77.40_{1.10}$	$38.87_{0.80}$	$23.93_{3.92}$	$66.89_{0.49}$	$36.28_{0.70}$	$18.72_{1.97}$
LSTM (Hochreiter, 1997)	$79.40_{0.50}$	$41.95_{1.09}$	$34.17_{1.97}$	$67.59_{0.30}$	$37.21_{0.49}$	$20.87_{1.32}$
Transformer (Vaswani et al., 2017)	$75.73_{2.43}$	$39.76_{2.64}$	$21.45_{10.75}$	$67.44_{0.21}$	$37.33_{0.11}$	$20.44_{0.40}$
IP-Net (Shukla and Marlin, 2019)	$77.58_{0.59}$	$46.02_{1.73}$	$36.8_{0.93}$	$68.20_{0.21}$	$38.4_{0.27}$	$20.62_{2.58}$
GRU-D (Che et al., 2018)	$53.72_{5.31}$	$17.54_{2.22}$	$8.99_{15.56}$	$50.53_{0.60}$	$23.14_{0.36}$	$9.42_{7.76}$
DGM-O (Wu et al., 2021)	$70.50_{10.43}$	$33.79_{7.38}$	$14.14_{12.42}$	$59.74_{2.34}$	$29.95_{2.05}$	$4.33_{4.12}$
mTAND (Shukla and Marlin, 2021)	$80.63_{0.33}$	$45.17_{0.47}$	$33.01_{4.57}$	$66.87_{0.14}$	$36.38_{0.18}$	$19.07_{1.28}$
SeFT (Horn et al., 2020)	$61.97_{0.96}$	$25.13_{0.42}$	-	$57.10_{0.04}$	$26.98_{0.09}$	-
UTDE (Zhang et al., 2023b)	$80.89_{0.10}$	$45.08_{0.30}$	$37.53_{2.03}$	$67.68_{0.05}$	$37.58_{0.19}$	$17.63_{0.44}$
Methods on CXR Image.						
Flat (Deznabi et al., 2021)	$62.37_{2.42}$	$23.97_{2.35}$	$23.38_{4.34}$	$66.10_{0.37}$	$36.04_{0.16}$	$40.40_{0.54}$
HierTrans (Pappagari et al., 2019)	$61.51_{1.43}$	$20.96_{1.62}$	$17.95_{15.54}$	$58.36_{3.28}$	$27.84_{2.16}$	$33.61_{5.74}$
T-LSTM (Baytas et al., 2017)	$53.62_{0.88}$	$16.82_{0.38}$	$21.40_{4.17}$	$58.56_{1.45}$	$28.12_{1.21}$	$33.06_{6.51}$
FT-LSTM (Zhang et al., 2020a)	$48.63_{4.47}$	$15.35_{0.90}$	$11.85_{13.75}$	$54.46_{3.15}$	$25.29_{2.42}$	$29.12_{7.55}$
GRU-D (Che et al., 2018)	$56.02_{1.05}$	$18.63_{0.76}$	$23.16_{6.54}$	$57.94_{0.39}$	$27.85_{0.47}$	$27.13_{3.73}$
mTAND (Shukla and Marlin, 2021)	$62.80_{3.42}$	$24.64_{4.22}$	$24.82_{9.43}$	$68.31_{0.68}$	$38.53_{0.93}$	$40.76_{1.14}$
Methods on Multiple modalities.						
MMTM (Joze et al., 2020)	$80.65_{0.79}$	$\underline{48.96}_{0.50}$	$47.40_{1.41}$	$70.28_{0.44}$	$39.61_{0.76}$	$43.32_{0.29}$
DAFT (Pölsterl et al., 2021)	$\underline{81.87}_{0.12}$	$47.79_{0.88}$	$\underline{48.91}_{1.98}$	$\underline{70.87}_{0.24}$	$40.22_{0.22}$	$\underline{44.10}_{0.26}$
MedFuse (Hayat et al., 2022)	$81.60_{0.28}$	$48.35_{1.35}$	$48.12_{0.75}$	$70.82_{0.37}$	$\underline{40.39}_{0.51}$	$44.03_{0.52}$
DrFuse (Yao et al., 2024)	$80.94_{0.44}$	$45.64_{0.44}$	$48.58_{1.35}$	$70.27_{0.22}$	$39.90_{0.23}$	$43.43_{0.09}$
CTPD (Ours)	$\mathbf{83.53}_{0.44}$	$\mathbf{49.94}_{0.23}$	$\mathbf{49.53}_{2.39}$	$\mathbf{71.96}_{0.40}$	$\mathbf{42.46}_{0.60}$	$\mathbf{44.85}_{0.61}$

Table 11: Ablation results on loss weights. The parameters λ_1 and λ_2 control the strength of the $\mathcal{L}_{\text{TPNCE}}$ and $\mathcal{L}_{\text{Recon}}$ losses, respectively.

λ_1	λ_2	48-IHM			24-PHE		
		AUROC	AUPR	F1	AUROC	AUPR	F1
0.1	0.1	$87.21_{0.36}$	$53.80_{0.28}$	$47.41_{9.17}$	$82.93_{0.05}$	$55.62_{0.22}$	$54.00_{0.20}$
0.1	0.5	$\mathbf{88.15}_{\mathbf{0.28}}$	$\mathbf{53.86}_{\mathbf{0.65}}$	$\mathbf{53.85}_{\mathbf{0.16}}$	$\mathbf{83.34}_{\mathbf{0.05}}$	$\mathbf{56.39}_{\mathbf{0.17}}$	$\mathbf{53.83}_{\mathbf{0.43}}$
0.5	0.5	$87.20_{0.47}$	$53.77_{0.47}$	$42.44_{7.45}$	$82.59_{0.09}$	$55.22_{0.04}$	$53.42_{0.04}$
1.0	0.5	$86.59_{2.19}$	$51.93_{5.31}$	$47.93_{3.13}$	$82.44_{0.03}$	$54.84_{0.18}$	$53.29_{0.29}$
1.0	1.0	$86.64_{2.10}$	$52.34_{5.79}$	$50.09_{4.34}$	$82.54_{0.48}$	$55.03_{0.39}$	$53.94_{1.31}$
1.0	2.0	$85.56_{1.19}$	$50.66_{4.17}$	$44.67_{3.88}$	$76.88_{1.84}$	$41.89_{3.05}$	$44.83_{2.31}$

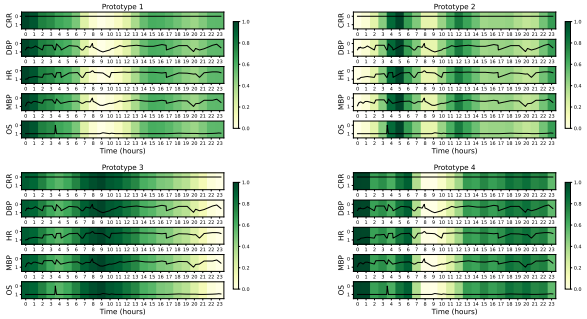


Figure 4: Visualization of the learned prototypes in our CTPD framework. Here we select 5 representative clinical variables to visualize the time series. 'CPR" denotes "Capillary Refill Rate", "DBP" denotes "Diastolic blood pressure", ', "HR" denotes "Heart Rate", "MBP" denotes "Mean blood pressure" and "OS" denotes "Oxygen saturation".

C.3 Ablation Results on Loss Weights

We analyze the effects of loss weights λ_1 and λ_2, shown in Table 11. For the loss weights, $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ consistently yield the best performance across all settings. In contrast, larger values (rows 4–6) lead to performance drops, likely due to the model overemphasizing alignment and reconstruction losses at the expense of prediction accuracy. Moderate changes to λ_1 and λ_2 (rows 1–3) yield stable results, particularly on the complex 24-PHE task, indicating robustness. However, the simpler 48-IHM task is more sensitive, possibly due to its limited temporal complexity and tendency to overfit. In practice, we adopt $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ to maintain a balanced loss scale without extensive tuning. Future work may explore finer-grained searches to better understand the trade-offs across tasks.

C.4 Visualization of Assignment Weights of Prototypes.

Fig. 4 presents the distribution of assignment weights across different time scales for a 48-IHM example. With time window sizes denoted by T,

our model utilizes three time scales with 20, 10, and 5 prototypes respectively. The variation in assignment weights across these scales, as showcased in the figure, underlines our model's proficiency in capturing and differentiating temporal patterns at varying scales. We will try to interpret these learned temporal pattern embeddings in our future work.