# LARGE LANGUAGE MODELS AS TUTORS OF CLINICAL SEMANTICS FOR LEARNING CARDIAC REPRESENTATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

ECG–language pretraining derives its primary advantage over signal-only self-supervised learning (SSL) by incorporating clinical reports, thereby grounding ECG representation learning in explicit clinical semantics. However, existing methods typically treat text as an auxiliary modality jointly learned with ECG signals, pursuing performance gains through increasingly elaborate cross-modal alignment mechanisms. This paper is driven by a fundamental yet largely overlooked question: does the quality of text representations itself constitute a bottleneck for ECG representation learning? To address this, we fundamentally redesign the ECG–language framework. Instead of relying on conventional text encoders, we employ a frozen large language model (LLM) to directly encode clinical reports, leveraging its extensive knowledge priors to generate superior text representations. Furthermore, we introduce targeted architectural and optimization mechanisms that enable effective ECG representation learning under these fixed textual supervisory signals. Extensive experiments across three widely used public datasets demonstrate that our proposed method outperforms existing approaches, despite relying solely on a simple global alignment mechanism.

## 1 INTRODUCTION

Electrocardiography (ECG) is one of the most commonly used cardiovascular diagnostic tests (Siontis et al., 2021). Because high-quality labeled ECG data are costly and time-consuming to obtain, ECG self-supervised learning (SSL) has been widely used to learn ECG representations from unlabeled signals, thereby reducing reliance on manual annotations (Oh et al., 2022; McKeen et al., 2025). However, as its supervision mainly comes from the ECG signal itself, SSL remains limited in learning disease-related high-level semantic representations.

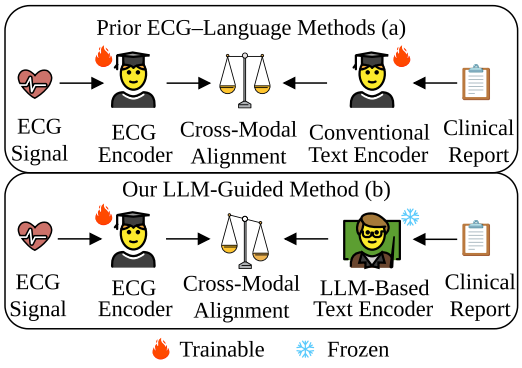



Figure 1: Comparison of (a) conventional joint ECG–text learning and (b) our LLM-guided framework with a frozen text encoder.

To address this limitation, ECG–language methods (Liu et al., 2024), inspired by CLIP-style cross-modal alignment frameworks, align ECG representations with paired report representations through cross-modal contrastive learning, thereby linking signal-level representations to clinical semantics. Although subsequent methods have improved cross-modal alignment mechanisms from different perspectives, their core idea remains unchanged: they treat clinical reports as an auxiliary modality, use a text encoder to obtain text representations from them, and jointly learn these representations with ECG representations during training (Figure 1(a)).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This observation prompts us to reexamine the role of text representations in ECG representation learning. The key distinction of ECG–language methods is that they use text representations to guide ECG representations toward a clinical semantic space. We therefore focus on the question: does the representational capacity of the text encoder itself constitute a bottleneck in ECG representation learning? Prior studies have observed that the choice of text encoder can affect training performance, yet it has primarily been treated as an experimental factor (Wang et al., 2025; Pham Hung et al., 2025). Motivated by this gap, we shift our focus from the design of alignment mechanisms to text representation quality.

Clinical reports in real-world settings commonly exhibit heterogeneity in expression. When the resulting text representations are not sufficiently robust to such heterogeneity, clinically equivalent diagnoses may be mapped to dispersed representations, thereby introducing semantic noise into ECG representation learning. The knowledge priors of large language models (LLMs) provide a possible way to alleviate this issue. For example, Cai et al. (2025) and Liu et al. (2025) use LLMs to extract and normalize diagnosis-related entities from free-text ECG reports, thereby constructing less noisy structured supervisory signals. Other studies have also shown the benefits of using LLMs to process clinical text (Hein et al., 2025). We therefore introduce an LLM into the ECG–language framework to obtain higher-quality text representations; rather than using the LLM as an additional language preprocessing tool, we directly leverage the resulting representations to supervise ECG representation learning (Figure 1(b)).

However, introducing an LLM did not readily translate into the expected performance gains, and instead brought new challenges. On the text side, LLM-derived text representations exhibit pronounced anisotropy, with a few dominant common components accounting for most of the variance (Timkey & van Schijndel, 2021; Ethayarajh, 2019). This obscures discriminative semantic nuances, hindering convergence during cross-modal alignment. On the signal side, effectively matching this high-quality semantic space requires robust ECG encoding; however, per-lead standardization attenuates scale cues across leads that may carry diagnostic information. Finally, from a computational perspective, performing the LLM forward pass during online training incurs substantial GPU memory overhead, constraining the feasible batch size and consequently the number of in-batch negatives available for contrastive learning.

To address these challenges, we introduce a set of targeted architectural and optimization mechanisms, including text-space rectification, scale-aware ECG encoding, offline LLM deployment, and out-of-batch negative set expansion. Together, these mechanisms enable effective ECG representation learning using only a simple global cross-modal alignment mechanism. Overall, our main contributions are as follows:

(1) We propose an ECG pretraining framework that leverages the knowledge priors encoded in a frozen LLM to construct high-quality text representations, which serve as direct and robust supervisory signals for ECG representation learning.

(2) We introduce targeted architectural and optimization mechanisms: text-space rectification to mitigate LLM representation anisotropy, and a scale-aware FiLM module to compensate for the attenuation of scale cues across ECG leads.

(3) We decouple LLM deployment from online training through an offline cache and introduce an out-of-batch negative set expansion strategy, thereby alleviating the memory bottleneck and improving the effectiveness of contrastive learning.

(4) Experiments on three widely used public datasets (PTB-XL, CSN, and CPSC2018) demonstrate that, relying solely on global alignment, our proposed method outperforms existing approaches.

## 2 Related Work

### 2.1 ECG Representation Learning

ECG SSL can exploit unlabeled ECG signals to capture morphological and temporal structures (Chen et al., 2020; Yang & Hong, 2022; Jin et al., 2025; Na et al., 2024), but its supervision is derived from the ECG modality itself, thereby lacking explicit constraints from clinical semantics. To address this limitation, recent ECG–language studies have adopted cross-modal align-

ment methods. Liu et al. (2024) proposed MERL, a dual-encoder architecture that separately extracts representations from ECG signals and clinical reports, and uses global contrastive learning to pull paired ECG–text samples closer while pushing unpaired samples apart. Subsequent studies, such as Wang et al. (2025) and Chen et al. (2025), further enhance cross-modal interaction through multi-granularity alignment and multi-stage alignment, respectively, achieving stronger performance. These methods refine cross-modal alignment mechanisms from different perspectives, but they pay limited attention to the representational capacity of the text encoder itself. In fact, these studies have suggested that performance is sensitive to the quality of text representations: more capable or ECG-domain-adapted text encoders tend to yield better performance (Pham Hung et al., 2025; Liu et al., 2024). Therefore, if future improvements remain focused solely on alignment mechanisms, the resulting gains may eventually be constrained by the representational ceiling of the text encoder.

### 2.2 Large Language Models for Clinical Text

Existing studies have used LLMs to process clinical reports with the goal of improving downstream task performance. Some methods use LLMs to rewrite or refine clinical reports, thereby alleviating expression heterogeneity and improving terminology normalization (Kugic et al., 2024; Hein et al., 2025). Others use LLMs for clinical knowledge enhancement, thereby providing additional reference information for diagnosis (Tang et al., 2026; Tian et al., 2024). This work also leverages the knowledge priors encoded in an LLM, but not by using it as an auxiliary language preprocessing tool. Instead, we use a frozen LLM to directly encode clinical reports and obtain higher-quality text representations.

## 3 Method

### 3.1 Overall Framework

The overall framework of the proposed method is illustrated in Figure 2. We adopt a dual-branch architecture that uses the semantic information in clinical reports to guide ECG representation learning. For the $i$-th ECG–report pair, its ECG record and clinical report are denoted by $x_i$ and $r_i$, respectively. The text and ECG branches independently encode the paired inputs into global representations, which are subsequently aligned through cross-modal contrastive learning. Unlike methods that rely on complex cross-modal alignment mechanisms, our framework performs alignment only at the global representation level, allowing us to more directly examine how text representation quality affects ECG representation learning.

### 3.2 Dual-Branch Encoders

#### 3.2.1 LLM-based Text Encoder

For the text branch, we use a frozen large language model to encode each clinical report. Because the representations from the final layer are typically more closely tied to the autoregressive generation objective, we use the token-level hidden states from the penultimate Transformer layer, denoted by $H_i$. We then apply average pooling over the token dimension of $H_i$ and perform $\ell_2$ normalization, yielding the initial global text representation $t_i$.

#### 3.2.2 Scale-Aware ECG Encoder

The ECG branch combines per-lead standardization with feature modulation based on scale statistics. To reduce differences in amplitude scale across ECG records, we standardize each lead separately. For the $c$-th lead of the $i$-th ECG record, we compute the mean $\mu_{i,c}$ and standard deviation $\sigma_{i,c}$ over its effective sampling interval, and standardize the signal value at the $n$-th sampling point as

$$\tilde{x}_{i,c,n} = \frac{x_{i,c,n} - \mu_{i,c}}{\sigma_{i,c} + \epsilon},$$

where $\epsilon$ is a constant used for numerical stability. Detailed computations of the mean and standard deviation are provided in the Supplementary Document.

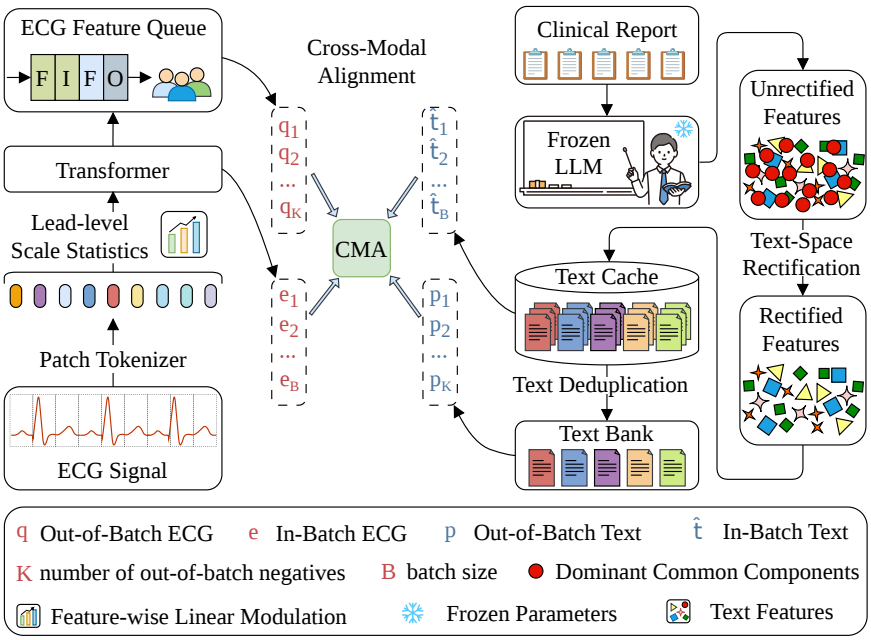



Figure 2: Overview of the proposed framework. The text branch uses a frozen LLM to encode clinical reports, and the rectified text representations are cached offline and deduplicated to construct a text bank. The ECG branch uses lead-level scale statistics before standardization to modulate the standardized ECG tokens through FiLM, which are then encoded by a Transformer. The ECG representations from the current batch and the out-of-batch representations from the ECG feature queue form the ECG candidate pool, while the text representations from the current batch and the out-of-batch representations from the text bank form the text candidate pool. Based on these two candidate pools, the model is optimized using the bidirectional cross-modal alignment (CMA) loss.

However, per-lead standardization attenuates scale cues across leads. To compensate for this potential information loss, we introduce a lightweight Feature-wise Linear Modulation (FiLM) branch to modulate the standardized ECG tokens. Specifically, we construct a scale vector $s_i$ from the pre-standardization lead-wise standard deviations, whose $c$-th component is $(s_i)_c = \log(\sigma_{i,c} + \epsilon)$. A lightweight multilayer perceptron then maps $s_i$ to feature-wise scaling and shift parameters, given by $[\gamma_i; \beta_i] = \mathrm{MLP}(s_i)$.

Meanwhile, the standardized multi-lead ECG is partitioned along the temporal axis into $T$ non-overlapping patches, each comprising the local signals from all leads within the same temporal window. A convolutional module then encodes these patches into the corresponding ECG tokens. Let $h_{i,j}$ denote the $j$-th token of the $i$-th ECG record. All tokens from the same record share the modulation parameters generated from $s_i$ and are modulated as

$$\tilde{h}_{i,j} = h_{i,j} \odot (1 + \gamma_i) + \beta_i, \qquad j = 1, \ldots, T,$$

where $\odot$ denotes element-wise multiplication. The modulated tokens are subsequently fed into a Transformer-based ECG encoder, followed by global average pooling to obtain the global ECG representation $e_i$.

### 3.3 Cross-Modal Alignment Strategy

#### 3.3.1 Text-Space Rectification

Before performing cross-modal alignment, we rectify the initial global text representations extracted by the LLM. Directly aligning these representations with ECG representations leads to poor training convergence, mainly because of anisotropy in the text representation space: a small number of dominant common components explain a large proportion of the variance and obscure more discrim-

inative semantic information. We therefore suppress these common components to obtain a more suitable text representation space for cross-modal alignment (Mu & Viswanath, 2018).

Specifically, we fit PCA using only the initial global text representations from the training split of the pretraining dataset, obtaining the training-set mean and the top four principal directions, which form the matrix $U$. Each text representation is centered using this mean to obtain $t_i^c$, after which its projection onto these directions is removed. Detailed computations are provided in the Supplementary Document. The rectified text representation is defined as:

$$\hat{t}_i = t_i^c - UU^\top t_i^c.$$

This operation suppresses dominant common components and makes the text representations more suitable for cross-modal alignment.

#### 3.3.2 Memory-Efficient Negative Expansion

Contrastive learning benefits from a sufficiently large negative candidate pool, whereas performing LLM inference during online training introduces substantial GPU memory and computational overhead. Because the LLM remains frozen throughout training and the set of pretraining reports is predetermined, we precompute the rectified representation $\hat{t}_i$ for every report in the training split of the pretraining dataset and store these representations in an offline text cache $\mathcal{M}_t$. This decouples text encoding from ECG representation learning. During training, the text representation paired with each ECG record is loaded directly from the cache, without loading the LLM or repeating its forward computation, thereby reducing the memory and computational costs of the text branch.

Building on this memory-efficient decoupling, we further extend the source of negative samples from the current batch to out-of-batch candidate sets. This reduces the dependence of the candidate pool size on the training batch size and enables a larger negative candidate pool under limited GPU memory (He et al., 2020; Li et al., 2021).

For the text branch, we construct a text bank $\mathcal{B}_t$ by deduplicating the offline cache $\mathcal{M}_t$ according to report content, retaining one rectified representation for each distinct report. Let $\mathcal{I}$ denote the index set of the current batch. Before sampling, we remove from $\mathcal{B}_t$ all representations whose reports either have content identical to a report in the current batch or are associated with the same ECG examination as any sample in the current batch. We then randomly sample a shared out-of-batch text negative set $\mathcal{N}^t$ from the remaining representations. The shared text candidate pool in the ECG-to-Text direction is

$$\mathcal{C}^t = \mathrm{concat}\big((\hat{t}_j)_{j\in\mathcal{I}}, \mathcal{N}^t\big)\,.$$

Here, $\hat{t}_j$ is the rectified text representation paired with ECG record $x_j$. For the ECG branch, we maintain a fixed-capacity online feature queue $\mathcal{Q}$ that stores ECG representations from recent batches after they have been detached from the computation graph. The queue is updated in a first-in, first-out (FIFO) manner: after each training step, the ECG representations from the current batch are appended to the queue, and the earliest representations are removed when its preset capacity is exceeded. Before sampling, we exclude queued representations from the same ECG examination as any sample in the current batch and randomly sample a shared out-of-batch ECG negative set $\mathcal{N}^e$ from the remaining representations. The shared ECG candidate pool in the Text-to-ECG direction is

$$\mathcal{C}^e = \mathrm{concat}\big((e_j)_{j\in\mathcal{I}}, \mathcal{N}^e\big)\,.$$

Here, $e_j$ is the global ECG representation of sample $x_j$ in the current batch. Overall, the offline text bank and the online ECG feature queue provide out-of-batch negatives for the two alignment directions, respectively. They expand the candidate space from the local in-batch set to a larger set jointly formed by the current batch and cached representations, thereby increasing the number of negative samples under limited GPU memory.

### 3.4 Training Objective

Based on the expanded cross-modal candidate pools defined above, the model is optimized using the cross-modal alignment (CMA) loss. Two learnable linear projection heads map the global ECG representation $e_i$ and the rectified text representation $\hat{t}_i$ into a shared alignment space. The projected representations are then $\ell_2$-normalized, yielding $z_i^e$ and $z_i^t$, respectively. Each representation in the

Table 1: Linear probing AUROC (%) on downstream datasets.

| Method | PTBXL-Sub | | | PTBXL-Super | | | PTBXL-Form | | | PTBXL-Rhythm | | | CSN | | | CPSC2018 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% |
| SimCLR | 60.84 | 68.27 | 73.39 | 63.41 | 69.77 | 73.53 | 54.98 | 56.97 | 62.52 | 51.41 | 69.44 | 77.73 | 59.02 | 67.26 | 73.20 | 59.78 | 68.52 | 76.54 |
| BYOL | 57.16 | 67.44 | 71.64 | 71.70 | 73.83 | 76.45 | 48.73 | 61.63 | 70.82 | 41.99 | 74.40 | 77.17 | 54.20 | 71.92 | 74.69 | 60.88 | 74.42 | 78.75 |
| Barlow Twins | 62.57 | 70.84 | 74.34 | 72.87 | 75.96 | 78.41 | 52.12 | 60.39 | 66.14 | 50.12 | 73.54 | 77.62 | 60.72 | 71.64 | 77.43 | 55.12 | 72.75 | 78.39 |
| MoCo-v3 | 55.88 | 69.21 | 76.69 | 73.19 | 76.65 | 78.26 | 50.32 | 63.71 | 71.31 | 51.38 | 71.66 | 74.33 | 54.61 | 74.26 | 77.68 | 62.13 | 76.74 | 75.29 |
| SimSiam | 62.52 | 69.31 | 76.38 | 73.15 | 72.70 | 75.63 | 55.16 | 62.91 | 71.31 | 49.30 | 69.47 | 75.92 | 58.25 | 68.61 | 77.41 | 58.35 | 72.89 | 75.31 |
| TS-TCC | 53.54 | 66.98 | 77.87 | 70.73 | 75.88 | 78.91 | 48.04 | 61.79 | 71.18 | 43.34 | 69.48 | 78.23 | 55.26 | 68.48 | 76.79 | 57.07 | 73.62 | 78.72 |
| CLOCS | 57.94 | 72.55 | 76.24 | 68.94 | 73.36 | 76.31 | 51.97 | 57.96 | 72.65 | 47.19 | 71.88 | 76.31 | 54.38 | 71.93 | 76.13 | 59.59 | 77.78 | 77.49 |
| CRT | 61.98 | 70.82 | 78.67 | 69.68 | 78.24 | 77.24 | 46.41 | 59.49 | 68.73 | 47.44 | 73.52 | 74.41 | 56.21 | 73.70 | 78.80 | 58.01 | 76.43 | 82.03 |
| ASTCL | 61.86 | 68.77 | 76.51 | 72.51 | 77.31 | 81.02 | 44.14 | 60.93 | 66.99 | 52.38 | 71.98 | 76.05 | 56.40 | 70.87 | 75.79 | 57.90 | 77.01 | 79.51 |
| ST-MEM | 54.12 | 57.86 | 63.59 | 61.12 | 66.87 | 71.36 | 55.71 | 59.99 | 66.07 | 51.12 | 65.44 | 74.85 | 59.77 | 66.87 | 71.36 | 56.69 | 63.32 | 70.39 |
| HeartLang | 64.68 | 79.34 | 88.91 | 78.94 | 85.59 | 87.52 | 58.70 | 63.99 | 80.23 | 62.08 | 76.22 | 90.34 | 57.94 | 68.93 | 82.49 | 60.44 | 66.26 | 77.87 |
| MERL | 64.90 | 80.56 | 84.72 | 82.39 | 86.27 | 88.67 | 58.26 | 72.43 | 79.65 | 53.33 | 82.88 | 88.34 | 66.60 | 82.74 | 87.95 | 70.33 | 85.32 | 90.57 |
| ECG2TOK | 72.31 | 79.42 | 84.52 | 81.23 | 85.68 | 87.99 | 55.88 | 73.78 | 85.28 | 81.16 | 89.41 | 91.41 | 72.66 | 85.14 | 92.84 | 68.42 | 83.41 | 91.04 |
| K-MERL | 68.22 | 81.54 | 88.00 | 84.19 | 87.71 | 89.83 | 60.11 | 73.71 | 81.48 | 63.72 | 84.16 | 91.04 | 69.51 | 83.53 | 93.71 | 71.91 | 86.13 | 91.26 |
| SuPreME | 66.30 | 74.20 | 84.84 | 73.58 | 79.07 | 87.67 | 58.94 | 58.93 | 74.06 | 56.92 | 76.27 | 84.42 | 65.42 | 75.08 | 89.16 | 58.28 | 70.51 | 86.74 |
| MELP | 79.22 | 84.40 | 87.46 | 85.82 | 87.61 | 87.87 | 63.41 | 76.71 | 83.30 | 88.83 | 94.65 | 96.91 | 78.25 | 84.83 | 90.17 | **88.54** | 91.75 | 94.32 |
| CG-DMER | 69.60 | 81.58 | 87.67 | 84.22 | 88.18 | 90.31 | 60.25 | 73.95 | 81.01 | 64.47 | 84.69 | 90.76 | 70.70 | 83.21 | 93.55 | 71.70 | 86.43 | 91.32 |
| CLEAR | 73.86 | 80.11 | 89.30 | 78.59 | 85.43 | 88.68 | 61.00 | 69.96 | 80.34 | 79.24 | 86.72 | 93.66 | 62.88 | 76.33 | 86.75 | 64.82 | 78.10 | 89.59 |
| Ours | **81.49** | **85.39** | **89.33** | **88.65** | **91.28** | **92.62** | **67.93** | **80.72** | **88.77** | **89.79** | **94.68** | **97.10** | **83.08** | **90.76** | **95.31** | 83.90 | **95.38** | **96.64** |

candidate pools $\mathcal{C}^t$ and $\mathcal{C}^e$ is processed by the corresponding projection head and normalized in the same manner, yielding the text candidate embedding pool $\mathcal{Z}^t$ and the ECG candidate embedding pool $\mathcal{Z}^e$. To avoid false negatives caused by duplicate reports or multiple candidates from the same ECG examination, for each $i \in \mathcal{I}$, we mask all in-batch text candidates indexed by $j \neq i$ whose report content is identical to that of $r_i$, while retaining the paired positive candidate $z_i^t$. The remaining valid text candidates are denoted by $\tilde{\mathcal{Z}}_i^t$. Similarly, in the Text-to-ECG direction, we mask all in-batch ECG candidates indexed by $j \neq i$ that originate from the same ECG examination as $x_i$, while retaining the paired positive candidate $z_i^e$. The remaining valid ECG candidates are denoted by $\tilde{\mathcal{Z}}_i^e$.

The contrastive losses in the ECG-to-Text ($E \to T$) and Text-to-ECG ($T \to E$) directions share the following form:

$$\mathcal{L}_{a\to b} = -\frac{1}{B}\sum_{i\in\mathcal{I}} \log \frac{\exp\big(\mathrm{sim}(z_i^a, z_i^b)/\tau\big)}{\sum_{z\in\tilde{\mathcal{Z}}_i^b}\exp(\mathrm{sim}(z_i^a, z)/\tau)}.$$

The choices $(a, b) = (e, t)$ and $(t, e)$ correspond to $\mathcal{L}_{E\to T}$ and $\mathcal{L}_{T\to E}$, respectively, using their direction-specific valid candidate pools. $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, $\tau$ is the temperature parameter, and $B = |\mathcal{I}|$ is the number of paired samples in the current batch. The final CMA loss is

$$\mathcal{L}_{\mathrm{CMA}} = \lambda_1\mathcal{L}_{E\to T} + \lambda_2\mathcal{L}_{T\to E},$$

where $\lambda_1$ and $\lambda_2$ are the loss weights for the ECG-to-Text and Text-to-ECG directions, respectively. We set $\lambda_1 : \lambda_2 = 20 : 1$, assigning greater weight to the ECG-to-Text direction to strengthen the supervision imposed by the text representations on ECG representation learning.

# 4 Experiments

## 4.1 Experimental Setup

**Pretraining Dataset.** Pretraining is conducted on the MIMIC-IV-ECG-v1.0 dataset (Gow et al., 2023). This dataset contains 800,035 ECG records from 161,352 unique subjects. Each record corresponds to a 10-s, 12-lead ECG recording sampled at 500 Hz. A single ECG record can be associated with up to 18 clinical text reports. To ensure data quality, we remove the following samples: (1) samples whose reports contain fewer than four words; (2) samples containing NaN or Inf values; and (3) samples for which more than 0.5% of the sampling points in a single ECG record have absolute amplitudes exceeding 10 mV. This leaves 761,064 ECG–report pairs.

**Implementation.** Meta Llama 3.1 8B is used as the frozen LLM in the text branch (Grattafiori et al., 2024), while a WavLM-style Transformer encoder is used in the ECG branch (Chen et al.,

Table 2: Cross-domain transfer AUROC (%) on downstream datasets.

| Source Domain | PTBXL-Super | | CPSC2018 | | CSN | |
|---|---|---|---|---|---|---|
| Target Domain | CPSC2018 | CSN | PTBXL-Super | CSN | PTBXL-Super | CPSC2018 |
| SimCLR | 69.62 | 73.05 | 56.65 | 66.36 | 59.74 | 62.11 |
| BYOL | 70.27 | 74.01 | 57.32 | 67.56 | 60.39 | 63.24 |
| Barlow Twins | 68.98 | 72.85 | 55.97 | 65.89 | 58.76 | 61.35 |
| MoCo-v3 | 69.41 | 73.29 | 56.54 | 66.12 | 59.82 | 62.07 |
| SimSiam | 70.06 | 73.92 | 57.21 | 67.48 | 60.23 | 63.09 |
| TS-TCC | 71.32 | 75.16 | 58.47 | 68.34 | 61.55 | 64.48 |
| CLOCS | 68.79 | 72.64 | 55.86 | 65.73 | 58.69 | 61.27 |
| ASTCL | 69.23 | 73.18 | 56.61 | 66.27 | 59.74 | 62.12 |
| CRT | 70.15 | 74.08 | 57.39 | 67.62 | 60.48 | 63.33 |
| ST-MEM | 76.12 | 84.50 | 62.27 | 75.19 | 73.05 | 64.66 |
| MERL | 85.84 | 85.13 | 81.81 | 82.81 | 77.51 | 83.58 |
| MELP | 85.77 | 87.83 | **84.59** | 90.25 | 77.26 | 86.06 |
| Ours | **88.33** | **91.29** | 82.22 | **91.76** | **80.13** | **90.12** |

Table 3: Zero-shot AUROC (%) on downstream datasets.

| Method | PTBXL-Sub | PTBXL-Super | PTBXL-Form | PTBXL-Rhythm | CSN | CPSC2018 |
|---|---|---|---|---|---|---|
| MERL | 75.05 | 73.76 | 58.38 | 78.25 | **73.61** | 77.14 |
| MELP | 73.78 | 71.50 | 63.32 | **83.42** | 71.69 | 76.35 |
| Ours | **77.47** | **79.61** | **64.27** | 78.99 | 72.02 | **83.01** |

2022). The model is optimized using AdamW with an initial learning rate of $3 \times 10^{-4}$, and pretraining is conducted for one epoch. Complete implementation details and training configurations are provided in the Supplementary Document.

**Downstream Datasets.** We evaluate the proposed model on three widely used public datasets: PTB-XL (Wagner et al., 2020), Chapman-Shaoxing-Ningbo (CSN) (Zheng et al., 2020), and CPSC2018 (Liu et al., 2018). Following prior work (Liu et al., 2024; Wang et al., 2025), we evaluate PTB-XL on four multilabel classification tasks: Superclass (5 classes), Subclass (23 classes), Form (19 classes), and Rhythm (12 classes). We evaluate CSN and CPSC2018 on multilabel classification tasks with 38 and 9 labels, respectively. Detailed dataset descriptions and statistics are provided in the Supplementary Document.

**Evaluation Protocol.** For all three downstream datasets, we follow the training/validation/test splits used in prior work (Liu et al., 2024; Wang et al., 2025). For linear probing, we freeze the pretrained ECG encoder and train a randomly initialized linear classifier on top of it using 1%, 10%, and 100% of the downstream training data. To assess generalization under distribution shift, we further conduct cross-domain transfer experiments by performing linear probing using the full source-domain training set and evaluating on the target-domain test set after label-space alignment. Following prior work (Liu et al., 2024; Wang et al., 2025; Jin et al., 2025), we use macro AUROC as the evaluation metric for all downstream tasks. Detailed randomization settings, checkpoint selection rules, baseline result sources, label mappings, and training configurations are provided in the Supplementary Document.

### 4.2 Quantitative Results

**Linear Probing.** Table 1 reports linear probing results for our method and a diverse set of baselines. The baselines cover both self-supervised and multimodal approaches, including SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), Barlow Twins (Zbontar et al., 2021), MoCo-v3 (Chen et al., 2021), SimSiam (Chen & He, 2021), TS-TCC (Eldele et al., 2021), CLOCS (Kiyasseh et al., 2021), CRT (Zhang et al., 2024), ASTCL (Wang et al., 2024), ST-MEM (Na et al., 2024), HeartLang (Jin et al., 2025), MERL (Liu et al., 2024), K-MERL (Liu

et al., 2025), SuPreME (Cai et al., 2025), ECG2TOK (Yuan et al., 2025), MELP (Wang et al., 2025), CG-DMER (Niu et al., 2026), and CLEAR (Pan et al., 2026).

The results show that, across the 18 linear probing settings, our method achieves the best performance in 17 settings and ranks second in the remaining one. Overall, our method attains an average AUROC of 88.49%, surpassing the second-best method, MELP, by 2.71 percentage points. Notably, on PTBXL-Super, our method reaches 91.28% using only 10% of the training data, outperforming all baselines under the 100% training-data setting. On CPSC2018, using only 10% of the training data, our method exceeds MELP, the best baseline under the 100% training-data setting, by 1.06 percentage points. These results indicate that using LLM-derived text representations from clinical reports as supervisory signals can effectively improve ECG representation learning.

**Robustness to Distribution Shift.** As shown in Table 2, our method achieves the best performance in five of the six cross-domain evaluation settings, with an average AUROC of 87.31%, surpassing MELP, the second-best method, by 2.02 percentage points. These results indicate that our method exhibits strong cross-domain generalization under domain shift.

**Zero-Shot Evaluation.** To evaluate the transferability of the learned ECG representations in the zero-shot setting, we compare our method with strong ECG–language baselines, MELP (Wang et al., 2025) and MERL (Liu et al., 2024), under a unified evaluation framework. To ensure a fair comparison, all methods use the same MEDCPT (Jin et al., 2023) text encoder as MERL and load its official pretrained weights to extract label-text representations. Newly initialized projection heads are then trained on MIMIC-IV-ECG only, and downstream prediction is performed by matching the ECG representations of each method with the label-text representations via cosine similarity. Detailed evaluation settings are provided in the Supplementary Document. As shown in Table 3, our method achieves an average AUROC of 75.90%, surpassing MELP (73.34%) and MERL (72.70%) by 2.55 and 3.20 percentage points, respectively. These results further demonstrate the superior transferability of the learned ECG representations.

### 4.3 Ablation Study

Following the ablation protocol adopted in MERL (Liu et al., 2024), we conduct all ablation experiments under linear probing using only 1% of the labeled training data. This low-resource setting limits the influence of downstream supervision, thereby providing a more stringent assessment of how each component contributes to the transferability of the pretrained ECG representations.

**Hyperparameter Analysis.** We examine batch-size and learning-rate configurations, as well as the weighting ratio of the bidirectional contrastive loss; complete results are provided in the Supplementary Document. Performance does not improve monotonically with batch size and reaches its optimum at 256. The 20:1 loss-weight ratio achieves the best performance; changing it to 1:1 decreases the average AUROC from 82.47% to 81.59%, while increasing it to 30:1 also leads to a performance drop. These results suggest that moderately strengthening the supervision from LLM-derived text representations benefits ECG representation learning. Subsequent component ablations use the standard 1:1 loss-weight setting as the baseline.

**Analysis of Text-Space Rectification.** This analysis examines the effect of text-space rectification (TSR) on downstream performance, optimization stability, and text representation geometry. As shown in Figure 4, removing TSR decreases the average AUROC from 81.59% to 75.29%, corresponding to a drop of 6.30 percentage points. As shown in Figure 3(a), without rectification, the training loss plateaus early after decreasing to around 6. In contrast, with TSR, the training loss continues to decrease and eventually stabilizes around 3, indicating that TSR improves optimization stability when aligning ECG representations with LLM-derived text representations.

Figures 3(b) and 3(c) further explain this effect from the text side. Without TSR, the mean pairwise cosine similarity remains consistently high, indicating that different text representations are overly similar to each other. Meanwhile, the fraction of variance explained by the top-4 principal components stays close to 1, suggesting that the representation space is dominated by a few common directions. After applying TSR, both quantities decrease substantially. These results provide

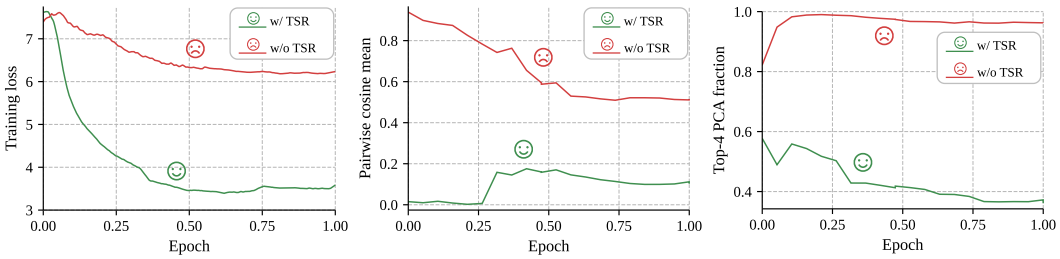



Figure 3: Analysis of text-space rectification (TSR). (a) Training loss curves with and without TSR. (b) Mean pairwise cosine similarity among text representations. (c) Fraction of variance explained by the top-4 principal components. In (b) and (c), all statistics are computed on text representations after the projection head.

evidence that TSR alleviates anisotropy in the LLM text representation space by reducing excessive similarity among text representations and suppressing dominant common components.

**Analysis of the Number of Out-of-Batch Negatives.** As shown in Figure 4, this analysis examines the effect of the number of out-of-batch negatives. The preceding experiments show that merely increasing the number of in-batch negatives does not yield sustained gains. We therefore vary the number of out-of-batch negatives among 0, 2,048, 4,096, 8,192, and 16,384, yielding average AUROCs of 80.98%, 81.28%, 81.59%, 81.13%, and 81.48%, respectively. Performance peaks at 4096 out-of-batch negatives and does not improve with larger negative sets. This suggests that more negatives are not always beneficial: additional negatives may provide limited contrastive information and can include semantically close false negatives, thereby weakening the effectiveness of the contrastive objective.

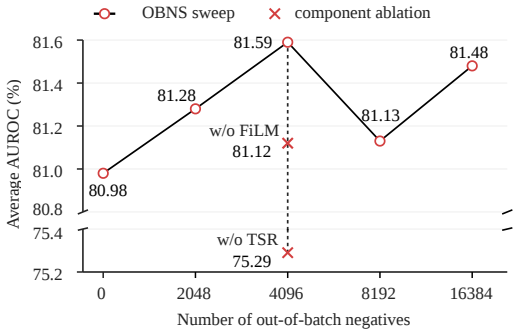



Figure 4: Component ablations and an ablation on the number of out-of-batch negatives. Circles show the sweep over the number of out-of-batch negatives; crosses show component ablations.

**Analysis of FiLM.** Feature-wise Linear Modulation (FiLM) is introduced to compensate for the attenuation of scale cues across leads caused by per-lead standardization. As shown in Figure 4, although FiLM adds only about 4.76M parameters, removing it decreases the average AUROC from 81.59% to 81.12%, corresponding to a drop of 0.47 percentage points. Per-lead standardization reduces differences in amplitude scale across ECG records, but also attenuates scale cues across leads. These cues may provide complementary discriminative information. By explicitly incorporating lead-level statistics, FiLM partially compensates for their attenuation and yields modest additional gains.

## 5 Conclusion and Limitations

In this work, we propose an ECG–language pretraining framework that uses a frozen LLM to directly encode clinical reports, with the resulting text representations serving as supervisory signals for ECG representation learning. Combined with scale-aware ECG encoding, text-space rectification, offline LLM text encoding, and out-of-batch negative set expansion, the proposed method enables effective ECG representation learning under a simple global cross-modal alignment mechanism. Despite the encouraging results, the current framework does not leverage the generative reasoning capability of the LLM for clinical diagnosis. In addition, its predictions are not yet accompanied by interpretable evidence from ECG signals. Addressing these limitations remains future work.

## AI USE STATEMENT

Generative AI tools were used to assist with language polishing, manuscript formatting, and code debugging. All AI-assisted work was reviewed and verified by the authors, who take full responsibility for the scientific content, experimental results, and final manuscript.

## ETHICS STATEMENT

This work uses only previously released research datasets and does not involve new participant recruitment or clinical data collection. We follow the access conditions, license terms, and citation requirements of each dataset.

## REPRODUCIBILITY STATEMENT

The Supplementary Document provides further details on implementation and data preprocessing, downstream evaluation protocols, hyperparameter configurations, and additional experimental results. It also lists access information and licensing terms for the datasets, pretrained models, and baseline resources used in this work.

## REFERENCES

Mingsheng Cai, Jiuming Jiang, Wenhao Huang, Che Liu, and Rossella Arcucci. SuPreME: A supervised pre-training framework for multimodal ECG representation learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11817–11844, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.633. URL https://aclanthology.org/2025.findings-emnlp.633/.

Jian Chen, Xiaoru Dong, Wei Wang, Shaorui Zhou, Lequan Yu, and Xiping Hu. DERI: Cross-modal ECG representation learning with deep ECG-report interaction. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 4824–4832. International Joint Conferences on Artificial Intelligence Organization, August 2025. doi: 10.24963/ijcai.2025/537. URL https://doi.org/10.24963/ijcai.2025/537. Main Track.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://doi.org/10.1109/JSTSP.2022.3188113.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/

html/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2352–2359. International Joint Conferences on Artificial Intelligence Organization, August 2021. doi: 10.24963/ijcai.2021/324. URL https://doi.org/10.24963/ijcai.2021/324. Main Track.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent – a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

David Hein, Alana Christie, Michael Holcomb, Bingqing Xie, AJ Jain, Joseph Vento, Neil Rakheja, Ameer Hamza Shakur, Scott Christley, Lindsay G. Cowell, James Brugarolas, Andrew R. Jamieson, and Payal Kapur. Iterative refinement and goal articulation to optimize large language models for clinical information extraction. *npj Digital Medicine*, 8(1):301, 2025. doi: 10.1038/s41746-025-01686-z. URL https://doi.org/10.1038/s41746-025-01686-z.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Hz1Ko087B.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651. URL https://doi.org/10.1093/bioinformatics/btad651.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

Amila Kugic, Stefan Schulz, and Markus Kreuzthaler. Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association*, 31(9):2040–2046, 2024. doi: 10.1093/jamia/ocae157. URL https://doi.org/10.1093/jamia/ocae157.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/505259756244493872b7709a8a01b536-Abstract.html.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31949–31963. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bg.html.

Che Liu, Cheng Ouyang, Zhongwei Wan, Haozhe Wang, Wenjia Bai, and Rossella Arcucci. Knowledge-enhanced multimodal ECG representation learning with arbitrary-lead inputs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7298–7316, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.385. URL https://aclanthology.org/2025.findings-emnlp.385/.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Yin Kwee Ng. 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. doi: 10.1166/jmihi.2018.2442. URL https://doi.org/10.1166/jmihi.2018.2442.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025. doi: 10.1093/jamiaopen/ooaf122. URL https://doi.org/10.1093/jamiaopen/ooaf122.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WcOohbsF4H.

Ziwei Niu, Hao Sun, Shujun Bian, Xihong Yang, Lanfen Lin, Yuxin Liu, and Yueming Jin. CG-DMER: Hybrid contrastive-generative framework for disentangled multimodal ECG representation learning. In *ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2026. IEEE. doi: 10.1109/ICASSP55912.2026.11461612. URL https://doi.org/10.1109/ICASSP55912.2026.11461612.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of ECG. In Gerardo Flores, George H. Chen, Tom Pollard, Joyce C. Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 338–353. PMLR, 7–8 Apr 2022. URL https://proceedings.mlr.press/v174/oh22a.html.

Tan Pan, Yixuan Sun, Chen Jiang, Qiong Gao, Rui Sun, Xingmeng Zhang, Zhenqi Yang, Limei Han, Yixiu Liang, Yuan Cheng, and Kaiyu Guo. Tracing the heart's pathways: ECG representation learning from a cardiac conduction perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2):890–898, 2026. doi: 10.1609/aaai.v40i2.37057. URL https://doi.org/10.1609/aaai.v40i2.37057.

Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting masked ECG-text auto-encoders as discriminative learners. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49277–49291. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pham-hung25a.html.

Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia, and Paul A. Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021. doi: 10.1038/s41569-020-00503-2. URL https://doi.org/10.1038/s41569-020-00503-2.

Jialu Tang, Hung Manh Pham, Ignace De Lathauwer, Henk S. Schipper, Yuan Lu, Dong Ma, and Aaqib Saeed. Interpretable multimodal zero shot ECG diagnosis via structured clinical knowledge alignment. *npj Cardiovascular Health*, 3(1):1, 2026. doi: 10.1038/s44325-025-00099-x. URL https://doi.org/10.1038/s44325-025-00099-x.

Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5(12):101875, 2024. doi: 10.1016/j.xcrm.2024.101875. URL https://doi.org/10.1016/j.xcrm.2024.101875.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Florian I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL: A large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6. URL https://doi.org/10.1038/s41597-020-0495-6.

Fuying Wang, Jiacheng Xu, and Lequan Yu. From token to rhythm: A multi-scale approach for ECG-language pretraining. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65059–65074. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wang25du.html.

Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13845–13859, 2024. doi: 10.1109/TNNLS.2023.3272153. URL https://doi.org/10.1109/TNNLS.2023.3272153.

Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25038–25054. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yang22e.html.

Xiaoyan Yuan, Wei Wang, Han Liu, Jian Chen, and Xiping Hu. ECG2TOK: ECG pre-training with self-distillation semantic tokenizers. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 9990–9998. International Joint Conferences on Artificial Intelligence Organization, August 2025. doi: 10.24963/ijcai.2025/1110. URL https://doi.org/10.24963/ijcai.2025/1110. AI and Social Good.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16129–16138, 2024. doi: 10.1109/TNNLS.2023.3292066. URL https://doi.org/10.1109/TNNLS.2023.3292066.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, 2020. doi: 10.1038/s41597-020-0386-x. URL https://doi.org/10.1038/s41597-020-0386-x.

# SUPPLEMENTARY DOCUMENT FOR LARGE LANGUAGE MODELS AS TUTORS OF CLINICAL SEMANTICS FOR LEARNING CARDIAC REPRESENTATIONS

**Anonymous authors**
Paper under double-blind review

## 1 ADDITIONAL METHOD DETAILS

### 1.1 DETAILS OF INPUT STANDARDIZATION

For the $c$-th lead of the $i$-th ECG record, let $L_i$ denote the number of valid, non-padded samples used for standardization. The per-lead mean and standard deviation are computed as

$$\mu_{i,c} = \frac{1}{L_i}\sum_{t=1}^{L_i} x_{i,c,t}, \qquad \sigma_{i,c} = \sqrt{\frac{1}{L_i}\sum_{t=1}^{L_i}(x_{i,c,t}-\mu_{i,c})^2},$$

where $c \in \{1,\dots,C\}$ and $t \in \{1,\dots,L_i\}$. Each valid signal value is then standardized as

$$\tilde{x}_{i,c,t} = \frac{x_{i,c,t}-\mu_{i,c}}{\sigma_{i,c}+\epsilon},$$

where $\epsilon = 10^{-5}$ is used for numerical stability.

To ensure a uniform model input length, all ECG recordings are processed to $5{,}000$ samples. Recordings exceeding this length are truncated to the first $5{,}000$ samples. For shorter recordings, the per-lead mean and standard deviation are computed using only the original valid samples, which are then standardized before zero-padding is applied at the end to reach $5{,}000$ samples. Therefore, padded samples are excluded from the computation of the standardization statistics.

### 1.2 DETAILS OF TEXT-SPACE RECTIFICATION

This subsection provides the detailed computation of the text-space rectification procedure used in the main text. The mean vector and principal directions used for text-space rectification are estimated exclusively from reports in the training split of the pretraining dataset. Let $\{t_i\}_{i=1}^{N}$ denote the initial global text representations used to estimate the common components, where $t_i \in \mathbb{R}^{d_h}$ and $N$ is the number of text representations used in the estimation. We first compute the mean vector:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} t_i. \tag{1}$$

Each text representation is then centered as:

$$t_i^c = t_i - \mu. \tag{2}$$

The covariance matrix of the centered text representations is computed as:

$$\Sigma_t = \frac{1}{N}\sum_{i=1}^{N} t_i^c (t_i^c)^\top = \frac{1}{N}(T^c)^\top T^c, \tag{3}$$

where $T^c = [t_1^c,\dots,t_N^c]^\top \in \mathbb{R}^{N\times d_h}$ denotes the matrix of centered text representations, and $\Sigma_t \in \mathbb{R}^{d_h\times d_h}$.

We then perform eigendecomposition on the covariance matrix:

$$\Sigma_t = U\Lambda U^\top, \tag{4}$$

Table 1: Hyperparameter settings for downstream linear probing.

| Hyperparameter | Value |
|---|---|
| Learning rate | 0.001 |
| Batch size | 16 |
| Epochs | 100 |
| Optimizer | AdamW |
| Learning rate scheduler | Cosine annealing |

where the columns of $U$ are the principal directions and $\Lambda$ contains the corresponding eigenvalues. The eigenvalues are sorted in descending order, and the top-$k$ principal directions are selected as:

$$U_k = [u_1, \ldots, u_k] \in \mathbb{R}^{d_h \times k}. \tag{5}$$

Here, $k$ denotes the number of dominant common components to remove, and we set $k = 4$ in our experiments. The projection of each centered text representation onto these dominant directions is removed:

$$\hat{t}_i = t_i^c - U_k U_k^\top t_i^c = (I_{d_h} - U_k U_k^\top) t_i^c, \tag{6}$$

where $I_{d_h}$ is the identity matrix. The resulting $\hat{t}_i$ is used as the rectified text representation for subsequent cross-modal alignment.

## 2 Additional Experimental Details

### 2.1 Implementation and Training Details

The pretraining configuration uses a weight decay of $1 \times 10^{-2}$ and a cosine learning rate schedule. Training is conducted on a single NVIDIA A40 GPU with 48 GB memory using bfloat16 (bf16) precision and a batch size of 256. The target number of out-of-batch negatives is set to 4,096 in both alignment directions, and the capacity of the online ECG feature queue is also set to 4,096.

For downstream evaluation, the hyperparameters used for linear probing are summarized in Table 1.

### 2.2 Downstream Dataset Descriptions and Statistics

We evaluate the proposed model on three widely used public downstream datasets: PTB-XL (Wagner et al., 2020), Chapman-Shaoxing-Ningbo (CSN) (Zheng et al., 2020), and CPSC2018 (Liu et al., 2018).

**PTB-XL.** PTB-XL contains 21,837 12-lead ECG records from 18,885 patients. Each record is sampled at 500 Hz with a duration of 10 s and includes annotations covering diagnostic, form, and rhythm categories. Following prior work (Liu et al., 2024; Wang et al., 2025), we evaluate the model on four multilabel classification tasks: Superclass (5 classes), Subclass (23 classes), Form (19 classes), and Rhythm (12 classes).

**CSN.** Following prior work (Liu et al., 2024; Wang et al., 2025), we use the processed CSN dataset containing 23,026 standard 12-lead ECG records with 38 diagnostic labels. Each record is sampled at 500 Hz with a duration of 10 s. After the data split, one training record could not be decoded and was therefore removed, resulting in 23,025 records used in our experiments.

**CPSC2018.** The CPSC2018 dataset contains 6,877 12-lead ECG records sampled at 500 Hz, with recording lengths ranging from 6 to 60 s, and includes 9 diagnostic labels. These variable-length recordings are processed according to the unified input preprocessing procedure described in Section 1.1.

Detailed task-specific train/validation/test split statistics used in our experiments are provided in Table 2.

Table 2: Statistics of downstream datasets used in linear probing evaluation.

| Dataset | Classes | Train | Validation | Test |
|---|---|---|---|---|
| PTBXL-Super | 5 | 17,084 | 2,146 | 2,158 |
| PTBXL-Sub | 23 | 17,084 | 2,146 | 2,158 |
| PTBXL-Rhythm | 12 | 16,832 | 2,100 | 2,098 |
| PTBXL-Form | 19 | 7,197 | 901 | 880 |
| CPSC2018 | 9 | 4,950 | 551 | 1,376 |
| CSN | 38 | 16,545 | 1,860 | 4,620 |

Table 3: Pairwise category correspondence between PTBXL-Super and CPSC2018 for cross-domain transfer. "None" indicates that no corresponding category is available in the other dataset.

| PTBXL-Super | CPSC2018 |
|---|---|
| HYP | None |
| NORM | NORM |
| CD | 1AVB, CRBBB, CLBBB |
| MI | None |
| STTC | STE, STD |

### 2.3 Linear Probing Evaluation Protocol

We adopt the same downstream randomization settings as the publicly released implementation of MERL (Liu et al., 2024): the PyTorch random seed is set to 42, while the Python `random` and NumPy random seeds are set to 0. For each linear-probing experiment, we select the checkpoint with the highest macro AUROC on the validation set. In the 1% and 10% low-resource settings, the corresponding subsets of the downstream training data are sampled using `random_state=42`.

The linear probing baseline results are taken from publicly reported results in prior ECG representation learning studies that use the same downstream datasets, data splits, and linear probing protocol. These benchmark results have been consistently adopted and reported in subsequent studies (Liu et al., 2024; Jin et al., 2025; Pham Hung et al., 2025). Results for all downstream tasks are reported as macro AUROC percentages to enable direct comparison with existing studies.

### 2.4 Cross-Domain Evaluation Protocol

The linear-probing hyperparameters summarized in Table 1 are also used for cross-domain evaluation. For MERL and MELP, we use their officially released pretrained ECG encoders and evaluate them under the same cross-domain linear-probing protocol as ours. Results for the other baselines are taken from previously published studies using full source-domain training data for linear probing (Liu et al., 2024; Wang et al., 2025). We follow the label-alignment protocol of MERL (Liu et al., 2024). Tables 3–5 provide the pairwise category correspondences used in the cross-domain transfer experiments and collectively cover the six directional transfer settings reported in the main paper. Because different downstream datasets use non-identical label sets, a separate aligned label space is constructed for each pairwise comparison. Specifically, labels from both datasets are mapped into this aligned label space according to these correspondences. Finer-grained categories are merged into their corresponding coarser categories, whereas unmatched categories, together with their associated samples, are removed before evaluation.

### 2.5 Zero-Shot Evaluation Protocol

For MELP and MERL, we freeze their officially released pretrained ECG encoders. The newly initialized linear projection heads are trained for one epoch on the upstream MIMIC-IV-ECG dataset (Gow et al., 2023). During downstream zero-shot evaluation, no parameters are updated. For each downstream diagnostic class, we follow the Clinical Knowledge Enhanced Prompt Engineering (CKEPE) protocol used in MERL (Liu et al., 2024) to construct class-level text prompts.

Table 4: Pairwise category correspondence between PTBXL-Super and CSN for cross-domain transfer.

| PTBXL-Super | CSN |
|---|---|
| HYP | RVH, LVH |
| NORM | SR |
| CD | 2AVB, 2AVB1, 1AVB, AVB, LBBB, RBBB, STDD |
| MI | MI |
| STTC | STTC, STE, TWO, STTU, QTIE, TWC |

Table 5: Pairwise category correspondence between CPSC2018 and CSN for cross-domain transfer.

| CPSC2018 | CSN |
|---|---|
| AFIB | AFIB |
| VPC | VPB |
| NORM | SR |
| 1AVB | 1AVB |
| CRBBB | RBBB |
| STE | STE |
| PAC | APB |
| CLBBB | LBBB |
| STD | STE, STTC, STTU, STDD |

# 3 Additional Results

## 3.1 LLM Backbone Comparison

To examine whether the proposed framework is sensitive to the choice of LLM backbone, we replace the default Meta Llama 3.1 8B with Qwen2-7B (Yang et al., 2024), while keeping all other pretraining and downstream evaluation settings unchanged. Both are representative general-purpose open-weight LLMs introduced during the same period in 2024. As shown in Table 6, Meta Llama 3.1 8B and Qwen2-7B achieve average AUROCs of 88.49% and 88.45%, respectively, across the 18 linear-probing settings, corresponding to a difference of only 0.04 percentage points. These results indicate that the proposed framework achieves comparable performance with different representative LLM backbones. Given their nearly identical overall performance, we use Meta Llama 3.1 8B as the default text encoder in the remaining experiments.

## 3.2 Controlled Comparison with Biomedical Text Encoders

To directly examine whether the quality of frozen text representations affects ECG representation learning, we replace the default Meta Llama 3.1 8B text encoder with three representative biomedical text encoders: MEDCPT (Jin et al., 2023), BioClinicalBERT (Alsentzer et al., 2019), and PubMedBERT (Gu et al., 2022). For this controlled comparison, only the frozen text encoder is changed, while the ECG encoder, FiLM module, text-space rectification procedure, negative-pool construction and capacity, optimization hyperparameters, pretraining schedule, and downstream linear-probing protocol remain unchanged.

As summarized in Table 8, replacing the frozen biomedical text encoders with Meta Llama 3.1 8B yields overall gains of 3.28–3.71 percentage points. The improvements are most pronounced under the 1% labeled-data setting, ranging from 4.87 to 5.49 percentage points, and gradually decrease as more downstream labels become available. Since all other components and training settings are held fixed, these results provide controlled evidence that stronger frozen text representations can offer more effective semantic supervision for ECG representation learning, particularly in low-resource scenarios.

Table 6: Comparison of different LLM backbones under the linear-probing evaluation. Results are reported as AUROC (%). Average denotes the mean AUROC across all 18 evaluation settings. Bold values indicate the better result for each setting; ties are both highlighted.

| | PTBXL-Sub | | | PTBXL-Super | | | PTBXL-Form | | | PTBXL-Rhythm | | | CSN | | | CPSC2018 | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM Backbone | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | |
| Llama 3.1-8B | **81.49** | **85.39** | **89.33** | **88.65** | **91.28** | **92.62** | 67.93 | 80.72 | 88.77 | **89.79** | 94.68 | 97.10 | 83.08 | **90.76** | 95.31 | 83.90 | 95.38 | 96.64 | **88.49** |
| Qwen2-7B | 80.50 | 84.15 | 88.99 | 88.03 | **91.28** | 92.60 | **68.11** | **81.61** | **88.80** | 88.65 | **94.98** | **97.25** | **83.41** | 90.04 | **95.68** | **85.75** | **95.40** | **96.88** | 88.45 |

Table 7: Controlled comparison with frozen biomedical text encoders under linear-probing evaluation. Only the frozen text encoder is changed, while all other model components, pretraining settings, and downstream evaluation protocols remain unchanged. Results are reported as AUROC (%). Average denotes the mean AUROC across all 18 evaluation settings. Bold values indicate the best result for each setting.

| | PTBXL-Sub | | | PTBXL-Super | | | PTBXL-Form | | | PTBXL-Rhythm | | | CSN | | | CPSC2018 | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Text Encoder | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | |
| Llama 3.1-8B | **81.49** | **85.39** | **89.33** | **88.65** | **91.28** | **92.62** | **67.93** | **80.72** | **88.77** | **89.79** | **94.68** | **97.10** | **83.08** | **90.76** | **95.31** | **83.90** | **95.38** | **96.64** | **88.49** |
| MEDCPT | 73.37 | 82.18 | 87.30 | 87.09 | 89.36 | 90.88 | 63.70 | 76.46 | 87.58 | 83.41 | 90.67 | 94.31 | 75.10 | 86.98 | 93.55 | 81.69 | 94.41 | 95.77 | 85.21 |
| BioClinicalBERT | 77.84 | 81.36 | 86.22 | 86.70 | 89.18 | 90.36 | 59.77 | 74.28 | 84.73 | 85.96 | 91.68 | 92.97 | 75.42 | 87.01 | 92.90 | 79.91 | 94.41 | 95.48 | 84.79 |
| PubMedBERT | 76.23 | 81.68 | 85.72 | 87.48 | 89.16 | 90.15 | 59.18 | 74.94 | 85.74 | 85.56 | 91.52 | 95.67 | 74.79 | 86.92 | 93.32 | 78.68 | 94.09 | 95.18 | 84.78 |

### 3.3 Statistical Analysis

Due to computational constraints, we conduct repeated-seed statistical analysis only for linear probing and cross-domain transfer evaluation, two key downstream evaluation protocols in ECG representation learning. Specifically, we repeat pretraining with three different random seeds and evaluate the resulting ECG encoders using the same downstream seed. This keeps the downstream evaluation fixed and allows us to examine how randomness in the pretraining framework affects the learned representations and downstream performance. Consequently, the reported variability mainly reflects randomness from the pretraining stage, including parameter initialization, mini-batch ordering, mini-batch-based optimization dynamics, and negative-sample construction. Since repeated runs of baseline methods are not available, we report statistical variability for our method only and do not perform significance tests against baselines.

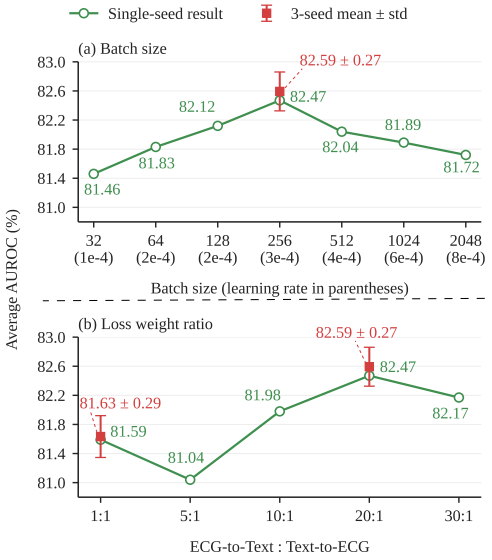



Figure 1: Hyperparameter analysis. Circles denote single-seed results, while squares with error bars denote the mean and standard deviation over three seeds. (a) Batch size, with the learning rate shown in parentheses. (b) ECG-to-Text:Text-to-ECG loss-weight ratio.

### 3.4 Additional Ablation Results

**Hyperparameter Analysis.** The batch size determines the number of in-batch negative samples and therefore affects the effectiveness of contrastive learning. We compare seven batch-size and learning-rate configurations, with batch sizes ranging from 32 to 2,048 and the learning rate adjusted accordingly. The complete results are provided in Figure 1(a) and Table 11. This analysis is consistent with prior studies showing that a larger batch size does not necessarily lead to stable performance gains, as it may introduce additional optimization and generalization issues (Goyal et al., 2017; Keskar et al., 2017). We further analyze the effect of the bidirectional contrastive loss weighting ratio, with the complete results provided in Figure 1(b) and Table 12.

Table 8: Average AUROC gains of Meta Llama 3.1 8B over each frozen biomedical text encoder, measured in percentage points. The 1%, 10%, and 100% columns are averaged across the six downstream tasks at the corresponding labeled-data fraction. Overall denotes the average gain across all 18 linear-probing settings.

| Compared Encoder | 1% | 10% | 100% | Overall |
|---|---|---|---|---|
| MEDCPT | +5.08 | +3.03 | +1.73 | +3.28 |
| BioClinicalBERT | +4.87 | +3.38 | +2.85 | +3.70 |
| PubMedBERT | +5.49 | +3.32 | +2.33 | +3.71 |

Table 9: Statistical analysis of linear probing results over three pretraining seeds. Results are reported as AUROC (%), with mean ± standard deviation.

| Dataset | Data | Run 1 | Run 2 | Run 3 | Mean ± Std. |
|---|---|---|---|---|---|
| PTBXL-Sub | 1% | **81.49** | **81.49** | 81.28 | 81.42 ± 0.12 |
| | 10% | **85.39** | 85.14 | 84.32 | 84.95 ± 0.56 |
| | 100% | 89.33 | **89.35** | 89.21 | 89.30 ± 0.08 |
| PTBXL-Super | 1% | **88.65** | 88.23 | 88.00 | 88.29 ± 0.33 |
| | 10% | **91.28** | 91.21 | 91.05 | 91.18 ± 0.12 |
| | 100% | **92.62** | 92.52 | 92.31 | 92.48 ± 0.16 |
| PTBXL-Form | 1% | 67.93 | 67.80 | **68.60** | 68.11 ± 0.43 |
| | 10% | **80.72** | 80.56 | 80.68 | 80.65 ± 0.08 |
| | 100% | 88.77 | 88.32 | **88.95** | 88.68 ± 0.32 |
| PTBXL-Rhythm | 1% | 89.79 | 89.65 | **90.07** | 89.84 ± 0.21 |
| | 10% | 94.68 | **95.51** | 95.44 | 95.21 ± 0.46 |
| | 100% | 97.10 | 97.10 | **97.24** | 97.15 ± 0.08 |
| CSN | 1% | 83.08 | 83.56 | **84.15** | 83.60 ± 0.54 |
| | 10% | **90.76** | 89.79 | 89.89 | 90.15 ± 0.53 |
| | 100% | 95.31 | 95.31 | **95.47** | 95.36 ± 0.09 |
| CPSC2018 | 1% | 83.90 | 83.73 | **85.29** | 84.31 ± 0.86 |
| | 10% | 95.38 | 95.37 | **95.45** | 95.40 ± 0.04 |
| | 100% | 96.64 | 96.56 | **96.82** | 96.67 ± 0.13 |

**Multi-Seed Analysis of Key Ablation Settings.** To assess whether the main ablation conclusions are robust to training randomness, we repeat three representative settings using three pretraining seeds: the selected configuration with 4,096 out-of-batch negatives, the removal of FiLM, and the removal of TSR. In Figure 2, circles denote the single-seed sweep over the number of out-of-batch negatives, crosses denote the single-seed component ablations (w/o FiLM and w/o TSR), and squares with error bars denote the mean and standard deviation over three pretraining seeds for the selected 4,096-negative configuration and the two component ablations. Their average AUROCs are 81.63 ± 0.29, 81.17 ± 0.35, and 74.77 ± 0.46, respectively. The multi-seed results preserve the qualitative ordering observed in the single-seed ablations: FiLM provides a modest improvement, whereas TSR is critical to effective ECG–text alignment. This consistency indicates that the main ablation conclusions are not driven by a particular random seed.

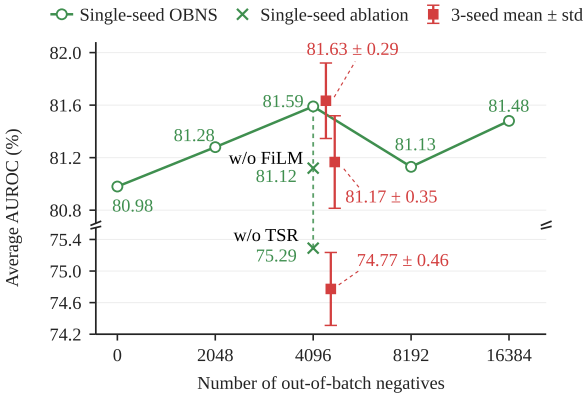



Figure 2: Multi-seed analysis of selected key ablation settings. Results are reported as average AUROC (%) under the 1% linear-probing setting.

Table 10: Statistical analysis of cross-domain transfer results over three pretraining seeds. Results are reported as AUROC (%), with mean ± standard deviation.

| Source domain | PTBXL-Super | | CPSC2018 | | CSN | |
|---|---|---|---|---|---|---|
| Target domain | CPSC2018 | CSN | PTBXL-Super | CSN | PTBXL-Super | CPSC2018 |
| Run 1 | 88.33 | **91.29** | 82.22 | 91.76 | **80.13** | 90.12 |
| Run 2 | **88.82** | 91.11 | 81.78 | **91.98** | 79.81 | **90.83** |
| Run 3 | 88.26 | 91.25 | **82.88** | 91.63 | 79.87 | 90.14 |
| Mean ± Std. | $88.47 \pm 0.31$ | $91.22 \pm 0.09$ | $82.29 \pm 0.55$ | $91.79 \pm 0.18$ | $79.94 \pm 0.17$ | $90.36 \pm 0.40$ |

Table 11: Analysis of batch size in contrastive pretraining. Results are reported as average AUROC under the 1% linear-probing setting.

| Batch size | Learning rate | Linear Probing (1%) |
|---|---|---|
| 32 | 1e-4 | 81.46 |
| 64 | 2e-4 | 81.83 |
| 128 | 2e-4 | 82.12 |
| 256 | 3e-4 | **82.47** |
| 512 | 4e-4 | 82.04 |
| 1024 | 6e-4 | 81.89 |
| 2048 | 8e-4 | 81.72 |

Table 12: Analysis of bidirectional contrastive loss weighting ratio. The ratio denotes the ECG-to-Text:Text-to-ECG loss weights, and all results are based on a batch size of 256. Results are reported as average AUROC under the 1% linear-probing setting.

| Ratio | Linear Probing (1%) |
|---|---|
| 30:1 | 82.17 |
| 20:1 | **82.47** |
| 10:1 | 81.98 |
| 5:1 | 81.04 |
| 1:1 | 81.59 |

## 4 Data, Licenses, and Asset Usage

**Public Datasets.** All datasets used in this paper are publicly available research datasets. The pretraining dataset is MIMIC-IV-ECG (Gow et al., 2023), which is available from PhysioNet under the Open Data Commons Open Database License v1.0. The downstream datasets include PTB-XL (Wagner et al., 2020), CPSC2018 (Liu et al., 2018), and the Chapman-Shaoxing-Ningbo (CSN) database (Zheng et al., 2020). PTB-XL and CSN are available from PhysioNet under the Creative Commons Attribution 4.0 International Public License. CPSC2018 is available from the official China Physiological Signal Challenge 2018 website. We follow the access terms and citation requirements of each dataset provider.

- **MIMIC-IV-ECG**: `https://physionet.org/content/mimic-iv-ecg/1.0/`
- **PTB-XL**: `https://physionet.org/content/ptb-xl/1.0.0/`
- **CPSC2018**: `http://2018.icbeb.org/Challenge.html`
- **CSN**: `https://physionet.org/content/ecg-arrhythmia/1.0.0/`

**Pretrained Models and Baseline Assets.** We use Meta Llama 3.1 8B (Grattafiori et al., 2024) as the frozen LLM text encoder and follow the license and Acceptable Use Policy of the official Meta release. The ECG branch uses a WavLM-style Transformer backbone (Chen et al., 2022). Following the backbone configuration adopted by MELP, we employ an 8-layer Transformer encoder with a hidden dimension of 768. The ECG backbone is trained from scratch without using the pretrained WavLM checkpoint. For standardized zero-shot evaluation, we use the official MEDCPT model (Jin et al., 2023) as the unified text encoder and follow the usage terms of its official release. For baseline

comparison, we use officially released checkpoints or evaluation resources from MERL (Liu et al., 2024) and MELP (Wang et al., 2025) when applicable, following the terms specified by the original authors.

## REFERENCES


Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909/.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://doi.org/10.1109/JSTSP.2022.3188113.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour, 2017. URL https://arxiv.org/abs/1706.02677.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Hz1Ko087B.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651. URL https://doi.org/10.1093/bioinformatics/btad651.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31949–31963. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bg.html.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Yin Kwee Ng. 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. doi: 10.1166/jmihi.2018.2442. URL https://doi.org/10.1166/jmihi.2018.2442.

Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting masked ECG-text auto-encoders as discriminative learners. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49277–49291. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pham-hung25a.html.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Florian I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL: A large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6. URL https://doi.org/10.1038/s41597-020-0495-6.

Fuying Wang, Jiacheng Xu, and Lequan Yu. From token to rhythm: A multi-scale approach for ECG-language pretraining. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65059–65074. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wang25du.html.

An Yang et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1):48, 2020. doi: 10.1038/s41597-020-0386-x. URL https://doi.org/10.1038/s41597-020-0386-x.