MedualTime: A Dual-Adapter Language Model for Medical Time Series-Text Multimodal Learning

Jiexia Ye[1]† , **Weiqi Zhang**[2]† , **Ziyue Li**[3*] , **Jia Li**[2*] , **Meng Zhao**[4] , **Fugee Tsung**[2]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]University of Cologne
[4]Columbia University

jye324@connect.hkust-gz.edu.cn, wzhangcd@connect.ust.hk, zlibn@wiso.uni-koeln.de, mz3048@cumc.columbia.edu, {jialee, season}@ust.hk

Abstract

The recent rapid advancements in language models (LMs) have garnered attention in medical time series-text multimodal learning. However, existing contrastive learning-based and prompt-based LM approaches tend to be biased, often assigning a primary role to time series modality while treating text modality as secondary. We classify these approaches under a temporal-primary paradigm, which may overlook the unique and critical task-relevant information embedded in text modality like clinical reports, thus failing to fully leverage mutual benefits and complementarity of different modalities. To fill this gap, we propose a novel textual-temporal multimodal learning paradigm that enables either modality to serve as the primary while being enhanced by the other, thereby effectively capturing modality-specific information and fostering cross-modal interaction. In specific, we design ***MedualTime***, a language model composed of dual adapters to implement temporal-primary and textual-primary modeling simultaneously. Within each adapter, lightweight adaptation tokens are injected into the top layers of LM to encourage high-level modality fusion. The shared LM pipeline by dual adapters not only achieves adapter alignment but also enables efficient fine-tuning, reducing computational resources. Empirically, MedualTime demonstrates superior performance on medical data, achieving notable improvements of 8% accuracy and 12% F1 in supervised settings. Furthermore, MedualTime's transferability is validated by few-shot transfer experiments from coarse-grained to fine-grained medical data.

1 Introduction

In the medical domain, time series and text modalities are pivotal in representing patient information. Time series, such as electrocardiograms (ECG), electroencephalograms (EEG), and vital signs, provide continuous measurements that reflect a patient's physiological states over time [Wang *et al.*, 2024].

On the other hand, text modalities encompass various clinical documentation, such as radiology reports and physicians' notes, which offer contextual insights for understanding the broader scope of a patient's clinical condition [Davuluri, 2024]. Traditional methods typically involve designing models tailored for single data types and tasks [Fries *et al.*, 2021; Xie *et al.*, 2022; Weimann and Conrad, 2021]. However, jointly modeling time series with text modalities offers richer insights for clinical decision-making. For instance, EEG signals capture electrical activity in the brain, while clinical reports usually provide health history. Analyzing only the symptom history might suggest epilepsy but struggles to specify seizure types, while EEGs detect abnormal brain activity but lack individual medical context. Integrating both modalities can improve diagnostic precision and rationality. A key challenge for jointly modeling is to effectively represent and exploit the complementarity and interactions between different modalities [Guo *et al.*, 2019].

Recently, large-scale pre-trained language models (LMs) have shown exceptional proficiency in understanding sequential data [Chang *et al.*, 2023; Gruver *et al.*, 2024], sparking interest in leveraging them to integrate time series and text modalities. Several contrastive learning-based works in medical domain utilize language models as textual encoders to extract meaningful representations from clinical knowledge, guiding the pre-training of time series encoder [Liu *et al.*, ; Yu *et al.*, 2024; King *et al.*, 2023]. However, the clinical knowledge is not present during the inference stage, leading to a potential loss of critical information. For instance, METS [Li *et al.*, 2024] utilizes a frozen LM to derive embeddings from clinical reports, aligning them with ECG embedding through contrastive learning to augment ECG representation. And only ECG encoder provides decision for inference. Other prompt-based works integrate the text modality into a prompt to guide LM to comprehend time series input [Jia *et al.*, 2024; Liu *et al.*, 2024]. Specifically, MedTsLLM [Chan *et al.*, 2024] augments the text prompt with patient-specific clinical information to adapt LLM for medical signal processing.

In these LM-based multimodal time series-text works, time series is typically considered as the primary modality, being more relevant for decision-making, while text serves as
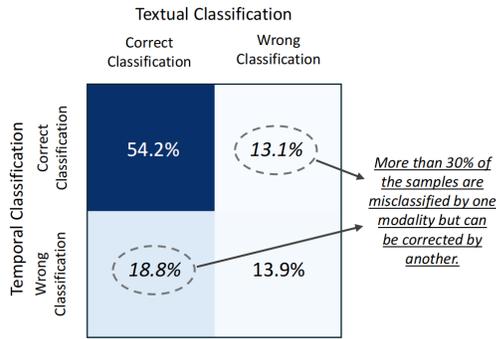
Figure 1: Unimodal classification results on the ECG multimodal dataset (PTB-XL 5 Classes), using LSTM for time series classification and BERT for text classification. The circled samples are misclassified by one modality but corrected by another, demonstrating the complementary information of different modalities.

an auxiliary modality to enhance the time series embedding, either by projecting textual knowledge into the time series encoder using contrastive learning [Liu *et al.*,] or by guiding LM with a textual prompt to comprehend temporal inputs [Chan *et al.*, 2024]. We classify these approaches as temporal-primary multimodal models (Figure 2 (a)). However, in some cases, the textual information is no less important than temporal information. As shown in Figure 1, we conduct a unimodal classification experiment on the PTB-XL ECG dataset and find that 18.8% of samples are correctly classified by the text modality but misclassified by the time series modality, while 13.1% shows the reverse. This highlights the complementarity of these two modalities and suggests that the text modality contains even more unique task relevant information in some cases. Therefore, viewing text modality as auxiliary may introduce bias and fail to capture essential information while a text-primary perspective could enable a more comprehensive understanding of the informative textual content.

To fully exploit the complementarity and mutual benefits of different modalities, we propose a novel textual-temporal multimodal learning paradigm to integrate both temporal-primary and textual-primary perspectives (Figure 2 (a)). However, to effectively construct LM-based approach of such paradigm is technically non-trivial. The most straightforward solution is to train two LM-based submodels separately. Nevertheless, there remain two-fold challenges: First, considering LMs involved, two separately trained submodels suffer non-negligible computational costs. Second, the integration and design of submodels should fuse the two modalities from different perspectives to capture both shared and unique information effectively. Note that the naive multimodal concatenation at LM input layer of existing works [Chan *et al.*, 2024] is difficult to extract the intertwined and high-level multimodal semantics.

To address the aforementioned challenges, we propose MedualTime, a language model for medical time series-text multimodal learning, consisting of a temporal-primary multimodal adapter and a textual-primary multimodal adapter to effectively explore the complementary information in multimodal medical input. Under dual adapter design, each modal-

ity has the chance to serve as the primary modality and get improved by the other modality. Within each adapter, multimodal fusion is achieved by injecting learnable adaptation tokens into the top layers to extract high-level multimodal semantics. Furthermore, both adapters share the same LM backbone to reduce computational resources. Meanwhile, we keep the majority of LM's parameters frozen to make different modalities benefit from its pre-trained knowledge. We update only a small portion of LM's parameters, adapting it to our task while enabling efficient fine-tuning. In addition, by pipeline sharing, the modality alignment of different adapters could be accomplished. Our main contributions are as below:

- We are the first to propose a textual-temporal multimodal learning paradigm that treats both time series and text modality equally in medical scenarios. This paradigm fully leverages the rich complementary semantics and captures the intricate interaction between different modalities.

- We propose **MedualTime**, a dual-adapter language model for medical time series-text multimodal learning. Each adapter performs the mutual integration of time series and text modalities by introducing learnable tokens into the top layers of the LM backbone, facilitating high-level multimodal semantic fusion. The frozen shared LM pipeline allows both adapters to leverage the pre-trained knowledge and achieve alignment in the same semantic space, largely reducing the computational resources with adapter tuning.

- **MedualTime** demonstrates superior performance on public medical multimodal datasets, showing its strong generalization and transferability in medical diagnosis tasks. Notably, it achieves an average improvement of **8%** in accuracy and **12%** in F1 score under supervised learning. The code is in https://github.com/start2020/MedualTime.

2 Related Work

In this section, we discuss LM based approaches for time series-text multimodal learning in both medical and general domains. We categorize them into two branches based on how they derive multimodal representation.

Contrastive Learning based Approaches adopt contrastive learning to align time series and text modalities into separate but coordinated spaces to enforce shared information between modalities [Liang *et al.*, 2024]. This branch includes METS [Li *et al.*, 2024], MERL [Liu *et al.*,], ESI [Yu *et al.*, 2024] and [King *et al.*, 2023]. They leverage LMs to obtain embedding representations of the text modality, which then guide the pre-training of time series encoder, enhancing the quality and robustness of time series representation. For instance, MERL employs contrastive learning to enhance ECG signal representations under the supervision of clinical reports. However, during training, contrastive learning tends to focus on capturing shared semantics between modalities, often overlooking modality-specific information. Furthermore, in the inference stage, only the time series modality is available, while the text modality is absent. As a result, this framework relies solely on time series data for decision-making,
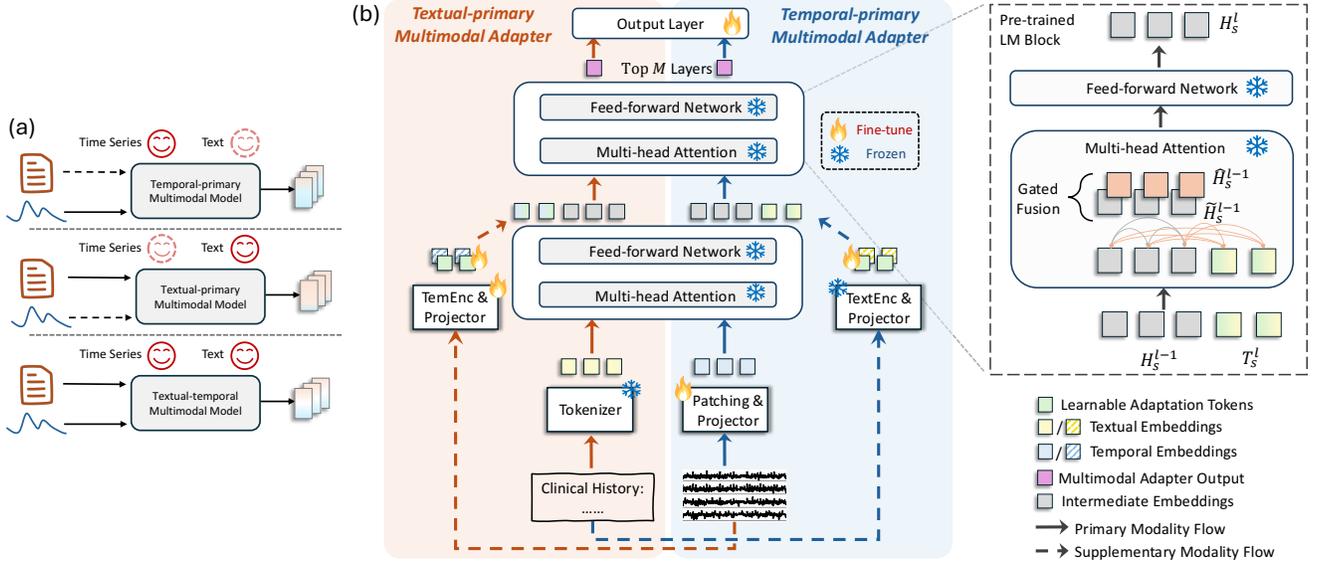
Figure 2: (a) Comparisons between different time series multimodal modeling paradigms. The solid line represents the primary modality and the dashed line for the secondary modality. (b) MedualTime architecture. It consists of dual adapters to model time series and text as primary modality respectively. Dual adapters share the same LM parameters to reduce computational cost and realize adapter alignment. The LM's pre-trained knowledge is preserved by adopting a zero-initialized gating strategy. The cross-modal fusion is achieved by injecting trainable adaptation tokens in the top layers of LM within each adapter.

potentially disregarding unique and task-relevant information embedded in the text. This limitation can lead to suboptimal model performance.

Prompt-based Approaches incorporate text modality into a textual prompt, which is processed by a language model to generate embeddings. Such embedding is then prepended as a prefix of temporal embedding to enhance LLM's reasoning capacities on time series modality. This branch includes Time-LLM [Jin *et al.*, 2023], UniTime [Liu *et al.*, 2024], GPT4MTS [Jia *et al.*, 2024], InstructTime [Cheng *et al.*, 2024], MedTsLLM [Chan *et al.*, 2024]. For instance, Time-LLM assembles dataset description, task instruction, and data statistics into a textual prompt. The prompt is turned to textual embedding by a frozen LLM, concatenated before the temporal embedding as a prefix to facilitate LLM's understanding of time series data. However, such sequential concatenation implies that the concatenated modalities are not equally important, making LLM focus more on time series.

All these LM-based multimodal works consider time series as the primary modality for decision-making, with text serving as an auxiliary modality to enhance time series modeling. In contrast, MedualTime allows each modality to act as a primary modality, which can comprehensively capture the unique and shared semantics in different modalities.

3 Methodology

In this work, we focus on sample-level medical time series-text data. Specifically, each sample is a time-text pair (e.g., ECG signal and its coupled clinical report). The whole dataset is denoted as $\mathcal{S} = \{(\boldsymbol{X_1}, \boldsymbol{S_1}), (\boldsymbol{X_2}, \boldsymbol{S_2}), ..., (\boldsymbol{X_N}, \boldsymbol{S_N})\}$, where $\boldsymbol{X_i} \in \mathbb{R}^{T \times d}$ denotes a d-dimension multivariate time series modality with

length T and $\boldsymbol{S_i}$ denotes the paired textual modality. For simplicity, we omit the sample indicator subscript in the following. To fully utilize the complementary information of different modalities, MedualTime consists of two multimodal adapters, namely a textual-primary multimodal adapter and a temporal-primary multimodal adapter. Each adapter treats one modality as the primary modality and enhances it with the other modality. Both adapters share the same frozen pre-trained language model with L layers. Each adapter implements multimodal fusion in the topmost M ($M \leq L$) transformer blocks of the language model. The shared language model backbone facilitates efficient fine-tuning and encourages the dual adapters' embedding space alignment.

3.1 Textual-primary Multimodal Adapter

Processed by the textual tokenizer, the text input can be modeled by I^s-length word tokens with embedding $\boldsymbol{E}_s \in \mathbb{R}^{I^s \times D}$, where D is the hidden dimension. For the first $L - M$ transformer layers, which are standard transformer layers, the forward process of layer-l is:

$$\tilde{\boldsymbol{H}}_s^{l-1} = \text{LN}\left(\text{MHA}\left(\boldsymbol{W}_q^l \boldsymbol{H}_s^{l-1}, \boldsymbol{W}_k^l \boldsymbol{H}_s^{l-1}, \boldsymbol{W}_v^l \boldsymbol{H}_s^{l-1}\right)\right) + \boldsymbol{H}_s^{l-1} \tag{1}$$

$$\boldsymbol{H}_s^l = \text{LN}\left(\text{MLP}\left(\tilde{\boldsymbol{H}}_s^{l-1}\right)\right) + \tilde{\boldsymbol{H}}_s^{l-1} \tag{2}$$

where $\boldsymbol{H}_s^l$ is the output of layer-l with $\boldsymbol{H}_s^0 = \boldsymbol{E}_s$, MHA, LN, MLP denote the multi-head attention, the layer normalization, and the multi-layer perception, respectively. To obtain the query, key, value matrics at layer-l, $\boldsymbol{W}_q^l, \boldsymbol{W}_k^l, \boldsymbol{W}_v^l$ are parameterized by the pre-trained language model. Meanwhile, the attention operation Attention is de-

fined by:

$$\text{Attention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^T / \sqrt{d_k}\right)\boldsymbol{V} \quad (3)$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are corresponding query, key, and value matrices, d_k is the dimension of key.

Furthermore, inspired by the adapter architecture in [Zhang *et al.*, 2023a], we utilize a lightweight adapter mechanism to achieve multimodal modeling at the topmost M transformer blocks. Specifically, we adopt learnable length-P adaptation tokens $\boldsymbol{T}_s^l$ at each multimodal fusion layer l $(L-M+1 \le l \le L)$, where the adaptation tokens $\boldsymbol{T}_s^l \in \mathbb{R}^{P \times D}$ have the same dimension as language model. As to the secondary temporal modality, a trainable temporal encoder and a cross-modal projector are utilized to transform the time series input into the language model embedding space:

$$\boldsymbol{Z}_s = \text{Projector}\left(\text{TemEncoder}\left(\boldsymbol{X}\right)\right) \quad (4)$$

The temporal encoder can be any time-series encoder that best fits the specific datasets, while the projector is a linear layer responsible for dimension transformation. For decreasing the computational cost, different multimodal fusion layers share the same temporal embedding. Thus, the adaptation tokens of textual-primary multimodal adapter is calculated by:

$$\tilde{\boldsymbol{T}}_s^l = \boldsymbol{T}_s^l + \boldsymbol{Z}_s \quad (5)$$

For the topmost M transformer layers, the multimodal forward process is formalized as:

$$\tilde{\boldsymbol{H}}_s^{l-1} = \text{LN}\left(\text{MHA}\left(\boldsymbol{W}_q^l \boldsymbol{H}_s^{l-1}, \boldsymbol{W}_k^l \boldsymbol{H}_s^{l-1}, \boldsymbol{W}_v^l \boldsymbol{H}_s^{l-1}\right)\right) + \boldsymbol{H}_s^{l-1} \quad (6)$$

$$\hat{\boldsymbol{H}}_s^{l-1} = \text{LN}\left(\text{MHA}\left(\boldsymbol{W}_q^l \boldsymbol{H}_s^{l-1}, \boldsymbol{W}_k^l \tilde{\boldsymbol{T}}_s^l, \boldsymbol{W}_v^l \tilde{\boldsymbol{T}}_s^l\right)\right) + \boldsymbol{H}_s^{l-1} \quad (7)$$

$$\boldsymbol{H}_s^l = \text{LN}\left(\text{MLP}\left(Gate^l \hat{\boldsymbol{H}}_s^{l-1} + \tilde{\boldsymbol{H}}_s^{l-1}\right)\right) + \left(Gate^l \hat{\boldsymbol{H}}_s^{l-1} + \tilde{\boldsymbol{H}}_s^{l-1}\right) \quad (8)$$

In particular, combined with the pre-trained projection matrices $\boldsymbol{W}_k^l, \boldsymbol{W}_v^l$, the learnable adaptation tokens serve as key, value matrices of the multi-head attention layer. In Equation (8), we perform a zero-initialized gating strategy to achieve multimodal adaptation token fusion [Zhang *et al.*, 2023a]. Gating parameter $Gate^l$ will be initialized as zero at the beginning of training, the multimodal adaptation tokens will be injected gradually, which can preserve the pre-trained knowledge and capacities of LMs.

3.2 Temporal-primary Multimodal Adapter

Considering the sequential property of time series, the temporal-primary multimodal adapter takes the time series data as the language model input. We utilize the common patching strategy for time series modeling in related works [Nie *et al.*, 2022; Zhou *et al.*, 2024]. Several adjacent timestamps will be assembled as a token, which can provide local semantic information within a time series. For a pre-defined patch size p and stride s, the time series input $\boldsymbol{X} \in \mathbb{R}^{T \times d}$ can be reorganized as $\tilde{\boldsymbol{X}} \in \mathbb{R}^{T_s \times (p \times d)}$, where $T_s = \left\lceil \frac{T-p}{s} \right\rceil + 1$ is the number of temporal tokens. Subsequently, we utilize a projector (i.e. linear layer) to adjust the dimension of temporal tokens. The adjusted temporal token can be denoted as

$\boldsymbol{E}_t$ $(\boldsymbol{E}_t \in \mathbb{R}^{T_s \times D})$. With $\boldsymbol{H}_t^0 = \boldsymbol{E}_t$ as the input of the first transformer layer, the model forward process will be similar to the ones introduced in Section 3.1, e.g., Equation (1-2) and Equation (5 - 8).

For the secondary text input, we use a frozen pre-trained language model (LM) as a text encoder (similar to the temporal encoder in Equation (4)) to extract textual information:

$$\boldsymbol{Z}_t = \text{Projector}\left(\text{LM}\left(\boldsymbol{S}\right)\right) \quad (9)$$

3.3 Pre-trained Language Model Parameters Sharing

Aided by our dual-adapter model design, most of the parameters in the pre-trained language model (e.g., the attention weight matrices $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$, and the MLP layer of each transformer block) can be shared between the textual-primary and temporal-primary multimodal adapters. On the one hand, freezing these parameters preserves the language model's knowledge and sequential modeling capabilities. On the other hand, since most of the parameters in our proposed adapters are shared, the increase in training parameters is minimal compared to using a single adapter. This ensures complementary modeling between the two modalities while maintaining efficient fine-tuning. Additionally, by sharing the same LM pipeline, the embedding spaces of the different adapters align easily, further facilitating the integration of dual adapters.

3.4 Training Loss

Supervised Learning. For supervised classification, we add the last transformer layer output of each adapter together to obtain the final multimodal representation. Then, an extra linear classifier and the cross-entropy loss are used for supervised training.

Unsupervised Representation Learning. For unsupervised representation learning, following previous works [Zhang *et al.*, 2023b], we adopt the contrastive learning paradigm. In particular, for data augmentation, we add random Gaussian noise to the original input. The noise-corrupted sample and its original sample are a positive pair within each adapter. We denote $\boldsymbol{H}_s^{\prime L}$ as the augmentation of $\boldsymbol{H}_s^L$, and $\boldsymbol{H}_t^{\prime L}$ as the augmentation of $\boldsymbol{H}_t^L$. The contrastive loss could be divided into two parts: within-adapter contrastive loss and cross-adapter contrastive loss.

Formally, by maximizing the agreement between positive pairs and minimizing the similarity between negative pairs (i.e., different input instances), in a mini-batch with size B, the within-adapter contrastive losses are

$$\mathcal{L}_s = -\sum_{i=1}^B \log \frac{\exp\left(\text{sim}\left(\boldsymbol{H}_{s,i}^L, \boldsymbol{H}_{s,i}^{\prime L}\right)/\tau\right)}{\sum_{k=1}^B \mathbb{1}_{[k \ne i]} \exp\left(\text{sim}\left(\boldsymbol{H}_{s,i}^L, \boldsymbol{H}_{s,k}^L\right)/\tau\right)} \quad (10)$$

$$\mathcal{L}_t = -\sum_{i=1}^B \log \frac{\exp\left(\text{sim}\left(\boldsymbol{H}_{t,i}^L, \boldsymbol{H}_{t,i}^{\prime L}\right)/\tau\right)}{\sum_{k=1}^B \mathbb{1}_{[k \ne i]} \exp\left(\text{sim}\left(\boldsymbol{H}_{t,i}^L, \boldsymbol{H}_{t,k}^L\right)/\tau\right)} \quad (11)$$

where $\mathbb{1}_{[k \ne i]}$ is the indicator function and τ is the temperature parameter, $\text{sim}(\cdot, \cdot)$ is the dot product between two ℓ_2-normalized vectors.

Modality	Model	PTB-XL								TUSZ								Average	
		4 Classes				5 Classes				2 Classes				4 Classes					
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	F1
Time	LSTM	0.68	0.60	0.48	0.48	0.67	0.63	0.50	0.52	0.76	0.53	0.54	0.54	0.58	0.44	0.27	0.26	0.67	0.45
	TimesNet	0.68	0.46	0.46	0.45	0.67	0.59	0.48	0.50	0.74	0.59	0.63	0.59	0.76	0.75	0.72	0.71	0.71	0.56
	LightTS	0.68	0.59	0.53	0.54	0.59	0.46	0.44	0.45	0.74	0.53	0.53	0.54	0.71	0.72	0.58	0.58	0.68	0.53
	Dlinear	0.68	0.58	0.50	0.49	0.61	0.46	0.41	0.41	0.78	0.52	0.52	0.52	0.71	0.62	0.60	0.59	0.70	0.50
	Pyraformer	0.76	0.66	0.59	0.58	0.66	0.56	0.49	0.51	0.84	0.47	0.50	0.47	0.75	0.77	0.67	0.72	0.75	0.57
	ETSformer	0.72	0.63	0.57	0.55	0.54	0.45	0.38	0.40	0.79	0.53	0.53	0.53	0.73	0.70	0.66	0.66	0.70	0.54
	Autoformer	0.72	0.56	0.56	0.54	0.62	0.47	0.44	0.44	0.79	0.52	0.51	0.51	0.70	0.64	0.64	0.61	0.71	0.53
	Crossformer	0.66	0.58	0.51	0.53	0.65	0.55	0.48	0.50	0.79	0.50	0.51	0.50	0.72	0.71	0.58	0.58	0.71	0.53
	FEDformer	0.67	0.57	0.50	0.51	0.65	0.53	0.47	0.49	0.76	0.57	0.58	0.57	0.68	0.48	0.54	0.48	0.69	0.51
	Informer	0.67	0.59	0.51	0.52	0.67	0.59	0.51	0.52	0.82	0.57	0.55	0.55	0.77	0.74	0.69	0.71	0.73	0.58
	Reformer	0.69	0.56	0.53	0.54	0.65	0.53	0.48	0.49	0.84	0.52	0.50	0.48	0.74	0.75	0.61	0.66	0.73	0.54
	iTransformer	0.56	0.42	0.36	0.37	0.54	0.39	0.31	0.29	0.80	0.50	0.50	0.49	0.73	0.75	0.59	0.61	0.66	0.44
	PatchTST	0.78	0.76	0.62	0.62	0.74	0.69	0.59	0.62	0.73	0.54	0.55	0.54	0.70	0.65	0.59	0.57	0.74	0.59
Time	GPT4TS	0.71	0.58	0.52	0.53	0.59	0.46	0.45	0.45	0.78	0.48	0.48	0.48	0.71	0.73	0.60	0.64	0.70	0.53
Text	GPT2	0.72	0.65	0.56	0.58	0.73	0.65	0.61	0.62	0.72	0.49	0.49	0.50	0.64	0.69	0.53	0.58	0.70	0.57
	BERT	0.70	0.64	0.51	0.53	0.73	0.65	0.59	0.62	0.72	0.49	0.49	0.49	0.59	0.45	0.39	0.40	0.69	0.51
	Llama 3	0.73	0.60	0.60	0.60	0.74	0.69	0.56	0.65	0.72	0.53	0.53	0.55	0.66	0.62	0.47	0.47	0.71	0.57
	ClinicalBERT	0.73	0.57	0.54	0.53	0.74	0.63	0.58	0.66	0.72	0.55	0.63	0.56	0.67	0.36	0.64	0.43	0.72	0.55
Time + Text	TimeLLM	0.69	0.60	0.48	0.47	0.67	0.59	0.46	0.48	0.75	0.51	0.51	0.51	0.69	0.70	0.50	0.47	0.70	0.48
	UniTime	0.67	0.33	0.42	0.37	0.64	0.54	0.43	0.44	0.79	0.54	0.53	0.53	0.77	0.78	0.71	0.71	0.72	0.51
	GPT4MTS	0.72	0.59	0.60	0.59	0.65	0.48	0.50	0.48	0.82	0.64	0.63	0.63	0.70	0.72	0.60	0.53	0.72	0.56
	MedTsLLM	0.74	0.64	0.63	0.64	0.68	0.64	0.60	0.56	0.81	0.66	0.64	0.64	0.72	0.74	0.59	0.60	0.74	0.61
	MedualTime (Time)	0.73	0.65	0.66	0.62	0.71	0.62	0.64	0.59	0.83	0.68	0.63	0.60	0.74	0.75	0.61	0.62	0.75	0.61
	MedualTime (Text)	0.79	0.73	0.71	0.70	0.75	0.66	0.67	0.66	0.82	0.67	0.64	0.63	0.75	0.76	0.68	0.65	0.78	0.66
	MedualTime	**0.83**	**0.78**	**0.75**	**0.76**	**0.81**	**0.75**	**0.75**	**0.75**	**0.87**	**0.75**	**0.65**	**0.68**	**0.79**	**0.82**	**0.74**	**0.75**	**0.83**	**0.73**

Table 1: **Supervised Learning**. MedualTime achieves an average improvement of **8%** in Acc. and **12%** in F1 compared to baselines. The best results are in **bold** while the second and third best are in underlined. "Acc.", "Pre.", and "Rec." represent accuracy, precision and recall respectively. All LM-based models are highlighted in grey.

The cross-adapter contrastive learning assumes that the embeddings from two adapters for one temporal-textual input pair should be similar. Concurrently, embeddings from different instances are considered negative pairs. In this vein, the cross-adapter contrastive loss is given by:

$$\mathcal{L}_{cross} = -\sum_{i=1}^{B} \left(\log \frac{\exp\left(\text{sim}\left(\boldsymbol{H}_{s,i}^{L}, \boldsymbol{H}_{t,i}^{L}\right)/\tau\right)}{\sum_{k=1}^{B} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}\left(\boldsymbol{H}_{s,i}^{L}, \boldsymbol{H}_{t,k}^{L}\right)/\tau\right)} \right.$$
$$\left. + \log \frac{\exp\left(\text{sim}\left(\boldsymbol{H}_{t,i}^{L}, \boldsymbol{H}_{s,i}^{L}\right)/\tau\right)}{\sum_{k=1}^{B} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}\left(\boldsymbol{H}_{t,i}^{L}, \boldsymbol{H}_{s,k}^{L}\right)/\tau\right)} \right) \quad (12)$$

The overall loss function of unsupervised representation learning is given by:

$$\mathcal{L}_{unsup} = \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_{cross} \quad (13)$$

Note that for the variants of MedualTime, namely MedualTime (Time) and MedualTime (Text), we only adopt the within-adapter contrastive loss for training.

4 Experiments

4.1 Experimental Setup

Datasets Since our approach focuses on medical tasks involving coupled time-series and text modalities, where each modality provides valuable information for decision-making, we conduct experiments using publicly available datasets that meet the necessary criteria above: (1) PTB-XL [1] [Wagner *et al.*, 2020]: This electrocardiogram (ECG) corpus consists of 12-lead ECG signals, which capture the electrical activity of the heart, along with clinical reports describing signal characteristics without diagnostic labels. PTB-XL provides two label sets: a coarse-grained label set for disease detection (four classes) and a fine-grained label set for specific disease classification (five classes). (2) TUSZ v1.5.2 [2] [Shah *et al.*, 2018]: The Temple University Seizure Corpus (TUSZ) is a large-scale corpus of EEG (electroencephalogram) signals that record the electrical activity of the brain. It includes 19-channel EEG recordings and the clinical history for each patient session. Similar to PTB-XL, TUSZ offers two label sets: a coarse-grained label set for distinguishing seizure and non-seizure EEG signals (two classes), and a fine-grained label set for seizure type classification (four classes). More details about the datasets, including the label sets, data splits, and preprocessing steps, are provided in **Appendix 1.1**.

Baselines Representative baselines are selected to ensure sufficient experiments. (1) **Unimodal LM-free methods**: MLP-based models (LightTS [Zhang *et al.*, 2022], DLinear [Zeng *et al.*, 2023]); RNN-based models (LSTM [Hochreiter and Schmidhuber, 1997]); CNN-based models (TimesNet [Wu *et al.*, 2022], TS2Vec [Yue *et al.*, 2022], TS-CoT [Zhang *et al.*, 2023b]); Transformer-based models (Pyraformer [Liu *et al.*, 2021], ETSformer [Woo *et al.*, 2022], Autoformer [Wu *et al.*, 2021], Crossformer [Zhang and Yan, 2022], FED-former [Zhou *et al.*, 2022], Informer [Zhou *et al.*, 2021], Reformer [Kitaev *et al.*, 2020], iTransformer [Liu *et al.*, 2023], PatchTST [Nie *et al.*, 2022], TS-TCC [Eldele *et al.*, 2021]). (2) **Unimodal LM-based methods**: BERT [Devlin *et al.*, 2018], GPT-2 [Radford *et al.*, 2019], Llama 3, Clini-

---

[1] https://physionet.org/content/ptb-xl/1.0.3/

[2] https://isip.piconepress.com/projects/nedc/html/tuh_eeg/

calBERT [Wang *et al.*, 2023], GPT4TS [Zhou *et al.*, 2024]. (3) **Multimodal LM-based methods**: TimeLLM [Jin *et al.*, 2023], UniTime [Liu *et al.*, 2024], GPT4MTS [Jia *et al.*, 2024], MedTsLLM [Chan *et al.*, 2024] for supervised learning; METS [Li *et al.*, 2024], MERL [Liu *et al.*,] for unsupervised learning. (4) **MedualTime variants**: *MedualTime (Time)* for temporal-primary multimodal adapter, *MedualTime (Text)* for textual-primary multimodal adapter. Note that for GPT-2, BERT, Llama 3, and ClinicalBERT, we use textual embeddings generated by them and then train a linear classifier from scratch for the downstream task. More introductions about baselines are in **Appendix 1.4**.

Implementations Inspired by [Liu *et al.*, 2024; Jia *et al.*, 2024; Zhou *et al.*, 2024], which have validated GPT-2's capability in handling time-series data, MedualTime adopts a frozen GPT-2 as the backbone. In the textual-primary multimodal adapter, the tokenizer is derived from GPT-2. To avoid heavy computational costs, we use a lightweight CNN-based model as the temporal encoder, consisting of three convolutional blocks, each with three CNN layers. This encoder is trained from scratch to adapt to the specific tasks. In the temporal-primary multimodal adapter, following METS [Li *et al.*, 2024], a frozen ClinicalBERT [Wang *et al.*, 2023] pre-trained on medical data as the textual encoder. All hidden dimensions are set to 768 to match the backbone (i.e., GPT-2). The time series patch size and stride are both set to 25. Adam is adopted as the optimizer [Kingma, 2014]. All experiments are implemented by PyTorch Framework with NVIDIA A6000 (48G) GPU.

4.2 Supervised Learning

A linear classifier is added as the output layer of MedualTime under supervised learning. As shown in Table 1, **(1)** Time-only models perform better than text-only models, achieving second best in most experiments. **(2)** Among multimodal approaches based on LMs, UniTime and GPT4MTS exhibit similar performance, outperforming TimeLLM by a 2% accuracy improvement. This performance gap may be due to the differences in their fine-tuning strategies. While TimeLLM relies on a frozen LLM, UniTime and GPT4MTS employ parameter-efficient fine-tuning techniques. MedTsLLM outperforms GPT4MTS, perhaps due to its tailored design for medical data. **(3)** MedualTime significantly outperforms these LM-based multimodal methods by a 9% accuracy improvement. This discrepancy likely arises from their temporal-primary paradigm, which overlooks critical information embedded in the text modality. In contrast, MedualTime utilizes a textual-temporal perspective, allowing for a more comprehensive understanding of the multimodal interactions. **(4)** MedualTime achieves the best performance, improving accuracy by **8%** and F1 by **12%** on average.

4.3 Few-shot Learning for Label Transfer

To evaluate the transferability of representations in a supervised setting, we propose a **Few-shot Label Transfer** setting for in-dataset transfer between coarse-grained and fine-grained label sets (Figure 3). It is common that in real-world applications coarse-grained labels, such as disease presence, are easier to obtain, while fine-grained labels, like specific
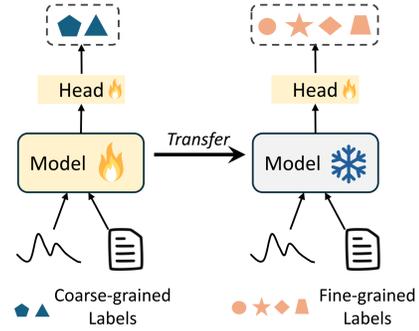

Figure 3: Illustration for **Few-shot Label Transfer**.
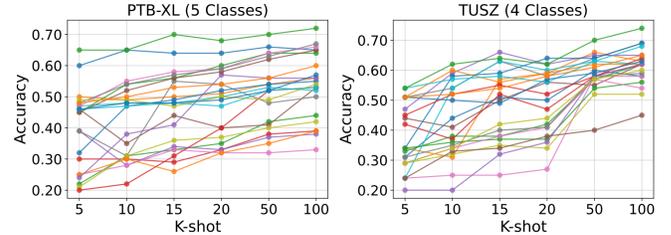

Figure 4: Few-shot experiment results on accuracy. The green line is MedualTime.

disease types, are expensive to obtain. In this setting, we pre-train the model on coarse-grained labels and fine-tune it on limited fine-grained labels by freezing the pre-trained parameters and training an additional classifier for few-shot learning. Specifically, we train the model on PTB-XL with four classes and fine-tune on five classes, and train on TUSZ with two classes and fine-tune on four classes. We conduct {5, 10, 15, 20, 50, 100}-shot experiments. Figure 4 shows the results on accuracy and more results are in **Appendix 1.3**. Generally speaking, all models' accuracy generally shows a continuous growth trend as the few-shot (K) increases. MedualTime consistently outperforms baselines on both datasets and achieves strong classification performance even with limited training samples, demonstrating its effectiveness and robust transferability powered by language models and multimodal inputs.

Modality	Model	PTB-XL				TUSZ				Average	
		4 Classes		5 Classes		2 Classes		4 Classes			
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Time	TSTCC	0.68	0.54	0.65	0.50	<u>0.74</u>	0.48	0.67	0.45	0.69	0.49
	TS2Vec	0.61	0.43	0.61	0.49	0.70	0.49	<u>0.70</u>	0.53	0.66	0.48
	TSCoT	0.73	<u>0.60</u>	<u>0.75</u>	0.63	0.67	0.53	0.69	<u>0.60</u>	0.71	<u>0.59</u>
	PatchTST	0.60	0.35	0.55	0.30	<u>0.73</u>	0.50	0.67	0.45	0.64	0.40
Text	GPT2	0.72	0.58	0.73	0.62	0.72	0.50	0.64	0.39	0.70	0.52
	Bert	0.70	0.53	0.73	0.62	0.72	0.49	0.59	0.40	0.69	0.51
	Llama 3	0.73	<u>0.60</u>	0.74	0.55	0.72	<u>0.63</u>	0.66	0.47	0.71	0.56
	ClinicalBERT	0.73	0.53	0.74	0.55	0.72	<u>0.66</u>	0.67	0.43	<u>0.72</u>	0.54
Time + Text	METS	0.74	0.58	0.71	0.60	0.65	0.53	0.57	0.20	0.67	0.48
	MERL	<u>0.75</u>	0.58	<u>0.75</u>	<u>0.66</u>	0.70	0.57	<u>0.70</u>	0.50	<u>0.73</u>	<u>0.58</u>
	MedualTime (Time)	0.71	0.47	0.65	0.47	0.69	0.52	0.69	0.44	0.69	0.48
	MedualTime (Text)	0.73	0.58	0.71	0.62	0.71	0.52	<u>0.71</u>	<u>0.62</u>	<u>0.72</u>	<u>0.59</u>
	MedualTime	**0.77**	**0.65**	**0.77**	**0.58**	**0.76**	**0.57**	**0.74**	**0.62**	**0.76**	**0.61**

Table 2: **Unsupervised Learning**. 100% of the labeled training data are used to train a linear classifier.

4.4 Unsupervised Learning

We follow the experimental setting in previous works [Zhang *et al.*, 2023b] to assess our model's ability to generate general representations without ground truth supervision. First, we conduct unsupervised experiments. Then, using the unsuper-

vised embeddings obtained, we train a linear classifier with varying proportions (from 10% to 100%) of labeled training data. Table 2 shows the results for the 100% labeled training data proportion.

As shown in Table 2, text-only models generally outperform time-only models, highlighting the importance of text data in the unsupervised setting. While the multimodal model MERL outperforms the best time-only model TSCoT, METS lags behind, suggesting that multimodal does not always outperform single modality. The effectiveness of multimodal fusion is crucial. MedualTime surpasses MERL in most experiments, emphasizing the advantages of our complementary textual-temporal multimodal design. Overall, MedualTime achieves an average accuracy improvement of 3% and F1 improvement of 2%, consistently outperforming other baselines.

4.5 More Explorations

(a) Ablation Study We ablate MedualTime into MedualTime (Time) and MedualTime (Text) under both supervised and unsupervised settings, as shown in Tables 1 and 2. Generally, MedualTime (Text) achieves better performance than MedualTime (Time) in both settings, suggesting that the backbone language model demonstrates a stronger understanding of text compared to time series. Overall, MedualTime consistently outperforms the single-adapter variants, highlighting the complementary nature of textual and temporal information.

Model	PTB-XL				TUSZ				Average	
	4 Classes		5 Classes		2 Classes		4 Classes			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
MedualTime (BERT)	0.83	0.76	0.81	0.74	0.84	0.69	0.79	0.78	0.82	**0.74**
MedualTime (RoBERTa)	0.83	0.76	0.80	0.73	0.87	0.61	0.79	0.74	0.82	0.71
MedualTime (GPT-2)	0.82	0.75	0.80	0.73	0.86	0.52	0.72	0.60	0.80	0.65
MedualTime (ClinicalBERT)	0.83	0.76	0.81	0.75	0.87	0.68	0.79	0.75	**0.83**	0.73

Table 3: Textual Encoders Analysis under supervised learning.

(b) Textual Encoder Analysis We investigate the impact of various textual encoders by examining the following options: BERT, RoBERTa [Liu, 2019], ClinicalBERT [Wang *et al.*, 2023], and GPT-2. The results of supervised learning are in Table 3. More results about unsupervised learning are in **Appendix 1.5**. As shown in Table 3, BERT-based textual encoders consistently outperform GPT-2. This is likely due to GPT-2's primary focus on text generation, while BERT and its variants excel in comprehending the entire textual input. ClinicalBERT, pre-trained on medical texts, achieves the highest accuracy on average, perhaps attributed to its domain-specific pre-training knowledge.
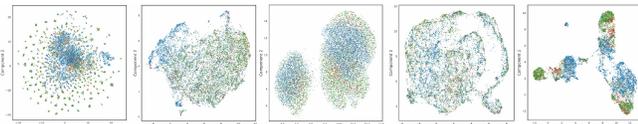

Figure 5: Embedding visualizations of different encoders on PTB-XL (4 Classes), i.e. ClinicalBERT, TS2Vec, MedualTime (Text), MedualTime (Time), MedualTime, from left to right. Different colors represent different labels.

(c) Visualization To better visualize the learned representations of unsupervised learning, we use UMAP [McInnes *et*

al., 2018] to project the unsupervised representation learning results into 2D plots. **(1)** Figure 5 displays the embeddings of various encoders on PTB-XL, with labels assigned to different categories. Both TS2Vec (time-only) and ClinicalBERT (text-only) cannot successfully identify different categories of ECGs. **(2)** Compared with ClinicalBERT, MedualTime (Text) can better distinguish abnormal ECG and normal ECG, indicating the effectiveness of two modalities over one modality. **(3)** Compared with MedualTime (Time), MedualTime (Text) has obviously better discriminative capacity, supporting the advantage of textual-primary modeling over temporal-primary modeling. **(4)** Overall, MedualTime provides the most distinct representations, attributed to the benefits of complementary multimodal modeling.
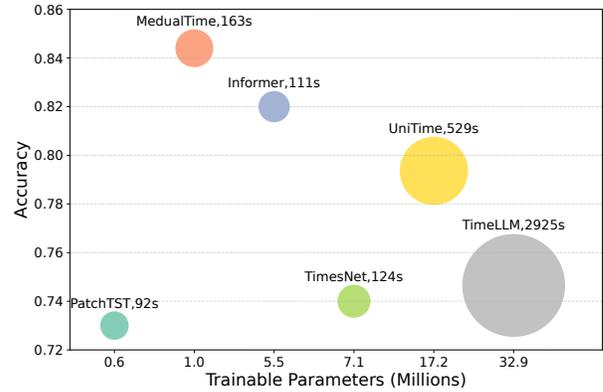

Figure 6: Efficiency comparison on the TUSZ dataset (2 Classes). The bubble size represents the relative training time per epoch.

(d) Efficiency Evaluation Due to limited space, we compare the computational costs in seconds on representative baselines and our model under supervised setting in Figure 6. MedualTime employs a frozen LLM backbone and introduces learnable adaptation tokens, enabling more efficient fine-tuning. It has approximately 1.0 million trainable parameters, even smaller than TimesNet with 7.1 M. Its training efficiency (163s per epoch) is close to TimesNet (124s per epoch) even though it has far larger total parameters. Compared with UniTime and TimeLLM, MedualTime has the shortest training time and highest accuracy, highlighting its efficiency and superior performance.

5 Conclusion

We propose a new textual-temporal paradigm for medical time series-text multimodal learning to explore the complementary modeling of different modalities. Our model leverages a dual-adapter design for temporal-primary and textual-primary modeling, with high-level multimodal fusion achieved through learnable token injection in the top layers of a shared frozen pre-trained LM pipeline, ensuring lightweight fine-tuning. Extensive experiments demonstrate that our model significantly enhances representation learning of medical data in both supervised and unsupervised settings. Few-shot label transfer experiments highlight its transferability. Our model leverages the capabilities of language models to equip clinicians with multimodal tools that enhance medical diagnosis.

Acknowledgments

Contribution Statement

† Co-first Authors, * Corresponding Authors.

References

[Chan *et al.*, 2024] Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. Medtsllm: Leveraging llms for multimodal medical time series analysis. *arXiv preprint arXiv:2408.07773*, 2024.

[Chang *et al.*, 2023] Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. LLM4TS: Two-Stage Fine-Tuning for Time-Series Forecasting with Pre-Trained LLMs. In *arXiv:2308.08469*, 2023.

[Cheng *et al.*, 2024] Mingyue Cheng, Yiheng Chen, Qi Liu, Zhiding Liu, and Yucong Luo. Advancing time series classification with multimodal language modeling. *arXiv preprint arXiv:2403.12371*, 2024.

[Davuluri, 2024] Manaswini Davuluri. An overview of natural language processing in analyzing clinical text data for patient health insights. *Research-gate journal*, 10(10), 2024.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2352–2359, 2021.

[Fries *et al.*, 2021] Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):2017, 2021.

[Gruver *et al.*, 2024] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

[Guo *et al.*, 2019] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Jia *et al.*, 2024] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23343–23351, 2024.

[Jin *et al.*, 2023] 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. Timellm: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[King *et al.*, 2023] Ryan King, Tianbao Yang, and Bobak J Mortazavi. Multimodal pretraining of medical time series and notes. In *Machine Learning for Health (ML4H)*, pages 244–255. PMLR, 2023.

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

[Kitaev *et al.*, 2020] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[Li *et al.*, 2024] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pages 402–415. PMLR, 2024.

[Liang *et al.*, 2024] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[Liu *et al.*,] Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In *Forty-first International Conference on Machine Learning*.

[Liu *et al.*, 2021] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

[Liu *et al.*, 2023] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[Liu *et al.*, 2024] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, 2024.

[Liu, 2019] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[McInnes *et al.*, 2018] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation

and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[Nie *et al.*, 2022] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

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

[Shah *et al.*, 2018] Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.

[Wagner *et al.*, 2020] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[Wang *et al.*, 2023] Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.

[Wang *et al.*, 2024] Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems*, 36, 2024.

[Weimann and Conrad, 2021] Kuba Weimann and Tim OF Conrad. Transfer learning for ecg classification. *Scientific reports*, 11(1):5251, 2021.

[Woo *et al.*, 2022] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv preprint arXiv:2202.01381*, 2022.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[Wu *et al.*, 2022] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The eleventh international conference on learning representations*, 2022.

[Xie *et al.*, 2022] Jin Xie, Jie Zhang, Jiayao Sun, Zheng Ma, Liuni Qin, Guanglin Li, Huihui Zhou, and Yang Zhan. A transformer-based approach combining deep learning network and spatial-temporal information for raw eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2126–2136, 2022.

[Yu *et al.*, 2024] Han Yu, Peikun Guo, and Akane Sano. Ecg semantic integrator (esi): A foundation ecg model

pretrained with llm-enhanced cardiological text. *arXiv preprint arXiv:2405.19366*, 2024.

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

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[Zhang and Yan, 2022] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2022.

[Zhang *et al.*, 2022] Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. *arXiv preprint arXiv:2207.01186*, 2022.

[Zhang *et al.*, 2023a] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[Zhang *et al.*, 2023b] Weiqi Zhang, Jianfeng Zhang, Jia Li, and Fugee Tsung. A co-training approach for noisy time series learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3308–3318, 2023.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.

[Zhou *et al.*, 2024] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36, 2024.