ProMedTS: A Self-Supervised, Prompt-Guided Multimodal Approach for Integrating Medical Text and Time Series

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown remarkable performance in vision-language tasks, but their application in the medical field remains underexplored, particularly for integrating structured time series data with unstructured clinical notes. In clinical practice, dynamic time series data such as lab test results capture critical temporal patterns, while clinical notes provide rich semantic context. Merging these modalities is challenging due to the inherent differences between continuous signals and discrete text. To bridge this gap, we introduce ProMedTS, a novel self-supervised multimodal framework that employs prompt-guided learning to unify these heterogeneous data types. Our approach leverages lightweight anomaly detection to generate anomaly captions that serve as prompts, guiding the encoding of raw time series data into informative embeddings. These embeddings are aligned with textual representations in a shared latent space, preserving fine-grained temporal nuances alongside semantic insights. Furthermore, our framework incorporates tailored self-supervised objectives to enhance both intra- and inter-modal alignment. We evaluate ProMedTS on disease diagnosis tasks using real-world datasets, and the results demonstrate that our method consistently outperforms state-of-the-art approaches.

1 Introduction

Recent advancements in natural language processing (NLP) have revolutionized healthcare by enabling deeper insights into electronic health records (EHRs). EHRs combine structured data, such as time series laboratory (lab) test results, with unstructured data, including clinical notes and medical images. While large language models (LLMs) excel at processing unstructured text (Nori et al., 2023; Singhal et al., 2023) and vision transformers have driven progress in medical image analysis (Wang et al., 2022; Chen et al., 2021), integrating

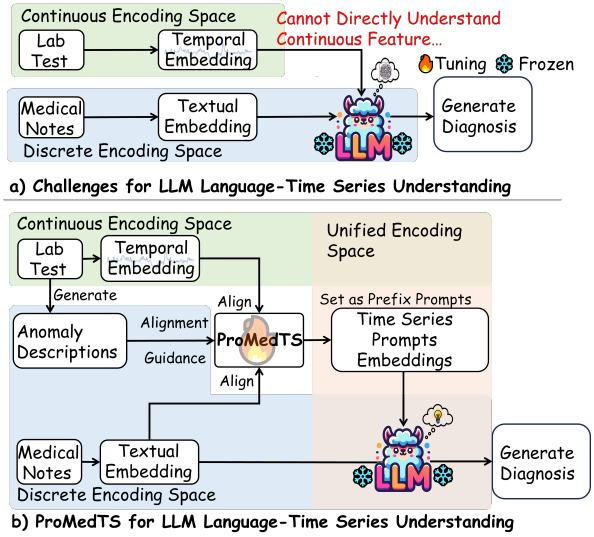

Figure 1: (a) LLMs struggle to process continuous time series data due to modality gaps with discrete textual representations. (b) ProMedTS bridges this gap by leveraging anomaly descriptions and time series prompts, aligning structured EHR data with clinical notes for improved multimodal understanding.

continuous time series data with text remains a challenge. Unlike text, which is composed of discrete tokens, time series data contain continuous signals with temporal dependencies (Jin et al., 2023) as illustrated in Figure 1(a).

Current multimodal learning approaches, especially contrastive learning methods (Radford et al., 2021; Li et al., 2023), have been effective in aligning vision and language. However, they are less suited to bridge the gap between time series and text. Time series data require fine-grained temporal representations in a high-dimensional space and are often irregularly sampled, exhibit diverse frequencies, and include missing values (Harutyunyan et al., 2019a). In addition, the lack of large-scale paired datasets that link raw time series with textual descriptions further hampers LLMs from incorporating structured information into clinical decision-

making (Niu et al., 2024). Without an effective fusion mechanism, LLMs cannot fully exploit the rich temporal patterns in structured EHR data.

To address these challenges, we propose ProMedTS, a self-supervised and prompt-guided framework designed to unify medical notes and time series lab test for naturally understood by LLMs. As shown in Figure 1(b), instead of feeding raw time series directly into LLMs, our framework introduce *anomaly descriptions*, capturing key patterns in lab test results for helping multimodal fusion between text and time series. These descriptions are generated using prompting with anomaly detection technology(Vinutha et al., 2018), converting continuous signals into human-readable summaries. The process involves two steps. First, anomaly descriptions establish a direct connection between time series EHRs and medical notes. Second, time series prompt embeddings are generated and added as prefix tokens to the LLM input. This method integrates structured time series information into the language modeling process without altering the LLM architecture, unifying both modalities within a same encoding space and enhancing clinical decision-making.

We optimize ProMedTS with three self-supervised learning objectives. A contrastive loss maps textual and time series modalities into a shared latent space. An anomaly-time series matching loss links lab test with their corresponding anomaly descriptions to reinforce consistency. Finally, an anomaly caption generation loss improves the fine-grained alignment between numeric time series lab test and time series prompt embeddings. Together, these objectives enable LLMs to process both structured and unstructured EHR data more effectively, addressing the gap between language and time series representations in healthcare applications.

- We propose ProMedTS, a self-supervised framework that integrates structured time series and unstructured textual EHR data into LLMs without changing their architectures.
- We introduce anomaly descriptions as a textual bridge to align time series data with clinical notes, supported by three self-supervised objectives.
- We demonstrate that ProMedTS significantly improves disease diagnosis on MIMIC-III and MIMIC-IV, setting a new benchmark for multimodal EHR learning.

2 Related Work

The increasing diversity of EHR data has led to significant advancements in multimodal learning for healthcare applications. MedCLIP (Wang et al., 2022) employs semantic contrastive learning to align medical images with textual reports, while RAIM (Qiao et al., 2019) and GLoRIA (Huang et al., 2021) integrate numerical and image data with text using attention mechanisms. LDAM (Niu et al., 2021a) further extends these approaches by leveraging cross-attention with disease labels to fuse features from lab tests and clinical notes. EHR-KnowGen (Niu et al., 2024) transforms structured lab data into text and incorporates external knowledge for improved modality fusion. Despite these advancements, achieving a unified latent embedding that effectively captures interactions across diverse modalities remains a key challenge in multimodal EHR processing.

Beyond multimodal learning, recent research has explored generative approaches to healthcare modeling. Conventional methods have primarily relied on discriminative models for disease risk assessment and diagnosis (Choi et al., 2016; Niu et al., 2021b; Qiao et al., 2019). However, generative models are increasingly being adopted, as demonstrated by Clinical CoT (Kwon et al., 2024), applying LLMs for disease diagnosis generation. Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and Chain-of-Thought (CoT) prompting (Wei et al., 2022) have further enhanced medical reasoning capabilities in models such as GatorTron (Yang et al., 2022), MedPalm (Singhal et al., 2023), and GPT4-Med (Nori et al., 2023). While these models excel in medical question-answering, they remain limited in real-world direct disease diagnosis and multimodal EHR processing. EHR-KnowGen (Niu et al., 2024) reframes disease diagnosis as a text-to-text generation problem but overlooks the crucial temporal details embedded in time series lab tests, underscoring the need for more effective and dedicated multimodal fusion strategies.

3 Methodology

In this section, we present the ProMedTS framework for unifying heterogeneous EHR data through prompt-guided learning. We begin by defining the problem and describing the model inputs, then provide a high-level overview of the architecture. In

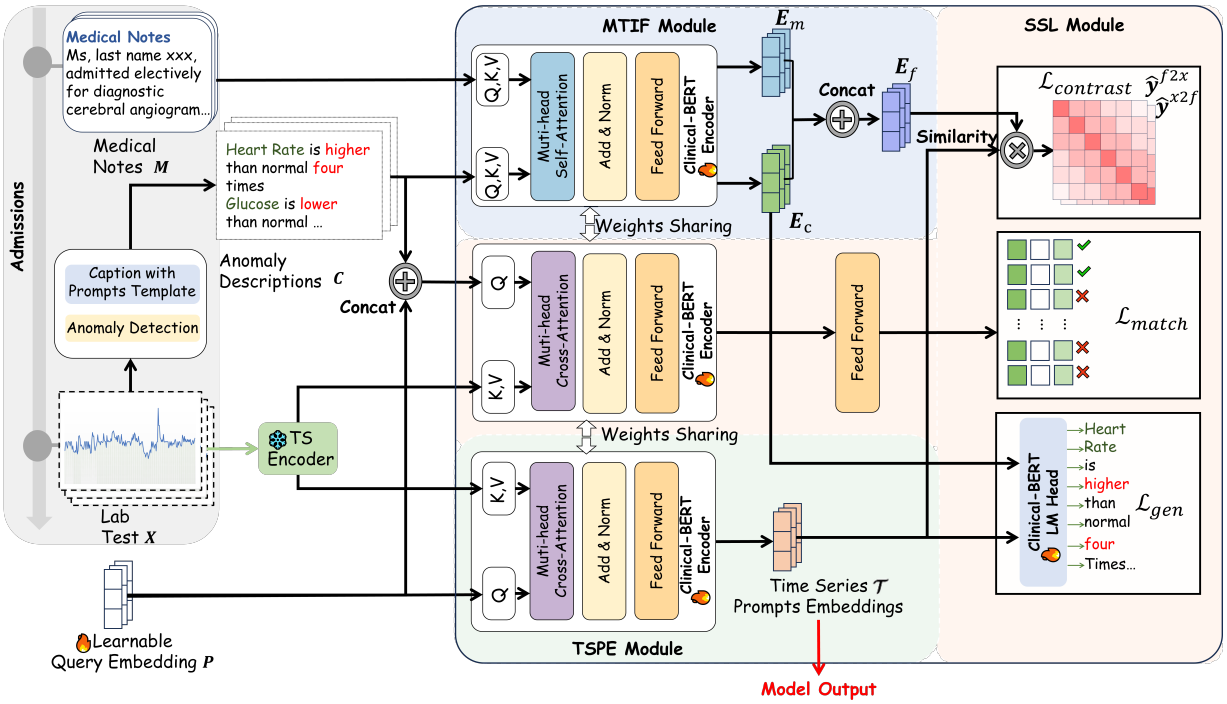

Figure 2: The ProMedTS model comprises three modules: the Time Series Prompt Embedding (TSPE) module, the Multimodal Textual Information Fusion (MTIF) module, and the Self-supervised Learning (SSL) module. The MTIF module utilizes Clinical-BERT to encode medical notes $\boldsymbol{M}$, lab test data $\boldsymbol{X}$, and anomaly descriptions $\boldsymbol{C}$ to generate time series prompt embeddings $\boldsymbol{\mathcal{T}}$.

subsequent sections, we detail each module and discuss how these components are applied to downstream tasks such as disease diagnosis.

3.1 Problem Definition

We introduce ProMedTS, aiming to reduce discrepancies between language and time series EHRs. Specifically, it leverages anomaly captions and generates time series prompt embeddings to unify both modalities in a shared latent space. The inputs to ProMedTS, denoted by $\{\boldsymbol{M}, \boldsymbol{X}\}$, include medical notes $\boldsymbol{M} \in \mathbb{R}^{B \times N_m}$ (where B is the batch size and N_m is the number of tokens) and numeric lab test data $\boldsymbol{X} \in \mathbb{R}^{B \times L \times N_x}$ (where L is the sequence length and N_x is the number of lab test variants). Additionally, a lightweight anomaly detection (Vinutha et al., 2018) is employed to generate textual descriptions of anomalies $\boldsymbol{C} \in \mathbb{R}^{B \times N_c}$ (details in Appendix A.2). ProMedTS also uses learnable time series query embeddings $\boldsymbol{P} \in \mathbb{R}^{B \times N_p \times D}$, which are transformed into time series prompt embeddings $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{B \times N_p \times D}$, where N_p is the query length and D is the hidden dimension.

3.2 Model Overview

Figure 2 illustrates the overview of ProMedTS, which comprises three main modules. Three modules share the same Clinical-BERT(Alsentzer et al., 2019) structured model and are extended to support cross-attention, self-attention, and prompt generation. The Time Series Prompt Embedding (TSPE) module applies a cross-attention mechanism to convert raw lab test data into prompt embeddings, preserving key temporal features. The Multimodal Textual Information Fusion (MTIF) module encodes and merges medical notes with anomaly captions in a unified latent space, facilitating the extraction of complementary semantic information. Finally, the Self-supervised Learning (SSL) module employs tailored loss functions to bridge the modality gap and maintain fine-grained temporal details in the learned representations. These modules work in tandem to achieve robust alignment and fusion of heterogeneous EHRs, and the following sections provide in-depth explanations of each component and their applications.

3.3 Time Series Prompt Embedding

The objective of the TSPE module is to extract and encapsulate the inherent information from time

series lab test data into a time series prompt embedding. Let $\{\boldsymbol{X}, \boldsymbol{P}\}$ represent the module inputs. The numeric lab test data $\boldsymbol{X}$ is first processed by a time series encoder (TSE) using PatchTST (Nie et al., 2022). In parallel, the learnable query embeddings $\boldsymbol{P}$, initialized using vectors extracted from the Clinical-BERT word embedding layer, serve as query tokens in the cross-attention mechanism, guiding the selection of relevant temporal features by attending to time series lab test results encoded by TSE. To generate the final prompt embedding $\boldsymbol{\mathcal{T}}$, we extend the multi-head self-attention encoder of Clinical-BERT to support a multi-head cross-attention mechanism, following a strategy similar to that adopted in (Li et al., 2023). We designate $\boldsymbol{X}$ as both key and value while $\boldsymbol{P}$ serves as the query:

$$\boldsymbol{\mathcal{T}} = \text{Clinical-BERT}\big(\boldsymbol{P}, \text{TSE}(\boldsymbol{X}), \text{TSE}(\boldsymbol{X})\big). \tag{1}$$

This design ensures that the rich temporal patterns in $\boldsymbol{X}$ are captured within $\boldsymbol{\mathcal{T}}$, enabling subsequent modules to leverage these features effectively.

3.4 Multimodal Textual Information Fusion

The MTIF module is designed to fuse medical notes and anomaly descriptions effectively. We use the anomaly captioning method to generate anomaly descriptions, as illustrated in Figure 2. The inputs to the MTIF module are medical notes $\boldsymbol{M}$ and lab test anomaly descriptions $\boldsymbol{C}$, which are encoded separately by Clinical-BERT via the multi-head self-attention mechanism:

$$\begin{aligned} \boldsymbol{E}_m &= \text{Clinical-BERT}(\boldsymbol{M}, \boldsymbol{M}, \boldsymbol{M}), \\ \boldsymbol{E}_c &= \text{Clinical-BERT}(\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}), \end{aligned} \tag{2}$$

where $\boldsymbol{E}_m \in \mathbb{R}^{B\times N_m\times D}$ and $\boldsymbol{E}_c \in \mathbb{R}^{B\times N_c\times D}$. The repeated inputs indicate that the key, query, and value matrices are identical for the self-attention mechanism. This structure enables the model to encode each type of textual information independently while capturing the inherent characteristics and context of each input. The combined textual representation is then derived from these encoded inputs:

$$\boldsymbol{E}_f = AVG([\boldsymbol{E}_m \oplus \boldsymbol{E}_c]), \tag{3}$$

where $\boldsymbol{E}_f \in \mathbb{R}^{B\times D}$, with $\oplus$ indicating concatenation, and AVG representing average pooling.

3.5 Self-Supervised Learning

This module addresses the modality gap between textual and time series EHR data using three specialized loss functions. By simultaneously aligning cross-modal representations and preserving fine-grained temporal details, the model learns to capture both semantic and temporal nuances.

3.5.1 Cross-Modal Contrastive Alignment

To promote cross-modal alignment, we design a contrastive loss that brings language and time series embeddings closer when they originate from the same patient and pushes them apart otherwise. We first compute similarity matrices by multiplying the fused text representation $\boldsymbol{E}_f$ with the time series prompt embeddings $\boldsymbol{\mathcal{T}}$:

$$\begin{aligned} \boldsymbol{g}_{(\boldsymbol{E}_f,\boldsymbol{X})} &= \max\Big(\big[\boldsymbol{E}_f\,\boldsymbol{\mathcal{T}}^{(1)T}, \ldots, \boldsymbol{E}_f\,\boldsymbol{\mathcal{T}}^{(N_p)T}\big]\Big), \\ \boldsymbol{g}_{(\boldsymbol{X},\boldsymbol{E}_f)} &= \max\Big(\big[\boldsymbol{\mathcal{T}}^{(1)}\,\boldsymbol{E}_f^{T}, \ldots, \boldsymbol{\mathcal{T}}^{(N_p)}\,\boldsymbol{E}_f^{T}\big]\Big), \end{aligned} \tag{4}$$

where the $\max$ operator performs max-pooling across N_p dimensions, yielding $\boldsymbol{g}_{(\boldsymbol{E}_f,\boldsymbol{X})}$ and $\boldsymbol{g}_{(\boldsymbol{X},\boldsymbol{E}_f)} \in \mathbb{R}^{B\times B}$. Note that $\boldsymbol{g}_{(\boldsymbol{E}_f,\boldsymbol{X})}$ measures text-to-time series similarity (by fixing $\boldsymbol{E}_f$ and iterating over $\boldsymbol{\mathcal{T}}$), while $\boldsymbol{g}_{(\boldsymbol{X},\boldsymbol{E}_f)}$ captures time-series-to-text similarity (by fixing $\boldsymbol{\mathcal{T}}$ and iterating over $\boldsymbol{E}_f$). This process is the same as that used in vision-language contrastive learning (Radford et al., 2021; Li et al., 2023). We then apply the SoftMax function to generate two distinct sets of logits:

$$\begin{aligned} \hat{\boldsymbol{y}}_c^{f2x} &= \text{SoftMax}\big(\boldsymbol{g}_{(\boldsymbol{E}_f,\boldsymbol{X})}\big), \\ \hat{\boldsymbol{y}}_c^{x2f} &= \text{SoftMax}\big(\boldsymbol{g}_{(\boldsymbol{X},\boldsymbol{E}_f)}\big). \end{aligned} \tag{5}$$

Let $\boldsymbol{y}_c^{f2x}$ and $\boldsymbol{y}_c^{x2f}$ denote the ground truth labels indicating whether the pairs correspond to the same patient in a training batch (1 if matched, 0 otherwise). We use cross-entropy $H(\cdot)$ to define the contrastive loss:

$$\begin{aligned} \mathcal{L}_{contrast} \;=\; \tfrac{1}{2}\,\mathbb{E}\Big[&H\big(\boldsymbol{y}_c^{f2x},\, \hat{\boldsymbol{y}}_c^{f2x}\big) \\ &+\; H\big(\boldsymbol{y}_c^{x2f},\, \hat{\boldsymbol{y}}_c^{x2f}\big)\Big]. \end{aligned} \tag{6}$$

3.5.2 Intra-Modal Matching

To further capture intra-modality consistency, we align lab tests with corresponding anomaly descriptions. This alignment is modeled as a binary classification task, distinguishing matched from unmatched pairs of lab tests and anomaly captions. Following Li et al. (2021), we employ a negative mining strategy to generate labels $\boldsymbol{y}_m$ by selecting the most similar pairs in a training batch as negative samples, where the top 1-ranked pair is labeled as 1 and the others as 0, based on the similarity computed in Equation 4. We employ Clinical-BERT's

cross-attention, where the concatenation of $\boldsymbol{C}$ and $\boldsymbol{P}$ serves as the query, and the encoded time series $\boldsymbol{X}$ is used as both key and value. A Multilayer Perceptron (MLP) classifier with softmax activation, denoted f_{match}, predicts the probability $\hat{\boldsymbol{y}}_m$:

$$\hat{\boldsymbol{y}}_m = f_{match}\Big(\text{Clinical-BERT}\big(f_{\mathcal{W}}(\boldsymbol{C}) \oplus \boldsymbol{P}, \text{TSE}(\boldsymbol{X}), \text{TSE}(\boldsymbol{X})\big)\Big), \quad (7)$$

where $f_{\mathcal{W}}$ is the word embedding layer in Clinical-BERT. We define the matching loss as:

$$\mathcal{L}_{match} = \mathbb{E}\big[H\big(\boldsymbol{y}_m, \hat{\boldsymbol{y}}_m\big)\big], \quad (8)$$

where $\boldsymbol{y}_m$ is the one-hot ground truth label.

3.5.3 Anomaly Description Reconstruction

To ensure the time series prompt embeddings encode both coarse anomaly descriptions and fine-grained temporal details, we reconstruct anomaly captions from the learned embeddings. This step helps unify language tokens and time series representations in a shared space. Specifically, we use Clinical-BERT with a language model head f_{head}, setting $\boldsymbol{E}_c$ as the query and $\boldsymbol{\mathcal{T}}$ as key and value:

$$\mathcal{L}_{gen} = \mathbb{E}\Big[H\big(\boldsymbol{C}, f_{head}(\text{Clinical-BERT}(\boldsymbol{E}_c, \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{T}}))\big)\Big]. \quad (9)$$

This objective is a standard language model generation loss, computed as cross-entropy between the predicted token distribution and the ground truth tokens, encouraging the model to generate accurate textual descriptions, thereby reinforcing alignment between time series prompts and language tokens.

Overall Loss: We combine these objectives into a single training loss:

$$\mathcal{L}_{total} = \alpha\,\mathcal{L}_{contrast} + \beta\,\mathcal{L}_{match} + \gamma\,\mathcal{L}_{gen}, \quad (10)$$

where α, β, and γ are hyperparameters balancing the three losses (see Appendix A.6). Our training algorithm aims to minimize $\mathcal{L}_{total}$ across all samples (details in Appendix A.1).

3.6 LLM-based Disease Diagnosis with ProMedTS

To illustrate the practical effectiveness of ProMedTS in unifying textual and time series data, we employ a pre-trained, frozen LLM model for disease diagnosis. As depicted in Figure 3, ProMedTS first converts numeric lab test results into time series prompt embeddings, which are then aligned via a fully connected layer to match the LLM's input dimensions. These embeddings serve as prefix soft prompts, concatenated with the medical notes so that the model can ingest structured signals from time series alongside unstructured clinical text. By bridging language and time series modalities, the LLM can process both inputs concurrently, leveraging complementary information for enhanced diagnostic accuracy.

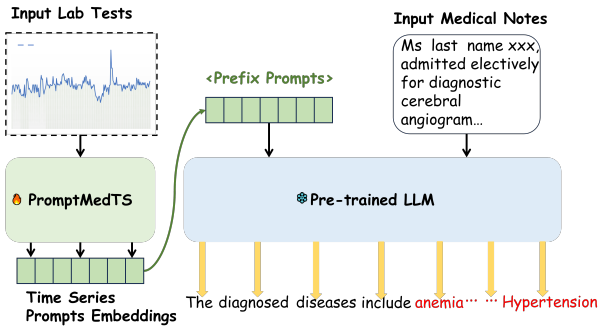

Figure 3: ProMedTS for empowering LLMs to in disease diagnosis.

4 Experiments

4.1 Datasets and Preprocessing

The MIMIC-III dataset (Johnson et al., 2016) is a publicly available EHR dataset containing de-identified patients who were admitted to ICUs between 2001 and 2012. It includes medical discharge summaries, lab test results, chest x-ray images and more. Our analysis focuses on EHR data from approximately 27,000 patients including complete medical discharge summaries and lab test results. The MIMIC-IV dataset (Johnson et al., 2023) comprises EHR data from 2008 to 2019. We utilize approximately 29,000 EHR records from MIMIC-IV, which include complete medical discharge summaries and lab test results. Our study targets 25 disease phenotypes as defined in the MIMIC-III benchmark (Harutyunyan et al., 2019a).

Data Pre-processing. For medical notes, we extract the brief course from medical discharge summaries, removing numbers, noise, and stopwords. Numerical lab test results are converted into time series data using the benchmark tools (Harutyunyan et al., 2019b), with missing values filled using the nearest available numbers. Time series anomaly descriptions are used the method defined in Appendix A.2. Data splitting follows the guidelines in (Harutyunyan et al., 2019b), using a 4:1 ratio for

Models	Size	Type		Modality		Micro			Macro		
		CLS	GEN	Lab	Note	Precision	Recall	F1	Precision	Recall	F1
MIMIC-III											
GRU	7.9M	✓		✓		$46.41_{(3.48)}$	$21.88_{(3.59)}$	$29.43_{(1.89)}$	$30.47_{(4.23)}$	$13.00_{(1.14)}$	$14.59_{(1.48)}$
PatchTST	19.2M	✓		✓		$32.64_{(3.59)}$	$42.72_{(5.01)}$	$36.02_{(1.09)}$	$26.86_{(3.51)}$	$29.71_{(4.78)}$	$19.25_{(3.50)}$
TimeLLM	78M	✓		✓		$37.43_{(1.17)}$	$54.93_{(6.56)}$	$36.59_{(1.17)}$	$10.18_{(2.30)}$	$35.21_{(6.47)}$	$15.16_{(2.17)}$
CAML	36.1M	✓			✓	$69.04_{(0.18)}$	$55.87_{(2.72)}$	$61.54_{(0.30)}$	$65.08_{(2.56)}$	$50.12_{(3.05)}$	$54.42_{(0.94)}$
DIPOLE	39M	✓			✓	$64.38_{(0.89)}$	$57.94_{(1.15)}$	$60.98_{(0.27)}$	$61.63_{(1.03)}$	$53.02_{(1.18)}$	$55.68_{(0.49)}$
Flan-T5	60M		✓		✓	$58.12_{(1.11)}$	$66.23_{(0.72)}$	$62.03_{(0.54)}$	$56.56_{(1.03)}$	$62.47_{(0.76)}$	$58.87_{(0.71)}$
PROMPTEHR	75.2M		✓		✓	$59.29_{(0.97)}$	$65.53_{(0.69)}$	$62.24_{(0.23)}$	$57.44_{(0.97)}$	$62.87_{(0.61)}$	$59.10_{(0.24)}$
LLaMA	7B		✓	✓	✓	$61.42_{(.2.08)}$	$65.98_{(1.53)}$	$63.64_{(0.41)}$	$61.08_{(1.54)}$	$61.64_{(1.27)}$	$60.55_{(0.44)}$
LDAM	41.3M	✓		✓	✓	$68.00_{(1.23)}$	$57.12_{(0.47)}$	$62.18_{(0.40)}$	$67.38_{(0.35)}$	$51.50_{(0.95)}$	$57.44_{(0.60)}$
FROZEN	265M		✓	✓	✓	$61.09_{(1.81)}$	$64.07_{(1.58)}$	$62.51_{(0.34)}$	$59.96_{(1.55)}$	$59.99_{(1.66)}$	$59.15_{(0.30)}$
EHR-KnowGen	76.9M		✓	✓	✓	$60.01_{(0.29)}$	$65.51_{(0.18)}$	$62.62_{(0.06)}$	$58.34_{(0.38)}$	$61.81_{(0.28)}$	$59.44_{(0.06)}$
ProMedTS	267.5M		✓	✓	✓	$61.32_{(0.54)}$	$66.65_{(0.51)}$	$\underline{63.67}_{(0.08)}$	$60.35_{(0.61)}$	$61.62_{(0.71)}$	$\underline{60.42}_{(0.18)}$
ProMedTS*	1B		✓	✓	✓	$60.62_{(0.22)}$	$67.83_{(0.18)}$	$\mathbf{64.02}_{(0.11)}$	$59.43_{(0.37)}$	$63.65_{(0.54)}$	$\mathbf{60.78}_{(0.13)}$
MIMIC-IV											
GRU	7.9M	✓		✓		$56.23_{(1.13)}$	$25.77_{(1.58)}$	$35.21_{(1.36)}$	$38.37_{(1.90)}$	$16.97_{(1.22)}$	$20.65_{(1.32)}$
PatchTST	19.2M	✓		✓		$27.26_{(0.03)}$	$57.42_{(0.41)}$	$36.97_{(0.10)}$	$20.59_{(2.76)}$	$43.72_{(0.25)}$	$21.78_{(2.83)}$
TimeLLM	78M	✓		✓		$30.30_{(1.78)}$	$60.46_{(1.98)}$	$40.31_{(1.20)}$	$24.61_{(2.21)}$	$47.26_{(2.37)}$	$25.56_{(1.60)}$
CAML	36.1M	✓			✓	$72.82_{(0.54)}$	$59.48_{(0.82)}$	$65.40_{(0.36)}$	$67.25_{(0.99)}$	$50.73_{(1.49)}$	$54.71_{(1.42)}$
DIPOLE	39M	✓			✓	$72.39_{(0.51)}$	$61.38_{(0.83)}$	$66.43_{(0.33)}$	$70.45_{(0.37)}$	$55.65_{(0.79)}$	$60.37_{(0.62)}$
Flan-T5	60M		✓		✓	$66.24_{(0.52)}$	$69.53_{(0.18)}$	$67.92_{(0.41)}$	$64.28_{(0.58)}$	$66.01_{(0.54)}$	$64.79_{(0.36)}$
PROMPTEHR	75.2M		✓		✓	$65.24_{(0.68)}$	$70.31_{(0.56)}$	$68.02_{(0.17)}$	$63.53_{(0.47)}$	$67.02_{(0.65)}$	$65.01_{(0.28)}$
LLaMA	7B		✓	✓	✓	$68.54_{(1.12)}$	$69.54_{(0.73)}$	$69.29_{(0.32)}$	$67.53_{(0.91)}$	$66.24_{(1.13)}$	$66.21_{(0.64)}$
LDAM	41.3M	✓		✓	✓	$72.01_{(0.85)}$	$62.74_{(0.62)}$	$66.91_{(0.20)}$	$69.77_{(0.18)}$	$56.72_{(0.69)}$	$60.77_{(0.48)}$
FROZEN	265M		✓	✓	✓	$67.81_{(0.78)}$	$69.08_{(0.94)}$	$68.42_{(0.08)}$	$66.27_{(1.00)}$	$65.21_{(0.97)}$	$65.30_{(0.05)}$
EHR-KnowGen	76.9M		✓	✓	✓	$65.80_{(0.64)}$	$70.85_{(0.45)}$	$68.16_{(0.11)}$	$63.82_{(0.53)}$	$67.24_{(0.55)}$	$65.11_{(0.13)}$
ProMedTS	267.5M		✓	✓	✓	$71.63_{(0.46)}$	$67.81_{(0.85)}$	$\underline{69.69}_{(0.18)}$	$70.12_{(0.47)}$	$63.58_{(0.79)}$	$\underline{66.21}_{(0.17)}$
ProMedTS*	1B		✓	✓	✓	$71.12_{(0.31)}$	$69.33_{(0.42)}$	$\mathbf{70.21}_{(0.05)}$	$70.97_{(0.42)}$	$65.51_{(0.64)}$	$\mathbf{67.56}_{(0.09)}$

Table 1: The performance of comparative methods in the disease diagnosis tasks on MIMIC-III and MIMIC-IV. Please note CLS - classification model, GEN -generative model, Lab - lab test result, and Note - medical notes.

training and testing.

4.2 Baseline Methods

We benchmark our approach against a range of methods: GRU (Cho et al., 2014), PatchTST (Nie et al., 2022), TimeLLM (Jin et al., 2023), CAML (Mullenbach et al., 2018), DIPOLE (Ma et al., 2017), Flan-T5 (Chung et al., 2024), PROMPTEHR (Wang and Sun, 2022), LLaMA-1-7B (Touvron et al., 2023) with anomalies input, LDAM (Niu et al., 2021a), FROZEN (Tsimpoukelli et al., 2021), and EHR-KnowGen (Niu et al., 2024). Detailed configurations of these baselines are provided in Appendix A.3. For the disease diagnosis task, we adopt two scales of Flan-T5 (Chung et al., 2024) as the frozen LLM to validate our model's effectiveness in understanding multimodal EHRs, primarily driven by resource considerations and ease of experimentation. The Flan-T5-Small-based model is denoted as PromMedTS, while the Flan-T5-Large-based model is denoted as ProMedTS*. In principle, any sufficiently LLM could be substituted to potentially achieve even stronger results.

To ensure a fair comparison, all baselines also employ Flan-T5 as their backbone. Reported results are averaged over five runs with different random seeds. The statistical significance determined at $p < 0.05$ by t-test. Implementation details for every model are described in Appendix A.4, and training instructions appear in Appendix A.5. Our code is publicly available at https://anonymous.4open.science/r/PromptMedTS-V1-5F51.

4.3 Disease Diagnosis Performance

Table 1 shows that ProMedTS achieves the highest overall performance, particularly in F1 scores on MIMIC-IV. In addition, replacing the LLM with a larger model improves F1 scores on both datasets, indicating our model's scalability and robustness across different LLMs. Furthermore, TimeLLM performs strongly with lab test, highlighting the value of time-series inputs for LLMs in disease diagnosis. Text-based methods (e.g., Flan-T5) generally outperform time-series approaches, suggesting that medical notes capture richer disease-related information. Multimodal models (e.g., EHR-KnowGen, LLaMA) exceed single-modality baselines (e.g., TimeLLM, PROMPTEHR), confirming

Models	Micro			Macro		
	Precision	Recall	F1	Precision	Recall	F1
MIMIC-III						
ProMedTS	$61.32_{(0.54)}$	$66.65_{(0.51)}$	$\mathbf{63.67}_{(0.08)}$	$60.35_{(0.61)}$	$61.62_{(0.71)}$	$\mathbf{60.42}_{(0.18)}$
***w/o* LAB**	$58.91_{(0.83)}$	$66.59_{(0.57)}$	$62.34_{(0.26)}$	$57.32_{(0.88)}$	$62.56_{(0.61)}$	$59.05_{(0.22)}$
***w/o* ANOMALY**	$60.09_{(0.32)}$	$65.03_{(0.98)}$	$62.44_{(0.22)}$	$59.13_{(0.43)}$	$60.46_{(1.15)}$	$59.11_{(0.25)}$
MIMIC-IV						
ProMedTS	$71.63_{(0.46)}$	$67.81_{(0.85)}$	$\mathbf{69.69}_{(0.18)}$	$70.12_{(0.47)}$	$63.58_{(0.79)}$	$\mathbf{66.21}_{(0.17)}$
***w/o* LAB**	$67.16_{(0.55)}$	$69.42_{(0.59)}$	$68.22_{(0.31)}$	$65.74_{(0.62)}$	$64.69_{(0.48)}$	$64.33_{(0.18)}$
***w/o* ANOMALY**	$70.94_{(0.37)}$	$66.44_{(1.37)}$	$68.47_{(0.12)}$	$68.95_{(0.79)}$	$62.45_{(0.96)}$	$65.13_{(0.12)}$

Table 2: Ablation studies on different modality input and alignment designs for disease diagnosis.

Models	Micro			Macro		
	Precision	Recall	F1	Precision	Recall	F1
MIMIC-III						
ProMedTS	$61.32_{(0.54)}$	$66.65_{(0.51)}$	$\mathbf{63.67}_{(0.08)}$	$60.35_{(0.61)}$	$61.62_{(0.71)}$	$\mathbf{60.42}_{(0.18)}$
***w/o* CONTRAST**	$60.24_{(0.25)}$	$66.00_{(0.39)}$	$62.99_{(0.07)}$	$59.92_{(0.60)}$	$61.41_{(0.51)}$	$59.73_{(0.07)}$
***w/o* MATCH**	$60.12_{(0.58)}$	$66.14_{(1.50)}$	$62.96_{(0.02)}$	$59.70_{(1.18)}$	$61.37_{(1.34)}$	$59.65_{(0.11)}$
***w/o* GEN**	$59.95_{(0.38)}$	$66.15_{(0.27)}$	$62.89_{(0.19)}$	$59.57_{(0.55)}$	$61.32_{(0.30)}$	$59.61_{(0.20)}$
MIMIC-IV						
ProMedTS	$71.63_{(0.46)}$	$67.81_{(0.85)}$	$\mathbf{69.69}_{(0.18)}$	$70.12_{(0.47)}$	$63.58_{(0.79)}$	$\mathbf{66.21}_{(0.17)}$
***w/o* CONTRAST**	$70.19_{(0.25)}$	$66.22_{(0.39)}$	$68.61_{(0.09)}$	$69.05_{(0.24)}$	$62.40_{(0.35}$	$65.21_{(0.12)}$
***w/o* MATCH**	$70.79_{(0.34)}$	$66.49_{(0.38)}$	$68.67_{(0.15)}$	$68.91_{(0.73)}$	$62.25_{(0.48)}$	$65.47_{(0.15)}$
***w/o* GEN**	$71.30_{(0.29)}$	$65.79_{(0.57)}$	$68.44_{(0.13)}$	$69.14_{(0.48)}$	$62.05_{(0.54)}$	$65.03_{(0.13)}$

Table 3: Ablation studies on the effectiveness of different loss functions of our model for disease diagnosis.

the benefits of integrating text and time series. Generative approaches (e.g., TimeLLM, LLaMA, EHR-KnowGen) also outperform classification-based methods. Although LLaMA performs well, its higher variance and parameter requirements reduce practicality. Notably, our ProMedTS and ProMedTS* surpass all baselines (especially a large improvement on Flan-T5) in F1 scores, highlighting its efficiency and effectiveness.

4.4 Ablation Studies

4.4.1 Effect of Modality Alignment in ProMedTS

This section presents ablation studies to evaluate each module in ProMedTS. ProMedTS *w/o* LAB excludes lab test, removing modality alignment with anomaly descriptions and medical notes. ProMedTS *w/o* ANOMALY removes alignment with anomaly descriptions while keeping alignment between lab test and medical notes to assess the impact of self-supervision. Table 2 summarizes the results, showing that ProMedTS *w/o* LAB suffers a significant drop in F1 scores, highlighting the importance of lab test. ProMedTS *w/o* ANOMALY also shows reduced performance, highlighting the challenges of aligning modalities from discrete and continuous encoding spaces and the adverse effects of misalignment on multimodal understanding.

4.4.2 Impact of Self-Supervised Loss Functions

Table 3 summarizes an ablation study on the loss functions in ProMedTS. Both ProMedTS *w/o* CONTRAST and ProMedTS *w/o* MATCH show slight declines in F1 scores, emphasizing the importance of $\mathcal{L}_{contrast}$ for aligning and unifying time series and textual inputs within a shared latent space. The results also underscore the role of $\mathcal{L}_{match}$ in intra-modal alignment, ensuring the distinctiveness of time series data by aligning lab test with time series prompt embeddings. Notably, ProMedTS *w/o* GEN exhibits a significant drop in F1 scores, highlighting the critical role of $\mathcal{L}_{gen}$ in refining prompt embeddings and integrating temporal information from time series data and anomaly descriptions.

4.5 Model Efficiency and Complexity

Figure 4 illustrates the parameter counts and computation times of baseline models on the two datasets. Our model, ProMedTS, matches the parameter counts and computation times of multimodal baselines such as LDAM and FROZEN, while using 25× fewer parameters and requiring one-third less training time than LLaMA, all while achieving superior diagnostic performance, highlighting its efficiency and effectiveness in language-time series multimodal alignment and fusion.

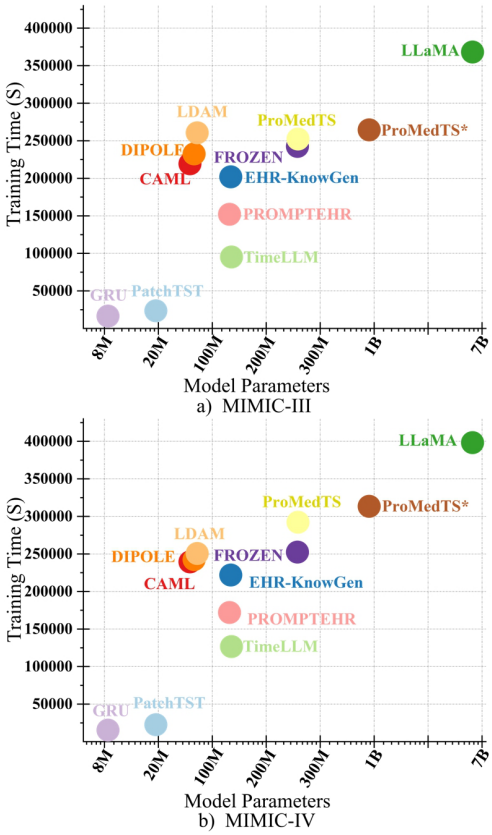

Figure 4: The model parameters and computation time of all baselines.

N_p	Micro F1	Macro F1
MIMIC-III		
12	$63.09_{(0.06)}$	$59.69_{(0.12)}$
24	$63.67_{(0.08)}$	$60.42_{(0.18)}$
36	$63.32_{(0.09)}$	$59.96_{(0.15)}$
MIMIC-IV		
12	$68.98_{(0.15)}$	$65.43_{(0.18)}$
24	$69.69_{(0.18)}$	$66.21_{(0.17)}$
36	$69.41_{(0.19)}$	$65.91_{(0.20)}$

Table 4: Sensitivity analysis on different length of time series prompt embedding.

Lab Test Input	Micro F1	Macro F1
MIMIC-III		
Numerical Values	$32.21_{(1.33)}$	$23.53_{(1.21)}$
Anomaly captions	$35.19_{(0.92)}$	$24.75_{(0.76)}$
Time series prompts	$36.11_{(1.14)}$	$25.47_{(1.02)}$
MIMIC-IV		
Numerical Values	$37.75_{(1.46)}$	$26.10_{(1.09)}$
Anomaly captions	$39.56_{(1.14)}$	$27.22_{(0.77)}$
Time series prompts	$40.14_{(1.05)}$	$28.43_{(0.91)}$

Table 5: Micro and Macro F1 Scores Across Various Lab Test Input Types on the LLM for Disease Diagnosis

4.6 Sensitivity Analysis of Time Series Prompt Length

We performed a sensitivity analysis to examine the impact of the time series prompt embedding length (N_p) on the performance of ProMedTS in disease diagnosis. Table 4 shows the F1 scores for embedding lengths of 12, 24, and 36. Slight fluctuations are observed in both micro and macro F1 scores across datasets. The optimal embedding length is 24 for both datasets, consistent with the configuration used in our experiments.

4.7 Evaluating the Role of Anomaly Descriptions

To highlight the advantages of using lab test anomaly captions over raw numerical time series values in LLMs, we evaluate Flan-T5-small with both input types. Table 5 presents the evaluation results on the MIMIC-III and MIMIC-IV datasets for disease diagnosis. The results show that Flan-T5 achieves over a 2% improvement in Micro F1 score when using anomaly captions, demonstrating that LLMs interpret anomaly captions more effectively than raw numerical values in time series lab test data. Additionally, the inclusion of time series prompts underscores the effectiveness of our model, ProMedTS, in capturing both fine-grained and coarse-grained temporal information from lab test results for disease diagnosis.

5 Conclusion and Future Work

In this paper, we introduce ProMedTS, a lightweight and effective modality fusion framework that leverages self-supervised prompt learning for multimodal EHR integration. By bridging the modality gap between medical notes and lab test results, ProMedTS enables LLMs to process structured and unstructured medical data more effectively. Its three key modules and self-supervised loss functions advance language–time series integration in healthcare, providing a scalable and adaptable approach for real-world clinical applications. Evaluation on two real-world EHR datasets demonstrates that ProMedTS significantly outperforms existing models in disease diagnosis, underscoring its potential to enhance clinical decision-making and improve patient care. In future work, we plan to extend our approach to larger and more diverse datasets, explore additional LLM architectures, and investigate further improvements in modality alignment techniques.

Limitations

While this study focuses on modality alignments and their application in downstream tasks, enhancing the explainability of disease diagnosis remains an area for future work, where we plan to incorporate the Chain-of-Thought rationale (Wei et al., 2022). Additionally, computational constraints required the use of a relatively compact LLM, limiting the amount of clinical text processed at once, which may impact the model's ability to leverage full medical histories. Expanding to more capable models will help address this challenge. Furthermore, our study primarily targets higher-level disease phenotypes in the International Classification of Diseases (ICD) codes (Slee, 1978), which could be expended to more downstream tasks. Future work will explore larger models, such as LLaMA (Touvron et al., 2023) and Mistral (Jiang et al., 2023), to improve diagnostic granularity and broaden coverage.

Ethics Statement

Data Privacy: While the datasets utilized in our research, such as MIMIC-III and MIMIC-IV, are publicly accessible and feature de-identified patient data, accessing these datasets still requires passing the CITI examination and applying for the data through PhysioNet.

References

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.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019a. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18.

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

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, 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.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18417–18425.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances*

in neural information processing systems, 34:9694–9705.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*.

Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Li Guo, and Xian Yang. 2024. Ehr-knowgen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion*, 102:102069.

Shuai Niu, Yin Qin, Yunya Song, Yike Guo, and Xian Yang. 2021a. Label dependent attention model for disease risk prediction using multimodal electronic health records. In *Proceedings of the IEEE conference on data mining*, pages 455–464.

Shuai Niu, Yunya Song, Yin Qin, Yike Guo, and Xian Yang. 2021b. Label-dependent and event-guided interpretable disease risk prediction using ehrs. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction*, 1:A1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Vergil N Slee. 1978. The international classification of diseases: ninth revision (icd-9).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

HP Vinutha, B Poornima, and BM Sagar. 2018. Detection of outliers using interquartile range technique from intrusion dataset. In *Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA*, pages 511–518. Springer.

Zifeng Wang and Jimeng Sun. 2022. Promptehr: Conditional electronic healthcare records generation with prompt learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2873 – 2885. Association for Computational Linguistics.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Algorithm 1 The ProMedTS Model

Input: Given lab test $\boldsymbol{X}$ and medical note $\boldsymbol{M}$ denote the EHRs input. $\boldsymbol{P}$ consists of a set of learnable prompt embeddings.
while not converge **do**
  for mini-batch B **do**
    Obtain the time series anomaly caption $\mathcal{T}$ using equation (1).
    Obtain multimodal textual embedding $\boldsymbol{E}_f$ using equations (2) and (3).
    Calculate the contrastive loss $\mathcal{L}_{contrast}$ between lab test, anomalies, and medical notes using equations (4), (5), and (6).
    Calculate the matching loss $\mathcal{L}_{match}$ between lab test and anomalies using equations (7) and (8).
    Calculate the generation loss $\mathcal{L}_{gen}$ between lab test and anomalies using equation (9).
  end for
  Update parameters by minimizing the total loss $\mathcal{L}_{total}$ defined in Equation (10) by using the AdamW optimizer (Loshchilov and Hutter, 2018) for patients in each batch.
end while

A Appendix

A.1 Algorithm

The training procedure to optimize ProMedTS by minimizing the loss defined in Equation (10) is shown in Algorithm 1.

A.2 Lab Test Anomaly Caption

Time series anomaly descriptions are generated using the IQR method (Vinutha et al., 2018) to identify anomalies, capturing their timing and polarity (above or below standard values) and describing them with handcrafted templates. To caption the lab test anomaly into textual format, we design several text templates to describe the lab test anomalies. All templates are illustrated in Table 6.

A.3 Baseline Models

- **GRU**: The Gated Recurrent Unit (GRU) (Cho et al., 2014), a variant of recurrent neural networks (RNNs), employs two gates to capture both long-term and short-term temporal features effectively.
- **PatchTST**: PatchTST (Nie et al., 2022) is a transformer-based time series encoder designed for long-term forecasting. It segments time series into subseries-level patches, treating each as a token within the transformer architecture.
- **TimeLLM**: TimeLLM (Jin et al., 2023) is an LLM-based time series prediction model that reprograms input time series into text prototypes before processing them with a frozen LLM. It achieves state-of-the-art performance in mainstream forecasting tasks, particularly in few-shot and zero-shot scenarios.
- **CAML**: The Convolutional Attention for Multi-Label classification (CAML) (Mullenbach et al., 2018) is a classical model for classifying medical notes, incorporating a cross-attention mechanism and label embeddings to enhance interpretability. For a fair comparison with more recent language models, its original embedding layer is replaced with one from T5.
- **DIPOLE**: DIPOLE (Ma et al., 2017) is a classic disease prediction model that utilizes two Bi-directional RNNs. It incorporates an attention mechanism to integrate information from both past and future hospital visits. For a fair comparison with more recent language models, its original embedding layer has been replaced with one from T5.
- **Flan-T5**: showcased within the scaling instruction-fine-tuning framework for language models (Chung et al., 2024). It benefits from training on a wide array of datasets geared toward tasks like summarization and question answering.
- **PROMPTEHR**: PROMPTEHR (Wang and Sun, 2022) introduces a novel approach in generative models for electronic health records (EHRs), implementing conditional prompt learning. In this study, the model is specifically geared towards disease diagnosis.
- **LLaMA**: LLaMA-7B (Touvron et al., 2023), one of the leading large language models, is enhanced by Reinforcement Learning with Human Feedback (RLHF) and instructive tuning. It is fine-tuned for disease diagnosis in this study, demonstrating its versatility in various NLP tasks.
- **LDAM**: LDAM (Niu et al., 2021a) leverages

multimodal inputs, combining laboratory testing results and medical notes for disease risk prediction. It utilizes label embedding to effectively integrate these two modalities.

- **FROZEN**: FROZEN (Tsimpoukelli et al., 2021) represents the cutting-edge multimodal vision-language models for few-shot learning. In our study, it is adapted to the disease diagnosis task using inputs from lab test results and medical notes.

- **EHR-KnowGen**: EHR-KnowGen (Niu et al., 2024), touted as the state-of-the-art in EHR multimodal learning models, focuses on disease diagnosis generation. For this study, external domain knowledge is excluded to ensure a fair comparison.

A.4 Implementation Details

In experiments, we utilized PyTorch framework version 2.0.1, operating on a CUDA 11.7 environment. We employed the AdamW optimizer with a starting learning rate of $1e^{-5}$ and a weight decay parameter of 0.05. Additionally, we implemented a warm-up strategy covering 10% of the training duration. Our experiments were conducted on high-performance NVIDIA Tesla V100 GPUs. Within the ProMedTS model, we used 24 time series prompt embeddings, each with a dimensionality of 768. The model's hidden layer size was maintained at 768 for modality alignment and adjusted to 512 for downstream tasks. To standardize the time series data input, we padded all lab test results to a uniform length of 1000 time steps, allowing us to divide the data into 125 patches, with each patch containing 8 time steps. The Flan-T5 is fine-tuned on two MIMIC datasets (Johnson et al., 2016, 2023) and then frozen for downstream tasks.

A.5 Training Instruction Template

Table 7 illustrates the training instruction template for our model ProMedTS for disease diagnosis on MIMIC-III and MIMIC-IV datasets.

A.6 Sensitivity Analysis of Varying Ratios in Loss Function Components

To examine the impact of different combinations of the three loss functions, $\mathcal{L}_{contrast}$, $\mathcal{L}_{match}$, and $\mathcal{L}_{gen}$, on the downstream performance, we perform a sensitivity analysis using three sets of loss ratios: 1:1:1, 1:2:2, and 1:2:1 on MIMIC-III and MIMIC-IV datasets. Since the value of $\mathcal{L}_{contrast}$ is typically larger than those of $\mathcal{L}_{match}$ and $\mathcal{L}_{gen}$, we assign greater weights to $\mathcal{L}_{match}$ and $\mathcal{L}_{gen}$. Figure 5 presents the results, where lines indicate the variation in the sum of the three loss functions on the testing dataset and bars represent the Micro and Macro F1 scores. The figure reveals that varying the weight ratios of the three loss functions has minimal impact on model convergence and the performance of downstream disease diagnosis tasks.

A.7 Discussion on LLMs Selection

In this section, we explain our choice of Flan-T5 as the base LLM for downstream tasks, focusing on three key aspects. 1). Instruction Tuning: Our method trains with instructions, as shown in Table 7. Flan-T5 (Chung et al., 2024) is trained using instruction tuning, which allows it to generalize better across various tasks compared to the supervised fine-tuning used for OPT (Zhang et al., 2022) and GPT-2 (Radford et al., 2019). 2). Rapid Proof-of-Concept and Practical Development: Training large language models, such as LLaMA-7B, requires extensive computational resources and time (see Figure 4). Models with fewer parameters can still demonstrate our method's effectiveness in understanding multimodal EHRs, as shown in our experiment in Section 4.3 (ProMedTS vs. Flan-T5). Moreover, LLMs with fewer than 1 billion parameters will be more efficient for real-world healthcare applications. 3). Scaling LLMs for Higher Performance: As shown in Section 4.3, scaling Flan-T5 to 1 billion parameters leads to a stable increase in disease diagnosis performance, as measured by both Micro and Macro F1 scores on both datasets. In future work, we will extend our approach to different models and tasks to further demonstrate its effectiveness.

If lab test value is not an abnormal value: *{Lab features} is normal all the time.*
If the lab test value is an abnormal value higher than the standard: *{Lab features} is higher than normal {number of times} times.*
If the lab test value is an abnormal value lower than the standard: *{Lab features} is lower than normal {number of times} times.*
If the lab test value is an abnormal that include both higher and lower than the standard value: *{Lab features} is higher than normal {number of times} times and lower than normal*{number of *times} times.*

Table 6: Lab test anomaly caption template.

Diagnose disease from the following medical notes and lab test: **Medical Notes:** ms woman signficant pmh atrial fibrillation lung cancer resection congestive heart hypertension worsening diarrhea dysuria hypotensive admitted unit presumed active issues hypotensive pronounced leukocytosis multiple potential sources ct scan peritoneal fluids free fluid started surgery ... **Lab Test:** $< prefix_1 >, < prefix_2 >, ..., < prefix_n >$
Diagnosis: Acute and unspecified renal failure, Fluid and electrolyte disorders, Septicemia (except in labor), Shock, Chronic obstructive pulmonary disease and bronchiectasis, Disorders of lipid metabolism, Cardiac dysrhythmias, Congestive heart failure; nonhypertensive, Diabetes mellitus with complications, Other liver diseases.

Table 7: Training Instruction Template

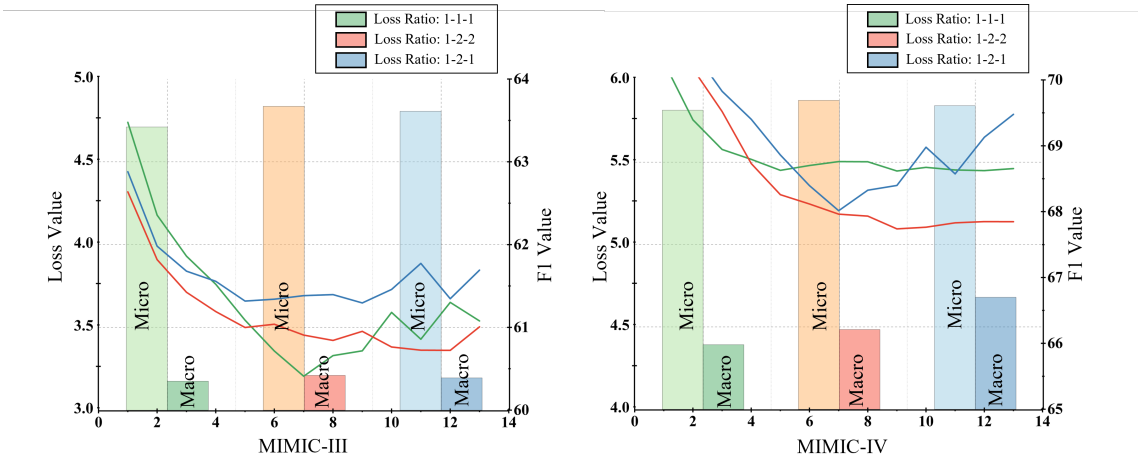

Figure 5: Sensitivity analysis of varying ratios in loss function components, showing Micro and Macro F1 scores for downstream tasks.