# Knowledge-enhanced Multimodal ECG Representation Learning with Arbitrary-Lead Inputs

**Anonymous ACL submission**


## Abstract

Recent advances in multimodal ECG representation learning center on aligning ECG signals with paired free-text reports. However, suboptimal alignment persists due to the complexity of medical language and the reliance on a full 12-lead setup, which is often unavailable in under-resourced settings. To tackle these issues, we propose **K-MERL**, a knowledge-enhanced multimodal ECG representation learning framework. **K-MERL** leverages large language models to extract structured knowledge from free-text reports and employs a lead-aware ECG encoder with dynamic lead masking to accommodate arbitrary lead inputs. Evaluations on six external ECG datasets show that **K-MERL** achieves state-of-the-art performance in zero-shot classification and linear probing tasks, while delivering an average **16%** AUC improvement over existing methods in partial-lead zero-shot classification[1].


## 1 Introduction

Recent advancements in deep learning have enabled automated classification of cardiovascular disease (CVD) using electrocardiograms (ECGs), one of the most crucial diagnostic tools. However, most methods are supervised, requiring large amounts of annotated data, which is costly and demands prohibitively extensive expert effort in annotation (Liu et al., 2023a; Huang and Yen, 2022). To address this challenge, self-supervised multimodal learning has recently emerged as an effective approach for learning representative ECG features from accompanied free-text clinical reports (Li et al., 2023; Pham et al., 2024; Liu et al., 2024). To this end, MERL (Liu et al., 2024) recently introduced the first comprehensive benchmark using the largest dataset MIMIC-ECG (Gow et al.) for pre-training, and six datasets (Wagner et al., 2020; Liu et al., 2018; Zheng et al., 2022, 2020) for evaluating downstream task performance, including zero-shot classification and linear probing.

Despite outperforming signal-only self-supervised approaches, multi-modal approaches, including MERL (Liu et al., 2024), still have notable drawbacks: They directly align ECG signals with reports, introducing unnecessary noise due to the free-text nature of the reports, and failing to fully exploit the rich cardiac knowledge contained within the text. Additionally, they encode ECG in a lead-agnostic manner, overlooking the unique spatial and temporal characteristics of the individual 12 ECG leads. Moreover, they require all 12 leads to be available as input, limiting their ability to generalize across different lead combinations. This raises important practical concerns since full 12-lead ECG data is not always available in clinical environments due to factors such as patient mobility issues, the need for rapid assessments in emergencies, and limited resource in pre-hospital care environments (Bray et al., 2021; Swor et al., 2006; Quinn et al., 2020; Nonogi et al., 2008; Kotelnik et al., 2021; Zhang and Frick, 2019; Nonogi et al., 2008).

To overcome the challenges listed above, we make the following contributions: **(1)** We propose a framework dubbed **K**nowledge-enhanced **E**CG **M**ultimodal **R**epresentation **L**earning (**K-MERL**), which extracts cardiac-related entities from free-text ECG reports, converting unstructured reports into structured knowledge to enhance self-supervised ECG multimodal learning. To the best of our knowledge, this is the first work to leverage structured cardiac entities extracted from clinical reports to improve ECG multimodal learning. **(2)** To effectively capture and leverage the lead-specific spatial and temporal characteristics of 12-lead ECGs, we explore various tokenization and positional embedding techniques. In particular, we design *lead-specific tokenization* and *lead-specific spatial positional embeddings*, enabling the frame-

[1] All data and code will be released upon acceptance.

work to capture the distinctiveness of each lead. **(3)** To enable our framework to handle arbitrary combinations of input leads, we introduce a *dynamic lead masking* strategy. In addition, we propose an *independent segment masking* strategy to further capture lead-specific temporal patterns. **(4)** Our K-MERL framework demonstrates superior performance in zero-shot classification and linear probing on multiple downstream datasets in various lead combinations, from a single lead to all 12 leads.

## 2 Method

### 2.1 Overview

To this end, we first utilize a general-purpose *open-source* large language model (LLM), such as Llama3.1 (AI@Meta, 2024), without domain-specific fine-tuning, to extract cardiac-related entities from free-text ECG reports.[2] This makes our approach adaptable and well-positioned to benefit from future advancements in LLMs. Additionally, we design a lead-aware ECG encoder with *lead and segment masking* strategies, allowing the model to handle arbitrary lead inputs while capturing lead-specific spatial-temporal patterns.

Our overall framework is illustrated in Fig 1(b), shown together with the previous state-of-the-art MERL that is based on naive cross-modal contrastive learning (Liu et al., 2024), in Fig 1(a). While both approaches utilize contrastive learning with an ECG signal encoder $\mathcal{F}_{\mathrm{E}}$ processing signal inputs and a text encoder $\mathcal{F}_{\mathrm{T}}$ processing reports, our method introduces substantial innovations, including lead-specific processing, dynamic masking strategies, and the extraction of cardiac-related entities from free-text reports, significantly enhancing ECG multimodal learning.

In the following sections, we introduce the model framework and lead-specific processing in Sec 2.2, followed by the proposed masking strategies in Sec 2.3. We then describe the pipeline for extracting cardiac-related entities as structured knowledge from ECG reports in Sec 2.4. Finally, in Sec 2.5, we explain the knowledge-enhanced ECG multimodal learning process, a synergy of the aforementioned components.

[2]Entity extraction is inherently simpler than high-level text comprehension in specialized domains, and has been shown effective with general-purpose LLMs (Zhang et al., 2023b).

### 2.2 Lead-specific Processing

To begin with, we define the symbols used in our framework: Given a training dataset $\mathcal{X}$ consisting of $N$ ECG-report pairs, we represent each pair as $(\mathbf{e}_i^l, \mathbf{t}_i)$, where $\mathbf{e}_i^l \in \mathcal{E}$ denotes the raw 12-lead ECG signals for lead $l \in \{1, 2, 3, \ldots, 12\}$ of the $i$-th subject ($i = 1, 2, 3, \ldots, N$), and $\mathbf{t}_i \in \mathcal{T}$ represents the associated free-text report. We then perform lead-specific processing, as illustrated in Fig 2.

**Lead-specific Tokenization.** Consider an input ECG signal $e_i^l$ with 12 leads and a signal length denoted by $S$. We split the time-series signal into $M$ non-overlapping segments, each segment of length $\frac{S}{M}$, and perform tokenization for them. In this way, each lead ECG is projected into a sequence of tokens:

$$e_i^l\left[p_1\right], e_i^l\left[p_2\right], e_i^l\left[p_3\right], \ldots, e_i^l\left[p_M\right] \tag{1}$$

where $e_i^l\left[p_m\right]$ corresponds to the ECG token for the $m$-th segment for lead $l$. For 12 leads, the total number of tokens is $12 \times M$. Unlike MERL (Liu et al., 2024), which generates a single token for a 12-lead ECG temporal segment, we produce tokens separately for each individual lead to capture the lead-specific nature.

**Lead-specific Spatial Positional Embedding.** We apply a learnable linear projection $\mathbf{W} \in \mathbb{R}^{p \times d}$ to each token $e_i^l\left[p_m\right]$. Then, we introduce learnable *lead embeddings* $[\texttt{lead}_1, \ldots, \texttt{lead}_{12}]$, where $\texttt{lead}_l \in \mathbb{R}^d$, to capture the characteristics of each lead. The resulting input sequence can be written as:

$$\begin{aligned} &\texttt{lead}_1 + \mathbf{W}e_i^l[p_1], \ldots, \quad \texttt{lead}_1 + \mathbf{W}e_i^l[p_M], \\ &\cdots\cdots \\ &\texttt{lead}_{12} + \mathbf{W}e_i^l[p_1], \ldots, \quad \texttt{lead}_{12} + \mathbf{W}e_i^l[p_M]. \end{aligned} \tag{2}$$

**Lead-agnostic Temporal Positional Embedding.** In line with lead-specific spatial positional embedding, we also incorporate *learnable lead-agnostic temporal embeddings* to retain the temporal information of ECG signals. These embeddings are denoted as $[\texttt{temp}_1, \ldots, \texttt{temp}_M]$, where $\texttt{temp}_m \in \mathbb{R}^d$. It is worth noting that these positional embeddings are shared across leads, enabling the model to recognize temporal properties across leads, as all leads originate from the same source and share the same temporal domain properties. The resulting input sequence can be written as:

$$\begin{aligned} &\texttt{temp}_1 + \texttt{lead}_1 + \mathbf{W}e_i^l[p_1], \\ &\quad \ldots, \quad \texttt{temp}_M + \texttt{lead}_1 + \mathbf{W}e_i^l[p_M], \\ &\quad \ldots, \quad \texttt{temp}_1 + \texttt{lead}_{12} + \mathbf{W}e_i^l[p_1], \\ &\quad \ldots, \quad \texttt{temp}_M + \texttt{lead}_{12} + \mathbf{W}e_i^l[p_M]. \end{aligned} \tag{3}$$

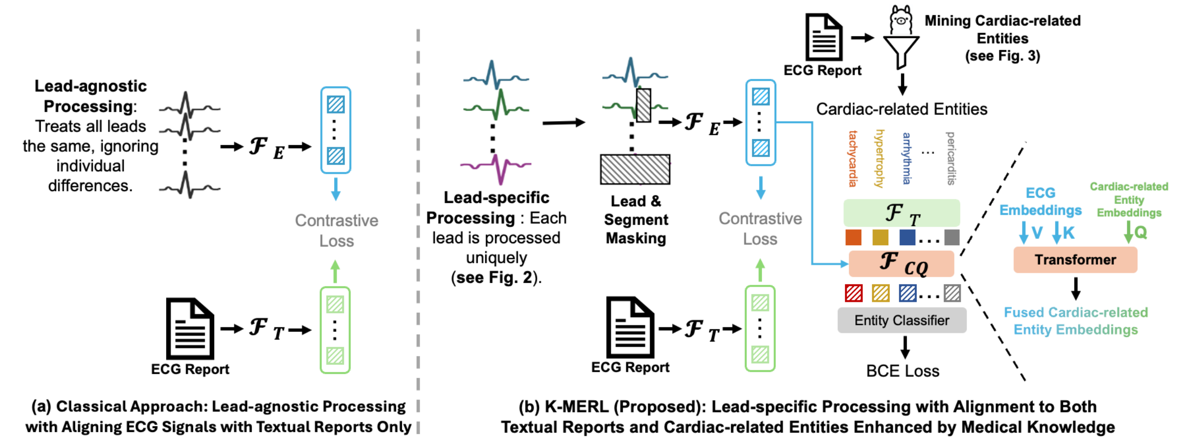



Figure 1: Comparison between classical ECG multimodal learning and our K-MERL framework. **(a):** The classical approaches (e.g., MERL (Liu et al., 2024)) are suboptimal: they processes all leads in a lead-agnostic manner and naively align ECG signals directly free-text reports. **(b):** K-MERL introduces lead-specific processing and lead & segment masking to capture spatial-temporal patterns unique to each lead. It also extracts cardiac-related entities from reports as structured knowledge and aligns them with ECG features to enhance multimodal learning, thereby reducing the complexity introduced by the grammatical structure of free-text reports.

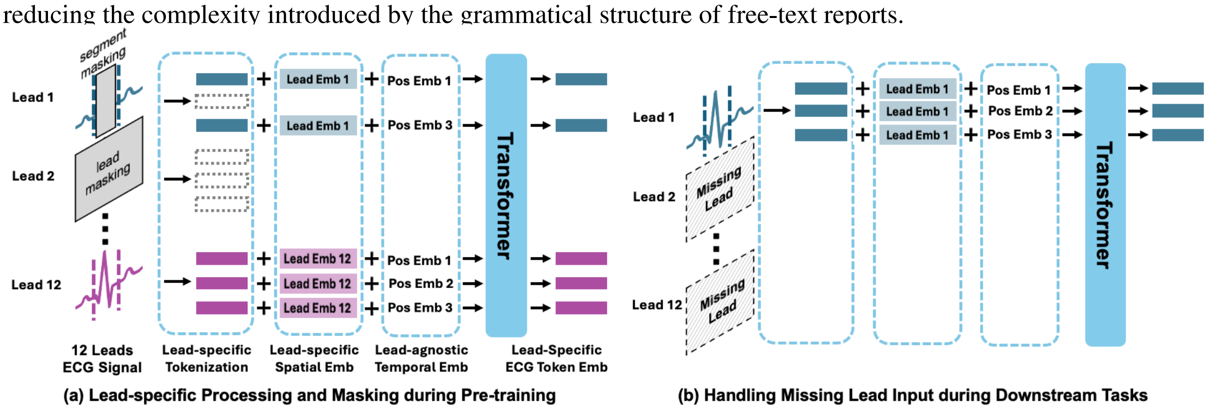



Figure 2: Illustration of our lead-specific processing and handling of partial leads input in K-MERL. **(a):** Lead-specific processing and masking during pre-training. The model employs lead-specific tokenization, spatial embeddings, and lead-agnostic temporal embeddings to capture spatial-temporal patterns for each lead (see Sec 2.2). Dynamic lead masking is used to simulate inputs with arbitrary combinations of leads, while segment masking encourage the framework to captures temporal patterns (see Sec 2.3). **(b):** Handling partial lead input during downstream tasks. When leads are missing, the model processes only the available leads using lead-specific embeddings, allowing maintained performance even with incomplete data.

## 2.3 Lead and Segment Masking

Using a fixed number of masked leads limits the model's flexibility in handling arbitrary lead inputs. To address this, we propose **D**ynamic **L**ead **M**asking **(DLM)**, enabling the model to handle varying lead combinations (Fig. 2 a). For an ECG signal $e_i^l$ with 12 leads, we first randomly sample a number from $\{9, 10, 11\}$, which determines how many leads will be *masked*. Then, we randomly select a set of unmasked lead indices, denoted as $\hat{l}$, and mask the remaining leads. This approach ensures the model is exposed to diverse combinations of unmasked and masked leads during pretraining. The resulting ECG signal with the selected unmasked leads is denoted as $e_i^{\hat{l}}$.

To better capture the temporal patterns of each ECG lead, we introduce **L**ead-independent **S**egment **M**asking **(LSM)** (Fig. 2 a). Applying masking across all tokens from an ECG signal could lead to imbalances, where some leads have more masked tokens than others. To avoid this, LSM applies masking separately to each lead, ensuring an equal number of masked tokens per lead. For each unmasked lead signal $e_i^{\hat{l}}$, we randomly select masked token indices $\mathcal{H}^{\hat{l}}$ based on a masking proportion of 0.25. The model then processes only the unmasked tokens, denoted as $\{e_i^{\hat{l}}[p_h]\}_{h \notin \mathcal{H}^{\hat{l}}}$. We ablate DLM or LSM to verify their effectiveness, as shown in Tab 2c and Fig 7.

## 2.4 Mining Cardiac-related Entities from Report

In this section, we introduce the structured knowledge extraction process for handling free-text ECG

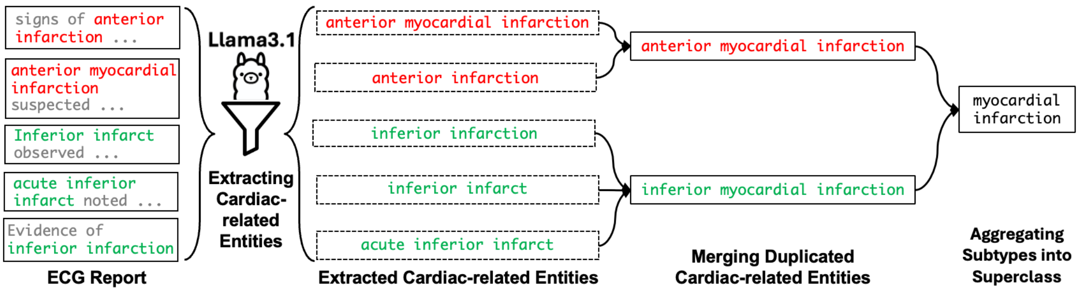



Figure 3: Illustration of mining structured knowledge from free-text reports (see Sec 2.4). First, cardiac-related entities are extracted from free-text ECG reports using an open-source LLM (e.g., Llama3.1-70B-Instruct). Next, we query the LLM to merge duplicated or synonymous cardiac-related entities into a list of unique names. Finally, the LLM detects and aggregates subtypes into their respective superclasses, creating a structured hierarchy of cardiac-related entities.

reports. The pipeline is illustrated in Fig 3. Since each ECG report provides descriptions of cardiac-related entities, as shown in the leftmost part of Fig. 3, our goal is to extract all positive cardiac-related entities mentioned in the report as structured knowledge to enhance the supervision signals for ECG multimodal learning.

**Extracting Cardiac-related Entities.** Unlike existing biomedical multimodal learning approaches from the radiology domain, which rely on knowledge graphs to extract structured knowledge from reports (Zhang et al., 2023b; Wu et al., 2023), we directly query an LLM with the following prompt: 'Please extract all positive Cardiac-related Entities from the given ECG report. Output format is [Entity1, Entity2, ...]'. There are two main reasons for this approach. First, there is no off-the-shelf knowledge graph (KG) specifically focused on ECG, making it impractical to use KG-based methods for extracting structured knowledge. Second, since we are only extracting existing terms from the free-text report, we can easily verify that the extracted cardiac-related entities are present and positive, ensuring no non-existent terms are generated by the LLM. Moreover, (Zhang et al., 2023b) has already demonstrated that a general-purpose LLM can effectively extract existing medical terms from free-text reports independently of any external knowledge database. To ensure accuracy, after each extraction operation, we query the LLM with: 'Please verify the extracted cardiac-related entities as existing and positive in the given report. Output format is YES or NO', and only retain the cardiac-related entities with a 'YES' response. After this stage, we obtain a total of 341 unique cardiac-related entities in the whole dataset..

**Merging Duplicated Cardiac-related Entities.** After extracting all cardiac-related entities from whole dataset, we observe that many names share the same semantics but are expressed differently, as shown in the second part of Fig 3. This variation arises because different clinical protocols generate ECG reports in different styles, even though they describe the same cardiac-related entities. To address this, we query the LLM with: 'Please merge the cardiac-related entities that have the same semantics but different expressions. Here are <all Cardiac-related Entities>. Output format is JSON, where the key is the original name and the value is the merged name.' After this stage, we obtain a total of 252 unique cardiac-related entities in the whole dataset..

**Aggregating Subtypes into Superclasses.** Since cardiac-related entities are organized in a clear hierarchical structure (Arnaout et al., 2016; Okshina et al., 2019), for example, as shown in the rightmost part of Fig 3, 'anterior myocardial infarction' and 'inferior myocardial infarction' are subtypes of the superclass 'Myocardial infarction' (Brieger et al., 2000), we query the LLM with the following prompt: 'Please detect all the superclasses present in <all Cardiac-related Entities>. Output format is JSON, where the key is the superclass name and the values are the cardiac-related entities that belong to this superclass.'

After this stage, we identify 25 superclasses of cardiac-related entities. By the end of the process, we obtain a list of 277 unique cardiac-related entities for the entire dataset. The list of these entities is represented as $\mathcal{Q} = \{q_1, q_2, \ldots, q_Q\}$, where

$Q = 277$. For each ECG report $\mathbf{t}_i$, we create a label vector of length 277, where the positions corresponding to present and positive cardiac-related entity are set to 1, and all other positions are set to 0. This results in a binary label vector for each report, which we denote as $\mathbf{y}_i \in \{0,1\}^{277}$.

### 2.5 Knowledge-enhanced ECG Multimodal Learning

**Aligning ECG and Reports.** In this framework, as shown in Fig 1 (b), two distinct encoders for ECG signals and text reports, symbolized as $\mathcal{F}_\mathrm{E}$ and $\mathcal{F}_\mathrm{T}$, transform the sample pair $(\mathbf{e}_i, \mathbf{t}_i)$ into the latent embedding space, represented as $(\mathbf{z}_{e,i}, \mathbf{z}_{t,i})$. The dataset at the feature level is then denoted as $\mathcal{X} = \{(\mathbf{z}_{e,1}, \mathbf{z}_{t,1}), (\mathbf{z}_{e,2}, \mathbf{z}_{t,2}), \ldots, (\mathbf{z}_{e,N}, \mathbf{z}_{t,N})\}$, where $\mathbf{z}_{e,i} = \mathcal{F}_\mathrm{E}(\mathbf{e}_i)$ and $\mathbf{z}_{t,i} = \mathcal{F}_\mathrm{T}(\mathbf{t}_i)$. Afterward, two non-linear projectors for ECG and text embeddings, denoted as $\mathcal{P}_e$ and $\mathcal{P}_t$, transform $\mathbf{z}_{e,i}$ and $\mathbf{z}_{t,i}$ into the same dimensionality $d$, with $\hat{\mathbf{z}}_{e,i} = \mathcal{P}_e(\mathrm{AvgPool}(\mathbf{z}_{e,i}))$ and $\hat{\mathbf{z}}_{t,i} = \mathcal{P}_t(\mathrm{AvgPool}(\mathbf{z}_{t,i}))$. Next, we compute the cosine similarities as $s^{e2t}_{i,i} = \hat{\mathbf{z}}_{e,i}^\top \hat{\mathbf{z}}_{t,i}$, representing the ECG-report similarities, and formulate the ECG-report contrastive loss $\mathcal{L}_\mathrm{contrast}$.

$$\mathcal{L}^{e2t}_{i,j} = -\log \frac{\exp(s^{e2t}_{i,j}/\tau)}{\sum_{k=1}^{L} \mathbb{1}_{[k \neq i]} \exp(s^{e2t}_{i,k}/\eta)},$$
$$\mathcal{L}_\mathrm{contrast} = \frac{1}{L} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{L}^{e2t}_{i,j}. \tag{4}$$

The temperature hyper-parameter, denoted as $\eta$, is set to 0.07 in our study. $L$ refers to the batch size per training step, which is a subset of $N$.

**Aligning ECG and Cardiac-related Entities.** To learn the knowledge from extracted cardiac-related entities, we design a cardiac query network, denoted as $\mathcal{F}_\mathrm{CQ}$. This network consists of four transformer layers concatenated with a linear classifier that predicts each ECG's corresponding cardiac entity labels $\mathbf{y}_i$. Given the set of cardiac-related entities $\mathcal{Q}$, we compute a corresponding set of cardiac query vectors using the text encoder, denoted as $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_Q\}$, where each query vector $\mathbf{q}_i$ is obtained as $\mathbf{q}_i = \mathcal{F}_\mathrm{T}(q_i)$. These query vectors are then used as inputs for the cardiac query network $\mathcal{F}_\mathrm{CQ}$. During pre-training, the ECG features $\mathbf{z}_{e,i}$ serve as the key and value inputs to the cardiac query network $\mathcal{F}_\mathrm{CQ}$. We use binary cross-entropy (BCE) loss to compute the predictions from $\mathcal{F}_\mathrm{CQ}$ and compare them to the existence labels $\mathbf{y}_i$. The total loss is defined as:

$$\mathcal{L}_\mathrm{CQ} = \frac{1}{L} \sum_{i=1}^{N} \mathrm{BCE}(\mathcal{F}_\mathrm{CQ}(\mathbf{Q}, \mathbf{z}_{e,i}), \mathbf{y}_i),$$
$$\mathcal{L}_\mathrm{total} = \mathcal{L}_\mathrm{contrast} + \mathcal{L}_\mathrm{CQ}. \tag{5}$$

## 3 Experiments

### 3.1 Pre-training Configurations

**MIMIC-ECG.** We pre-train K-MERL using the MIMIC-ECG dataset (Gow et al.), comprising 800,035 ECG-report pairs. Each sample includes a raw ECG signal recorded at 500Hz over a 10-second duration, along with its corresponding report. For fair comparison with the MERL framework (Liu et al., 2024), we adhere to their preprocessing protocol, available in the official GitHub repository[3]. After preprocessing, we obtain 771,693 samples for model pre-training.

**Implementation.** For pre-training, we inherit the settings from MERL (Liu et al., 2024), using a ViT-tiny model as the ECG encoder and Med-CPT (Jin et al., 2023) as the text encoder. The key differences in our approach are the proposed lead-specific tokenizer and spatial-temporal positional embeddings. For extracting cardiac-related entities from the ECG reports, we utilize Llama3.1-70B-Instruct (AI@Meta, 2024) as our main extractor. While we are aware that smaller LLMs, such as those at the 7B scale, can also acceptable good results, we select the 70B model to maximize extraction quality. Additionally, entity extraction is performed only once prior to pretraining, so the computational cost of using a larger model has minimal impact on overall efficiency. Ablation results comparing different LLM extractors are provided in Table 7c. Pre-training configuration details are provided in Sec B.

### 3.2 Downstream Tasks Configurations

We evaluate our framework on both zero-shot classification and linear probing, using full and partial lead ECGs across multiple public datasets covering over 100 cardiac conditions. We adhere to the data split and preprocessing provided by MERL (Liu et al., 2024). The tasks are implemented on the following datasets: **(1) PTBXL:** The PTBXL dataset (Wagner et al., 2020) includes 21,837 ECG signals from 18,885 patients, sampled at 500 Hz for 10 seconds. It provides four subsets for multi-label classification: **Superclass** (5 categories), **Subclass** (23 categories), **Form** (19 categories), and

[3]https://github.com/cheliu-computation/MERL-ICML2024/tree/main

**Rhythm** (12 categories), with varying sample sizes. **(2) CPSC2018:** The CPSC2018 dataset (Liu et al., 2018) contains 6,877 12-lead ECG records, sampled at 500 Hz, annotated with 9 distinct labels. **(3) CSN:** The Chapman-Shaoxing-Ningbo (CSN) dataset (Zheng et al., 2020, 2022) comprises 45,152 ECG records sampled at 500 Hz for 10 seconds. After excluding records with 'unknown' annotations, the final curated dataset includes 23,026 ECG records with 38 labels. Detailed information about the downstream datasets is presented in Tab 3. We introduce the details of the compared methods in Sec. C.

In the downstream tasks, we implement three scenarios: zero-shot classification, linear probing, and partial lead analysis. The implementation details are provided in Sec D.3.

### 3.3 State-of-the-art on Zero-shot Classification

We first evaluate K-MERL on zero-shot classification using 12-lead input across all downstream datasets. The results for each dataset, along with the average AUC score across six datasets, are shown in Fig 4. Our framework significantly outperforms MERL with both backbone architectures, demonstrating the superiority of K-MERL when using the original disease names as text prompts. **State-of-the-art on Unseen Disease Prediction.** Additionally, since we extract cardiac-related entities from reports during pre-training, there may be overlap with categories in downstream tasks. This could provide our model with prior knowledge of certain categories, leading to an unfair comparison with MERL (Liu et al., 2024). To address this, we use Med-CPT (Jin et al., 2023), the text encoder, to extract embeddings for all 277 cardiac-related entities and for all category names in the downstream datasets. We compute the similarity between these embeddings, and if the similarity exceeds 0.95, we consider them overlapped. We identify 35 out of 277 extracted cardiac-related entities that overlap with downstream categories, as listed in Tab 5. We label these as '*Seen Classes*,' while the remaining downstream categories are labeled as '*Unseen Classes*.'

The average F1 score are depicted in Fig 5(b). K-MERL outperforms MERL in both seen and unseen categories. Notably, both K-MERL and MERL exhibit performance drops on unseen classes compared to seen classes, demonstrating that we successfully detected an overlap of approximately 12.7% between the extracted cardiac-related entities from MIMIC-ECG and downstream categories, effectively separating the tasks into 'seen' and 'unseen' groups. The results show that K-MERL performs well not only on categories present during pre-training but also on unseen categories, demonstrating its generalizability. Since the original MERL (Liu et al., 2024) framework relies on manual prompt engineering (PE) at inference time to enhance performance, we also evaluate MERL with customized prompts, as detailed in Sec. F, to provide a comprehensive comparison. Notably, our method outperforms MERL with PE while being entirely independent of prompt engineering.

### 3.4 Performance of Linear Probing

As shown in Tab 1, K-MERL consistently outperforms multimodal methods, including MERL (Liu et al., 2024) with both ResNet and ViT backbones, as well as all eSSL methods across datasets and data ratios. This highlights K-MERL's robust performance and the quality of its learned ECG features, which not only improve multimodal tasks but also significantly enhance single-modality tasks.

### 3.5 Performance with Partial Leads Input

We explore the robustness of K-MERL to missing leads by simulating missing-lead scenarios through progressively adding visible leads starting from a single lead, as there is no publicly available ECG dataset explicitly designed to evaluate missing-lead scenarios.[4] As shown in Fig 6 (a) and (b), K-MERL consistently outperforms MERL across all lead combinations from 1 to 12 in both zero-shot classification and linear probing. Impressively, K-MERL with just a single lead surpasses MERL's performance using all 12 leads. Additionally, K-MERL shows a stable performance trend as the number of leads increases, unlike MERL, which exhibits fluctuations in Fig 6 (a). This demonstrates the effectiveness of our dynamic lead masking strategy, lead-specific processing, and spatial-temporal positional embeddings, contributing to K-MERL's superior results.

## 4 Analysis

This section provides ablation studies on the key components of K-MERL and reports zero-shot classification results for single-lead and 12-lead inputs

[4] In most cases, publicly released datasets undergo strict curation, with samples containing missing leads often excluded as corrupted data, despite their clinical relevance.

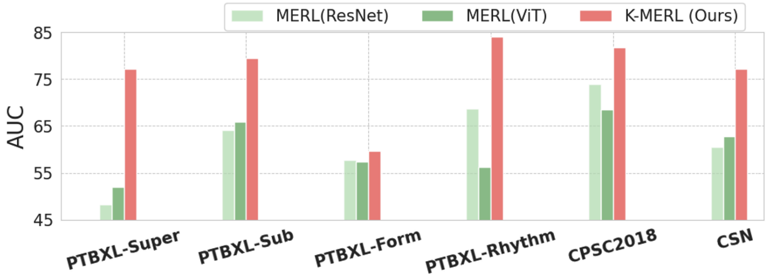
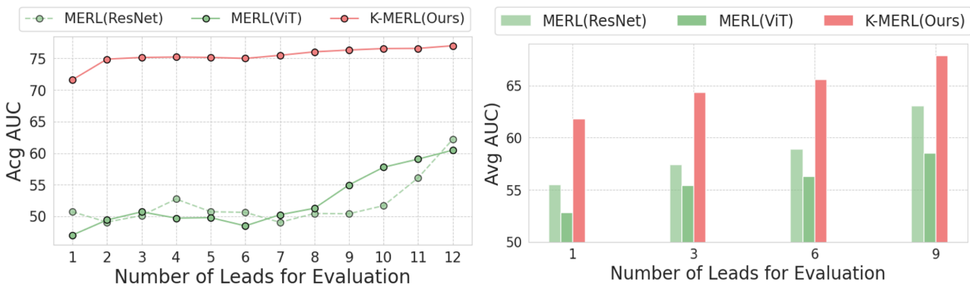



Figure 4: Performance on zero-shot classification across six datasets, comparing K-MERL with previous ECG multimodal learning methods. Notably, we use the original disease category names as prompts for both K-MERL and MERL to ensure a fair comparison.


Figure 5: Comparison of K-MERL and MERL on seen and unseen classes, reporting **(a)** Average AUC and **(b)** Average F1 scores. Definitions are in Sec 3.3.

Table 1: Linear probing results of K-MERL and other ECG learning methods, with best results **bolded**.

| | 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% |
| ***From Scratch*** | | | | | | | | | | | | | | | | | | |
| Random Init (CNN) | 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 |
| Random Init (Transformer) | 70.31 | 75.27 | 77.54 | 53.36 | 67.56 | 77.43 | 53.47 | 61.84 | 72.08 | 45.36 | 60.33 | 77.26 | 52.93 | 68.0 | 77.44 | 45.55 | 60.23 | 71.37 |
| ***ECG only SSL*** | | | | | | | | | | | | | | | | | | |
| 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 |
| ST-MEM | 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 |
| ***Multimodal Methods*** | | | | | | | | | | | | | | | | | | |
| MERL (ResNet) | 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 |
| MERL (ViT) | 78.64 | 83.90 | 85.27 | 61.41 | 77.55 | 82.98 | 56.32 | 69.11 | 77.66 | 52.16 | 78.07 | 81.83 | 69.25 | 82.82 | 89.44 | 63.66 | 78.67 | 84.87 |
| **K-MERL (Ours)** | **84.19** | **87.71** | **89.83** | **68.22** | **81.54** | **88.00** | **60.11** | **73.71** | **81.48** | **63.72** | **84.16** | **91.04** | **71.91** | **86.13** | **91.26** | **69.51** | **83.53** | **93.71** |




**(a) Performance of Zero-shot Classification with Partial Leads Input** **(b) Performance of 1% Data Linear Probing with Partial Lead Input**

Figure 6: Performance comparison of K-MERL and MERL with partial lead inputs. **(a)** Zero-shot classification shows K-MERL consistently outperforming MERL with two backbones across all lead combinations from 1 to 12. **(b)** Linear probing with 1% data demonstrates K-MERL's superior performance and robustness, even with limited data and varying lead inputs.

across all downstream datasets. Due to the page limit, we show more ablation studies in Sec H.

**Loss Ablation.** Tab 2a shows the effect of removing $\mathcal{L}_{\text{contrast}}$ and $\mathcal{L}_{\text{CQ}}$ during pre-training. Removing $\mathcal{L}_{\text{CQ}}$, which excludes structured knowledge from cardiac-related entities, leads to a significant performance drop. While removing $\mathcal{L}_{\text{contrast}}$ also reduces performance, the impact is less severe. This indicates that both losses are necessary, with cardiac-related entities alignment providing a larger benefit for pre-training.

**Tokenization Size.** In Fig 7 (a), we ablate the token size $p$ and find the optimal length to be 100. Larger token sizes (e.g., 200) have a more negative impact than smaller sizes (e.g., 25), likely due to convert multiple segments to one token, which introduces ambiguity. Across all token sizes, K-MERL consistently outperforms MERL (Liu et al., 2024), demonstrating the robustness and effectiveness of our method.

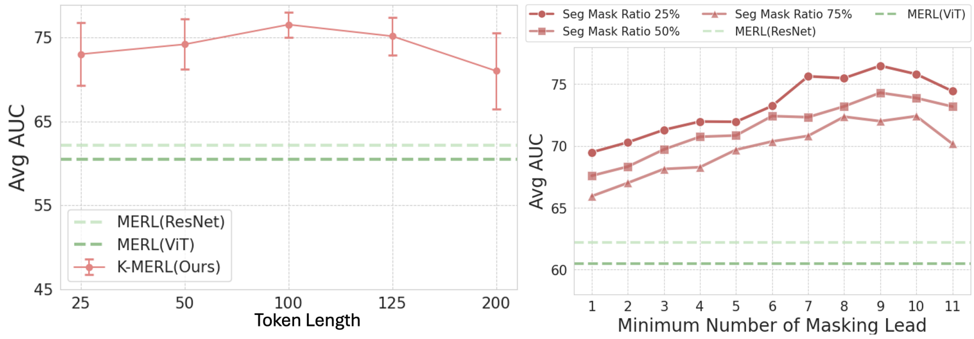



Figure 7: Ablation study on zero-shot classification with 12 leads. **Left:** Performance of K-MERL across varying token lengths, showing optimal results with a token length of 100, consistently outperforming MERL. **Right:** Impact of different segment masking ratios (25%, 50%, 75%) and the minimum number of masked leads. K-MERL outperforms MERL, with the best performance at a 25% mask ratio and a minimum of 9 masked leads.

Table 2: Results of various ablation experiments. The best results are **bolded**.

(a) Ablating Loss Function.

| Loss | 1 Lead | 12 Leads |
|---|---|---|
| K-MERL (Ours) | **71.61** | **76.52** |
| – ECG-Text Alignment ($\mathcal{L}_{\text{contrast}}$) | 69.23 | 73.98 |
| – ECG-Condition Alignment ($\mathcal{L}_{\text{CQ}}$) | 65.44 | 68.95 |

(b) Effects of Entities Processing.

| Methods | 1 Lead | 12 Leads |
|---|---|---|
| K-MERL (Ours) | **71.61** | **76.52** |
| – Subtype Aggregation | 70.11 | 74.62 |
| – Merging Duplicated Patterns | 70.54 | 74.93 |

(c) Effects of Masking Strategy.

| Masking Strategy | 1 Lead | 12 Leads |
|---|---|---|
| K-MERL (Ours) | **71.61** | **76.52** |
| – Lead-independent Segment Masking | 70.32 | 75.21 |
| – Segment Masking | 68.93 | 74.74 |
| – Dynamic Lead Masking | 67.84 | 72.11 |
| – Lead Masking | 65.41 | 69.10 |

**Cardiac-related Entities Processing.** As shown in Tab 2b, both subtype aggregation and merging duplicate entity names improve K-MERL's performance. However, the best results are achieved when both procedures are applied together, indicating they complement each other.

**Masking Strategy and Ratio.** Tab 2c shows the results of various masking strategies, where all approaches enhance K-MERL's performance. Removing dynamic lead masking and using a fixed number of masked leads degrades performance, highlighting its importance. Similarly, omitting lead masking during pre-training causes a sharp drop in zero-shot classification, indicating its role in capturing lead-specific features. Fig 7 (b) explores mask ratios and lead masking. An optimal configuration is identified with a mask ratio of 25% and a minimum of 9 masked leads. Increasing the mask ratio beyond this or using more than 9 leads as the minimum for masking leads to a decrease in performance.

## 5 Related Work

Recent ECG self-supervised learning (eSSL) approaches have explored contrastive (Kiyasseh et al., 2021; Wang et al., 2023) and generative (Zhang et al., 2022; Na et al., 2023; Jin et al., 2024) objectives, but they remain unimodal and often lack clinical context. Multimodal ECG methods (Liu et al., 2024; Li et al., 2023; Yu et al., 2024) typically align signals with text using simplistic prompts or assume fixed lead configurations, limiting generalizability. Drawing inspiration from knowledge-enhanced radiograph-language models (Zhang et al., 2023b; Wu et al., 2023), we introduce structured cardiovascular entities extracted from ECG reports to provide fine-grained supervision without relying on radiology-specific ontologies, preserving clinical distinctions such as myocardial infarction subtypes (Thygesen et al., 2018). Additionally, we address the practical challenge of partial-lead ECG inputs (Alizadeh Meghrazi et al., 2020) by designing dynamic masking and lead-specific processing, enabling our model to operate robustly under varying lead configurations. A more comprehensive review and comparison with prior work are provided in Appendix A.

## 6 Conclusion

We present K-MERL, a knowledge-enhanced ECG multimodal learning framework capable of processing arbitrary lead inputs. First, we mine cardiac-related entities as structured knowledge from ECG free-text reports using a general LLM, without relying on external domain-specific resources. Next, we align ECG features with these cardiac-related entities to integrate this knowledge into the ECG multimodal learning. Additionally, we introduce lead-specific processing and lead&segment masking strategies to capture the spatial-temporal patterns unique to each ECG lead, enabling the model to handle varying lead inputs. Our experiments on six downstream ECG classification tasks, along with extensive ablation studies, demonstrate K-MERL's superior performance over existing ECG representation learning methods.

## Limitation

While K-MERL demonstrates promising results in handling arbitrary lead inputs and integrating knowledge from ECG reports, there are some limitations to consider. The framework's reliance on LLMs for mining cardiac-related entities, though effective, may be limited by the model's ability to capture highly specialized domain knowledge. Due to the lack of publicly available datasets explicitly designed for missing-lead ECG scenarios, we simulate this setting by progressively adding ECG leads starting from a single lead. While this strategy enables controlled evaluation, we hope future research will explore more realistic clinical benchmarks to better validate performance under naturally occurring lead dropouts. Additionally, while our experiments show strong zero-shot and linear probing performance, further evaluation is needed to assess K-MERL's effectiveness in real-world clinical settings, where data quality and noise levels can be more challenging. Future work will focus on enhancing the robustness of knowledge extraction and developing more adaptive strategies for handling diverse ECG data sources.

## References


AI@Meta. 2024. Llama 3 model card.

Milad Alizadeh Meghrazi, Yupeng Tian, Amin Mahnam, Presish Bhattachan, Ladan Eskandarian, Sara Taghizadeh Kakhki, Milos R Popovic, and Milad Lankarany. 2020. Multichannel ecg recording from waist using textile sensors. *BioMedical Engineering OnLine*, 19:1–18.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Rima Arnaout, Gregory Nah, Gregory M Marcus, Zian H Tseng, Elyse Foster, Ian Harris, Punag Divanji, Jeffrey Olgin, Liviu Klein, Juan Gonzalez, et al. 2016. Peripartum cardiomyopathy and hypertensive pregnancy subtypes among 1.6 million pregnancies in california independently predict subsequent incident myocardial infarction, heart failure, and stroke. *Circulation*, 134(suppl_1):A14574–A14574.

Jonathan James Hyett Bray, Elin Fflur Lloyd, Firdaus Adenwalla, Sarah Kelly, Kathie Wareham, and Julian PJ Halcox. 2021. Single-lead ecgs (alivecor) are a feasible, cost-effective and safer alternative to 12-lead ecgs in community diagnosis and monitoring of atrial fibrillation. *BMJ open quality*, 10(1):e001270.

David B Brieger, Koon-Hou Mak, David P Miller, Robert M Califf, Eric J Topol, GUSTO-I Investigators, et al. 2000. Hierarchy of risk based on history and location of prior myocardial infarction in the thrombolytic era. *American Heart Journal*, 140(1):29–33.

Bhavin Chamadiya, Kunal Mankodiya, Manfred Wagner, and Ulrich G Hofmann. 2013. Textile-based, contactless ecg monitoring for non-icu clinical settings. *Journal of Ambient Intelligence and Humanized Computing*, 4:791–800.

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

X Chen, S Xie, and K He. 2021. An empirical study of training self-supervised vision transformers. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629.

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

Seokmin Choi, Sajad Mousavi, Phillip Si, Haben G Yhdego, Fatemeh Khadem, and Fatemeh Afghah. 2023. Ecgbert: Understanding hidden language of ecgs with self-supervised representation learning. *arXiv preprint arXiv:2306.06340*.

Ming Dai, Xueliang Xiao, Xin Chen, Haoming Lin, Wanqing Wu, and Siping Chen. 2016. A low-power and miniaturized electrocardiograph data collection system with smart textile electrodes for monitoring of cardiac function. *Australasian physical & engineering sciences in medicine*, 39:1029–1040.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. 2024. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12902–12915.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.

Piero Fontana, Neusa R Adão Martins, Martin Camenzind, René M Rossi, Florent Baty, Maximilian Boesch, Otto D Schoch, Martin H Brutsche, and Simon Annaheim. 2019. Clinical applicability of a textile 1-lead ecg device for overnight monitoring. *Sensors*, 19(11):2436.

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.

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. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Xin Huang, Yu Fang, Mingming Lu, Fengqi Yan, Jun Yang, and Yilu Xu. 2020. Dual-ray net: automatic diagnosis of thoracic diseases using frontal and lateral chest x-rays. *Journal of Medical Imaging and Health Informatics*, 10(2):348–355.

Yu Huang and Yen. 2022. Snippet policy network v2: Knee-guided neuroevolution for multi-lead ecg early classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Mary Jahrsdoerfer, Karen Giuliano, and Dean Stephens. 2005. Clinical usefulness of the easi 12-lead continuous electrocardiographic monitoring system. *Critical care nurse*, 25(5):28–37.

Jiarui Jin, Haoyu Wang, Jun Li, Sichao Huang, Jiahui Pan, and Shenda Hong. 2024. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

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

Valdery Moura Junior, Matthew Reyna, Shenda Hong, Aditya Gupta, Manohar Ghanta, Reza Sameni, Jonathan Rosand, Aaron Aguirre, Li Qiao, Gari Clifford, and M. Brandon Westover. 2023. Harvard-emory ecg database (version 1.0).

Talmadge E King Jr. 2017. Approach to the adult with interstitial lung disease: Diagnostic testing. *UpToDate. Waltham, MA. Accessed on: March*, 25.

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

Vladimir Kotelnik, Kevin Pesce, William M Masterton, Robert T Marshall, Gregson Pigott, Nathaniel Bialek, Jason Winslow, and Lauren M Maloney. 2021. 12-lead electrocardiograms acquired and transmitted by emergency medical technicians are of diagnostic quality and positively impact patient care. *Prehospital and Disaster Medicine*, 36(1):47–50.

Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. 2023. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications*, 14(1):3741.

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar A, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, Rakesh Barve, and Melwin Babu. 2023. ECG representation learning with multi-modal EHR data. *Transactions on Machine Learning Research*.

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. 2023. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*.

Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. 2019. Attention based glaucoma detection: A large-scale database and cnn model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580.

Che Liu, Sibo Cheng, Weiping Ding, and Rossella Arcucci. 2023a. Spectral cross-domain neural network with soft-adaptive threshold spectral enhancement. *arXiv preprint arXiv:2301.10171*.

Che Liu, Cheng Ouyang, Sibo Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2023b. G2d: From global to dense radiography representation learning via vision-language pre-training. *arXiv preprint arXiv:2312.01522*.

Che Liu, Zhongwei Wan, Sibo Cheng, Mi Zhang, and Rossella Arcucci. 2023c. Etp: Learning transferable ecg representations via ecg-text pre-training. *arXiv preprint arXiv:2309.07145*.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, 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.

John E Madias. 2003. A comparison of 2-lead, 6-lead, and 12-lead ecgs in patients with changing edematous states: implications for the employment of quantitative electrocardiography in research and clinical applications. *Chest*, 124(6):2057–2063.

Sidharth Maheshwari, Amit Acharyya, Pachamuthu Rajalakshmi, Paolo Emilio Puddu, and Michele Schiariti. 2014. Accurate and reliable 3-lead to 12-lead ecg reconstruction methodology for remote health monitoring applications. *IRBM*, 35(6):341–350.

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. 2024. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2023. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *The Twelfth International Conference on Learning Representations*.

Hiroshi Nonogi, Hiroyuki Yokoyama, Yoritaka Otsuka, Yoichiro Kasahara, Yu Kataoka, Nobuaki Kokubu, and Kazuhiro Sase. 2008. Abstract p182: Usefulness of mobile telemedicine system in real-time transmission of out-of-hospital 12-lead ecg.

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

EY Okshina, MM Lukiyanov, OM Drapkina, VG Klyashtorny, EV Kudryashov, EN Belova, AD Deev, AN Makoveeva, and SA Boytsov. 2019. P5352 the main characteristics and multimorbidity in patients with history of stroke, myocardial infarction and their combination (hospital registry data). *European Heart Journal*, 40(Supplement_1):ehz746–0319.

Manh Pham, Aaqib Saeed, and Dong Ma. 2024. C-melt: Contrastive enhanced masked auto-encoders for ecg-language pre-training. *arXiv preprint arXiv:2410.02131*.

Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. 2024. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501.

T Quinn, A Watkins, C Hampton, M Halter, C Weston, CP Gale, L Gavalova, T Driscoll, G Davies, and HA Snooks. 2020. Has the proportion of patients diagnosed with myocardial infarction that receives a 12 ecg in the prehospital setting in the uk changed over time? *European Heart Journal*, 41(Supplement_2):ehaa946–1650.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. 2020. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760.

Veer Sangha, Akshay Khunte, Gregory Holste, Bobak J Mortazavi, Zhangyang Wang, Evangelos K Oikonomou, and Rohan Khera. 2024. Biometric contrastive learning for data-efficient deep learning from electrocardiographic images. *Journal of the American Medical Informatics Association*, 31(4):855–865.

Shinnosuke Sawano, Satoshi Kodera, Hirotoshi Takeuchi, Issei Sukeda, Susumu Katsushika, and Issei Komuro. 2022. Masked autoencoder-based self-supervised learning for electrocardiograms to detect left ventricular systolic dysfunction. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*.

Robert Swor, Stacey Hegerberg, Ann McHugh-McNally, Mark Goldstein, and Christine C McEachin. 2006. Prehospital 12-lead ecg: efficacy or effectiveness? *Prehospital Emergency Care*, 10(3):374–377.

Kristian Thygesen, Joseph S Alpert, Allan S Jaffe, Bernard R Chaitman, Jeroen J Bax, David A Morrow, Harvey D White, and Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. 2018. Fourth universal definition of myocardial infarction (2018). *Circulation*, 138(20):e618–e651.

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

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

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*.

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

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training in radiology. *arXiv preprint arXiv:2301.02228*.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. 2021. Chest imagenome dataset. *Physio Net*.

Han Yu, Peikun Guo, and Akane Sano. 2024. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text. *arXiv preprint arXiv:2405.19366*.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. 2022. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15.

Qingxue Zhang and Kyle Frick. 2019. All-ecg: A least-number of leads ecg monitor for standard 12-lead ecg tracking during motion. In *2019 IEEE Healthcare Innovations and Point of Care Technologies,(HI-POCT)*, pages 103–106. IEEE.

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

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542.

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

J Zheng, H Guo, and H Chu. 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)*.

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. 2020. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1):2898.

Shuang Zhou, Xiao Huang, Ninghao Liu, Wen Zhang, Yuan-Ting Zhang, and Fu-Lai Chung. 2024. Open-world electrocardiogram classification via domain knowledge-driven contrastive learning. *Neural Networks*, 179:106551.

## A Related Work

### A.1 ECG Representation Learning

Recently, ECG self-supervised learning (eSSL) has shown promise in learning ECG representations from unannotated signals (Lai et al., 2023; Chen et al., 2020; Sangha et al., 2024). Contrastive methods such as CLOCS (Kiyasseh et al., 2021) and ASTCL (Wang et al., 2023) explore temporal and spatial invariance, while generative techniques (Zhang et al., 2022; Sawano et al., 2022; Na et al., 2023; Jin et al., 2024; Choi et al., 2023; Oh et al., 2022; McKeen et al., 2024) focus on masked segment reconstruction. However, both approaches often lack clinical domain knowledge and are limited to single-modality settings, restricting the quality of learned representations.

Multimodal learning has shown success in multiple biomedical applications (Wan et al., 2023; Liu et al., 2023b; Wu et al., 2023). However, ECG signals pose unique challenges due to their complex spatial-temporal structure, necessitating well-tailored modeling. As a result, few studies have explored multimodal ECG learning. (Lalam et al., 2023; Yu et al., 2024) demonstrated the effectiveness of combining ECG and EHR data using large language models (LLMs) to rewrite textual reports. However, their work is restricted to private datasets, making reproducing and comparisons challenging. Other works such as (Li et al., 2023; Liu et al., 2023c; Zhou et al., 2024) explored multimodal ECG learning for zero-shot classification. However, their methods were over simplistic: They align signals with text without sufficiently capturing the distinctiveness of individual ECG leads, and rely on naive category names as prompts, which fail to capture relative patterns, leading to suboptimal performance. Their limited evaluations on small datasets also fall short of fully assessing multimodal ECG learning in real-world scenarios. Additionally, works such as (Zhao et al., 2024; Wan et al., 2024) focus on ECG-to-text generation tasks, but their results are not publicly accessible, making reproducing and comparisons difficult.

MERL (Liu et al., 2024) is the first open-source study to demonstrate the potential of ECG multimodal learning in zero-shot classification and linear probing across diverse datasets. Therefore, we mainly compare our work to MERL. However, like other methods, MERL relies on all 12 ECG leads as input and cannot handle arbitrary lead combinations, limiting its applicability in real-world clinical scenarios where all 12 leads may not always be available (Jahrsdoerfer et al., 2005; Madias, 2003; Fontana et al., 2019; Maheshwari et al., 2014)

### A.2 Knowledge Enhanced Medical Multimodal Learning

Leveraging medical knowledge to improve medical multimodal learning has advanced significantly, particularly in the radiograph domain, with methods like MedKLIP, KAD, and MAVL (Zhang et al., 2023b; Wu et al., 2023; Phan et al., 2024). These approaches focus on extracting structured knowledge, such as clinical entities from free-text radiology reports, and using this information as an additional supervisory signal to guide multimodal learning. Many models mimic radiological practices or modify structures based on diagnostic routines (Li et al., 2019; Huang et al., 2020; Zhang et al., 2023b; Wu et al., 2023). However, they rely heavily on well-annotated knowledge graphs, such as RadGraph (Delbrouck et al., 2024) and Chest ImaGenome (Wu et al., 2021), which require substantial human annotation and are limited to the radiology domain. Due to the distinct nature of ECG signals compared to radiographs, the above pipelines cannot be directly adapted for ECG multimodal learning. Furthermore, CVD has a clear hierarchical structure because conditions can have multiple subtypes, such as myocardial infarction, which can be further classified as inferior or anterior myocardial infarction (Thygesen et al., 2018). Unlike lung diseases, typically categorized by morphological or pathological patterns rather than distinct region based subtypes (King Jr, 2017), directly using only the entity from an ECG report can lead to information loss by ignoring the superclass or subtypes.

### A.3 Challenge in Partial Leads ECG Input

Currently, full 12 leads ECG data dominates publicly accessible ECG datasets (Gow et al.; Ribeiro et al., 2020; Junior et al., 2023). However, in real clinical scenarios, obtaining a standard 12 leads ECG can be excessive and often requires advanced clinical knowledge, which may not always be readily available (Chamadiya et al., 2013; Alizadeh Meghrazi et al., 2020; Dai et al., 2016). This makes partial-lead ECG data both crucial and common for practical applications. Despite its importance, partial leads issue is often overlooked and remain unaddressed in existing ECG multimodal representation learning studies. To handle partial lead inputs across various downstream tasks, in this work, we design lead-specific processing and dynamic lead masking strategies that enable our model to accept any combination of ECG leads as input. adaptable to various clinical scenarios (Jahrsdoerfer et al., 2005; Madias, 2003; Fontana et al., 2019; Maheshwari et al., 2014). We evaluate our model on extensive downstream tasks with partial lead inputs, demonstrating its ability to recognize and adapt to the lead-specific nature of ECG signals.

# B Pre-training Configuration

Following MERL (Liu et al., 2024), we employ the AdamW optimizer with a learning rate of $2 \times 10^{-4}$ and a weight decay of $1 \times 10^{-5}$. Pre-training runs for 50 epochs, with a cosine annealing scheduler for learning rate adjustments. We use a batch size of 512 per GPU, with all experiments conducted on eight NVIDIA A100-80GB GPUs.

# C Baseline Details

In this work, we compare against both multimodal and unimodal ECG representation learning methods for a comprehensive evaluation. For multimodal methods, we compare with MERL (Liu et al., 2024), as it is the first multimodal ECG representation learning framework that publicly released both code and pretrained models. For unimodal methods, following the setup in (Liu et al., 2024), we compare with general self-supervised learning (SSL) methods such as SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), Barlow Twins (Zbontar et al., 2021), MoCo v3 (Chen et al., 2021), and SimSiam (Chen and He, 2021), as well as TS-TCC (Eldele et al., 2021), a time-series-specific SSL baseline. Furthermore, we include comparisons with ECG-specific SSL methods including CLOCS, ASTCL, CRT, and ST-MEM (Kiyasseh et al., 2021; Wang et al., 2023; Zhang et al., 2023a; Na et al., 2023). All results are directly taken from the original MERL (Liu et al., 2024) paper to ensure a fair comparison.

# D Downstream Task Details

## D.1 Downstream Task Data Split

We detail the data splits in Tab. 3. For all datasets, we follow the splits provided by MERL[5]. The preprocessing for all datasets is also done using MERL's official codebase[6].

[5]https://github.com/cheliu-computation/MERL-ICML2024/tree/main/finetune/data_split

[6]https://github.com/cheliu-computation/MERL-ICML2024/tree/main/finetune

## D.2 Downstream Task Configuration

We detail the key hyperparameters used across all downstream tasks in Tab. 4. For each dataset (PTBXL-Super, PTBXL-Sub, PTBXL-Form, PTBXL-Rhythm, CPSC2018, and CSN), we maintain consistency in the learning rate, batch size, number of epochs, and optimizer configuration with MERL (Liu et al., 2024).

## D.3 Downstream Tasks Implementation

**Zero-shot Classification.** For zero-shot classification, we freeze the entire model and use the original category names from the dataset as entity queries $\mathcal{Q}$ for input to the cardiac query network, $\mathcal{F}_{\mathrm{CQ}}$. The ECG signals are converted into ECG feature with $\mathcal{F}_{\mathrm{E}}$, serving as the key and value inputs for $\mathcal{F}_{\mathrm{CQ}}$. The output of $\mathcal{F}_{\mathrm{CQ}}$ provides the predicted probabilities for each category.
**Linear Probing.** For linear probing, we keep the ECG encoder $\mathcal{F}_{\mathrm{E}}$ frozen and only update the parameters of a randomly initialized linear classifier. We conduct linear probing with $\{1\%, 10\%, 100\%\}$ of the training data. This configuration is used consistently across all linear probing tasks. Further implementation details are provided in the Tab 4.
**Partial Lead Setting.** In the partial lead setting, we follow the lead order from the MIMIC-ECG dataset (Gow et al.): [I, II, III, aVF, aVR, aVL, V1, V2, V3, V4, V5, V6], progressively expanding the input from a single lead to all 12 leads in sequence. In contrast, since MERL (Liu et al., 2024) requires a full 12-lead input, we pad the missing leads with zeros to maintain the 12-lead format.

## D.4 Overlapped Categories

As described in Sec 3.3 and Fig 5, we observe that 35 categories are present in both the pre-training and downstream datasets, and we list all the class names in Tab 5.

# E Standard Deviation Comparison with MERL

We provide standard deviations in Tab. 6 to support the performance claims more rigorously. Notably, we report all results using the original disease

Table 3: Details on Data Split.

| Dataset | Number of Categories | Train | Valid | Test |
|---|---|---|---|---|
| PTBXL-Super (Wagner et al., 2020) | 5 | 17,084 | 2,146 | 2,158 |
| PTBXL-Sub (Wagner et al., 2020) | 23 | 17,084 | 2,146 | 2,158 |
| PTBXL-Form (Wagner et al., 2020) | 19 | 7,197 | 901 | 880 |
| PTBXL-Rhythm (Wagner et al., 2020) | 12 | 16,832 | 2,100 | 2,098 |
| CPSC2018 (Liu et al., 2018) | 9 | 4,950 | 551 | 1,376 |
| CSN (Zheng et al., 2022, 2020) | 38 | 16,546 | 1,860 | 4,620 |

Table 4: Hyperparameter settings on downstream tasks.

| | PTBXL-Super | PTBXL-Sub | PTBXL-Form | PTBXL-Rhythm | CPSC2018 | CSN |
|---|---|---|---|---|---|---|
| Learning rate | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| Batch size | 16 | 16 | 16 | 16 | 16 | 16 |
| Epochs | 100 | 100 | 100 | 100 | 100 | 100 |
| Optimizer | AdamW | AdamW | AdamW | AdamW | AdamW | AdamW |
| Learing rate scheduler | Cosine anealing | Cosine anealing | Cosine anealing | Cosine anealing | Cosine anealing | Cosine anealing |
| Warump steps | 5 | 5 | 5 | 5 | 5 | 5 |

Table 5: Overlap of cardiac-related entities between downstream tasks and the pretraining dataset.

| | |
|---|---|
| prolonged qt interval | normal |
| arrhythmia | first degree av block |
| anterior myocardial infarction | ventricular premature complex |
| conduction disturbance | second degree av block |
| hypertrophy | st depression |
| atrial premature complex | prolonged pr interval |
| t wave abnormalities | premature complex |
| atrial fibrillation | sinus tachycardia |
| sinus arrhythmia | sinus bradycardia |
| atrial flutter | supraventricular tachycardia |
| atrial premature complex | abnormal q wave |
| av block | left bundle branch block |
| myocardial infarction | right bundle branch block |
| st elevation | st-t changes |
| t wave changes | ventricular bigeminy |
| ventricular premature complex | sinus tachycardia |
| atrial flutter | supraventricular tachycardia |
| atrial tachycardia | |

names as prompts to ensure fair and consistent evaluation across models. The results are directly taken from MERL (Liu et al., 2024).

Table 6: Comparison of AUC scores and standard deviations between MERL and K-MERL under zero-shot and 1% linear probing settings.

| Methods | Zero-shot (AUC) | 1% Linear Probe (AUC) |
|---|---|---|
| MERL (ResNet) | $61.5 \pm 1.6$ | $65.96 \pm 2.1$ |
| MERL (ViT) | $63.7 \pm 2.0$ | $63.53 \pm 2.6$ |
| **K-MERL (Ours)** | $\mathbf{76.5 \pm 1.8}$ | $\mathbf{69.61 \pm 1.7}$ |

## F State-of-the-Art Without Prompt Engineering

It is important to note that MERL heavily relies on prompt engineering (PE), which requires tailoring the text prompt of each possible disease at *inference time*, querying external knowledge bases using LLM, which is inefficient (Liu et al., 2024). To fully showcase the our method's capabilities, we compare K-MERL with the PE-enhanced version of MERL in Fig 8. Unlike MERL, K-MERL does not depend on any customized disease prompts at inference time, as it has better leveraged cardiac knowledge contained in the reports during pre-training. Despite being free from PE, K-MERL still surpasses MERL with PE, demonstrating the superiority of our approach.

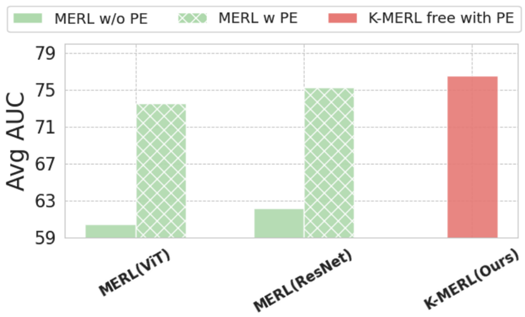



Figure 8: Comparison of K-MERL and MERL with prompt engineering (PE). Notably, even though MERL with PE uses customized disease prompts with human effort, K-MERL, **free with PE**, still surpasses both versions of MERL, demonstrating its generalizability and effectiveness.

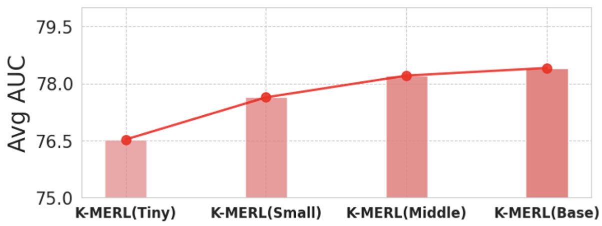



Figure 9: Reported performance of zero-shot classification with scaled ECG encoders. As the model size increases from K-MERL(Tiny) to K-MERL(Base), the performance improves, demonstrating the scalability of the model.

Table 7: Additional Ablation Studies.

(a) Effects of Lead-specific Processing.

| Methods | 1 Lead | 12 Leads |
|---|---|---|
| K-MERL (Ours) | **71.61** | **76.52** |
| – Lead-specific Tokenization | 68.47 | 74.23 |
| – Lead-specific Spatial Positional Embedding | 69.12 | 75.35 |
| – Lead-agnostic Temporal Positional Embedding | 70.84 | 75.10 |

(b) Effects of Text Encoder.

| Text Encoder | 1 Lead | 12 Leads |
|---|---|---|
| BioClinicalBERT | 68.25 | 73.21 |
| Med-KEBERT | 69.62 | 74.59 |
| Med-CPT | **71.61** | **76.52** |

(c) Effects of LLM on Processing Cardiac-related Entities.

| Methods | 1 Lead | 12 Leads |
|---|---|---|
| Llama3.1-8B-Instruct | 68.52 | 74.19 |
| Gemma-2-9B | 68.94 | 74.47 |
| Gemma-2-27B | 70.54 | 75.81 |
| Llama3.1-70B-Instruct | **71.61** | **76.52** |

(d) Effects of the Number of Transformer Layers in the Cardiac Query Network $\mathcal{F}_{CQ}$

| Num of Layers | 1 Lead | 12 Leads |
|---|---|---|
| 1 | 69.92 | 72.96 |
| 2 | 70.14 | 73.13 |
| 3 | 70.31 | 74.40 |
| 4 | **71.61** | **76.52** |
| 5 | 69.25 | 74.94 |

(e) Effects of the Number of Heads in the Cardiac Query Network $\mathcal{F}_{CQ}$.

| Num of Heads | 1 Lead | 12 Leads |
|---|---|---|
| 1 | 68.76 | 74.89 |
| 2 | 70.25 | 74.23 |
| 3 | 70.27 | 75.36 |
| 4 | **71.61** | **76.52** |
| 5 | 71.23 | 75.48 |

# G Scalability

We scale our ECG encoder using ViT-Tiny, ViT-Small, ViT-Middle, and ViT-Base, as shown in Fig. 9. K-MERL consistently improves as model size increases, demonstrating its scalability for ECG multimodal learning.

# H Additional Ablation Studies

**Lead-specific Processing.** In Tab 7a, we ablate the effects of lead-specific tokenization, lead-specific spatial positional embedding, and lead-agnostic temporal embedding. he results show each component enhances K-MERL's performance, with the full combination yielding the best results. The results demonstrate that lead-specific processing is crucial for enabling the ECG multimodal model to recognize lead uniqueness.

**Text Encoder.** Tab. 7b shows Med-CPT (Jin et al., 2023) outperforms BioClinicalBERT (Alsentzer et al., 2019) and Med-KEBERT (Zhang et al., 2023b), due to contrastive pretraining on a large medical corpus, suggesting contrastive pretraining improves text encoder performance for this task.

**Entity Extractor and Query Network.** Tab. 7c, 7d, and 7e present ablation results on the LLM-based entity extractor and the Cardiac Query Network $\mathcal{F}_{CQ}$. Tab. 7c shows that larger LLMs improve the extraction of cardiac-related entities, with Llama3.1-70B-Instruct achieving the best performance across both 1-lead and 12-lead settings. Tab. 7d examines the effect of the number of transformer layers in $\mathcal{F}_{CQ}$, where performance improves and saturates at 4 layers. Tab. 7e analyzes the number of attention heads, with 4 heads yielding the optimal results.