DERI: Cross-Modal ECG Representation Learning with Deep ECG-Report Interaction

Jian Chen[1,2], **Xiaoru Dong**[3], **Wei Wang**[1,4,*], **Shaorui Zhou**[2,*], **Lequan Yu**[3], **Xiping Hu**[1,4,*]
[1]Artificial Intelligence Research Institute, Shenzhen MSU-BIT University
[2]School of Intelligent System Engineering, Sun Yat-sen University
[3]School of Computing and Data Science, The University of Hong Kong
[4]School of Medical Technology, Beijing Institute of Technology
chenj589@mail2.sysu.edu.cn, xrdong,lqyu@cs.hku.hk, ehomewang@ieee.org,
zhoushr5@mail.sysu.edu.cn, huxp@bit.edu.cn

Abstract

Electrocardiogram (ECG) is widely used to diagnose cardiac conditions via deep learning methods. Although existing self-supervised learning (SSL) methods have achieved great performance in learning representation for ECG-based cardiac conditions classification, the clinical semantics can not be effectively captured. To overcome this limitation, we proposed to learn cross-modal ECG representations that contain more clinical semantics via a novel framework with **D**eep **E**CG-**R**eport **I**nteraction (**DERI**). Specifically, we design a novel framework combining multiple alignments and mutual feature reconstructions to learn effective representation of the ECG with the clinical report, which fuses the clinical semantics of the report. An RME-Module inspired by masked modeling is proposed to improve the ECG representation learning. Furthermore, we extend ECG representation learning to report generation with a language model, which is significant for evaluating clinical semantics in the learned representations and even clinical applications. Comprehensive experiments with various settings are conducted on various datasets to show the superior performance of our DERI. Our code is released on https://github.com/cccccj-03/DERI.

1 Introduction

Electrocardiogram (ECG) is a widely used data for reflecting heart electrical activity [Attia *et al.*, 2019], which is of great importance for cardiac conditions classification. Supervised learning methods have obtained effective performance in ECG signal classification with high-quality annotations [Huang *et al.*, 2022; Chen *et al.*, 2024a]. However, there are a large number of unlabeled ECG signals in the real world, and supervised learning methods have difficulty utilizing this resource effectively. To reduce the dependence on labeled data, ECG representation learning methods based on self-supervised learning (SSL) have demonstrated their powerful performance [Oh *et al.*, 2022]. Compared to supervised learning methods, SSL methods aim to learn effective representations from ECG signal data without labels and thus tend to be more generalizable and adaptable to different downstream tasks, showing great potential.

ECG self-supervised learning (SSL) methods are mainly generative or contrastive. Generative models reconstruct inputs via masked modeling, while contrastive methods distinguish between similar and dissimilar samples. However, their single-modal nature limits the capture of rich clinical semantics. Generative methods focus on reconstructing low-level signal patterns, often missing clinical semantics [Zhang *et al.*, 2023], while contrastive methods rely on input-level augmentations, which risk distorting ECG semantics [Na *et al.*, 2024]. Multi-modal learning has emerged as a promising solution for these limitations due to its great learning ability with multiple data sources [Chen *et al.*, 2024b; Chen *et al.*, 2024c]. Compared to ECG signals, clinical reports offer direct high-level semantic insights. Inspired by advances in medical imaging and radiology reports [Liu *et al.*, 2023a], Liu et al. proposed a multi-modal representation learning approach called MERL by aligning ECG signals with clinical reports [Liu *et al.*, 2024b]. However, their method aligns ECG features with report features in the feature space, drawing inspiration from CLIP [Radford *et al.*, 2021], but the interaction between modalities is relatively shallow. Furthermore, although the ECG representations learned by MERL perform well in classification tasks, they fail to effectively convey the underlying semantics of ECG recordings, which are crucial for understanding cardiac conditions.

To overcome these limitations, we proposed a novel **D**eep **E**CG **R**eport **I**nteraction (**DERI**) framework for cross-modal representation learning. To better capture the clinical semantics of ECG signals, we design an encoder-decoder structure to conduct multiple cross-modal alignments and mutual feature reconstruction. Specifically, ECG signals and the corresponding clinical reports are first encoded and projected into a shared alignment space to achieve an initial alignment. To enhance interaction, two specialized decoders are employed to reconstruct features by decoding the aligned representations into the other modality. This reconstruction process captures the latent semantics in both modalities, enabling the learning of richer cross-modal representations. Subsequently, the de-

* Corresponding authors.

coded features are fused with the modality-specific aligned features to create mixed representations incorporating both ECG signals and clinical reports semantics. These mixed representations are further utilized for a second alignment. Additionally, we introduce a **R**andom **M**asked **E**nhancement Module (RME-Module) to improve ECG representation learning. Furthermore, the proposed DERI framework is integrated with language models to generate reports, providing a way to assess the learned clinical semantics. Extensive experiments across various settings and datasets are conducted to demonstrate the effectiveness of DERI. The main contributions are summarized as follows:

- To learn effective ECG representation from reports, we propose a novel cross-modal framework of ECG-Report via multiple feature alignment and mutual feature reconstruction. An RME-Module is also designed for ECG representation learning enhancement.
- To better illustrate the clinical semantics learned by DERI, we combine it with a language model for report generation. The pre-trained model provides effective ECG representation and a language model is used to generate clinical reports for clinical semantics.
- Comprehensive experiments on downstream datasets are conducted to evaluate the proposed DERI method, including zero-shot classification, linear probing, and even report generation. Experimental results illustrate that our DERI method surpasses all SOTA methods.

2 Related Work

Single-modal ECG Representation Learning. There are various SSL methods for ECG representation learning. Most of these methods are single-modal, which conduct generative learning or contrastive learning on unannotated ECG signals. CLOCS [Kiyasseh *et al.*, 2021] and ASTCL [Wang *et al.*, 2023] are the SOTA single-modal contrastive learning methods that explore the spatial-temporal correlation of ECG signals. Similarly, ST-MEM [Na *et al.*, 2024] proposes to learn ECG representation by spatial-temporal masking modeling and reconstruction of 12-lead ECG signals. Although all these unimodal methods have achieved good performance, they still fall short in learning the clinical semantics of ECG signals [Liu *et al.*, 2024b]. Single-modal contrastive and generative methods extract representations only from ECG signals without diagnostic reports.

Multi-modal ECG Representation Learning. Several works conduct ECG multi-modal learning for better classification. Ref. [Raghu *et al.*, 2022] proposes to learn representations from ECG signals and structured data from labs and vitals by contrastive learning. Ref. [Lalam *et al.*, 2023] combines ECG signals with structured Electronic Health Records (EHRs) to conduct contrastive learning. BPNet fuses ECG signals with PPG signals to better conduct blood pressure estimation [Long and Wang, 2023]. However, these methods do not use diagnostic report data, making it difficult to learn the clinical semantics effectively. To learn the clinical semantics of ECG signals, Liu et al. proposed to align ECG features with clinical reports inspired by multi-modal learning in medical images and radiology reports [Liu *et al.*, 2024a; Liu *et al.*, 2024b]. Introducing corresponding diagnostic reports for ECG representation learning greatly improves their performance on the downstream cardiac condition classification tasks, but these multimodal approaches only achieve shallow modal interaction. The learned representations can not efficiently incorporate the semantics in the reports. Therefore, we design our DERI framework to conduct deep cross-modal interaction and then expand the model to report generation with great meaning for clinical diagnosis.

Clinical Report Generation. Clinical report generation in radiology has obtained great performance inspired by imaging captioning [You *et al.*, 2021]. R2Gen [Chen *et al.*, 2020b] uses a memory-driven Transformer to generate a radiology report directly with the representation of the medical image. CvT2DistilGPT2 [Liu *et al.*, 2023b] demonstrates that pre-trained NLP models can provide benefits for radiology report generation as well. X-REM is proposed to fuse the image-text multi-modal representation and then used retrieval-based methods to generate the predicted reports from the retrieval corpus [Jeong *et al.*, 2024]. Inspired by these image-based clinical report generation methods, we extend our ECG representation learning method to ECG-based clinical report generation, which can help understand the clinical semantics of the cardiac conditions from the ECG signals.

3 Methodology

3.1 Overview

Our DERI is a dual-encoder framework for learning effective multimodal representations from ECG signals and clinical reports. As shown in Fig. 1, it leverages feature alignment and reconstruction for cross-modal interaction, with an RME module further enhancing representation quality.

3.2 Multiple ECG-Report Alignment

The Multiple ECG-Report Alignment in DERI contains two strategies: modal-specify and mix-modal feature alignment. Given an ECG signal recording e_i with corresponding clinical report r_i, we construct an ECG-Report pair as (e_i,r_i), with $i = 1, 2, 3, ..., N$ where N is the number of recordings. Two distinct encoders $\mathcal{F}_e$ and $\mathcal{F}_t$ are used to learn the latent encoding of ECG signals and report texts respectively, represented as $z_{e,i}$ and $z_{t,i}$. Specifically, the latent encoding is obtained by $z_{e,i} = \mathcal{F}_e(e_i)$ and $z_{t,i} = \mathcal{F}_t(r_i)$. To align the ECG encoding and text encoding, we use two linear projectors $\mathcal{P}_e$ and $\mathcal{P}_t$ to map them into an alignment space of the same dimension, which can be represented as $\hat{A}_{e,i} = \mathcal{P}_e(z_{e,i})$ and $\hat{A}_{t,i} = \mathcal{P}_t(z_{t,i})$. The align loss $\mathcal{L}_{align}$, which is inspired by the CLIP loss, is used to close the distance between the representations of ECG signals and reports in the alignment space. Specifically, each ECG signal and the corresponding report are regarded as a positive pair and others as negative pairs. The loss function $\mathcal{L}_{align}$ is shown as Eq. 1 to Eq. 3.

$$\mathcal{L}_{align} = \frac{1}{2B} \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathcal{L}_{i,j}^{e,t} + \mathcal{L}_{i,j}^{t,e}), \tag{1}$$

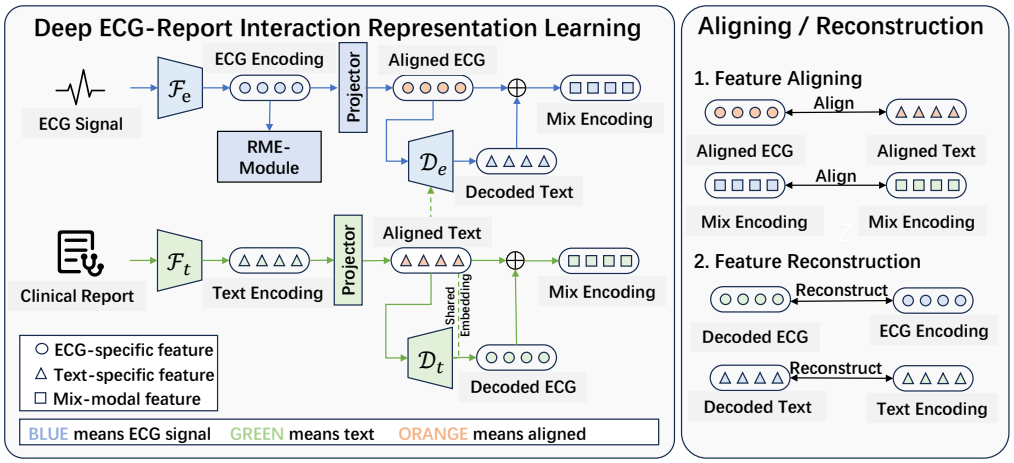

Figure 1: Framework of proposed DERI for ECG-Report multi-modal representation learning.

$$\mathcal{L}_{i,j}^{x,y} = -log\frac{exp(s_{i,j}^{x,y}/\tau)}{\sum_{k=1}^{B}\mathbb{I}_{[k\neq i]}exp(s_{i,j}^{x,y}/\tau)}, x, y \in [e,t], \quad (2)$$

$$s_{i,j}^{x,y} = \frac{\hat{A}_{x,i}^{\top}\hat{A}_{y,j}}{\parallel \hat{A}_{x,i} \parallel \parallel \hat{A}_{y,j} \parallel}, \quad (3)$$

where B is the batch size, $x \in [e,t], y \in [e,t]$ represents the ECG modal or the text modal, τ is the temperature coefficient which is set as 0.7, and $\mathbb{I}$ is the indicator function. Through calculating the mutual align loss of ECG-report $\mathcal{L}_{i,j}^{e,t}$ and report-ECG $\mathcal{L}_{i,j}^{t,e}$ respectively, the model can perform better feature alignment.

Furthermore, to conduct deep ECG-report interaction, we adopt Cross-modal Mutual Reconstruction (depicted in Section 3.3) to decode the representation of another mode (D_e and D_t represent decoded ECG feature and decoded text feature) from the aligned modal representation. Then the decoded features are added to the aligned modal representation to obtain the mix-modal encoding. By combining alignment and reconstruction, mixed-modal can better achieve deep ECG-Report interaction. Specifically, we will use the $\hat{A}_{e,i}$ and $\hat{A}_{t,i}$ as the core to obtain the mix-modal encoding $\hat{M}_i^e = \hat{A}_{e,i} \oplus D_t$ and $\hat{M}_i^t = \hat{A}_{t,i} \oplus D_e$, and then conduct the second encoding alignment in mixed space, thus obtaining the final multimodal representation. The mixed alignment loss can be calculated as Eq. 4.

$$\mathcal{L}_{mixed} = \frac{1}{2B}\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathcal{L}'^{e,t}_{i,j} + \mathcal{L}'^{t,e}_{i,j}), \quad (4)$$

where $\mathcal{L}'^{e,t}_{i,j}$ and $\mathcal{L}'^{t,e}_{i,j}$ is calculated by using $\hat{M}_i^e$ and $\hat{M}_i^t$ to replace $\hat{A}_{e,i}$ and $\hat{A}_{t,i}$ as Eq. 2 and Eq. 3. Therefore, our proposed method can effectively extract mixed modal representations with report characteristics by using only ECG signals after completing the pre-training stage. This can help the model better complete the task of zero-shot classification and report generation.

In conclusion, the whole loss for multiple ECG-report alignment can be written as Eq. 5:

$$\mathcal{L}_{ERA} = \mathcal{L}_{align} + \mathcal{L}_{mixed}. \quad (5)$$

3.3 Cross-modal Mutual Reconstruction

To better guide the model in achieving deeper modal interactions between ECG signals and diagnostic reports, we introduce cross-modal mutual reconstruction. Specifically, after we obtained the aligned ECG feature $\hat{A}_{e,i}$ and the aligned text features $\hat{A}_{t,i}$, we aim to facilitate modal interactions by reconstructing the target modality while bringing them closer to each other in space. We introduce transformers as decoders to decode the representations of one modal in the alignment space to another modal. Considering that reports offer intuitive semantic information valuable for heart state classification but are often unavailable without cardiologists, we introduce a shared embedding derived from the textual modality decoder for better learning textual semantics during pre-training. This shared embedding is combined with the ECG features, enriching them with additional textual features to enhance cardiac condition classification. After completing the pre-training in this manner, the final representation obtained from inputting only the ECG data effectively encapsulates the semantic information of the corresponding report text. This process is represented as Eq. 6:

$$\hat{D}_{e,i} = \mathcal{D}_e(\hat{A}_{t,i}), \quad \hat{D}_{t,i} = \mathcal{D}_t(Concat[\hat{A}_{e,i}, SE_t]), \quad (6)$$

where $\mathcal{D}_e$ and $\mathcal{D}_t$ are the decoder transformers to obtain ECG encoding $\hat{D}_{e,i}$ and report encoding $\hat{D}_{t,i}$ respectively, and SE_t is the shared embedding. Then we use standard contrastive loss on the original feature embeddings and the decoded embeddings for cross-modal mutual reconstruction as Eq. 7:

$$\mathcal{L}_{CMR} = \mathcal{L}_D^e + \mathcal{L}_D^t = \frac{1}{2B}\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathcal{L}_{i,j}^{ze,de} + \mathcal{L}_{i,j}^{zt,dt}), \quad (7)$$

where $\mathcal{L}_D^e$ and $\mathcal{L}_D^t$ represents the loss of ECG and report feature reconstruction respectively, $\mathcal{L}_{i,j}^{ze,de}$ and $\mathcal{L}_{i,j}^{zt,dt}$ represent to use $\mathcal{D}_e$ and $\mathcal{D}_t$ with the original features $z_{e,i}$ and $z_{t,i}$ to calculate the similarity as the same of Eq. 3.

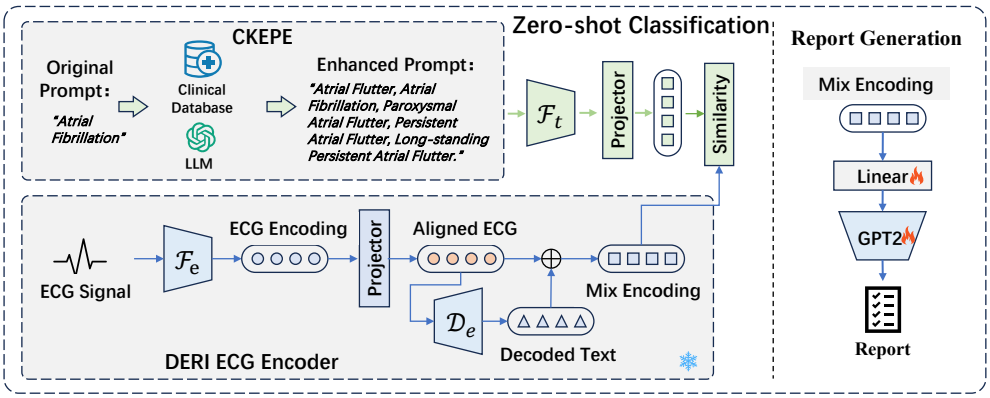

Figure 2: Pipeline of zero-shot classification and report generation of our proposed DERI.

3.4 Latent Random Masking Enhancement

We further conduct an RME-Module on latent ECG encoding to facilitate representation learning. Considering that augmentation directly at the data level entails the loss of semantic information about the signal, we propose an RME-Module conducting on the ECG encoding. Specifically, as the encoder extracts encodings, it tends to focus on the local features of the signal to form an encoding sequence. Rather than applying global average pooling to the ECG encoding sequence and using two separate dropout operations to create augmented views, we instead randomly mask the sequence twice, independently. Then, a multi-head attention mechanism is employed to aggregate the sequence, producing two augmented views of the encoding as a positive pair. This random masking approach helps preserve sequence-level semantic features while enabling the model to learn global features more effectively. We then use standard contrastive loss on these two augmented encoding views. The whole process can be illustrated in Eq. 8:

$$\begin{aligned} \mathcal{L}_{RME} &= -\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{N} log\frac{exp(s_{i,j}\tau)}{\sum_{k=1}^{L}\mathbb{I}_{[k\neq i]}exp(s_{i,j}^{x,y}/\tau)}, \\ z_{e,i}^{1} &= MHA(\text{Mask}(z_{e,i})) = MHA(\mathcal{M}_1 \times z_{e,i}), \\ z_{e,i}^{2} &= MHA(\text{Mask}(z_{e,i})) = MHA(\mathcal{M}_2 \times z_{e,i}), \end{aligned} \tag{8}$$

where $s_{i,i} = z_{e,i}^{1\top} z_{e,i}^{2}$, MHA is multi-head attention, Mask is the random mask strategy, which generates random mask sets $\mathcal{M}_1$ and $\mathcal{M}_2$. $\mathcal{M}_1$ and $\mathcal{M}_2$ with each entry independently sampled with masking ratio $p = 0.1$ are in $\mathbb{R}^{b\times n}$ where b is the batch size and n is the length of the embedding sequence. Each item in $\mathcal{M}$ is either 0 or 1, indicating whether the corresponding patch should be masked. We add a global average on the MHA to obtain the global representation of the masked embedding. Importantly, the random masks are generated by two independent random noises.

In summary, our proposed DERI learns representative ECG features with the help of clinical reports by jointly minimizing $\mathcal{L}_{ERA}$, $\mathcal{L}_{CMR}$ and $\mathcal{L}_{RME}$. The overall training loss of pre-training can shown as Eq. 9:

$$\mathcal{L}_{total} = \mathcal{L}_{ERA} + \mathcal{L}_{CMR} + \mathcal{L}_{RME}. \tag{9}$$

3.5 Downstream Tasks on DERI Framework

After training the proposed DERI model, we can obtain an effective representation of ECG signals that contains clinical report information. Then we can use the representation to conduct zero-shot classification and report generation. Considering the quality of the category prompts for zero-shot classification will have a great impact on the performance [Maniparambil *et al.*, 2023], we adopt the CKEPE prompts which are constructed by combining large language model (LLM) and clinical knowledge [Liu *et al.*, 2024b]. The whole process of these two tasks is illustrated in Fig. 2.

Zero-shot Classification. We adopt CKEPE as the category prompts and use the trained report encoder $\mathcal{F}_t$ and the projector to obtain the prompt embeddings of all categories. We then use the trained DERI model to obtain the Mix Encoding with ECG signals alone, which contains both ECG signal features and the corresponding clinical report features. Finally, we calculate the similarity between the Mix Encoding and the prompt encoding and then conduct an optimal classification threshold search, all categories above this threshold are considered to be predicted. Importantly, all the parameters of the proposed DERI are frozen in this process.

Report Generation. After we obtain the final representation of ECG, we adopt GPT-2 as the text decoder to construct an encoder-decoder structure since DistilGPT2 [Sanh, 2019] has shown its great performance on report generation [Wang *et al.*, 2024]. We adopt a trainable linear layer to transform the input encoding dimension to meet the dimension of the GPT-2 and perform fine-tuning on the GPT-2 to minimize a cross-entropy loss $\mathcal{L}_{CE}$ between the generated report and ground truth reports. After fine-tuning, we can generate corresponding diagnostic reports with ECG signals alone.

4 Experiment

4.1 Datasets

MIMIC-ECG. The pre-training process of our proposed DERI model is conducted on the MIMIC-ECG dataset [Gow *et al.*, 2023] with 800,035 paired ECG-report from 161,352 subjects. We removed all the samples without reports containing more than 3 words and replaced 'NaN' or 'Inf' in the

	PTBXL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
SimCLR	63.41	69.77	73.53	60.84	68.27	73.39	54.98	56.97	62.52	51.41	69.44	77.73	59.78	68.52	76.54	59.02	67.26	73.20
BYOL	71.70	73.83	76.45	57.16	67.44	71.64	48.73	61.63	70.82	41.99	74.40	77.17	60.88	74.42	78.75	54.20	71.92	74.69
BarlowTwins	72.87	75.96	78.41	62.57	70.84	74.34	52.12	60.39	66.14	50.12	73.54	77.62	55.12	72.75	78.39	60.72	71.64	77.43
MoCo-v3	73.19	76.65	78.26	55.88	69.21	76.69	50.32	63.71	71.31	51.38	71.66	74.33	62.13	76.74	75.29	54.61	74.26	77.68
SimSiam	73.15	72.70	75.63	62.52	69.31	76.38	55.16	62.91	71.31	49.30	69.47	75.92	58.35	72.89	75.31	58.25	68.61	77.41
TS-TCC	70.73	75.88	78.91	53.54	66.98	77.87	48.04	61.79	71.18	43.34	69.48	78.23	57.07	73.62	78.72	55.26	68.48	76.79
CLOCS	68.94	73.36	76.31	57.94	72.55	76.24	51.97	57.96	72.65	47.19	71.88	76.31	59.59	77.78	77.49	54.38	71.93	76.13
ASTCL	72.51	77.31	81.02	61.86	68.77	76.51	44.14	60.93	66.99	52.38	71.98	76.05	57.90	77.01	79.51	56.40	70.87	75.79
CRT	69.68	78.24	77.24	61.98	70.82	78.67	46.41	59.49	68.73	47.44	73.52	74.41	58.01	76.43	82.03	56.21	73.70	78.80
STMEM	61.12	66.87	71.36	54.12	57.86	63.59	55.71	59.99	66.07	51.12	65.44	74.85	56.69	63.32	70.39	59.77	66.87	71.36
MERL	82.39	86.27	88.67	64.90	80.56	84.72	58.26	72.43	79.65	53.33	82.88	88.34	70.33	85.32	90.57	66.60	82.74	87.95
Random Init	78.10	84.33	89.47	69.96	78.70	84.01	60.82	67.14	70.08	60.12	79.34	83.98	64.67	75.57	91.45	67.42	74.54	79.97
DERI (Ours)	**85.46**	**89.84**	**90.52**	**73.50**	**80.60**	**85.52**	**62.53**	**72.51**	**84.37**	**65.44**	**83.66**	**92.34**	**79.45**	**89.40**	**93.45**	**77.93**	**87.86**	**91.93**

Table 1: Linear probing results (AUC). We bold the best results and grey represents the second highest.

ECG signal with mean interpolation. Finally, the dataset used for pre-training has 771,693 samples.

PTBXL [Wagner *et al.*, 2020] contains 21,837 12-lead ECG signals from 18,885 patients at a sampling rate of 500 Hz with a duration of 10s. It contains four multi-label classification tasks: Superclass, Subclass, Form, and Rhythm.

CPSC2018 [Liu *et al.*, 2018] contains 6,877 12-lead ECG recordings with a sampling rate of 500 Hz. The duration of these signals ranges from 6 to 60 seconds with one corresponding label within nine categories.

Chapman-Shaoxing-Ningbo (CSN) [Zheng *et al.*, 2022] contains 45,152 12-lead ECG recordings with a sampling rate of 500 Hz. Each recording has a duration of 10 seconds, and signals with "unknown" annotation are removed. 23,026 ECG records with 38 categories are used for classification.

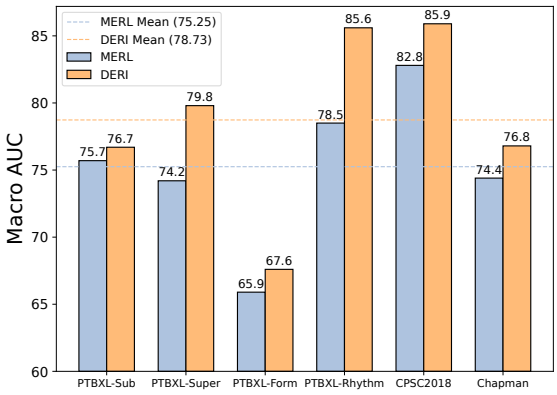

Figure 3: Zero-shot Classification Comparison (AUC).

4.2 Experimental Setup

Pre-training. For the encoders used for ECG signals and reports, we adopt a randomly initialized 1D-ResNet18 and the Med-CPT [Jin *et al.*, 2023], respectively. For decoders, we adopt two transformers with 8 attention heads, a depth of 2, and a hidden size of 256, respectively, for ECG encoding reconstruction and report encoding reconstruction. We use the AdamW optimizer with a learning rate of 1e-3 and a weight decay of 1e-8. The epoch for pre-training is set as 50 with a cosine annealing scheduler to adjust the learning rate. We conduct all the pre-trained experiments on 4 NVIDIA GeForce RTX 4090 GPUs with a batch size of 512.

Classification. We freeze the whole DERI and conduct zero-shot classification as illustrated in Section 3.5. For linear probing, we add a new linear classifier and freeze all other parameters in our DERI. We adopt three different settings, which utilize 1%, 10%, and 100% of the training data. Since these tasks are all classifications that contain many categories, we adopt the macro AUC as the evaluated metric. We conduct these experiments on one NVIDIA GeForce RTX 4090 GPU. The baselines we compared include **SimCLR [Chen *et al.*, 2020a]**, **BYOL [Grill *et al.*, 2020]**, **BarlowTwins [Zbontar *et al.*, 2021]**, **MoCo-v3 [Chen *et al.*, 2021]**, **SimSiam [Chen and He, 2021]**, **TS-TCC [Eldele *et al.*, 2021]**, **CLOCS [Kiyasseh *et al.*, 2021]**, **ASTCL [Wang *et al.*, 2023]**, **CRT [Zhang *et al.*, 2023]**, **STMEM [Na *et al.*, 2024]**, and **MERL [Liu *et al.*, 2024b]**. More details about implementation settings and baselines can be found in the Supplementary.

Report generation. We conduct report generation on the MIMIC-ECG dataset first. We adopt the Natural Language Generation (NLG) metrics which include BLEU-n and ROUGE-L as metrics. We further integrated Clinical Efficiency (CE) metrics inspired by zero-shot classification, as seen in our Supplementary. We conduct the experiments on 2 NVIDIA GeForce RTX 4090 GPUs.

4.3 Experimental Results

Classification. Since most of the existing ECG representation learning methods are proposed without a text encoder for zero-shot learning, we compared our proposed DERI with MERL to verify the cross-modal ECG representations learned from clinical reports. The comparison results are illustrated in Fig. 3. It is evident that our proposed method, DERI, significantly outperforms MERL across all tasks. The dotted line in the figure indicates the average performance of the two zero-sample methods on the six classification tasks. DERI achieves an average macro AUC of 78.73, while MERL attains only 75.25. This underscores DERI's capability to learn clinically relevant representations through deep cross-modal interactions. Compared to MERL which just aligns the ECG and report encoding, our proposed DERI achieves deep cross-modal interaction by multiple alignment and feature reconstruction, which enables the model to learn more effective representation for zero-shot clinical classification.

We evaluate our DERI framework using linear probing, a widely adopted protocol in SSL [Wang *et al.*, 2023], and compare it with existing ECG SSL methods. As shown in Table 1,

Source Domain / Target Domain	Zero-shot	Training Data Ratio	PTBXL-Super → CPSC2018	PTBXL-Super → CSN	CPSC2018 → PTBXL-Super	CPSC2018 → CSN	CSN → PTBXL-Super	CSN → CPSC2018
SimCLR	×	100%	69.62	73.05	56.65	66.36	59.67	62.11
BYOL	×		70.27	74.01	57.32	67.54	60.39	63.24
BarlowTwins	×		68.98	72.85	55.97	65.89	58.76	61.35
MoCo-v3	×		69.41	73.29	56.54	66.12	59.82	62.07
SimSiam	×		70.06	73.92	57.21	67.48	60.23	63.09
TS-TCC	×		71.32	75.16	58.47	68.34	61.55	64.48
CLOCS	×		68.79	72.64	55.86	65.73	58.69	61.27
ASTCL	×		69.23	73.18	56.61	66.27	59.74	62.12
CRT	×		70.15	74.08	57.39	67.62	60.48	63.33
STMEM	×		76.12	**84.50**	62.27	75.19	73.05	64.66
MERL	√	**0%**	88.21	78.01	76.77	76.56	74.15	82.86
DERI (Ours)	√	**0%**	**88.78**	78.83	**79.50**	**81.02**	**76.70**	**85.84**

Table 2: Distribution shift results (AUC). We bold the best results and grey represents the second highest.

DERI outperforms all baselines, including the multi-modal method MERL and state-of-the-art single-modal methods, across all datasets and training data ratios. "Random Init" denotes using DERI in a purely supervised setting without pre-training. Notably, DERI achieves the largest performance gain when trained with only 1% of labeled data, indicating strong generalization with minimal supervision. On the PTBXL-Super task, DERI, with just 1% data even surpasses all single-modal methods trained with 100% data. Moreover, both MERL and DERI exhibit strong performance across settings, confirming the benefit of integrating clinical reports. DERI's consistent advantage over MERL highlights its superior capability in cross-modal ECG representation learning for cardiac condition classification.

To assess the robustness of learned representations under distribution shift, we conduct linear probing with 100% training data on three classification tasks: PTBXL-Super, CPSC2018, and CSN. For zero-shot capable models like MERL and DERI, we reclassify their fixed representations using trained classifiers. Specifically, models are trained on one dataset (source domain) and evaluated on another (target domain). As shown in Table 2, DERI consistently outperforms MERL across all six transfer settings. Compared to single-modal SSL baselines, only STMEM achieves better performance than DERI in the PTBXL-to-CSN setting, while DERI outperforms all others in the remaining cases. MERL generally ranks second, demonstrating the benefit of incorporating clinical reports. The consistent improvements of DERI over MERL further validate the effectiveness of its enhanced cross-modal interaction in improving generalization and robustness across domains.

Report Generation. Beyond cardiac condition classification, DERI enables deep cross-modal interactions that allow the extracted ECG representations to support diagnostic report generation. To evaluate this capability, we conduct report generation experiments on the MIMIC-ECG dataset. We use pre-trained DERI and MERL as encoders and integrate them with DistilGPT2 in an encoder-decoder framework. To isolate the contribution of our cross-modal reconstruction strategy, we introduce a variant of DERI-align, which uses only the aligned ECG representations. As shown in Table 3, DERI outperforms MERL in both NLG and clinical efficacy (CE) metrics, indicating that it captures richer clinical semantics. Moreover, DERI also surpasses DERI-align, demonstrating the effectiveness of our cross-modal feature reconstruction

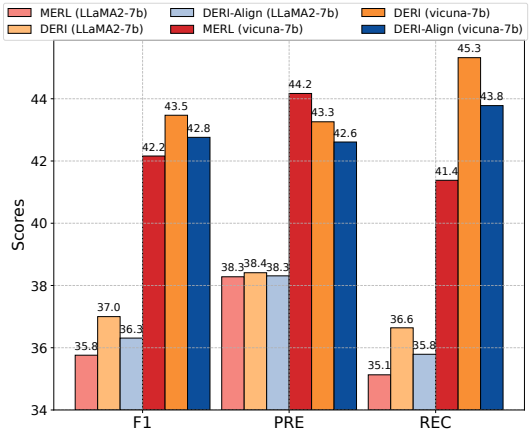

Figure 4: Report Generation CE Metrics with LLMs on MIMIC.

in incorporating diagnostic report information and enhancing semantic fidelity in generated reports.

To better verify the CE metrics calculation method, we also adopt large language models LLaMA2-7b [Touvron *et al.*, 2023] and vicuna-7b [Zheng *et al.*, 2023] to conduct report classification. Specifically, we feed the original reports and generated reports to the LLMs respectively and then ask the LLMs to choose the best class from six given categories: *Normal ECG, Myocardial Infarction, ST/T Change, Conduction Disturbance, Hypertrophy, and Others*. The answers of the original reports are regarded as ground truth and the answers of the generated reports are predicted labels. We then calculate the CE metrics as Fig. 4.

We can observe that the results of report classification using LLMs are basically the same as the results of our zero-shot categorization method: DERI is the best and DERI-align is the second best, while both methods outperform MERL on F1. Meanwhile, the calculation of CE using vicuna performs better results than LLaMA2. We also provide example-generated reports and report-generation tasks on PTB-XL compared with MEIT [Wan *et al.*, 2024] and ECG-Chat [Zhao *et al.*, 2024] in Supplementary.

4.4 Ablation Study

To better verify the performance of the key components/design choices of our DERI, we conduct comprehensive ablation studies on zero-shot classification and linear

Encoder	NLG					CE		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	F1	PRE	REC
MERL	59.68	54.09	49.58	46.52	69.52	23.03	25.01	23.05
DERI	**62.48**	**57.23**	**52.90**	**49.85**	**71.45**	**24.90**	**26.74**	**24.88**
DERI-Align	61.33	55.98	51.75	48.50	70.59	23.33	24.96	23.60

Table 3: Report generation results on MIMIC-ECG dataset. We bold the best results and grey represents the second highest.

probing with 1% training data across different classification datasets. All results are proposed with average AUC on six downstream datasets for ECG classification.

$\mathcal{L}_{align}$	$\mathcal{L}_{mixed}$	$\mathcal{L}_D^e$	$\mathcal{L}_D^t$	$\mathcal{L}_{RME}^e$	Zero-shot	Linear Probing (1%)
√		√	√	√	76.34	72.36
√	√		√	√	77.41	69.20
√	√	√		√	78.06	69.61
√	√	√	√		78.22	68.31
√	√	√	√	√	**78.73**	**74.05**

Table 4: Ablating Multiple Alignment and Feature Reconstruction.

Multiple Alignment and Feature Reconstruction. We realize the validation of the effect of these compositions by ablating the corresponding loss functions separately, and the experimental results are reported in Table 4. Table 4 shows that for zero-shot classification, $\mathcal{L}_{mixed}$ brings the best improvement while $\mathcal{L}_{RME}^e$ for linear probing. This suggests that the second alignment of mixed encoding can better fuse the clinical information from reports because zero-shot classification learning is conducted to calculate the similarity between the learned representations and prompt representations. The effect of $\mathcal{L}_{RME}^e$ on linear probing exemplifies the effectiveness of the RME-Module we designed to improve the learning ability of ECG representations. The completed DERI obtains the best performance, which illustrates the effectiveness of our DERI method.

	Latent Dropout	RME-Linear	RME-Module
Zero-shot	76.88	78.05	**78.73**
Linear Probing (1%)	71.86	72.40	**74.05**

Table 5: Ablating RME-Module.

RME-Module. We compare the effect of the RME-Module and its variant that uses linear projectors instead of the attention mechanism (which is set as RME-Linear) and the Latent Dropout strategy used by MERL. The experimental results are reported in Table 5. We can observe that the random masking strategy obtains better performance than dropout while using the multi-head attention mechanism instead of global meaning can achieve the best performance, enhancing the model's ability to learn ECG representation for classification. Masking encourages the model to learn context-aware representations, focusing on understanding relationships within the input. Random dropout removes (i.e., zeroes out) a fraction of the neurons (units) or edges in a network layer during training, but it does not apply this to the input itself. In dropout, neurons are randomly dropped independently at each forward pass. The designed RME-Module is used to enhance the global feature of the ECG signals, so we adopt random masking rather than a random dropout.

We then further explore the impact of the masking ratio p on the performance by changing it from 0.1 to 0.5 with a step of 0.1. The experimental results are shown in Table 6. We observe that the masking ratio of 0.1 obtains the best performance among other masking ratios in both zero-shot classification and linear probing. Therefore, we adopt the masking ratio of 0.1 in our RME-Module.

Mask-ratio	0.1	0.2	0.3	0.4	0.5
Zero-shot	**78.73**	77.67	78.05	77.62	76.35
Linear Probing (1%)	**74.05**	68.69	71.26	73.55	72.21

Table 6: Ablating Masking Ratio.

Shared Embedding and Mix Encoding. We also conduct experiments to verify the effect of the shared embedding used in cross-modal reconstruction. We remove the shared embedding from $\mathcal{D}t$ to $\mathcal{D}e$ as a variant of DERI. In addition, we use the pre-trained aligned ECG features to conduct zero-shot classification and linear probing to verify whether the mix encoding performs better than the aligned encoding for classification. This means that we adopt the same pre-training model but use the aligned ECG encoding instead of the mixed encoding for downstream tasks.

	Without SE	DERL-Align	DERL
Zero-shot	77.25	78.03	**78.73**
Linear Probing (1%)	70.22	73.59	**74.05**

Table 7: Ablating Shared Embedding and Mix Encoding.

The experimental results are shown in Table 7. It can be observed that removing the shared embedding from $\mathcal{D}t$ to $\mathcal{D}e$ during text encoding reconstruction leads to a decline in model performance for both zero-shot and linear probing tasks. Furthermore, using the mixed encoding, which includes the decoded report features, outperforms using only the aligned ECG features. These findings underscore the strong representation learning capability of DERI.

5 Conclusion

In this study, we proposed DERI, an innovative deep ECG-Report interaction framework for cross-modal representation learning. To obtain deep ECG-Report interaction for better representation learning, we design multiple alignments and cross-modal mutual reconstruction. Besides, an RME-Module is conducted on the ECG latent encoding for representation learning enhancement. Moreover, we extended ECG representation learning to clinical diagnostic report generation, aiming to deliver more intuitive ECG clinical insights. Our extensive experiments demonstrate the DERI's capability to learn the clinical semantics of ECG signals with the help of reports, which achieves the best performance on ECG classification and report generation.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported in part by the Sustainable Development Science and Technology Project of Shenzhen Science and Technology Innovation Commission under Grant KCXFZ20201221173411032.

References

[Attia *et al.*, 2019] Zachi I Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M McKie, Dorothy J Ladewig, Gaurav Satam, Patricia A Pellikka, Maurice Enriquez-Sarano, Peter A Noseworthy, Thomas M Munger, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine*, 25(1):70–74, 2019.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[Chen *et al.*, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2020b] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November 2020.

[Chen *et al.*, 2021] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[Chen *et al.*, 2024a] Jian Chen, Yuzhu Hu, Lalit Garg, Thippa Reddy Gadekallu, Gautam Srivastava, and Wei Wang. Graph enhanced low-resource ecg representation learning for emotion recognition based on wearable internet of things. *IEEE Internet of Things Journal*, 2024.

[Chen *et al.*, 2024b] Jian Chen, Yuzhu Hu, Qifeng Lai, Wei Wang, Junxin Chen, Han Liu, Gautam Srivastava, Ali Kashif Bashir, and Xiping Hu. Iifdd: Intra and intermodal fusion for depression detection with multi-modal information from internet of medical things. *Information Fusion*, 102:102017, 2024.

[Chen *et al.*, 2024c] Jian Chen, Wei Wang, Yuzhu Hu, Junxin Chen, Han Liu, and Xiping Hu. Tgca-pvt: Topic-guided context-aware pyramid vision transformer for sticker emotion recognition. In *ACM Multimedia*, 2024.

[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. *arXiv preprint arXiv:2106.14112*, 2021.

[Gow *et al.*, 2023] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[Huang *et al.*, 2022] Yu Huang, Gary G Yen, and Vincent S Tseng. Snippet policy network for multi-class varied-length ecg early classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6349–6361, 2022.

[Jeong *et al.*, 2024] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, pages 978–990. PMLR, 2024.

[Jin *et al.*, 2023] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

[Kiyasseh *et al.*, 2021] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[Lalam *et al.*, 2023] Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023.

[Liu *et al.*, 2018] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

[Liu *et al.*, 2023a] Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on*

Medical Image Computing and Computer-Assisted Intervention, pages 637–647. Springer, 2023.

[Liu *et al.*, 2023b] Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5361–5372, 2023.

[Liu *et al.*, 2024a] Che Liu, Zhongwei Wan, Sibo Cheng, Mi Zhang, and Rossella Arcucci. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8230–8234. IEEE, 2024.

[Liu *et al.*, 2024b] 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. *arXiv preprint arXiv:2403.06659*, 2024.

[Long and Wang, 2023] Weicai Long and Xingjun Wang. Bpnet: A multi-modal fusion neural network for blood pressure estimation using ecg and ppg. *Biomedical Signal Processing and Control*, 86:105287, 2023.

[Maniparambil *et al.*, 2023] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 262–271, 2023.

[Na *et al.*, 2024] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

[Oh *et al.*, 2022] Jungwoo Oh, Hyunseung Chung, Joonmyoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pages 338–353. PMLR, 2022.

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

[Raghu *et al.*, 2022] Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Contrastive pre-training for multimodal medical time series. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

[Sanh, 2019] V Sanh. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[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.

[Wan *et al.*, 2024] Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Zhenwu Peng, Jie Fu, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. Electrocardiogram instruction tuning for report generation. *arXiv preprint arXiv:2403.04945*, 2024.

[Wang *et al.*, 2023] Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Wang *et al.*, 2024] Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. *arXiv preprint arXiv:2407.15158*, 2024.

[You *et al.*, 2021] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer, 2021.

[Zbontar *et al.*, 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[Zhang *et al.*, 2023] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Zhao *et al.*, 2024] Yubao Zhao, Tian Zhang, Xu Wang, Puyu Han, Tong Chen, Linlin Huang, Youzhu Jin, and Jiaju Kang. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. *arXiv preprint arXiv:2408.08849*, 2024.

[Zheng *et al.*, 2022] J Zheng, H Guo, and H Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022Available online httpphysionet orgcontentecg arrhythmia10 0accessed on*, 23, 2022.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.