Fine-Grained ECG-Text Contrastive Learning with Waveform Feature Awareness

Anonymous submission

Abstract

Electrocardiograms (ECGs) are crucial for diagnosing cardiovascular diseases. While existing ECG-text contrastive learning methods improve representation transferability and enable zero-shot prediction by aligning paired ECGs with their reports and distinguishing them from unpaired reports, they still exhibit certain limitations. Specifically, ECG reports primarily focus on global information, such as diagnosis and rhythm, while often neglecting fine-grained details like waveform features. This omission arises because waveform features, which serve as an intermediate step in physicians' diagnostic reasoning process, are typically not recorded in the report once a diagnosis is provided. However, these features are critical for understanding the diagnostic process. Consequently, This coarse-grained alignment limits the model's ability to capture local ECG features and understand diagnostic reasoning based on waveform characteristics. To address this, we propose FG-CLEP (Fine-Grained Contrastive Language ECG Pre-training), which involves training a CLEP model using contrastive learning on ECG-report pairs; generating fine-grained reports with waveform features with the help of LLMs and CLEP; and further training the CLEP model with this fine-grained report to obtain the final FG-CLEP model. Furthermore, considering the frequent false negatives caused by the long-tail distribution of ECGs and the multi-label nature of ECG downstream tasks, we adopt a sigmoid-based loss function to accommodate multi-label requirements and introduced a semantic similarity matrix to guide contrastive learning. Experiments demonstrate that FG-CLEP outperforms state-of-the-art methods in both zero-shot prediction and linear probing across six ECG datasets.

Introductin

Electrocardiograms (ECGs) are vital, non-invasive diagnostic tools extensively used in clinical settings to detect and manage cardiac arrhythmic diseases (Sahoo et al. 2020; Rath et al. 2021; Ayano et al. 2022). However, traditional supervised methods require annotated data, which demands significant expertise and limits their accessibility and scalability. Recently, the emergence of self-supervised learning using unlabeled data has effectively addressed this issue.

Self-supervised learning in ECG analysis has been primarily explored through two paradigms: comparative self-supervision and generative self-supervision. Comparative self-supervision (Chen et al. 2020; Chen, Xie, and He 2021; Wang et al. 2023; Eldele et al. 2021) focuses on learning discriminative ECG features by differentiating between augmented positive and negative samples. Generative self-supervision (Zhang et al. 2022a; Hu, Chen, and Zhou 2023; Na et al. 2024; Zhang et al. 2022b) aims to reconstruct original signals from their masked versions. However, these methods often require annotated samples and struggle with classes not present in the fine-tuning stages.

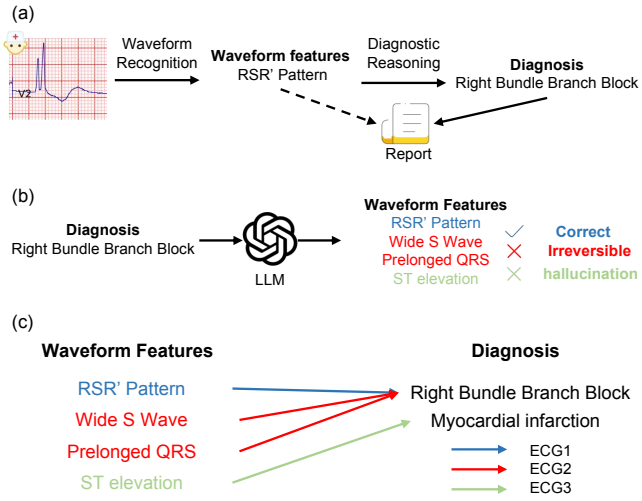

Figure 1: (a) Doctors typically make diagnoses based on waveform features, which, as intermediate products, are often not recorded. (b) Using LLMs to recover waveform features from diagnostic results faces two major challenges: model hallucinations and irreversibility illustrated in c. (c) There is a non-bijective relationship between waveform features and diagnoses, as the same disease may present different waveform characteristics across individuals.

Recent advancements in vision-text contrastive learning (Radford et al. 2021; Huang et al. 2021; Zhang et al. 2022c; Wang et al. 2022), have inspired research into ECG-report contrastive learning (Li et al. 2024; Liu et al. 2024b,a; Yu, Guo, and Sano 2024; Lalam et al. 2023). Methods like (Li et al. 2024; Liu et al. 2024a; Lalam et al. 2023) learn ECG representations by aligning embeddings of paired ECG signals and their corresponding reports while distinguish-

ing them from unpaired reports, enabling zero-shot prediction. MERL (Liu et al. 2024b) further enhances these representations with uni-modal alignment and employs the CK-EPE pipeline during inference to generate more descriptive prompts using LLMs. However, enhancing textual prompts only at the inference stage can't fully leverage the capabilities of LLMs. In contrast, ESI (Yu, Guo, and Sano 2024) enhances ECG reports during training by integrating a retrieval-augmented generation (RAG) pipeline that leverages LLMs and external medical knowledge to produce detailed textual descriptions. However, none of the aforementioned methods focus on the missing fine-grained waveform features in ECG reports. Although these features are essential for diagnostic reasoning, they are often omitted from reports as intermediate results once the diagnoses are established, as shown in Figure 1(a). Our objective is to reconstruct these omitted features by leveraging the capabilities of LLM.

It is important to note that simply using LLMs to generate detailed information (Liu et al. 2024b; Yu, Guo, and Sano 2024) like Figure 1(b) is untrusted for the following reasons: (1) the hallucination problem in medical LLMs (Huang et al. 2023; Günay, Öztürk, and Yiğit 2024) undermines the reliability of their output; and (2) the diagnostic process used by physicians involves observing waveform features and reasoning to arrive at a diagnosis, which is not reversible (i.e., inferring waveform features from only the diagnosis is not feasible). This difficulty stems from the non-bijective relationship between ECG waveform features and diagnoses (Jin 2018), where a single disease may exhibit different waveform characteristics across individuals as illustrated in Figure 1(c). Thus, methods like (Liu et al. 2024b; Yu, Guo, and Sano 2024), which rely solely on feeding ECG reports into LLMs to infer potential waveform features, often yield incomplete or imprecise results. However, these inferences can still serve as valuable references for identifying potential waveform features. To address this, we need to further compare the features generated by the LLM with the ECG signals to verify whether these features are genuinely present in the ECG data. Specificaly, we propose a two-stage training strategy. In the first stage, we train a base model using traditional contrastive learning methods (Radford et al. 2021; Liu et al. 2024a). This model is then utilized to evaluate the waveform features inferred by LLMs from ECG reports. High-probability features are treated as ground truth and incorporated into the reports, which are subsequently used in the second stage of training to refine the model. By doing so, we not only addresses the aforementioned irreversibility issue but also provides a mechanism for error correction when LLMs exhibits hallucination problems.

Additionally, the ECG data exhibit several unique characteristics, as illustrated in Figure 3. First, ECG data exhibit a long-tail distribution (Thai et al. 2017; Yogarajan et al. 2021), where the majority of ECGs are normal, leading to similar semantics in ECGs and reports from different patients. However traditional multimodal contrastive learning (Radford et al. 2021; Liu et al. 2024b,a) assumes that only the ECG and report from the same patient are positive samples for each other which makes the false negative problem frequent in the pre-training phase. Second, there is the challenge of multi-label prediction in downstream tasks. Most ECGs exhibit multi-label characteristics; however, traditional multimodal contrastive learning methods like CLIP (Radford et al. 2021) employ the info-NCE loss (Oord, Li, and Vinyals 2018), which utilizes a softmax operation that is incompatible with the multi-label nature of downstream tasks. To address these issues, we abandoned the traditional Info-NCE loss function and adopted a sigmoid-based loss function to meet the requirements of multi-label scenarios. Additionally, we introduced a semantic similarity matrix to measure the semantic similarity between different reports, enabling the identification of false negatives to guide contrastive learning. Benefiting from the proposed two-stage pretraining strategy, which utilizes fine-grained reports with waveform features and the model architecture tailored to ECG characteristics, FG-CLEP achieves higher accuracy compared to state-of-the-art methods on the PTB-XL, CPSC2018, and CSN datasets. This result highlights the superiority of our approach.

Overall, our contributions are threefold:

- We propose a two-stage pre-training process to develop a model capable of capturing fine-grained local waveform features, where stage 1 involves traditional training, and stage 2 leverages fine-grained reports with local waveform features, augmented with the help of LLMs and the stage 1 model.
- Considering the multi-label nature of ECG downstream tasks and the frequent false negatives caused by the long-tail distribution in ECG data, we implemented a sigmoid-based loss function and introduced a semantic similarity matrix to measure the similarity of reports from different patients, guiding contrastive learning and effectively mitigating the issue of false negatives.
- Experimental results indicate that the model, pre-trained on MIMIC-ECG using our FG-CLEP framework, surpasses state-of-the-art methods in both zero-shot prediction and linear probing across six datasets, including PTB-XL, CPSC2018, and CSN.

Related Work

ECG Self-Supervised Learning Self-supervised learning in ECG analysis has primarily been explored through two paradigms: contrastive and generative ways. Contrastive self-supervision (Chen et al. 2020; Chen, Xie, and He 2021; Wang et al. 2023; Eldele et al. 2021) typically involves augmenting the same ECG signal into two different views as positive samples, while different ECG signals serve as negative samples. This approach heavily relies on the quality of data augmentation. However, these complex augmentation methods (Wang et al. 2023) obscure the characteristics of the augmented samples, making it difficult to assess whether any critical features have been altered.

Generative self-supervision (Zhang et al. 2022a; Hu, Chen, and Zhou 2023; Na et al. 2024; Zhang et al. 2022b) first masks a portion of the ECG signal and then attempts to recover the masked part using the unmasked portion. This method highly emphasizes the local features of the ECG but

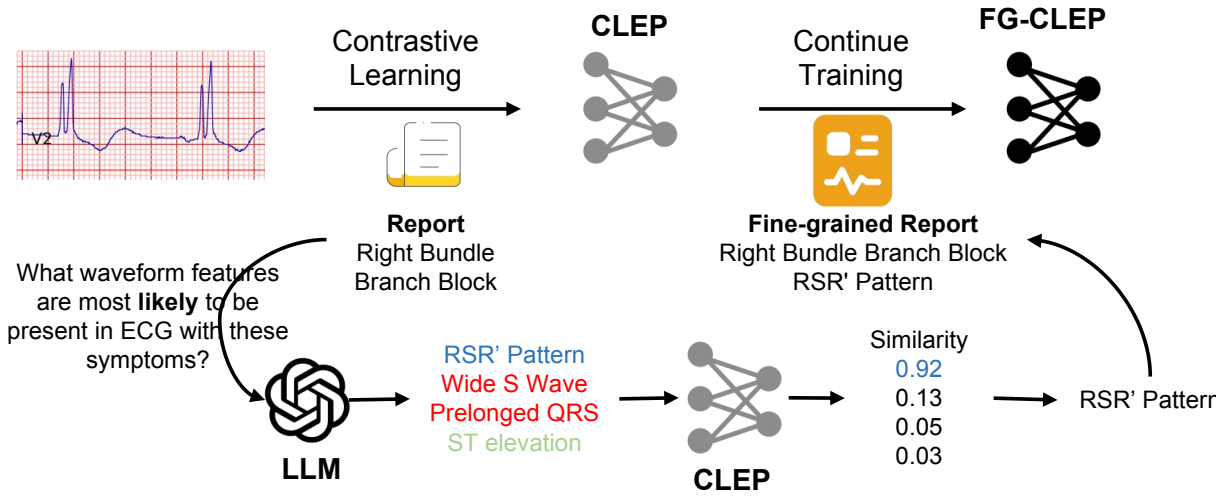

Figure 2: **Overview of the Two-Stage Pretraining Strategy.** (1) Stage 1: Train the CLEP model using original ECG-report pairs; (2) Obtaining Fine-Grained Reports: Query a LLM to obtain potential waveform features related to the ECG. Use the CLEP model from Stage 1 to compute the similarity, and incorporate the validated waveform features into the original report to produce fine-grained reports. (3) Stage 2: Continue training the CLEP model using the ECG and fine-grained report pairs to obtain FG-CLEP(Fine-Grain CLEP).

neglects the global features. However, both ECG diagnosis and rhythm analysis heavily depend on the effective extraction of global features.

Unlike previous ECG self-supervised learning methods, which often require annotated samples and struggle with classes not present in the fine-tuning stages, our ECG-text contrastive learning approach enables direct zero-shot prediction on downstream tasks.

Multi-modal Contrastive Learning Recently, in addition to the two aforementioned single-modal self-supervised learning methods, multi-modal contrastive learning has also gained significant attention, epitomized by the emergence of vision-text contrastive learning (Radford et al. 2021; Huang et al. 2021; Zhang et al. 2022c; Wang et al. 2022). CLIP (Radford et al. 2021) treats images and their corresponding captions as positive samples, and images with unrelated captions as negative samples for contrastive learning. This approach not only enhances the transferability of learned representations but also possesses a unique capability absent in the previous single-modal self-supervised methods: zero-shot ability.

The CLIP method (Radford et al. 2021) has been rapidly adapted to various fields, including ECG analysis (Li et al. 2024; Liu et al. 2024b,a; Yu, Guo, and Sano 2024; Lalam et al. 2023). Similar to the vision-text domain, (Li et al. 2024; Liu et al. 2024a; Lalam et al. 2023) learns ECG representations by aligning ECGs with their paired reports while distancing them from unpaired reports. In addition to this cross-modal alignment, MERL (Liu et al. 2024b) introduces a uni-modal alignment to enhance the learned representation. Furthermore, MERL employs the CKEPE pipeline during test time to generate more descriptive prompts via large language models (LLMs). However, enhancing textual prompts only at the inference stage introduces a distribution mismatch between training and testing text, which is suboptimal. In contrast, ESI (Yu, Guo, and Sano 2024) enhances ECG reports during training by integrating a retrieval-augmented generation (RAG) pipeline that leverages LLMs and external medical knowledge to produce detailed textual descriptions.

Similar to MERL (Liu et al. 2024b) and ESI (Yu, Guo, and Sano 2024), our method also utilizes LLMs to enhance ECG reports. However, none of these methods specifically focus on recovering waveform characteristics during the report enhancement process. Given that the same disease can exhibit different waveform features in different patients (Jin 2018) and the hallucination problem of medical LLMs (Huang et al. 2023; Günay, Öztürk, and Yiğit 2024), simply relying on LLMs to enhance reports cannot perfectly restore waveform characteristics. Instead, it only provides potential waveform feature references. Treating all LLM outputs as ground truth could lead to inaccuracies. Therefore, we further validate the potential waveform features suggested by the LLM by referencing the ECG signals to calculate their degree of alignment, and only consider high-confidence waveform features as ground truth.

False Negatives in Contrastive Learning The traditional multi-modal contrastive learning method (Radford et al. 2021) assumes that only the images and captions from the same record are positive samples for each other. However, this assumption is not always valid, especially when applied to the ECG domain, which has a long-tailed distribution where most ECGs are normal, and even among abnormal ECGs, many are due to common diseases, as illustrated in Figure 3(1). Under this assumption, what we call 'false negatives' will frequently occur. There have been several attempts to address this issue (Lavoie et al. 2024; Jiang et al. 2023b; Sun et al. 2023; Li et al. 2023; Wang et al. 2022).

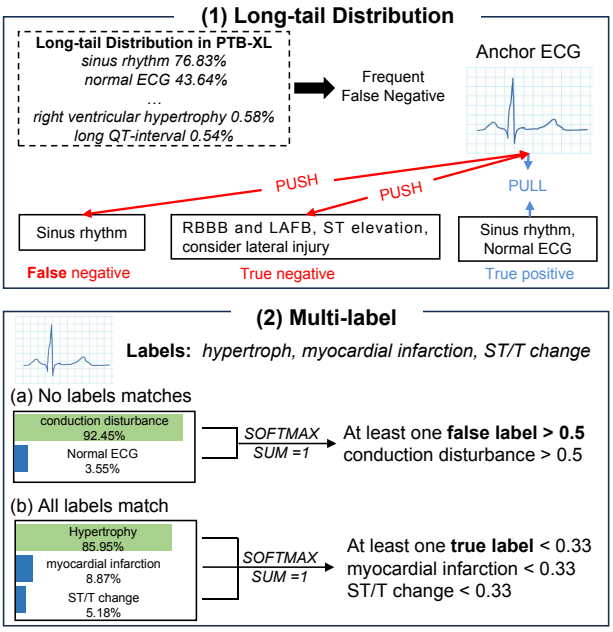

Figure 3: **Two critical properties of ECGs must be considered when designing the model architecture:** (1) Long-tail distribution: Most ECGs share similar reports, making false negatives frequent when pushing ECGs away from the reports from different records. (2) Multi-label: The softmax operation used in InfoNCE loss is not suitable for ECG multi-label classification downstream tasks. (a) and (b) illustrate two extreme cases.

Some approaches (Jiang et al. 2023b; Li et al. 2023) attempt to add a regularization term to mitigate false negatives. Others (Sun et al. 2023; Wang et al. 2022) introduce a matrix to measure the similarity between different reports, guiding contrastive learning to identify and address false negatives. In this paper, we explore the application of the latter approach in the ECG multi-modal contrastive learning domain.

Method

In this section, we present the technical details of the FG-CLEP framework, which includes a two-stage pre-training strategy, as illustrated in Figure 2, and a specific model architecture, as shown in Figure 4, considering the two properties of ECGs depicted in Figure 3. The architecture comprises the following components: (1) ECG and text encoders that extract embeddings, (2) a Semantic Similarity Matrix that measures the similarity of reports from different patients, and (3) L_{sig} and L_{fnm} that train the entire model.

Model Architecture

As illustrated in Figure 4, FG-CLEP consists of one ECG encoder and one text encoder.

ECG Encoder. We encode ECG signals into embeddings $\mathbf{e} \in \mathbb{R}^D$ using an ECG encoder E_{ecg}. A projection head then maps raw embeddings to $\mathbf{e}_p \in \mathbb{R}^P$.

$$\mathbf{e} = E_{\text{ecg}}(x_{\text{ecg}}) \tag{1a}$$
$$\mathbf{e}_p = f_e(\mathbf{e}) \tag{1b}$$

where f_v is the projection head of the ECG encoder.

Text Encoder We create clinically meaningful text embeddings $\mathbf{t} \in \mathbb{R}^M$ by a text encoder. We project them to $\mathbf{t}_p \in \mathbb{R}^P$ as

$$\mathbf{t} = E_{\text{txt}}(x_{\text{txt}}) \tag{2a}$$
$$\mathbf{t}_p = f_t(\mathbf{t}) \tag{2b}$$

where f_t is the projection head and E_{txt} denotes the text encoder. This gives the same embedding dimension P as the ECG encoder, suitable for contrastive learning.

Semantic Similarity Matrix False negatives in the pre-training phase arise from the assumption that ECGs and reports from different patients are unmatched. However, due to the long-tail distribution, many ECGs exhibit similar symptoms, leading to reports with similar semantics. To address this, we introduce a Semantic Similarity Matrix similar to (Sun et al. 2023; Wang et al. 2022) to measure the similarity of reports from different patients during the pre-training phase.

We denote an ECG-report dataset as $D = \{(x_{\text{ecg}_i}, x_{\text{txt}_i}) \mid i \in [0, n)\}$, where $(x_{\text{ecg}_i}, x_{\text{txt}_i})$ represents a sample with paired ECG-report content. The ECG and text signals are encoded into $(\mathbf{e}_p, \mathbf{t}_p)$ as disscussed above, and the semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ is defined as follows:

$$\mathbf{S}_{ij} = \text{sim}(\mathbf{t}_{pi}, \mathbf{t}_{pj}) \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity function.

$$\text{sim}(\mathbf{t}_{pi}, \mathbf{t}_{pj}) = \frac{\mathbf{t}_{pi} \cdot \mathbf{t}_{pj}}{\|\mathbf{t}_{pi}\| \|\mathbf{t}_{pj}\|} \tag{4}$$

This results in a similarity matrix $\mathbf{S}$, where each element $\mathbf{S}_{ij}$ indicates the similarity between the i-th and j-th report embeddings. The Semantic Similarity Matrix guides the contrastive learning process and reduces false negatives by considering the semantic similarities across different pairs. We give an intuitive visualization of the semantic similarity matrix in Section Analysis.

Loss Function The loss function of our FG-CLEP framework consists of two components: L_{sig} and L_{fnm}. L_{sig} represents a sigmoid-based loss function tailored to adapt to downstream ECG multi-label classification tasks. L_{fnm} addresses the issue of false negatives commonly observed in ECG datasets due to the long-tail distribution.

Sigmoid-based Contrastive Loss L_{sig} Traditional multi-modal contrastive learning often relies on the InfoNCE loss (Oord, Li, and Vinyals 2018), which involves a softmax operation. However, for ECG downstream tasks, which are mostly multi-label classification problems, using the softmax operation can lead to various issues, as illustrated in Figure 3(2). To address this problem, we propose using a sigmoid-based loss during the pre-training phase. Specifically, we leverage the sigmoid loss (Zhai et al. 2023), which

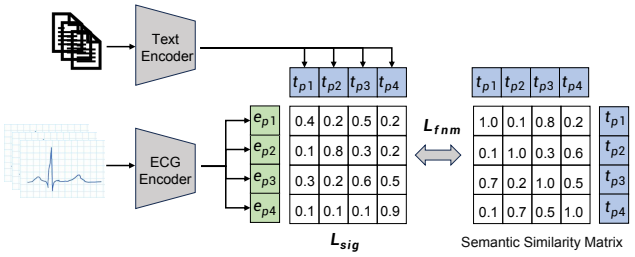

Figure 4: **The Overview of Model Architecture.** (1) A sigmoid-based loss function was used to suit multi-label downstream tasks; (2) The semantic similarity matrix was used to guide contrastive learning to mitigate false negatives.

is known for its memory efficiency compared to InfoNCE loss (Oord, Li, and Vinyals 2018). Here, we use it for its disuse of the softmax operation, making it more suitable for downstream ECG multi-label classification tasks.

The L_{sig} is defined as:

$$L_{\text{sig}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\log\left(\frac{1}{1+e^{-z_{ij}(-t\cdot\text{sim}(\mathbf{e}_{pj},\mathbf{t}_{pj})+b)}}\right) \tag{5}$$

Where B represents the batch size, z_{ij} denotes the match between a given ECG and report input, equaling 1 if they are paired and -1 otherwise. At initialization, the significant imbalance caused by numerous negatives dominates the loss, resulting in large initial optimization steps aimed at correcting this bias. To mitigate this issue, an additional learnable bias term b, similar to the temperature t, is used. This ensures that the training starts approximately close to the prior, avoiding the need for substantial over-correction.

False Negative Mitigation Loss L_{fnm} Due to the long-tail distribution of ECG labels, most electrocardiograms have similar reports. Simply aligning the embeddings of paired electrocardiograms with their corresponding reports while distinguishing them from unpaired reports results in false negatives. To address this issue, we incorporate the semantic similarity matrix S obtained above into the loss function to guide contrastive learning, similar to (Sun et al. 2023). This matrix captures the semantic similarity of reports from different patients, allowing us to identify and correct false negative samples.

The L_{fnm} is defined as:

$$L_{\text{fnm}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}|\text{sim}(\mathbf{e}_{pi},\mathbf{t}_{pj}) - \mathbf{S}_{ij}| \tag{6}$$

where B represents the batch size, $\mathbf{S}_{ij}$ is the semantic similarity between the i-th and j-th report embeddings, and the $|\cdot|$ represents the L1-distance. We alleviates the false negative problem by reducing the L1 distance between $\text{sim}(\mathbf{e}_{pi},\mathbf{t}_{pj})$ and $\mathbf{S}_{ij}$.

Combined Loss Function The final loss function combines L_{sig} and L_{fnm}:

$$L = L_{\text{sig}} + \lambda L_{\text{fnm}} \tag{7}$$

where λ is a hyperparameter that controls the trade-off between the two loss components.

In conclusion, by integrating the semantic similarity matrix $\mathbf{S}$ and using a sigmoid-based contrastive loss, our pretraining loss function better adapts to downstream ECG multi-label tasks when performing zero-shot prediction and mitigates the frequent false negatives caused by the long-tail distribution of ECG data, leading to improved learning of meaningful representations of ECG signals.

Inference During the inference phase, given an ECG and a text description, we calculate their similarity as described above and use the sigmoid function to determine the probability. The probability score P for the matching between the ECG embedding $\mathbf{e}_p$ and the text embedding $\mathbf{t}_p$ is computed as follows:

$$P = \sigma(\text{sim}(\mathbf{e}_p, \mathbf{t}_p)) \tag{8}$$

where σ denotes the sigmoid function. This probability score indicates how likely it is that the ECG and the text description are related. The higher the score, the more confident we are in the match.

Two-stage Pre-trinaing

As illustrated in Figure 2, we propose a two-stage pretraining framework. In the first stage, we use the original ECG and report pairs to train our model as discussed above. In the second stage, we use the fine-grained reports with local waveform features to continue training our model while keeping the model architecture unchanged. We call the model obtained from the first stage CLEP, and the model from the second stage is called FG-CLEP (Fine-Grained CLEP). To make the whole process clearer, we show a pseudo code on Algorithm including how to obtain fine-grained reports.

To obtain fine-grained reports, we first need to recap the annotation process of a clinical doctor, as illustrated in Figure 2(2)(a). Given an ECG signal, the doctor initially observes the waveform characteristics and then identifies key waveform features, such as the RSR' pattern, to infer a possible diagnosis, such as Right Bundle Branch Block (RBBB). However, waveform features are typically not recorded when writing the report; only the final diagnosis is documented.

We aim to restore and complete this waveform information to generate a fine-grained report. Given a report, we query a Large Language Model (LLM) with the question, 'What waveform features are most likely to be present in electrocardiograms with these symptoms?' to identify potential overlooked waveform features. To format the results, we further instruct, 'Organize these waveform features into a Python list, with each item representing a distinct waveform feature.' Using this explicit chain-of-thought (Wei et al. 2022) instruction, we can get a list of potential waveform features. However, the relationship between waveform features and diagnosis is not a simple one-to-one correlation;

it involves complex logical reasoning. Therefore, not every predicted waveform feature necessarily appears in the ECG.

To validate these features, we need to use the model trained in the first stage, CLEP. Specifically, we calculate the similarity between these possible waveform features and the ECG, select features with a similarity greater than a threshold, and incorporate these validated features into the original report to create a fine-grained report. We then use this fine-grained report to continue training CLEP to obtain FG-CLEP.

Algorithm 1: FG-CLEP: two-stage pretraining

Require: $D = \{(ECG_i, report_i) : i = 1 \text{ to } n\}$

(1) Stage1
Train CLEP on ECG-report pairs D
(2) Obtaining Fine-grained Reports
for each (ECG, report) in D **do**
  waveform features = LLM(report, prompt)
  for each waveform feature in waveform features **do**
    similarity = CLEP(ECG, waveform feature)
    if similarity >threshold **then**
      report = report + waveform feature
    end if
  end for
end for
(3) Stage2
Continue training CLEP on ECG-fine-grained report pairs to obtain FG-CLEP

Experiments

Datasets

We pre-train the FG-CLEP framework using the MIMIC-ECG (Gow et al.) dataset and test it on the PTB-XL (Wagner et al. 2020), CPSC2018 (Liu et al. 2018), and CSN (Zheng, Guo, and Chu 2022) datasets, following the benchmark proposed by (Liu et al. 2024b). The statistics of the datasets used are presented in Table 1. All the ECGs in the datasets are 12-lead recordings. The MIMIC-ECG dataset contains nearly 800,000 ECG-report pairs. To improve data quality, we excluded samples with an empty report or reports containing fewer than three words, removed reports without useful information, and discarded ECGs with unexpected situations. The PTB-XL dataset can be further divided into four subsets, and we follow the official train:validation:test split. For CPSC2018 and CSN, we split the dataset as 70%:10%:20% for the train:validation:test split.

Implementation Details

Pre-training Implementation: In the pre-training stage, we utilize a randomly initialized 1D-ResNet50 model (He et al. 2016) as the ECG encoder and BioClinicalBERT (Alsentzer et al. 2019) for text encoding. The AdamW optimizer is selected with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-4}. FG-CLEP is pre-trained for 10 epochs in stage1 and 3 epochs in stage2, using a cosine annealing scheduler for learning rate adjustments and a warmup phase for the first 10% of training steps. A batch size of 100 is maintained. The temperature parameters t and b are initialized to $\log 10$ and -10, respectively. We use LLaMa3-8B (AI@Meta 2024) as our LLM to query potential waveform features and use vLLM (Kwon et al. 2023) to speed up inference. All experiments were conducted on two NVIDIA A800-80GB GPUs, except for the ablation study of Llama3-70B, which was conducted on four NVIDIA A800-80GB GPUs.

Downstream Task Implementation: We evaluated the downstream tasks using both zero-shot and linear probe settings. For the zero-shot setting, we froze the entire model and used the text of each category as the prompt. We computed the similarity between the ECG embedding and the category text embedding as the classification probability. Additionally, we employed an ensemble method to enhance zero-shot performance. Specifically, in addition to using the category as text, we also added 'category in lead x' (x represents any of the 12 leads) as text to compute the probability and used the highest probability as the final probability for that category. For linear probing, we kept the ECG encoder frozen and updated only the parameters of a newly initialized linear classifier. We conducted linear probing for each task using 1%, 10%, and 100% of the training data, maintaining these configurations across all linear probing classification tasks. For all downstream tasks, we used macro AUC as the evaluation metric.

Table 1: The statistics of used datasets.

Pretrain	**# ECGs**	**# Reports**		
MIMIC-ECG	773,268	773,268		
Evaluation	**# Train**	**# Valid**	**# Test**	**# Classes**
PTB-XL Super	17,084	2,146	2,158	5
PTB-XL Sub	17,084	2,146	2,158	23
PTB-XL Form	7,197	901	880	19
PTB-XL Rhythm	16,832	2,100	2,098	12
CPSC2018	4,800	684	1,383	9
CSN	31,606	4,515	9,031	51

Zero-Shot Ability

We conducted a zero-shot ECG classification evaluation on four PTB-XL subsets, CPSC2018, and CSN. The learned ECG-text encoders were used to support zero-shot prediction by matching the encoded ECG embeddings with the embeddings of generated prompts for each label class. The results are illustrated in Table 2.

Both the first-stage model, CLEP, and the second-stage fine-grained model, FG-CLEP, performed well. A detailed examination of the data reveals that FG-CLEP significantly outperforms CLEP on PTBXL-Form, demonstrating that the second-stage training substantially enhanced the model's ability to capture local ECG waveform features. This improvement is particularly evident when using the ensemble method, which extends the label text to 12 leads ('label in lead x', where x represents any of the 12 leads). This further indicates FG-CLEP's fine-grained capture capability. However, the ensemble inference method often proves detrimen-

Table 2: Results of zero-shot classification. CLEP: stage1 model, FG-CLEP: stage2 model, ENS: ensemble inference

macro AUC	PTB-XL-Super	PTBXL-Sub	PTBXL-Form	PTBXL-Rhythm	CPSC2018	CSN
MERL	74.20	75.70	65.90	78.50	82.80	74.40
CLEP	77.50	81.85	66.29	88.60	85.15	80.10
CLEP_{ENS}	75.64	82.55	64.74	88.67	83.91	81.24
FG-CLEP	79.28	83.57	67.77	**92.31**	**88.24**	82.46
$\text{FG-CLEP}_{\text{ENS}}$	**79.68**	**83.65**	**70.79**	91.52	87.08	**84.60**

Table 3: Results of Linear Evaluation.

Method	PTB-XL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random Init	70.45	77.09	81.61	55.82	67.60	77.91	55.82	62.54	73.00	46.26	62.36	79.29	54.96	71.47	78.33	47.22	63.17	73.13
SimCLR(Chen et al. 2020)	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(Grill et al. 2020)	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(Zbontar et al. 2021)	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(Chen, Xie, and He 2021)	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(Chen and He 2021)	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(Eldele et al. 2021)	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(Kiyasseh, Zhu, and Clifton 2021)	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(Wang et al. 2023)	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(Zhang et al. 2023)	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
ST-MEM(Na et al. 2024)	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(Liu et al. 2024b)	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
CLEP	84.04	88.79	89.82	69.09	**86.08**	92.50	67.89	72.35	82.59	61.79	**91.86**	90.18	83.12	93.42	96.56	63.00	80.03	93.35
FG-CLEP	**84.89**	**89.51**	**90.77**	**69.96**	85.75	**92.62**	**68.91**	**74.80**	**85.42**	**68.99**	91.35	**94.08**	**83.35**	**93.60**	**96.65**	62.59	79.35	**93.46**

tal to the first-stage model, CLEP, as seen in PTBXL-Super, PTBXL-Form, and CPSC2018.

Linear Evaluation

We aim to evaluate the learned model transferability to downstream supervised tasks. We froze the ECG encoder and fine-tuned a randomly initialized linear classification head on the training data with binary cross-entropy loss. We compared a series of contrastive and generative self-supervised learning methods. Results in Table 3 show that FG-CLEP still achieves the best performances across all methods in most scenarios.

Furthermore, when comparing the linear probe result in Table 3 with the zero-shot result in Table 2, we surprisingly find that FG-CLEP's zero-shot predictions are comparable to Linear Probe evaluations using 10% of the data in PTBXL-Sub, PTBXL-Form, CPSC2018, and CSN. Additionally, the zero-shot performance in PTBXL-Form is comparable to the full 100% Linear Probe evaluation. This further confirms the robustness and generalizability of our framework.

Embedding Visualization

We also demonstrate the effectiveness of our representation learning framework by plotting t-SNE visualizations of ECG embeddings produced for PTB-XL ECGs in five classical waveform features. As shown in Fig. 5, our model produces well-clustered representations. Furthermore, as expected, the model obtained in stage two, FG-CLEP, learns more fine-grained local waveform features of ECGs. Specifically, FG-CLEP clusters 'prolonged PR interval' much better than the stage one model, CLEP.

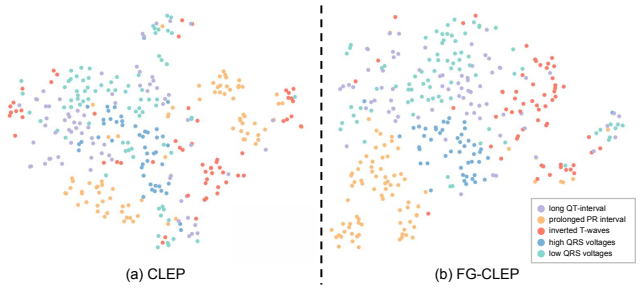

Figure 5: Embeddings visualization of PTB-XL ECGs in 5 waveform features by (a)CLEP and (b)FG-CLEP. Dimension reduced by t-SNE.

Analysis

In this section, we visualize the semantic similarity matrix to identify false negatives, conduct an ablation study to validate FG-CLEP's effectiveness, and assess its performance across various LLMs, text encoders, and ECG encoders. The reported metrics reflect the average zero-shot AUC across the aforementioned six datasets.

Semantic Similarity Matrix

We visualize the semantic similarity matrix in Figure 6. The left side shows the semantic similarity matrix from a random batch. As illustrated, ECGs and reports from different records may share similarities to some extent. Ignoring these similarities would result in a diagonal matrix with ones on the diagonal and zeros elsewhere, which is obviously wrong. The right side displays a semantic similarity matrix where the first 16 entries are normal ECGs and the last 16 are abnormal ECGs. The matrix effectively captures the semantic similarities of the normal ECGs.

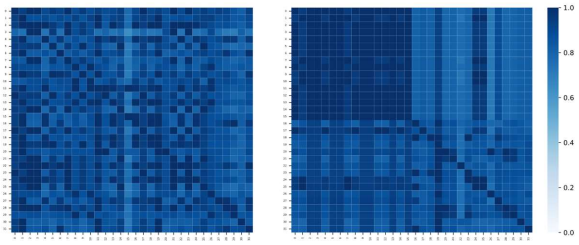

Figure 6: **The Heatmap of Semantic Similarity Matrix.** Left: from a random batch; Right: with the first 16 as normal ECG and the last 16 as abnormal ECG.

Ablation Study

We conducted an ablation study to verify the effectiveness of our two-stage pre-training strategy and the model architecture modifications designed for long-tail and multi-label tasks. The results in Table 4 indicate that the two-stage training strategy and our introduced semantic similarity matrix, along with sigmoid-based loss significantly enhance model performance.

Table 4: Results of ablation study.

	AUC
FG-CLEP	82.27
w/o fine-grained reports	79.92
w/o architecture modification	80.78

Different LLMs

In the stage2 pre-training, we used LLMs to obtain potential waveform features related to an ECG report. To assess our framework's feasibility, we conducted experiments on various LLMs, including general models as well as domain-specific medical LLMs. As shown in Table 5, our framework performed well across different LLMs. Meanwhile, the results suggest that larger models offer some improvement, though not significantly. Surprisingly, the domain-specific model underperformed compared to its base model. This may be because, while domain-specific LLMs possess more medical knowledge, their general instruction-following ability might decline. Since our framework depends on LLMs' capacity to organize lists from paragraphs, the domain-specific LLMs struggled with this task.

Table 5: Results on different LLMs.

LLM	AUC
Phi-3-mini-4k-instruct (Abdin et al. 2024)	81.51
Mistral-7B-Instruct-v0.2 (Jiang et al. 2023a)	81.89
Llama3-8B-Instruct (AI@Meta 2024)	82.27
Llama3-70B-Instruct (AI@Meta 2024)	82.80
BioMistral-7B (Labrak et al. 2024)	81.67
Llama3-OpenBioLLM-8B (Ankit Pal 2024)	82.36

Different Text Encoders

We conducted extensive experiments on various text encoders to verify the effectiveness of our framework. The results in Table 6 demonstrate that all text encoders benefited from the second stage of fine-grained training.

Table 6: Results on different text encoders.

Text Encoder	CLEP	FG-CLEP ↑
BioClinicalBERT (Alsentzer et al. 2019)	79.92	82.27
PubMedBERT (Gu et al. 2021)	80.10	81.87
Med-CPT (Jin et al. 2023)	78.25	80.87
BioBert (Lee et al. 2020)	78.33	80.20

Different ECG Encoders

The results in Table 7 demonstrate that various ECG encoders benefit from our fine-grained pretraining. Larger ResNet models (He et al. 2016) achieve superior performance. However, the ViT structure (Dosovitskiy et al. 2020) is not quite suitable for capturing ECG features.

Table 7: Results on different ECG encoders.

ECG Encoder	CLEP	FG-CLEP ↑
resnet18	79.70	81.26
resnet50	79.92	82.27
resnet152	80.35	82.98
ViT	76.20	78.12

ECG-Text Retrieval

We attempted to use FG-CLEP to retrieve electrocardiograms (ECGs) from the MIMIC-ECG dataset (Gow et al.) through text. To test our model's ability to capture fine-grained waveform features, we tested a series of typical waveform features such as 'RSR' Pattern,' 'Inverted T-waves,' and 'Low QRS voltages.' Figure 7 shows the Top 3 retrieved ECGs with probabilities all greater than 0.99. Our model demonstrated strong capability in retrieving ECGs through waveform feature text.

Conclusion

In this work, we present the Fine-Grained Contrastive Language Electrocardiogram Pre-training (FG-CLEP) framework. We utilized a two-stage pre-training strategy, leveraging a large language model and Stage 1 model to obtain fine-grained reports with waveform features. These fine-grained reports were then used to continue training the model, enabling it to capture fine-grained waveform features. Moreover, we introduced a sigmoid-based loss function and a semantic similarity matrix to better adapt to ECG multi-label downstream tasks and mitigate the frequent false negatives caused by long-tail distributions. Extensive experiments demonstrated FG-CLEP's superior performance in zero-shot prediction and linear evaluation tasks, confirming its robustness and effectiveness.

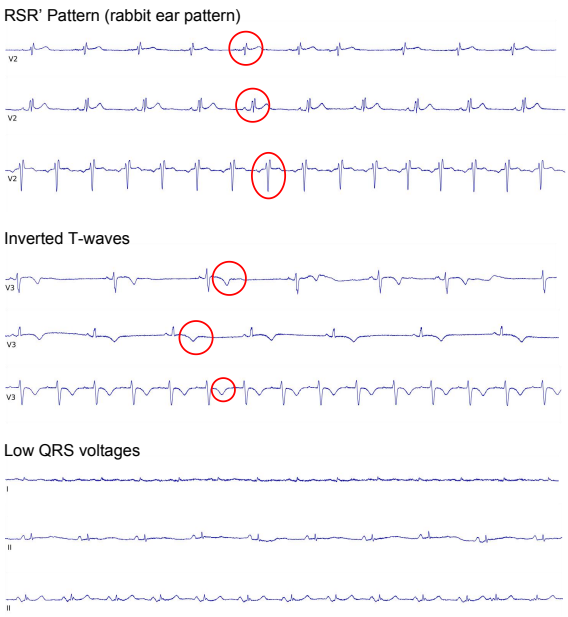

Figure 7: Top 3 retrieved ECG using FG-CLEP.

References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. Llama 3 Model Card.

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Ankit Pal, M. S. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Ayano, Y. M.; Schwenker, F.; Dufera, B. D.; and Debelee, T. G. 2022. Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review. *Diagnostics*, 13(1): 111.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9640–9649.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.

Gow, B.; Pollard, T.; Nathanson, L. A.; Johnson, A.; Moody, B.; Fernandes, C.; Greenbaum, N.; Berkowitz, S.; Moukheiber, D.; Eslami, P.; et al. ???? MIMIC-IV-ECG-Diagnostic Electrocardiogram Matched Subset.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Günay, S.; Öztürk, A.; and Yiğit, Y. 2024. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists. *The American journal of emergency medicine*, 84: 68–73.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, R.; Chen, J.; and Zhou, L. 2023. Spatiotemporal self-supervised representation learning from multi-lead ECG signals. *Biomedical Signal Processing and Control*, 84: 104772.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Huang, S.-C.; Shen, L.; Lungren, M. P.; and Yeung, S. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3942–3951.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023a. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, C.; Ye, W.; Xu, H.; Zhang, S.; Zhang, J.; Huang, F.; et al. 2023b. Vision Lanauge Pre-training by Contrastive Learning with Cross-Modal Similarity Regulation. *arXiv preprint arXiv:2305.04474*.

Jin, J. 2018. Screening for cardiovascular disease risk with ECG. *Jama*, 319(22): 2346–2346.

Jin, Q.; Kim, W.; Chen, Q.; Comeau, D. C.; Yeganova, L.; Wilbur, W. J.; and Lu, Z. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11): btad651.

Kiyasseh, D.; Zhu, T.; and Clifton, D. A. 2021. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, 5606–5615. PMLR.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Lalam, S. K.; Kunderu, H. K.; Ghosh, S.; Kumar, H.; Awasthi, S.; Prasad, A.; Lopez-Jimenez, F.; Attia, Z. I.; Asirvatham, S.; Friedman, P.; et al. 2023. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*.

Lavoie, S.; Kirichenko, P.; Ibrahim, M.; Assran, M.; Wildon, A. G.; Courville, A.; and Ballas, N. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, J.; Liu, C.; Cheng, S.; Arcucci, R.; and Hong, S. 2024. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, 402–415. PMLR.

Li, Z.; Guo, C.; Feng, Z.; Hwang, J.-N.; and Du, Z. 2023. Integrating language guidance into image-text matching for correcting false negatives. *IEEE Transactions on Multimedia*.

Liu, C.; Wan, Z.; Cheng, S.; Zhang, M.; and Arcucci, R. 2024a. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8230–8234. IEEE.

Liu, C.; Wan, Z.; Ouyang, C.; Shah, A.; Bai, W.; and Arcucci, R. 2024b. Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement. *arXiv preprint arXiv:2403.06659*.

Liu, F.; Liu, C.; Zhao, L.; Zhang, X.; Wu, X.; Xu, X.; Liu, Y.; Ma, C.; Wei, S.; He, Z.; et al. 2018. 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.

Na, Y.; Park, M.; Tae, Y.; and Joo, S. 2024. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. *arXiv preprint arXiv:2402.09450*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rath, A.; Mishra, D.; Panda, G.; and Satapathy, S. C. 2021. Heart disease detection using deep learning methods from imbalanced ECG samples. *Biomedical Signal Processing and Control*, 68: 102820.

Sahoo, S.; Dash, M.; Behera, S.; and Sabut, S. 2020. Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey. *Irbm*, 41(4): 185–194.

Sun, W.; Zhang, J.; Wang, J.; Liu, Z.; Zhong, Y.; Feng, T.; Guo, Y.; Zhang, Y.; and Barnes, N. 2023. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6420–6429.

Thai, N.; Nghia, N.; Binh, D.; Hai, N.; and Hung, N. 2017. Long-tail effect on ECG classification. In *2017 International Conference on System Science and Engineering (ICSSE)*, 34–38. IEEE.

Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Kreiseler, D.; Lunze, F. I.; Samek, W.; and Schaeffter, T. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*, 7(1): 154.

Wang, N.; Feng, P.; Ge, Z.; Zhou, Y.; Zhou, B.; and Wang, Z. 2023. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yogarajan, V.; Pfahringer, B.; Smith, T.; and Montiel, J. 2021. Improving Predictions of Tail-end Labels using Concatenated BioMed-Transformers for Long Medical Documents. *arXiv preprint arXiv:2112.01718*.

Yu, H.; Guo, P.; and Sano, A. 2024. ECG Semantic Integrator (ESI): A Foundation ECG Model Pretrained with LLM-Enhanced Cardiological Text. *arXiv preprint arXiv:2405.19366*.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, 12310–12320. PMLR.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, H.; Liu, W.; Shi, J.; Chang, S.; Wang, H.; He, J.; and Huang, Q. 2022a. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–15.

Zhang, W.; Yang, L.; Geng, S.; and Hong, S. 2022b. Self-Supervised Time Series Representation Learning via Cross Reconstruction Transformer. *arXiv preprint arXiv:2205.09928*.

Zhang, W.; Yang, L.; Geng, S.; and Hong, S. 2023. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022c. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25. PMLR.

Zheng, J.; Guo, H.; and Chu, H. 2022. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022Available online: http://physionet.org/content/ecg-arrhythmia/1.0. 0/(accessed on 23 November 2022)*.