Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement

Che Liu [1 2] **Zhongwei Wan** [3] **Cheng Ouyang** [4 5] **Anand Shah** [6 7] **Wenjia Bai** [1 8 9] **Rossella Arcucci** [1 2]

Abstract

Electrocardiograms (ECGs) are non-invasive diagnostic tools crucial for detecting cardiac arrhythmic diseases in clinical practice. While ECG Self-supervised Learning (eSSL) methods show promise in representation learning from unannotated ECG data, they often overlook the clinical knowledge that can be found in reports. This oversight and the requirement for annotated samples for downstream tasks limit eSSL's versatility. In this work, we address these issues with the **M**ultimodal **E**CG **R**epresentation **L**earning (**MERL**) framework. Through multimodal learning on ECG records and associated reports, MERL is capable of performing zero-shot ECG classification with text prompts, eliminating the need for training data in downstream tasks. At test time, we propose the **C**linical **K**nowledge **E**nhanced **P**rompt **E**ngineering (**CKEPE**) approach, which uses Large Language Models (LLMs) to exploit external expert-verified clinical knowledge databases, generating more descriptive prompts and reducing hallucinations in LLM-generated content to boost zero-shot classification. Based on MERL, we perform the first benchmark across six public ECG datasets, showing the superior performance of MERL compared against eSSL methods. Notably, MERL achieves an average AUC score of 75.2% in zero-shot classification (**without training data**), 3.2% higher than linear probed eSSL methods with 10% annotated training data, averaged across all six datasets.

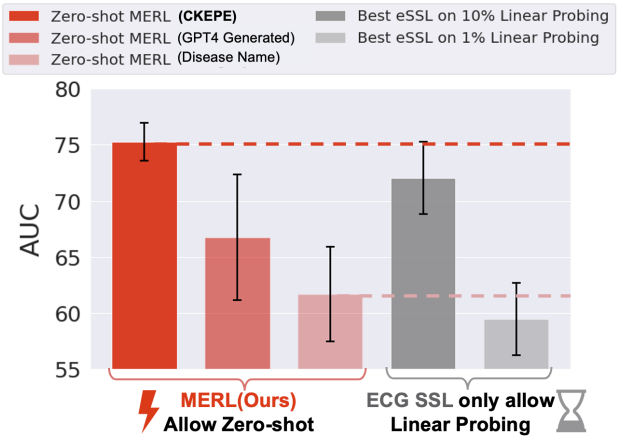

Figure 1. We demonstrate MERL, even without training samples and prompt engineering, surpasses the best-performing eSSL with 1% data linear probing from Tab. 2. Additionally, zero-shot MERL enhanced with our CKEPE outperforms the best eSSL results obtained from 10% data linear probing.

1. Introduction

Supervised learning methods effectively classify cardiac conditions using Electrocardiogram (ECG), a common clinical data for monitoring heart electrical activity (Liu et al., 2023b; Huang et al., 2023; Huang & Yen, 2022). However, these methods require large-scale data with high-quality annotations and expert review. To reduce dependence on annotations, ECG self-supervised learning (eSSL) has made significant strides by utilizing the rich resource of unlabeled ECG records. Current eSSL techniques predominantly fall into two categories: contrastive and generative (Eldele et al., 2021; Kiyasseh et al., 2021; Wang et al., 2023; Zhang et al., 2023; Na et al., 2023). Contrastive eSSL (C-eSSL) focuses on learning discriminative ECG features by differentiating between augmented positive and negative samples (Chen et al., 2020; 2021; Wang et al., 2023; Eldele et al., 2021; Kiyasseh et al., 2021; Wang et al., 2023), whereas generative eSSL (G-eSSL) aims at reconstructing original signals from their masked versions (Na et al., 2023; Zhang et al., 2023).

These methods, however, encounter two primary challenges:

[1]Data Science Institute, Imperial College London, UK [2]Department of Earth Science and Engineering, Imperial College London, UK [3]Ohio State University, Columbus, US [4]Department of Engineering Science, University of Oxford, Oxford, UK [5]Institute of Clinical Sciences, Imperial College London, UK [6]Department of Infectious Disease Epidemiology, Imperial College London, UK [7]Royal Brompton and Harefield Hospitals, UK [8]Department of Computing, Imperial College London, UK [9]Department of Brain Sciences, Imperial College London, UK. Correspondence to: Che Liu <che.liu21@imperial.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Semantic Distortion from Input-Level Augmentation in C-eSSL. Recent developments in C-eSSL for ECG representation learning often create two augmented views from the *same* ECG signal to build *positive* pairs (*i.e.*, aligning their features to be identical) and consider views from different ECG records as *negative* pairs (*i.e.*, with differing features) within a contrastive framework (Chen et al., 2020; 2021; Wang et al., 2023; Eldele et al., 2021; Kiyasseh et al., 2021; Wang et al., 2023). However, current ECG augmentation strategies, such as cutout and drop (Na et al., 2023), could distort semantic information in ECG signals (Anonymous, 2023), as shown in Fig. 2 (a). Consequently, the use of ECG with distorted semantics in positive and negative pairs compromises the quality of ECG representations learned through C-eSSL approaches.

Limited High-level Semantics in G-eSSL. As shown Fig. 2 (c), G-eSSL methods (Zhang et al., 2023; Na et al., 2023) learn to restore low-level signal patterns (e.g., local signal intensities and waveforms) while overlooking high-level semantics such as the diseases behind (Liu et al., 2023j;i; He et al., 2022). However, high-level semantics are essential for downstream ECG classification tasks. Therefore, the lack of high-level semantics in ECG representation can limit the performance of pre-trained models in these tasks.

Besides issues with distorted semantic information and missing high-level semantics, eSSL approaches are incapable of zero-shot classification as they only focus on extracting signal patterns, agnostic to the clinical concepts behind, limiting their versatility and risking distribution shifts in downstream tasks.

Multimodal learning has emerged as a promising approach for learning high-level semantics from other modalities, such as clinical reports (Radford et al., 2021b; Liu et al., 2023f;c; Wan et al., 2023). This strategy has achieved significant progress in the field of medical imaging and radiology reports (Liu et al., 2023a; Chen et al., 2023; Liu et al., 2023e;d). However, as these approaches are mostly designed for image-language domains, their efficacy in ECG and their associated reports remains underexplored. The inherent differences between data modalities (signals vs. images) and the distinct nature of ECG report versus radiology reports pose challenges in directly applying existing multimodal methods from radiography to ECG records. Furthermore, the text prompt for zero-shot classification, a new capability enabled by multimodal learning, requires a dynamic approach to generate more descriptive prompts at test time. It can also leverage external knowledge databases verified by clinical experts to ensure the quality and reliability of the generated prompts, moving beyond crude category names or fixed templates. Additionally, benchmarks are needed to thoroughly assess the influence of large-scale data on ECG representation learning and evaluate the performance and robustness of these methods across a range of cardiac conditions in public datasets.

To address these challenges, this work has four contributions:

- We propose a straightforward yet effective **M**ultimodal **E**CG **R**epresentation **L**earning framework (MERL) for ECG signals and associated reports. Unlike eSSL, MERL is capable of *zero-shot* classification. Zero-shot MERL even outperforms linear probed eSSL with 10% data, as averaged across six datasets. Furthermore, linear probed MERL outperform eSSL across all downstream datasets and data ratios.
- At training time, we introduce **C**ross-**M**odal **A**lignment (**CMA**) and **U**ni-**M**odal **A**lignment (**UMA**) for multimodal representation learning with ECG records and paired clinical reports, with augmentation at the latent level rather than the naive signal level to avoid semantic distortion.
- At test time, we design **C**linical **K**nowledge **E**nhanced **P**rompt **E**ngineering (**CKEPE**), utilizing LLMs to dynamically generate customized prompts for zero-shot classification by extracting and restructuring knowledge from customer-provided knowledge databases verified by clinical experts.
- To facilitate future research, we build the first benchmark by pre-training MERL and 10 eSSL methods on the largest publicly available ECG dataset, evaluating their performance on six diverse datasets covering over 100 cardiac conditions. This benchmark, covering zero-shot, linear probing, and data distribution transfer scenarios, assesses the quality and robustness of learned ECG representations.

2. Related Work

2.1. Representation Learning with Multimodal Medical Data

Various studies have explored medical multimodal learning, but mostly in radiography (Liu et al., 2023a;c; Wan et al., 2023; Liu et al., 2023f;e;d; Chen et al., 2023), with a focus on aligning global and local image features with radiology reports. Compared with images, ECG signals pose a unique challenge due to their global temporal and spatial structures, which span the entire prolonged signal period and are difficult to characterize at a local level. (Lalam et al., 2023) demonstrated the effectiveness of ECG and EHR multimodal learning, although their approach was limited to a private dataset. (Li et al., 2023; Liu et al., 2023g) attempt multimodal ECG learning for zero-shot classification but fall short: Their methods, crudely aligning signals

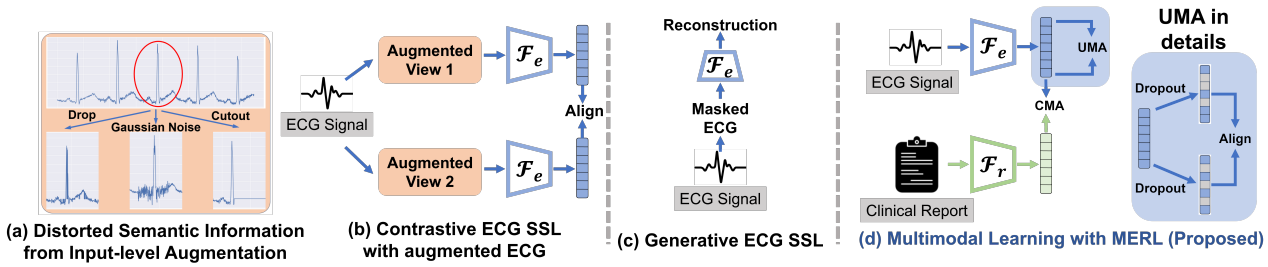

Figure 2. **(a)** Commonly used naive input-level data augmentation distorts semantics of ECG records, leading to sub-optimal representation learning performance. **(b)** Illustration of existing eSSL approaches. Their contrastive learning framework necessitates these naively augmented ECG signals. **(c)** Existing generative eSSL employs signal reconstruction as self-supervision task while being agnostic to the semantic meaning of ECG. **(d)** The proposed **MERL**, designed for multimodal ECG learning, leverages both ECG records and clinical reports for representation learning through **C**ross-**M**odal **A**lignment (**CMA**). MERL addresses the drawbacks of naive input-level augmentation by opting for latent augmentation (dropout) to prevent pattern corruption, and it enhances ECG learning through **U**ni-**M**odal **A**lignment (**UMA**). $\mathcal{F}_e$ denotes the ECG encoder, and $\mathcal{F}_r$ represents the text report encoder.

with text, overlook distinct signal patterns, and their reliance on simple cardiac condition names as prompts misses critical clinical attributes, leading to sub-optimal performance. Furthermore, their limited evaluations on small datasets is inadequate for assessing the potential of multimodal ECG learning in complicated real-world scenarios.

2.2. Self-supervised Learning for ECG on Signal Domain

Recently, ECG self-supervised learning (eSSL) has proven beneficial for learning transferrable representations directly from unannotated ECG signals (Lai et al., 2023; Chen et al., 2020; Sangha et al., 2024). Among them contrastive eSSL methods such as CLOCS (Kiyasseh et al., 2021) and ASTCL (Wang et al., 2023) have advanced eSSL by exploring temporal and spatial invariance and employing adversarial learning, respectively. Generative eSSL techniques, as discussed in (Zhang et al., 2022; Sawano et al., 2022; Na et al., 2023), learn ECG representations through pretext tasks involving masked segment reconstruction. However, they face challenges in attaining high-level semantic representations. Both contrastive and generative eSSL methods are often agnostic of high-level clinical domain knowledge. Therefore, there is still an unmet need for an effective unsupervised approach for learning semantically rich, transferable ECG representations.

2.3. Customizing Prompt for Zero-shot Classification

In zero-shot classification, the prompt's quality, often limited in traditional methods that use basic category names, is crucial for effective performance and classification (Radford et al., 2021b). To improve this, (Menon & Vondrick, 2022; Pratt et al., 2023) uses LLMs to generate attribute-rich prompts, boosting performance. Yet, in medical field, where terminologies are highly specialized, prompts generated by non-specialist LLMs might be inaccurate or untrustable, leading to performance degradation and safety concerns. Instead of relying on limited knowledge encoded in non-specialized LLMs, we re-purpose LLMs for extracting and re-formatting specialized clinical knowledge from trustable, external sources such as online and local clinical knowledge databases. By this mean, we can efficiently create clinically relevant, structured prompts without additional annotations.

3. Method

3.1. Overview

Our MERL framework learns transferable ECG representations directly from ECG signals and associated text reports. These learned representation can be then directly applied for zero-shot classification of unseen diseases. To achieve this, our framework is a synergy of a train-time ECG-report multimodal representation learning strategy and a test-time clinical knowledge enhanced prompt engineering approach. Specifically, as depicted in Fig. 2 (d), the representation learning strategy comprises **C**ross-**M**odal **A**lignment (**CMA**), detailed in Sec. 3.2, and **U**ni-Modal **A**lignment (**UMA**), described in Sec. 3.3. Additionally, **C**linical **K**nowledge **E**nhanced **P**rompt **E**ngineering (**CKEPE**) is introduced in Sec. 3.4.

3.2. Cross-Modal Alignment

The **C**ross-**M**odal **A**lignment (**CMA**) aims to learn ECG features informed with clinical knowledge under report supervision. Specifically, given a training dataset $\mathcal{X}$ consisting of N ECG-report pairs, we represent each pair as $(\mathbf{e}_i, \mathbf{r}_i)$, where $\mathbf{e}_i \in \mathcal{E}$ denotes the raw ECG records and $\mathbf{r}_i \in \mathcal{R}$ denotes the associated text report, respectively, with $i = 1, 2, 3, ..., N$. In this framework, as shown in Fig. 2 (d), two distinct encoders for ECG signals and text reports, symbolized as $\mathcal{F}_e$ and $\mathcal{F}_r$ respectively, transform the sample pair $(\mathbf{e}_i, \mathbf{r}_i)$ into the latent embedding space, represented as $(\mathbf{z}_{e,i}, \mathbf{z}_{r,i})$. Then dataset at feature-level is then denoted as $\mathcal{X} = \{(\mathbf{z}_{e,1}, \mathbf{z}_{r,1}), (\mathbf{z}_{e,2}, \mathbf{z}_{r,2}), \ldots, (\mathbf{z}_{e,N}, \mathbf{z}_{r,N})\}$, where

$\mathbf{z}_{e,i} = \mathcal{F}_e(\mathbf{e}_i)$ and $\mathbf{z}_{r,i} = \mathcal{F}_r(\mathbf{z}_{r,i})$. After that, two non-linear projectors for ECG and text embedding, denoted as $\mathcal{P}_e$ and $\mathcal{P}_r$ respectively, convert $\mathbf{z}_{e,i}$ and $\mathbf{z}_{r,i}$ into the same dimensionality d, with $\hat{\mathbf{z}}_{e,i} = \mathcal{P}_e(\mathbf{z}_{e,i})$, $\hat{\mathbf{z}}_{r,i} = \mathcal{P}_r(\mathbf{z}_{r,i})$. Then, we compute the cosine similarities as $s^{e2r}_{i,i} = \hat{\mathbf{e}}_i^\top \hat{\mathbf{r}}_i$, representing the ECG-report similarities. The loss function, $\mathcal{L}_{\text{CMA}}$, is then expressed as:

$$\mathcal{L}^{e2r}_{i,j} = -\log \frac{\exp(s^{e2r}_{i,j}/\tau)}{\sum_{k=1}^{L} \mathbb{1}_{[k\neq i]} \exp(s^{e2r}_{i,k}/\tau)}, \quad (1)$$

$$\mathcal{L}_{\text{CMA}} = \frac{1}{L} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{L}^{e2r}_{i,j}. \quad (2)$$

Here, $\mathcal{L}^{e2r}_{i,j}$ represent the ECG-report cross-modal contrastive losses, respectively. The temperature hyper-parameter, denoted as τ, is set to 0.07 in our study. Additionally, L signifies the batch size per step, being a subset of N.

3.3. Uni-Modal Alignment

On top of CMA, we further employ **U**ni-**M**odal **A**lignment (**UMA**) to facilitate representation learning. UMA is formulated as contrastive learning operating on the signal domain only. To circumvent the distortion of semantic information caused by naive input-level data augmentation (shown in Fig. 2 (a)), we use our proposed latent augmentation on the ECG embedding $\mathbf{z}_{e,i}$ to construct positive pairs for contrastive learning, which is illustrated in the blue shaded block of in Fig. 2 (d). Inspired by (Gao et al., 2021), to generate the positive pair $(\mathbf{z}^1_{e,i}, \mathbf{z}^2_{e,i})$, we adopt two independent dropout operations on the ECG embeddings separately. Then, we use standard contrastive loss on the positive pair and treat other unpaired combinations as negative pairs. The loss function of $\mathcal{L}_{\text{UMA}}$ can be denoted as:

$$\begin{aligned} \mathcal{L}_{\text{UMA}} = -\frac{1}{L} \sum_{i=1}^{N} \sum_{j=1}^{N} \log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{L} \mathbb{1}_{[k\neq i]} \exp(s_{i,j}/\tau)}, \\ \text{where } s_{i,i} = \mathbf{z}^{1\top}_{e,i} \mathbf{z}^2_{e,i}, \\ \mathbf{z}^1_{e,i} = \mathbf{z}_{e,i} \odot M^1,\ M^1 \sim \text{ Bernoulli } (p), \\ \mathbf{z}^2_{e,i} = \mathbf{z}_{e,i} \odot M^2,\ M^2 \sim \text{ Bernoulli } (p). \end{aligned} \quad (3)$$

M^1 and M^2 represent the dropout masks, which have the same sizes as $\mathbf{z}_{e,i}$'s, with each entry independently sampled with dropout ratio p, which is set to 0.1. $\odot$ denotes element-wise multiplication. The ablation study for p is detailed in Tab 8. Importantly, as the two dropout operations are independent, the ECG embeddings post-dropout will not be identical, thus avoiding a trivial solution.

In summary, our model learns representative ECG features by jointly minimizing $\mathcal{L}_{\text{UMA}}$ and $\mathcal{L}_{\text{CMA}}$, and the overall training loss can be written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CMA}} + \mathcal{L}_{\text{UMA}}, \quad (4)$$

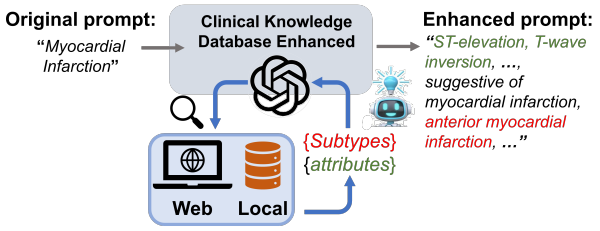

Figure 3. At test time, we design the CKEPE pipeline for generating more descriptive prompts via LLM for zero-shot classification. In particular, we leverage the capability of LLM to extract clinical knowledge from trustworthy external knowledge databases verified by clinicians, then restructure this knowledge (e.g., subtypes or attributes of cardiac conditions) for prompt generation, with less hallucination from LLM.

3.4. Enhancing Zero-Shot Prompts with External Clinical Knowledge Databases

The quality of text prompts, as part of input to a multimodality model, have been found to have significant impact on the zero-shot classification performance (Menon & Vondrick, 2022; Pratt et al., 2023; Maniparambil et al., 2023). The conventional method (Radford et al., 2021a; Zhang et al., 2020) merely use names of cardiac conditions or a fixed template as text prompts for zero-shot classification. This approach, however, often perform poorly due to a lack of sufficient attributes (*e.g.,* detailed description of signal patterns) and/or possible sub-categories at a finer level (*e.g.,* possible subtypes of a clinical condition) for distinguishing the target semantic class. While a powerful LLM can generate possible relevant attributes and subtypes for these categories as text prompts (Pratt et al., 2023), it incurs the risk of factual error/hallucination, often due to a lack of knowledge for the specialized domain (Liu et al., 2023h; Hyland et al., 2023; Umapathi et al., 2023). This risk renders this naive LLM-based paradigm unacceptable for medical applications. To address this, we introduce **C**linical **K**nowledge **E**nhanced **P**rompt **E**ngineering (**CKEPE**). Instead of directly, sourcing specialized knowledge from (possibly non-specialist) LLM, we leverage LLM to query and extract clinical knowledge from trustable, external knowledge bases, and reformat extracted knowledge as structured prompts. These clinical-knowledge-informed prompts contains descriptive attributes and possible finer-level labels (disease subtypes) of the targeted semantic class.

Web and Local Clinical Knowledge Databases. To implement CKEPE, we initially prepare two databases rich in precise, expert-evaluated clinical knowledge. The first is the Systemized Nomenclature of Medicine – Clinical Terms (SNOMEDCT) [1], a comprehensive, web-based database in-

[1] https://bioportal.bioontology.org/ontologies/SNOMEDCT

ternationally validated for recording clinical information in structured clinical vocabulary (Stearns et al., 2001).

The second database is focused on Standard Communications Protocol (SCP) statements for ECG (Rubel et al., 2016), which describes ECG states. As there is no publicly accessible database encompassing the entire SCP statement, we have constructed a local database by collecting relevant trustable sources from the internet.

Searching, Thinking, and Generation. After preparing the web and local databases, we initially query GPT-4, the only current LLM featuring original web browsing capabilities, with: '*Which attributes and subtypes does* `<cardiac condition>` *have? If this condition specifically describes symptoms or a subtype, please refrain from answering; otherwise, generate all possible scenarios.*' Following GPT-4's response, we enable its web browsing function towards the designated web database and we upload the local database file. We then instruct GPT-4 to search for relevant terms in both the web and local databases, ensuring the generated results exist in the clinical knowledge database and are relevant to the provided cardiac condition. Terms that are not found in either database are discarded. This verification step is automated by GPT-4 thinking.

Subsequently, we ask GPT-4 to reformulate the remaining terms into a structured text prompt for zero-shot classification, stylized like an ECG statement containing possible *diseases subtypes* and descriptive *attributes* describing signal patterns. As shown in Fig. 4, our prompts, unlike the original ones with only category names and fixed templates, are dynamic and tailored to specific cardiac conditions without the need for handcrafting. Moreover, the reformulated prompts adhere to clinically structured expressions, as all terms are cross-verified with the web and local clinical knowledge databases.

4. Experiments

4.1. Pre-training Configuration

MIMIC-ECG. In our study, we pre-train the MERL framework on the MIMIC-ECG dataset (Gow et al.). This dataset contains 800,035 paired samples from 161,352 unique subjects. Each sample is composed of a raw ECG signal and its associated report, with every ECG recording sampled at 500Hz for a duration of 10 seconds. To prepare the pre-training dataset, we executed the following procedures: **(1)** Exclude samples with an empty report or reports containing fewer than three words. **(2)** Substitute 'NaN' and 'Inf' values in ECG recordings with the average of the six neighboring points. After these curation steps, our tailored dataset for training MERL contains 771,693 samples. Each sample includes an ECG record and its corresponding report.

Implementation. In pre-training stage, we employ a random initialized 1D-ResNet18 as the ECG encoder. For text encoding, we employ Med-CPT (Jin et al., 2023) by default. The impact of various text encoders on downstream performance is discussed in Sec 5. We select the AdamW optimizer, setting a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. We pre-train MERL for 50 epochs, applying a cosine annealing scheduler for learning rate adjustments. We maintain a batch size of 512 per GPU, with all experiments conducted on eight NVIDIA A100-40GB GPUs. All code can be accessed at [2]

4.2. Downstream Tasks Configuration

We evaluate our framework on both zero-shot classification and linear probing, on three widely-used public datasets listed as follows, covering over 100 cardiac conditions. The details of the data split are shown in *Appendix*.

PTBXL. This dataset (Wagner et al., 2020) encompasses 21,837 ECG signals that were accumulated from 18,885 patients. The collected data consists of 12-lead ECG, each sampled at a rate of 500 Hz with a duration of 10 seconds. Based on ECG annotation protocol, there are four subsets with multi-label classification tasks: **Superclass** (5 categories), **Subclass** (23 categories), **Form** (19 categories), and **Rhythm** (12 categories). Notably, these four subsets have different number of samples. We follow the official data split (Wagner et al., 2020) for the train:val:test split.

CPSC2018. This publicly accessible dataset (Liu et al., 2018) comprises 6,877 standard 12-lead ECG records, each sampled at a rate of 500 Hz, and the duration of these records ranges from 6 to 60 seconds. The dataset is annotated with 9 distinct labels. We split the dataset as 70%:10%:20% for the train:val:test split.

Chapman-Shaoxing-Ningbo (CSN). This publicly accessible dataset (Zheng et al., 2020; 2022) comprises 45,152 standard 12-lead ECG records, each sampled at a rate of 500 Hz with a duration of 10 seconds. We drop the ECG signals with 'unknown' annotation, the curated version dataset has 23,026 ECG records with 38 distinct labels. The dataset is split into 70%:10%:20%.

Implementation. For linear probing, we keep the ECG encoder frozen and update only the parameters of a newly initialized linear classifier. We conduct linear probing for each task, utilizing $1\%, 10\%, 100\%$ of the training data. These configurations are consistent across all linear probing classification tasks. For zero-shot classification, we freeze the whole model, and use CKEPE to customize the prompt for each category. We compute the similarity between the ECG embedding and prompt embedding as the classification probability for the category associated with the prompt.

[2] https://github.com/cheliu-computation/MERL

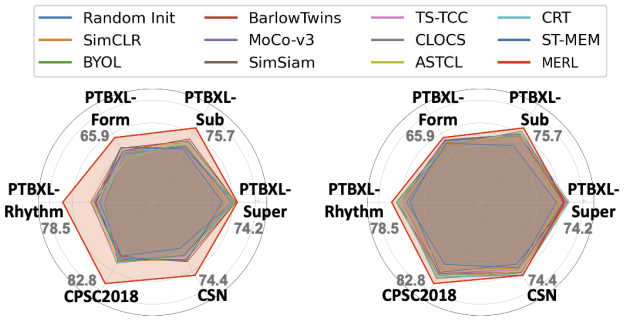

Figure 4. **Left**: Zero-shot MERL vs. linear probed eSSL with 1% Data. **Right**: Zero-shot MERL vs. linear probed eSSL with 10% Data. All performance are reported in the AUC score.

For all downstream tasks, we use macro AUC as the metric. Further details, including specifics of the implementation, are provided in *Appendix*.

4.3. Evaluation on zero-shot learning

Zero-shot classification using text prompts is a common task for assessing representation learning quality (Radford et al., 2021a). However, most research in ECG representation learning evaluates only linear probing. This limitation arises as these eSSL approaches merely operates on ECG signals only, without having a text encoder for receiving text prompts. We motivate zero-shot classification as a way to measure the quality and versatility of cross-modal ECG representations learned from clinical reports.

In our comprehensive analysis across six different datasets, we assess the performance of zero-shot MERL compared to eSSL approaches on linear probing, as shown in Fig. 1, 4. Fig. 1 demonstrates how zero-shot MERL with three types of prompts outperforms eSSL methods that are linear probed with additional annotated data. 'Disease Name' means the category name is used as the prompt for zero-shot classification. 'GPT4 Generated' indicates that we use GPT4 to create a detailed prompt based on the category name for zero-shot classification. Notably, zero-shot MERL, even without prompt enhancement, surpasses the top eSSL method linear probed with 1% additional training data, achieving higher average AUC across six datasets. Furthermore, with CKEPE, zero-shot MERL exceeds the best eSSL method probed with additional 10% data, underscoring the effectiveness of CKEPE and learned ECG representations from MERL.

We also present the performances of zero-shot MERL and eSSL performance on individual datasets, as shown on the left of Fig. 4. Remarkably, zero-shot MERL demonstrates superior performance to eSSL with linear probing across all downstream datasets, even without additional training samples. This underlines MERL's ability to learn robust, transferable cross-modal ECG features with clinically relevant knowledge from report supervision.

Interestingly, despite the significant overall performance gain, our method demonstrates a lesser advantage on the PTBXL-Super dataset. This behavior may be attributed to the 'simpler' nature of the PTBXL-Super dataset, which has only 5 broad categories (e.g., *myocardial infarction*), compared to the 9–38 detailed categories (e.g., *inferior myocardial infarction* or *inferolateral myocardial infarction*) in other datasets. Nevertheless, MERL notably outperforms other eSSL methods linear probed on 1%-10% additional annotated training data in the remaining five more challenging datasets. These results demonstrate that clinical reports are a valuable supervision signal for ECG representation learning.

Table 1. Comparing MERL to other multimodal ECG representation learning methods.

	Zero-shot	Linear Probing (1%)
METS (Li et al., 2023)	58.84±4.6	61.87±3.4
MERL (Ours)	**75.24±1.7**	**65.96±2.1**

We compare our method with the only existing multimodal ECG representation learning-based method, METS (Li et al., 2023), as shown in Tab 1. Our method significantly outperforms METS on zero-shot tasks, benefiting from UMA and CKEPE. This illustrates the benefits of learning unimodal representations and shows that employing a knowledge-enhanced detailed prompt positively influences zero-shot learning.

4.4. Evaluation as ECG Representations

While our MERL highlights its zero-shot classification capability, assessing ECG representations via uni-modal tasks after pretraining is more common. We select linear probing for our evaluation protocol because of its standardized procedures for eSSL methods (Kiyasseh et al., 2021; Zhang et al., 2023; Wang et al., 2023; Na et al., 2023).

Tab. 2 presents the results of linear probing for MERL and existing eSSL methods. Our MERL consistently outperform eSSL methods across a 1%-100% of training data ratio and six datasets. Notably, MERL's performance with just 1% data in linear probing on the PTBXL-Super dataset surpasses those of all eSSL methods using 100% data. Furthermore, MERL with 10% data in linear probing outperforms all eSSL methods with 100% data on the remaining five datasets. This demonstrates that clinical report supervision enables MERL, to learn discriminative ECG representations with richer semantics. Interestingly, the second-highest performance on four subsets of the PTBXL with 100% data linear probing is achieved by a randomly initialized ResNet18, rather than by any eSSL methods. We speculate this is due to two reasons: (1) for contrastive eSSL approaches, the quality of the learned representation decreases because the positive/negative pairs, generated through naive signal-level

Table 2. Linear probing results of MERL and eSSL methods. The best results are **bolded**, with gray indicating the second highest.

Method	PTBXL-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	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
MERL (Ours)	**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**

Table 3. Results under data distribution shift: 'Source Domain' denotes the dataset used for linear probing with the frozen pre-trained ECG encoder. 'Target Domain' refers to the corresponding test set. We include only those target domain samples that match categories from the source domain. The top results are highlighted in bold, while the gray color marks the second-highest achievements.

Source Domain	**Zero-shot**	**Training Data Ratio**	PTBXL-Super		CPSC2018		CSN	
Target Domain			CPSC2018	CSN	PTBXL-Super	CSN	PTBXL-Super	CPSC2018
Random Init	✗	100%	68.62	75.31	55.74	68.92	56.57	61.16
SimCLR (Chen et al., 2020)	✗		69.62	73.05	56.65	66.36	59.74	62.11
BYOL (Grill et al., 2020)	✗		70.27	74.01	57.32	67.56	60.39	63.24
BarlowTwins (Zbontar et al., 2021)	✗		68.98	72.85	55.97	65.89	58.76	61.35
MoCo-v3 (Chen et al., 2021)	✗		69.41	73.29	56.54	66.12	59.82	62.07
SimSiam (Chen & He, 2021)	✗		70.06	73.92	57.21	67.48	60.23	63.09
TS-TCC (Eldele et al., 2021)	✗		71.32	75.16	58.47	68.34	61.55	64.48
CLOCS (Kiyasseh et al., 2021)	✗		68.79	72.64	55.86	65.73	58.69	61.27
ASTCL (Wang et al., 2023)	✗		69.23	73.18	56.61	66.27	59.74	62.12
CRT (Zhang et al., 2023)	✗		70.15	74.08	57.39	67.62	60.48	63.33
ST-MEM (Na et al., 2023)	✗		76.12	**84.50**	62.27	75.19	73.05	64.66
METS (Li et al., 2023)	✓	0%	74.42	69.78	61.13	72.31	59.63	63.12
MERL (Ours)	✓	**0%**	**88.21**	78.01	**76.77**	**76.56**	**74.15**	**82.86**

augmentation, introduce semantic distortion; and (2) for generative eSSL methods, the ECG representation learned through reconstruction tasks lacks discriminative and high-level semantic information (Liu et al., 2023j;i).

4.5. Robustness to Distribution Shift

Distribution shift refers to scenarios where the test set's ECGs come from a different distribution (often caused by different data sources) than the training set. Among them we focus on the most common distribution shift in healthcare data: domain shift (covariate shift), where the label space are shared but input distributions vary. To evaluate the generalizability and robustness of the learned ECG representation across different sources, we conduct linear probing with eSSL methods and zero-shot MERL under domain shifts: training on one dataset (the 'source domain') and testing on another (the 'target domain'), which has categories in common with the source domain.

We prepare the target domain similarly to CLIP (Radford et al., 2021a). The details of preparation can be found in the *Appendix*. After preparing the target domain samples, we compare zero-shot MERL with all eSSL methods using 100% data for linear probing across six target domains. The results are outlined in Tab. 3. Remarkably, zero-shot MERL outperforms all eSSL methods that are linear probed with 100% data, except in the *PTBXL-Super→CSN* setting. We also observe that ST-MEM (Na et al., 2023) achieves the second-highest overall results. This may be attributed to ST-MEM being pre-trained on a reconstruction task without involving naive signal level data augmentation for constructing positive/negative pairs. This behavior supports our postulation that naive data augmentation in eSSL could impair the robustness of the learned ECG representation. Overall, these findings underscore that the ECG features learned via MERL are both representative and robust.

5. Ablation Studies

In this section, we perform comprehensive ablation studies on the key components/design choices, and report the average performance of zero-shot classification and linear probing with 1% data across six ECG classification datasets.

Loss Function. Tab. 4 shows that training with the combination of $\mathcal{L}_{\text{CMA}}$ and $\mathcal{L}_{\text{UMA}}$ improves performance compared to solely using CMA. This suggests that UMA enhances the model's ability to learn ECG representation in the latent space, benefiting downstream tasks.

Text Encoder. Tab. 5 shows the effects of various text

Table 4. Ablating Loss Function. '-' indicates inability to perform zero-shot classification because it is not aligned with text.

$\mathcal{L}_{\text{CMA}}$	$\mathcal{L}_{\text{UMA}}$	Zero-shot	Linear Probing (1%)
	✓	-	57.39±4.5
✓		60.84±3.8	64.25±2.6
✓	✓	61.67±4.2	**65.96±2.1**

encoders. Med-CPT (Jin et al., 2023) [3] achieves the highest performance in both tasks. The other two text encoders yield suboptimal outcomes. We attribute Med-CPT's superior performance to its discriminative and representative text embeddings: Med-CPT is pre-trained on a text contrastive learning task [4], differing from other encoders that are pre-trained on masked language modeling tasks.

Table 5. Effects of Text Encoder Choices.

	Zero-shot	Linear Probing (1%)
BioClinicalBERT (Alsentzer et al., 2019)	72.36±4.5	63.81±1.8
PubMedBERT (Gu et al., 2021)	71.84±3.2	64.29±2.5
Med-CPT (Jin et al., 2023)	**75.24±1.7**	**65.96±2.1**

Clinical Knowledge Database. We further explore the effects of web and local clinical knowledge databases in Tab. 6. Eliminating either database results in decreased performance. Specifically, removing the web database, SNOMEDCT , leads to a notable reduction in performance, attributable to its larger scale compared to the local database. This underscores that both clinical knowledge databases are beneficial for zero-shot classification, with the larger-scale database providing more improvements.

Table 6. Benefits of Clinical Knowledge Database

Database	Zero-shot
w/o SNOMEDCT (web)	72.17±2.3
w/o SCP Statement (local)	73.62±1.9
Ours	**75.24±1.7**

Data Augmentation Strategies. We implement four augmentation strategies and report the results in Tab. 7. As shown in Fig. 2 (a), naive data augmentation on ECG signal domain distorts semantic information and reduces the quality of the representation. Instead, the proposed latent space augmentation demonstrates superior performance compared to other strategies applied to raw ECG signals.

Table 7. Effect of Diverse ECG Augmentation Strategies.

Augmentation	Zero-shot	Linear Probing (1%)
Cutout	73.24±3.2	62.24±2.7
Drop	72.79±2.5	61.14±2.2
Gaussian noise	72.61±2.7	64.17±1.6
Latent Dropout (Ours)	**75.24±1.7**	**65.96±2.1**

Dropout Ratio. Finally, we implement latent augmentation using various dropout ratios and report the results in Tab. 8. A dropout ratio of 0.1 yields the best results, while both higher and lower ratios lead to decreased performance. Therefore, we opt for using a dropout ratio of 0.1 for latent augmentation in our method.

Table 8. Effect of Difference Dropout Ratios

Dropout Ratio	Zero-shot	Linear Probing (1%)
0.05	74.79±1.4	65.23±1.8
0.1	**75.24±1.7**	**65.96±2.1**
0.15	74.53±2.6	64.19±1.7
0.2	74.25±2.1	64.43±2.5

Feature Extractors for ECG. Tab. 9 outlines the ablation study for two ECG feature extractor networks (*i.e.,* backbones): the CNN-based ResNet18(He et al., 2016) and the transformer-based ViT-Tiny(Dosovitskiy et al., 2020). Results show that performance of MERL with CNN backbone surpasses that of ViT, suggesting CNN are better suited for capturing ECG patterns. The degraded efficacy of the transformer backbone may stem from its tokenization strategy, which discretizes continuous signals, possibly leading to information loss.

Table 9. Effects of ECG Feature Extractor Choices

Augmentation	Zero-shot	Linear Probing (1%)
ViT(Dosovitskiy et al., 2020)	73.54±2.3	63.53±2.6
ResNet(He et al., 2016)	**75.24±1.7**	**65.96±2.1**

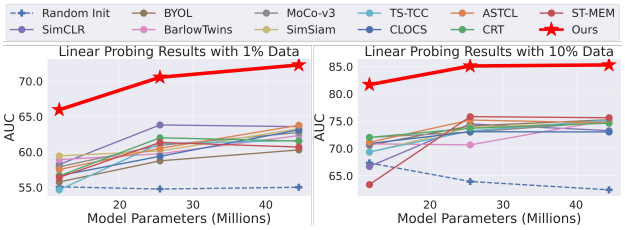

Figure 5. Average linear probing performance on six datasets of MERL and other eSSL methods with scaled ECG backbones. For ST-MEM, a transformer-based method, we use ViT-Tiny, ViT-Small, and ViT-Base.

Scalability. We scale up the backbone size using three models, ResNet18, ResNet50, and ResNet101, for MERL and other eSSL methods, and report their linear probing performance with 1-10% of data across six datasets in Fig. 5. Across these scaled models, MERL consistently outperforms the other eSSL methods. Significantly, even the smallest ECG backbone in MERL exceeds the performance of the largest backbones in other eSSL methods. This highlight the value of clinical reports for representation learning for ECG. Moreover, as the backbone size increases, MERL's performance improves, whereas eSSL methods experience varied performance fluctuation with scaled ECG encoder sizes. This findings demonstrate that MERL is a scalable, effective, and data-efficient strategy for ECG representation learning.

6. Conclusion

We introduce MERL, a scalable and effective multimodal ECG learning framework that incorporates CMA and UMA

[3] https://huggingface.co/ncbi/MedCPT-Query-Encoder

[4] This task is implemented on the query-article pair from PubMed search log, and has no overlap with the pre-training dataset in this work.

alignment strategies during training, and CKEPE, a strategy for customizing prompts, during testing. CKEPE leverages the capabilities of LLMs to extract and restructure clinical knowledge from a provided database, boosting zero-shot MERL to outperform eSSL with linear probing in classification tasks. Additionally, we establish the first benchmark that includes 10 eSSL methods and MERL, all pre-trained on the largest public ECG datasets and evaluated across a broad range of classification tasks. MERL's superior performance in both zero-shot and linear probing tasks underscores the advantages of multimodal ECG learning with report supervision over eSSL methods that only learn representations in the signal domain. We hope that both MERL and this benchmark will benefit the research community in ECG representation learning.

Impact Statement

Our MERL framework significantly advances ECG classification by utilizing a zero-shot approach that eliminates the need for annotated data in downstream ECG tasks. It outperforms linearly probed self-supervised learning methods through innovative prompt engineering. To support future research, we build the first comprehensive ECG representation learning benchmark, which covers six datasets and introduces domain transfer scenarios. Our research primarily uses LLMs to generate descriptive prompts. Although we compel the LLM to extract and reformat knowledge from clinical expert-verified databases, the generation process remains not fully controlled and transparent. Therefore, the field of safe, controllable, and trusted generation in clinical applications recognizes the need for further development and exploration. In the future, we aim to align ECG records with diverse medical data modalities, such as electronic health records, cardiac imaging, and cardiologist reports, to enhance multimodal medical data understanding.

References

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

Anonymous. Towards enhancing time series contrastive learning: A dynamic bad pair mining approach. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=K2c04ulKXn. under review.

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

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

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

Chen, Y., Liu, C., Huang, W., Cheng, S., Arcucci, R., and Xiong, Z. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023.

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

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

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

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

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

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

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

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

Huang, Y., Yen, G. G., and Tseng, V. S. Snippet policy network for multi-class varied-length ecg early classification. *IEEE Transactions on Knowledge & Data Engineering*, 35(06):6349–6361, 2023.

Hyland, S. L., Bannur, S., Bouzid, K., Castro, D. C., Ranjit, M., Schwaighofer, A., Pérez-García, F., Salvatelli, V., Srivastav, S., Thieme, A., et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

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

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

Lai, J., Tan, H., Wang, J., Ji, L., Guo, J., Han, B., Shi, Y., Feng, Q., and Yang, W. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications*, 14(1):3741, 2023.

Lalam, S. K., Kunderu, H. K., Ghosh, S., A, H. K., Awasthi, S., Prasad, A., Lopez-Jimenez, F., Attia, Z. I., Asirvatham, S., Friedman, P., Barve, R., and Babu, M. ECG representation learning with multi-modal EHR data. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=UxmvCwuTMG.

Li, J., Liu, C., Cheng, S., Arcucci, R., and Hong, S. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*, 2023.

Liu, C., Cheng, S., Chen, C., Qiao, M., Zhang, W., Shah, A., Bai, W., and Arcucci, R. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 637–647. Springer, 2023a.

Liu, C., Cheng, S., Ding, W., and Arcucci, R. Spectral cross-domain neural network with soft-adaptive threshold spectral enhancement. *arXiv preprint arXiv:2301.10171*, 2023b.

Liu, C., Cheng, S., Shi, M., Shah, A., Bai, W., and Arcucci, R. Imitate: Clinical prior guided hierarchical vision-language pre-training. *arXiv preprint arXiv:2310.07355*, 2023c.

Liu, C., Ouyang, C., Chen, Y., Quilodrán-Casas, C. C., Ma, L., Fu, J., Guo, Y., Shah, A., Bai, W., and Arcucci, R. T3d: Towards 3d medical image understanding through vision-language pre-training. *arXiv preprint arXiv:2312.01529*, 2023d.

Liu, C., Ouyang, C., Cheng, S., Shah, A., Bai, W., and Arcucci, R. G2d: From global to dense radiography representation learning via vision-language pre-training. *arXiv preprint arXiv:2312.01522*, 2023e.

Liu, C., Shah, A., Bai, W., and Arcucci, R. Utilizing synthetic data for medical vision-language pre-training: Bypassing the need for real images. *arXiv preprint arXiv:2310.07027*, 2023f.

Liu, C., Wan, Z., Cheng, S., Zhang, M., and Arcucci, R. Etp: Learning transferable ecg representations via ecg-text pre-training. *arXiv preprint arXiv:2309.07145*, 2023g.

Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Liu, Q., Hyland, S., Bannur, S., Bouzid, K., Castro, D. C., Wetscherek, M. T., Tinn, R., Sharma, H., Pérez-García, F., Schwaighofer, A., et al. Exploring the boundaries of gpt-4 in radiology. *arXiv preprint arXiv:2310.14573*, 2023h.

Liu, Y., Zhang, S., Chen, J., Chen, K., and Lin, D. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*, 2023i.

Liu, Y., Zhang, S., Chen, J., Yu, Z., Chen, K., and Lin, D. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5361–5372, 2023j.

Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., and O'Connor, N. E. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 262–271, 2023.

Menon, S. and Vondrick, C. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15691–15701, 2023.

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

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

Rubel, P., Pani, D., Schloegl, A., Fayn, J., Badilini, F., Macfarlane, P. W., and Varri, A. Scp-ecg v3. 0: An enhanced standard communication protocol for computer-assisted electrocardiography. In *2016 Computing in Cardiology Conference (CinC)*, pp. 309–312. IEEE, 2016.

Sangha, V., Khunte, A., Holste, G., Mortazavi, B. J., Wang, Z., Oikonomou, E. K., and Khera, R. Biometric contrastive learning for data-efficient deep learning from electrocardiographic images. *Journal of the American Medical Informatics Association*, 31(4):855–865, 2024.

Sawano, S., Kodera, S., Takeuchi, H., Sukeda, I., Katsushika, S., and Komuro, I. 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*, 2022.

Stearns, M. Q., Price, C., Spackman, K. A., and Wang, A. Y. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, pp. 662. American Medical Informatics Association, 2001.

Umapathi, L. K., Pal, A., and Sankarasubbu, M. Medhalt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

Wan, Z., Liu, C., Zhang, M., Fu, J., Wang, B., Cheng, S., Ma, L., Quilodrán-Casas, C., and Arcucci, R. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*, 2023.

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

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

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

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

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zheng, J., Chu, H., Struppa, D., Zhang, J., Yacoub, S. M., El-Askary, H., Chang, A., Ehwerhemuepha, L., Abudayyeh, I., Barrett, A., et al. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1):2898, 2020.

Zheng, J., Guo, H., and Chu, H. 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)*, 2022.

A. Downstream Task Details

A.1. Downstream Task Data Split

We detail the data split in Tab. 10. For the four subsets of PTBXL, we adhere to the official split from the original work of (Wagner et al., 2020). For CPSC2018 (Liu et al., 2018) and CSN (Zheng et al., 2022; 2020), we randomly split the data, and all data split information will be released post-acceptance.

Table 10. 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

A.2. Downstream Task Configuration

Since not all categories of the target domain are covered by the source domain, we merge target domain categories into the most similar source domain category similarly to CLIP (Radford et al., 2021a). For example, 'ST wave tilt up' and 'ST wave drop down' are merged into 'ST-T wave change' when the target domain has a broader range of categories than the source domain. If the target domain includes categories not present in the source domain (e.g., 'Wolf-Parkinson-White syndrome' from the CSN dataset), we remove these distinct categories and their associated samples from the target domain. The category relations for the data distribution transfer scenario can be found in Tab. 12, 13, and14. We also show hyperparameters for all downstream tasks are listed in Tab. 11.

Table 11. 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 12. Domain transfer category matching for PTBXL-Super to CPSC2018, 'None' indicates that there is no category from target domain belongs source domain.

Source Domain	Target Domain
HYP	None
NORM	NORM
CD	1AVB, CRBBB, CLBBB
MI	None
STTC	STE, STD

A.3. Performance on all datasets with scaled ECG backbone

We illustrate the linear probing performance of scaled ECG backbones across six downstream datasets at varying data ratios in Fig. 6. Notably, MERL's performance on PTBXL-Super, PTBXL-Sub, PTBXL-Rhythm, and CSN shows a positive correlation with increased ECG backbone sizes across all data ratios. However, this performance trend reaches a plateau on the CPSC2018 and PTBXL-Form datasets. This saturation point could be attributed to the inherent limitations of the datasets or the ceiling effect of MERL's learning capabilities, warranting further investigation. Despite these variations, MERL exhibits consistent superiority over other eSSL approaches in all assessed scenarios, encompassing various backbone

Table 13. Domain transfer category matching for PTBXL-Super to CSN, 'None' indicates that there is no category from target domain belongs source domain.

Source Domain	Target Domain
HYP	RVH, LVH
NORM	SR
CD	2AVB, 2AVB1, 1AVB, AVB, LBBB, RBBB, STDD
MI	MI
STTC	STTC, STE, TWO, STTU, QTIE, TWC

Table 14. Domain transfer category matching for CPSC2018 to CSN.

Source Domain	Target Domain
AFIB	AFIB
VPC	VPB
NORM	SR
1AVB	1AVB
CRBBB	RBBB
STE	STE
PAC	APB
CLBBB	LBBB
STD	STE, STTC, STTU, STDD

sizes, datasets, and data ratios. This comprehensive outperformance underscores MERL's robustness and versatility in diverse settings.

Additionally, we observe that other eSSL approaches demonstrate inconsistent performance shifts when increasing the ECG backbone size. Such fluctuations in performance metrics, hint at potential challenges these approaches face in capturing representative features in scenarios involving extensive data and larger model architectures. This observation raises important questions about the scalability and adaptability of current eSSL methodologies, especially in large-scale data and diverse downstream tasks typical of medical applications.

B. Visualization on Learned ECG Representation

To further investigate the learned ECG representation, we visualize the last layer output of the ECG encoder using three methods: MERL (multimodal), SimCLR (contrastive), and ST-MEM (reconstructive) on the CSN test set. As Fig. 7 shows, MERL distinguishes the samples with different categories even without supervised learning, while both SimCLR and ST-MEM struggle with mixed ECG features from unique categories, even though the pre-training target of SimCLR aims to learn the distinctiveness of each sample. This demonstrates that clinical report supervision benefits the ECG encoder in learning more discriminative ECG features. Additionally, the visualization in Fig. 7 indicates the flaw of the reconstructive method in learning high-level discriminative semantics from ECG, because the pre-training target focuses only on low-level signal patterns (e.g., signal intensity). It also highlights the flaw of the contrastive method, which learns representation from semantically distorted samples, as shown in Fig. 2 (a).

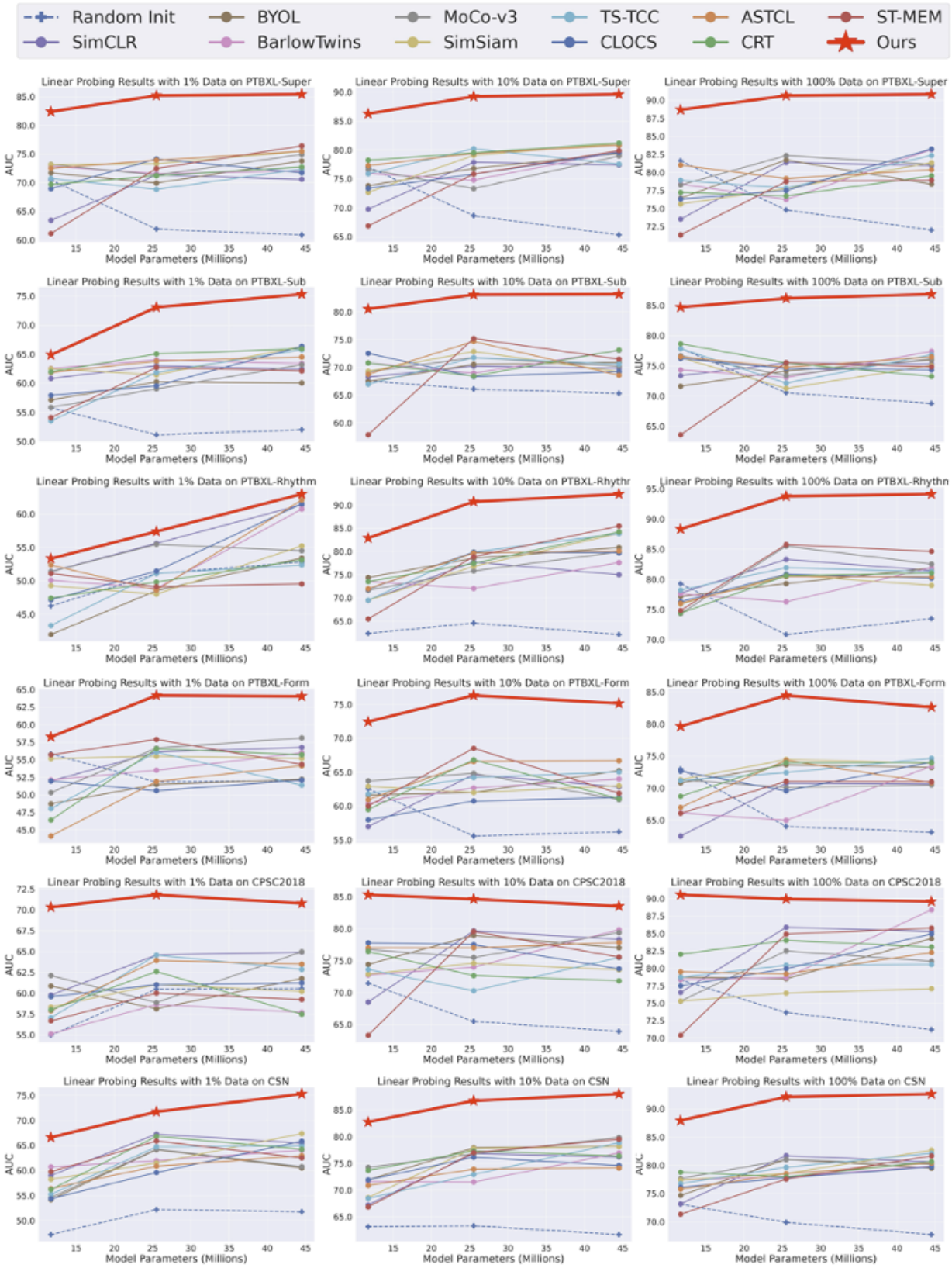

Figure 6. Linear probing performance on six datasets of MERL and other eSSL methods with scaled ECG backbones across [1%, 10%, 100%] data ratios.

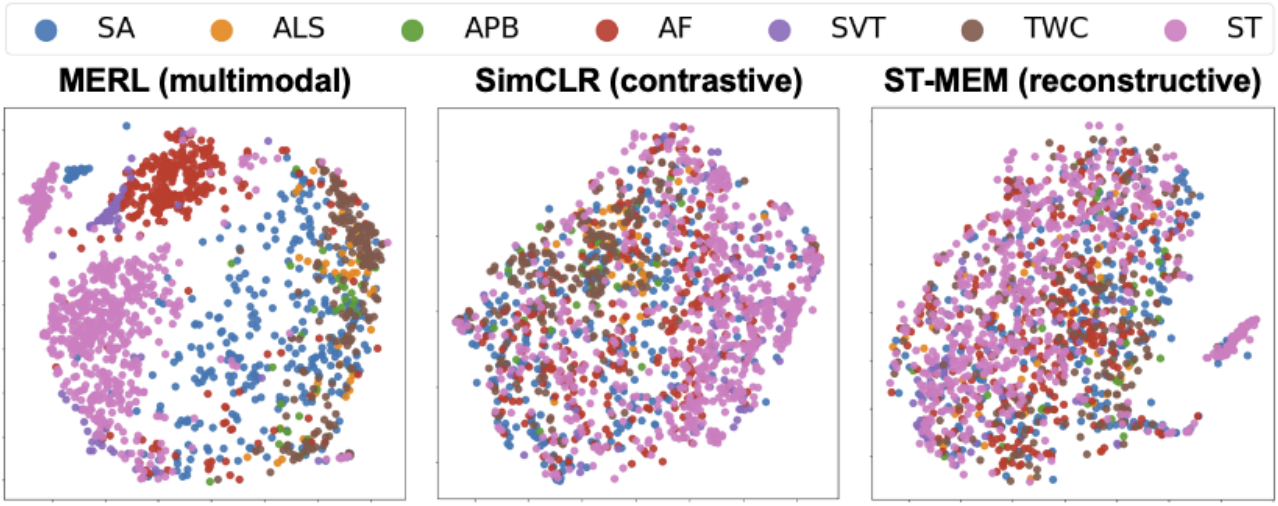

Figure 7. The t-SNE visualization of the learned ECG representation after pre-training. We utilize the test set of the CSN dataset, which includes only samples with unique categories and remove categories that have fewer than 50 samples for better visualization.