C-MELT: CONTRASTIVE ENHANCED MASKED AUTO-ENCODERS FOR ECG-LANGUAGE PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate interpretation of Electrocardiogram (ECG) signals is pivotal for diagnosing cardiovascular diseases. Integrating ECG signals with their accompanying textual reports holds immense potential to enhance clinical diagnostics through the combination of physiological data and qualitative insights. However, this integration faces significant challenges due to inherent modality disparities and the scarcity of labeled data for robust cross-modal learning. To address these obstacles, we propose C-MELT, a novel framework that pre-trains ECG and text data using a contrastive masked auto-encoder architecture. C-MELT uniquely combines the strengths of generative with enhanced discriminative capabilities to achieve robust cross-modal representations. This is accomplished through masked modality modeling, specialized loss functions, and an improved negative sampling strategy tailored for cross-modal alignment. Extensive experiments on five public datasets across diverse downstream tasks demonstrate that C-MELT significantly outperforms existing methods, achieving an average AUC improvement of 15% in linear probing with only one percent of training data and 2% in zero-shot performance without requiring training data over state-of-the-art models. These results highlight the effectiveness of C-MELT, underscoring its potential to advance automated clinical diagnostics through multi-modal representations.

1 INTRODUCTION

Electrocardiograms (ECGs), obtained through non-invasive electrode placement, provide a critical window into the heart's electrical activity by measuring voltage differences across specific anatomical regions. The standard 12-lead ECG, which captures unique electrical potential differences from each lead, plays a vital role in diagnosing a wide spectrum of cardiac conditions, like arrhythmias. In recent years, significant progress has been made in leveraging deep learning techniques for automated ECG interpretation (Yan et al., 2019; Ebrahimi et al., 2020; Siontis et al., 2021). However, these supervised deep learning approaches often necessitate large volumes of expertly annotated data, which are frequently scarce and expensive to acquire. Self-supervised learning (SSL) has emerged as a compelling alternative, offering the potential to learn robust representations from abundant unlabeled ECG data. These learned representations can be effectively utilized for zero-shot learning on novel tasks and adapted via fine-tuning to specific downstream applications, thereby mitigating the reliance on extensive labeled datasets.

Numerous studies have explored the potential of SSL in the ECG domain, demonstrating its efficacy in learning representations from vast quantities of unlabeled data. These efforts generally fall into two main tracks: contrastive and generative approaches. Contrastive methods, exemplified by works such as (Chen et al., 2020; 2021; Chen & He, 2021; Grill et al., 2020; Kiyasseh et al., 2021; Oh et al., 2022; McKeen et al., 2024), aim to learn discriminative representations by maximizing the similarity between positive pairs (e.g., different augmentations of the same ECG signal) and minimizing the similarity between negative pairs (e.g., ECGs from different patients) within the embedding space. Conversely, generative approaches (Hu et al., 2023; Zhang et al., 2022a; 2023) focus on reconstructing the input data, typically by predicting masked or missing segments of the ECG signal, thereby learning to capture the underlying data distribution. Therefore, integrating both contrastive and generative approaches within a unified framework could leverage their complementary strengths, leading to a more powerful method for learning robust representations (Kim et al., 2021; Li et al., 2022b; Song et al., 2024).

Despite advancements, existing ECG-based SSL approaches have largely overlooked the valuable information embedded within clinical text reports, which offer key insights into underlying cardiac conditions and have the potential to significantly enhance a model's diagnostic accuracy (Zhang et al., 2022c; Chen et al., 2022). This oversight highlights a critical gap in the field: the lack of emphasis on jointly learning ECG-text cross-modal representations. While some recent efforts (Liu et al., 2024; Lalam et al., 2023; Li et al., 2024) have attempted to bridge this gap by integrating ECG signals and clinical reports through cross-modal contrastive learning, the potential of learning unified representations that capture the intricate interplay between ECG signals and their corresponding textual descriptions shown in generative approaches remains largely unexplored. Moreover, the prevailing reliance on these contrastive methods presents inherent limitations. They depend on the availability of negative samples and often struggle to capture cross-modal relationships effectively due to difficulties in defining appropriate negative pairings across different modalities.

In this work, we depart from the reliance on either solely contrastive learning or stand-alone generative approaches for cross-modal representation learning. We introduce C-MELT, a novel hybrid framework that synergistically integrates both learning paradigms to effectively capture fine-grained input details and discriminative ECG-text features. Our approach employs a transformer-based encoder specifically for ECG signals and a well-pre-trained language model for clinical text encoder in a masked auto-encoder architecture, together with tailored loss functions that promote the joint learning of robust cross-modal representations. Additionally, we introduce a nearest-neighbor negative sampling strategy, a crucial refinement often overlooked in previous methods, to ensure that negative samples are contextually selected and thereby, enhance the discriminative capability of the learned representations. To rigorously evaluate the efficacy of C-MELT, we conduct extensive experiments on various public ECG datasets and demonstrate that our method significantly outperforms recent state-of-the-art baselines across all evaluation settings and datasets.

2 Related Work

ECG Self-supervised Learning. Self-supervised learning (SSL) has been shown to work effectively across various modalities, including vision (Li et al., 2022a; Han et al., 2021), language (Devlin, 2018; He et al., 2020; Chung et al., 2024), and time-series data (Tonekaboni et al., 2021; Zhang et al., 2022b; Saeed et al., 2019). Particularly, recent advances in applying SSL to ECG signals have demonstrated that models can learn meaningful representations from large amounts of unlabeled data, which is crucial in medical domains where labeled datasets are often limited and expensive to acquire. Here, we mainly discuss two common SSL approaches: generative and contrastive, which have seen notable progress in ECG representation learning in recent years.

Early contrastive methods such as SimCLR (Chen et al., 2020), MoCo (Chen et al., 2021), SimSiam (Chen & He, 2021), and BYOL (Grill et al., 2020) introduced the concept of maximizing agreement between augmented views of the same data sample by employing augmentation strategies to create challenging positive and negative pairs. In the context of ECG signals, recent approaches like 3KG (Gopal et al., 2021) apply physiologically inspired spatial and temporal augmentations, using vectorcardiogram (VCG) transformations to capture the three-dimensional spatiotemporal characteristics of the heart's electrical activity. Similarly, CLOCS (Kiyasseh et al., 2021) developed Contrastive Multi-Segment Coding (CMSC), which enhances the model's ability to handle varying ECG signal characteristics across different axes—space, time, and patients. Building on this, (Oh et al., 2022) incorporates Wav2vec 2.0 (Baevski et al., 2020), CMSC, and random lead masking to simulate different global and local lead configurations during training, thereby improving model robustness and achieving impressive results on ECG downstream tasks.

On the other hand, generative approaches (Hu et al., 2023; Zhang et al., 2022a; Na et al., 2024) are less prevalent, but play a crucial role in ECG SSL. These methods focus on capturing the underlying structure of the data by training auto-encoder models to generate or reconstruct masked input data, enabling the model to understand and represent key features and patterns. For instance, ST-MEM (Na et al., 2024) utilizes a masked auto-encoder with a spatio-temporal patchifying technique to model relationships in 12-lead ECG signals. Additionally, the Cross-Reconstruction Transformer (CRT) (Zhang et al., 2023) employs frequency-domain and temporal masking to reconstruct missing ECG segments, demonstrating the innovative use of generative SSL in ECG analysis.

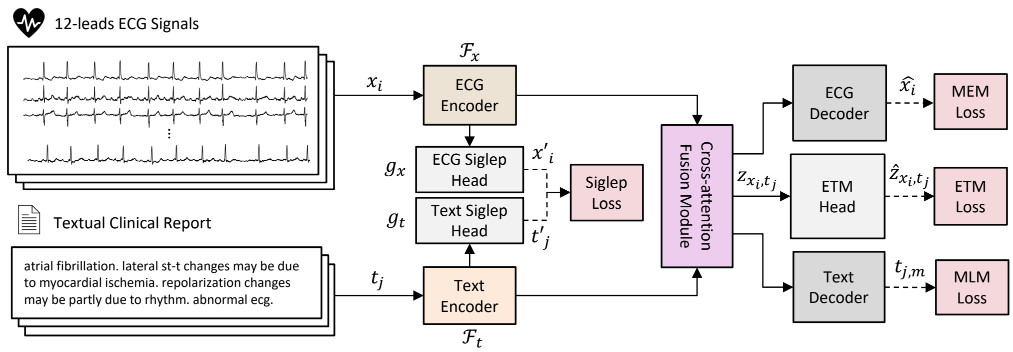

Figure 1: Illustration of our contrastive masked ECG-language modeling technique.

ECG-Text Multi-modal Representation Learning. Multi-modal representation learning combines information from different data types, shown to effectively improve model performance (Lin et al., 2024; Du et al., 2023). Particularly, pioneering works like CLIP-based models (Radford et al., 2021; Rasheed et al., 2023; Zhai et al., 2023) have proven the power of contrastive learning in aligning visual and textual modalities, achieving strong generalizations across a broad range of tasks. Applying similar ideas to the ECG domain, MERL (Liu et al., 2024) leverages cross-modal and uni-modal alignment techniques to generalize ECG and text-based medical classification tasks. However, it overlooks the critical role of negative sample selection for contrastive learning and lacks exploring generative approaches for fine-grained multi-modal learning, limiting performance in end tasks.

3 Method

We propose C-MELT, a framework designed to learn generalizable cross-modal representations by aligning electrocardiogram (ECG) signals and corresponding medical text reports. C-MELT leverages masked language modeling (MLM) and masked ECG modeling (MEM) to reconstruct randomly masked segments within the input text and ECG signals, respectively. This encourages the model to learn fine-grained features within each modality. Furthermore, we introduce Siglep (Sigmoid language ECG pre-training) loss, which is based on SigLIP Zhai et al. (2023), and a nearest-neighbor negative sampling strategy. These directly promote discriminative representation learning and enhance cross-modal alignment, besides the ECG-text matching (ETM) learning task.

Figure 1 depicts the overall architecture of C-MELT, which comprises two main branches: an ECG encoder and a text encoder. The ECG encoder utilizes a transformer-based architecture (Vaswani et al., 2023) to process the input ECG signals and generate corresponding representations, denoted as $\mathbf{H}_x \in \mathbb{R}^{L_x \times d}$, where L_x represents the sequence length of the ECG signal and d represents the embedding dimension. The text encoder utilizes the recent pre-trained Flan-T5 model (Chung et al., 2024) which, to our knowledge, has not been previously applied to this task, to extract high-level semantic embeddings from the clinical text, denoted as $\mathbf{H}_t \in \mathbb{R}^{L_t \times d}$, where L_t represents the sequence length of the text. These encoder outputs are then passed through a fusion module, which employs a cross-attention mechanism to integrate information from both modalities, generating fused representations denoted as $\mathbf{H}_f \in \mathbf{R}^{(L_x+L_t)\times d}$. The model subsequently employs three distinct heads: two decoders, responsible for reconstructing the masked portions of the ECG signal ($\hat{X}$) and text (T_m), respectively, and a contrastive prediction head for ECG-text matching. Additionally, we introduce two projection heads, g_x and g_t, following the ECG and text encoders, respectively. These projection heads, along with the Siglep loss, facilitate learning discriminative representation between these modalities. The model is trained by jointly optimizing four loss functions: masked language modeling loss ($\mathcal{L}_{MLM}$), masked ECG modeling loss ($\mathcal{L}_{MEM}$), ECG-text matching loss ($\mathcal{L}_{ETM}$), and the Siglep loss ($\mathcal{L}_{Siglep}$). The subsequent subsections provide a detailed description of each component within the C-MELT framework.

3.1 Multi-Modal masked auto-encoders.

ECG Encoder. We implement the ECG encoder (denoted as $\mathcal{F}_x$) based on a transformer architecture, which was originally developed for efficiently processing sequential data in parallel (Vaswani et al., 2023). We first follow (Oh et al., 2022) to apply a masking strategy to the ECG input $\mathbf{X} \in \mathbb{R}^{L \times C}$ to encourage robust feature learning, where L is the length of the signal and C is the number of channels. Specifically, we leverage random lead masking as an on-the-fly augmentation where each lead randomly masked with a probability of $p = 0.5$ during pre-training. Furthermore, we use a dropout layer on the input with $p = 0.1$ to enable masking modeling. We then pass the masked input into a series of convolutional layers, each followed by GELU activation functions and group normalization. The extracted features are subsequently projected into a 768-dimensional space. Following that, we add a convolutional positional encoding layer to preserve the temporal order of the ECG sequence. Next, we employ eight transformer encoder layers, each including a multi-head self-attention mechanism that allows the model to attend to different parts of the input sequence simultaneously. We conduct an experiment exploring the effects of different numbers of transformer layers in Section 4.

Language (Text) Encoder. For our text encoder, we utilize the Flan-T5-base encoder (denoted as $\mathcal{F}_t$), which outputs 768-dimensional embeddings. The input to the encoder consists of token indices generated by the Flan-T5 tokenizer, represented as $\mathbf{T} \in \mathbb{Z}^M$, where M is the maximum sequence length. Flan-T5 is an advanced version of the T5 model (Raffel et al., 2023), which has been pre-trained on a massive and diverse text dataset covering numerous tasks, such as summarization and question answering. Note that our text encoder is fine-tuned during the pre-training stage. We also conduct an ablation with various text encoders in Section 4 to support our choice of Flan-T5.

Fusion Module. The fusion module begins with linear projections that map the outputs of the ECG and language encoders to a 768-dimensional space. We apply modality-specific embeddings to the projected features to distinguish between ECG and text data. Importantly, we employ cross-attention to integrate the ECG and textual information, allowing each modality to inform the other by learning the relevant features. This cross-attention mechanism is crucial as it enables the model to leverage the complementary strengths of both ECG and text data more effectively.

Decoders and Loss Functions. After the fusion module, three distinct network heads are introduced, each associated with a specific loss function: masked language modeling (MLM), masked ECG modeling (MEM), and ECG-text matching (ETM). MLM and MEM are designed for reconstruction tasks, while ETM adopts a contrastive learning approach to align the different modalities. We detail each head and its corresponding loss function below:

Masked Language Modeling (MLM). The MLM head consists of a dense layer that outputs a probability distribution over the vocabulary. The MLM head focuses on predicting the masked tokens in the input text sequence, encouraging the model to learn contextualized word embeddings through a reconstruction task. We use the cross-entropy (CE) loss for MLM, as shown in Equation 1:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} \sum_{m \in \mathcal{M}_j} \log P(t_{j,m} | \mathbf{t}_{j \setminus \mathcal{M}_j}; \theta), \tag{1}$$

where $\mathcal{B}$ represents the batch size, $\mathcal{M}_j$ is the set of masked positions in the j^{th} sequence, $t_{j,m}$ is the masked token at position m in the j^{th} sequence, $\mathbf{t}_{j \setminus \mathcal{M}_j}$ represents the j^{th} input sequence with masked tokens removed, and θ represents the model parameters.

Masked ECG Modeling (MEM). Similar to MLM, the MEM head aims to reconstruct the masked ECG inputs. It consists of a linear embedding layer that maps the input sequence to a lower-dimensional space (384), followed by learnable mask tokens that represent the missing portions of the sequence. We apply positional encodings to preserve the temporal structure of the ECG data. Subsequently, we use a multi-layer transformer decoder to model the dependencies within the sequence. Finally, a linear projection layer outputs the predicted ECG features. We train the MEM head using the mean squared error (MSE) loss between the predicted ECG signal $\hat{\mathbf{x}}_i$ and the ground truth ECG signal $\mathbf{x}_i$, as shown in Equation 2:

$$\mathcal{L}_{\text{MEM}} = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} ||\hat{\mathbf{x}}_i - \mathbf{x}_i||_2^2 \tag{2}$$

ECG-Text Matching (ETM). Finally, we use ETM to promote alignment between ECG signals and their corresponding text reports. This is formulated as a binary classification task, where the ETM head consists of a single dense layer that outputs a scalar $\hat{z}_{\mathbf{x}_k, \mathbf{t}_k}$ representing the predicted probability. The ETM loss is defined as the binary cross-entropy loss:

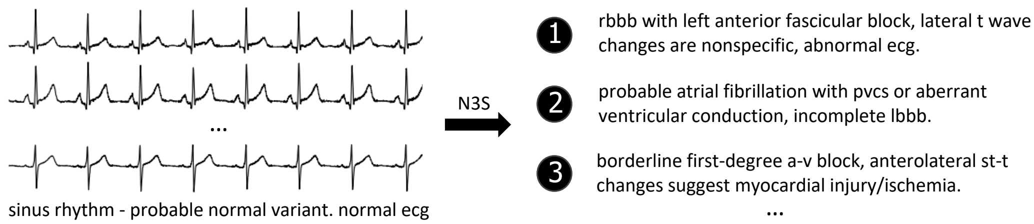

Figure 2: Example of ECG-Text pair (left) and its corresponding negative text samples (right).

$$\mathcal{L}_{\text{ETM}} = -\frac{1}{\mathcal{B}} \sum_{k=1}^{\mathcal{B}} \left[y_k \log \sigma(\hat{z}_{\mathbf{x}_k, \mathbf{t}_k}) + (1 - y_k) \log(1 - \sigma(\hat{z}_{\mathbf{x}_k, \mathbf{t}_k})) \right], \tag{3}$$

where σ is the sigmoid function, $y_k = 1$ if $(\mathbf{x}_k, \mathbf{t}_k)$ is a positive pair, and $y_k = 0$ otherwise.

3.2 Improving Contrastive Learning

Siglep Loss Function. In multi-modal masked auto-encoder architectures such as (Chen et al., 2022), contrastive learning's effectiveness can be limited by the inherent tension between the reconstruction-focused generative tasks of autoencoders and the discriminative nature of contrastive learning. They are more biased for learning to reconstruct masked inputs in generative manners. This can hinder the model's capability to learn discriminative features useful for downstream tasks, such as zero-shot inference or linear probing. Furthermore, although the ETM loss in such architectures can serve as a form of contrastive loss, it may not be sufficient for building a robust ECG encoder. Specifically, the ETM module is primarily designed for binary classification based on fused features rather than directly enhancing the discriminative power of individual encoders. This limitation can restrict the model's ability to produce high-quality multimodal embeddings.

Therefore, we propose strengthening contrastive learning in multi-modal masked auto-encoder architectures using Siglep loss function. Specifically, we adapt the SigLIP implementation Zhai et al. (2023), originally proposed for text-image pairs, to the text-ECG domain (Formula 4). This approach avoids the costly global normalization of softmax-based contrastive losses by operating independently on each ECG-text pair, improving memory efficiency and scalability. We introduce two additional network heads to the ECG and text encoders, respectively. Each head consists of a pooling layer, a Tanh activation function, and a dense layer, enabling them to output 768-dimensional embeddings (denoted as $\mathbf{x'}_i \in \mathbb{R}^{768}$ for the i^{th} ECG sample and $\mathbf{t'}_j \in \mathbb{R}^{768}$ for the j^{th} text report).

$$\mathcal{L}_{\text{Siglep}} = -\frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} \log \left(\frac{1}{1 + e^{-y_{ij} \mathbf{x'}_i^\top \mathbf{t'}_j}} \right), \tag{4}$$

where $y_{ij} = 1$ for positive (matching) ECG-text pairs, and $y_{ij} = -1$ otherwise.

Nearest-neighbor-based Negative Sampling (N3S). In contrastive learning, the selection of negative samples significantly impacts the training process (Xu et al., 2022). Conventional methods often employ random sampling, where negative text reports are chosen randomly to replace positive texts. However, this approach may lead to false negative selection, especially in medical datasets, where randomly chosen reports might share substantial similarities with the positive reports, hindering effective contrastive learning. This is discussed more in the *Appendix* A.2.

Therefore, we propose nearest-neighbor negative sampling (N3S), which selects negative samples based on their dissimilarity in the Flan-T5's feature space, ensuring they are sufficiently distinct from the positive samples while remaining semantically related to the domain. Specifically, we first utilize pre-trained Flan-T5 (small) to generate vector representations, denoted as $\mathbf{v}_t \in \mathbb{R}^{512}$, for each text report t in the training dataset $\mathcal{D}_{train}$. These embeddings capture the semantic meaning of the reports. During training, for a given ECG and its corresponding positive text report (x_k, t_k^+) in half of the training batch $\mathcal{B}$, the negative report t_k^- is selected as one of the top 64 largest cosine distance reports from the positive report's embedding $\mathbf{v}_{t_k^+}$. As the training progresses with batches being updated randomly, this ensures that the negative samples are continually changed, introducing variability while maintaining domain relevance.

To efficiently perform this process, we employ FAISS (Facebook AI Similarity Search) (Douze et al., 2024), a high-performance library designed for indexing and searching large collections of dense vectors. FAISS allows us to apply the N3S technique to large-scale datasets in a computationally tractable manner. Figure 2 shows one example of an ECG-text pair with its potential negative texts in the training dataset.

4 EXPERIMENTS

Table 1: Performance for 5 lead combinations in diagnosis classification (Dx., by CinC scores scaled by 100) and patient identification (Id., by %). P-N-lead indicates N zero-padded unavailable leads.

Methods	Tasks	# Leads				
		12-lead	P-6-lead	P-3-lead	P-2-lead	P-1-lead
W2V (Baevski et al., 2020)	Dx.	71.4	64.3	67.6	61.1	52.5
	Id.	49.2	41.1	47.0	41.4	24.7
CMSC (Kiyasseh et al., 2021)	Dx.	62.5	52.2	57.5	50.7	40.6
	Id.	51.3	39.2	51.0	37.8	22.7
3KG (Gopal et al., 2021)	Dx.	60.0	51.5	56.3	50.5	41.8
	Id.	40.7	32.0	36.7	31.0	19.8
SimCLR(RLM) (Chen et al., 2020)	Dx.	57.8	49.7	53.5	48.4	39.3
	Id.	35.3	28.9	36.8	30.4	19.2
W2V+CMSC (Oh et al., 2022)	Dx.	71.7	61.6	65.6	58.6	48.2
	Id.	55.0	43.7	46.6	41.0	28.0
W2V+CMSC+RLM (Oh et al., 2022)	Dx.	73.2	66.2	71.4	65.6	55.4
	Id.	57.7	45.9	54.8	45.7	31.3
C-MELT	**Dx.**	**85.7**	**81.1**	**84.2**	**81.9**	**76.5**
	Id.	**65.4**	**57.3**	**60.5**	**57.7**	**41.1**

4.1 IMPLEMENTATION DETAILS

4.1.1 PRE-TRAINING TASK.

Pre-train Dataset. In the pre-training stage, we utilize the MIMIC-IV-ECG v1.0 database (Gow et al., 2023), which includes 800,035 paired samples derived from 161,352 unique subjects. This dataset contains numerous 10-second ECG recordings sampled at 500 Hz and the corresponding text reports. Each ECG recording will have several reports, and we simply merge them into one single report (diagnosis). We apply some necessary processing steps to prepare the custom dataset for training (e.g., remove empty or containing NaN ECG recordings and clean text by using lowercase, strip, and punctuation removal), which eventually yields a training size of 779891 samples. We provide representative examples of ECG-text pairs in *Appendix* A.1.

Experimental Configurations. Our proposed model is developed based on the fairseq-signals * framework in our work. We select the Adam optimizer with a learning rate of 5×10^{-5} and use a tri-stage scheduler with ratios of 0.1, 0.4, and 0.5 for learning rate adjustments. The optimizer is configured with $\beta_1 = 0.9$, $\beta_2 = 0.98$, an epsilon value of 1×10^{-6}, and a weight decay of 0.01. We pre-train the proposed model for 300000 steps, maintaining a batch size of 128. The quantitative experiments are conducted on a single NVIDIA H100-80GB GPU.

4.1.2 DOWNSTREAM TASKS.

Downstream Datasets. We evaluate our pre-trained encoders on five widely-used public datasets: PhysioNet 2021 (Reyna et al., 2021), PTB-XL (Wagner et al., 2020), CSN (Zheng et al., 2022), CPSC2018 (Liu et al., 2018), and CODE-test (Ribeiro et al., 2020). We summarize the key information of each dataset as follows:

PhysioNet 2021. This dataset contains ECG samples (500 Hz) ranging between 5 and 144 seconds. We process and fine-tune the subsets as described in (Oh et al., 2022) to validate the pre-trained

*https://github.com/Jwoo5/fairseq-signals

Table 2: Performance comparison (AUC in %) across multiple methods and datasets. The results are shown for different percentages of training data used (1%, 10%, 100%).

Methods	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%**
SimCLR (Chen et al., 2020)	63.41	69.77	73.53	60.84	68.27	73.39	54.98	56.97	62.52	51.41	69.44	77.73	59.78	68.52	76.54	59.02	67.26	73.20
BYOL (Grill et al., 2020)	71.70	73.83	76.45	57.16	67.44	71.64	48.73	61.63	70.82	41.99	74.40	77.17	60.88	74.42	78.75	54.20	71.92	74.69
BarlowTwins (Zbontar et al., 2021)	72.87	75.96	78.41	62.57	70.84	74.34	52.12	60.39	66.14	50.12	73.54	77.62	55.12	72.75	78.39	60.72	71.64	77.43
MoCo-v3 (Chen et al., 2021)	73.19	76.65	78.26	55.88	69.21	76.69	50.32	63.71	71.31	51.38	71.66	74.33	62.13	76.74	75.29	54.61	74.26	77.68
SimSiam (Chen & He, 2021)	73.15	72.70	75.63	62.52	69.31	76.38	55.16	62.91	71.31	49.30	69.47	75.92	58.35	72.89	75.31	58.25	68.61	77.41
TS-TCC (Eldele et al., 2021)	70.73	75.88	78.91	53.54	66.98	77.87	48.04	61.79	71.18	43.34	69.48	78.23	57.07	73.62	78.72	55.26	68.48	76.79
CLOCS (Kiyasseh et al., 2021)	68.94	73.36	76.31	57.94	72.55	76.24	51.97	57.79	72.65	47.19	71.88	76.31	59.59	77.78	77.49	54.38	71.93	76.13
ASTCL (Wang et al., 2023)	72.51	77.31	81.02	61.86	68.77	76.51	44.14	60.93	66.99	52.38	71.98	76.05	57.90	77.01	79.51	56.40	70.87	75.79
CRT (Zhang et al., 2023)	69.68	78.24	77.24	61.98	70.82	78.67	46.41	59.49	68.73	47.44	73.52	74.41	58.01	76.43	82.03	56.21	73.70	78.80
ST-MEM (Na et al., 2024)	61.12	66.87	71.36	54.12	57.86	63.59	55.71	59.99	66.07	51.12	65.44	74.85	56.69	63.32	70.39	59.77	66.87	71.36
MERL (Liu et al., 2024)	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
C-MELT	**83.15**	**88.36**	**90.11**	**77.74**	**82.92**	**85.15**	**70.10**	**78.91**	**83.98**	**86.61**	**92.83**	**96.71**	**85.46**	**91.35**	**94.92**	**80.04**	**87.36**	**90.71**

Table 3: Zero-shot performance (AUC in %) comparison across multiple datasets.

Methods	PTBXL-Super	PTBXL-Sub	PTBXL-Form	PTBXL-Rhythm	CPSC2018	CSN	Average
MERL	74.2	75.7	65.9	78.5	**82.8**	74.4	75.3
C-MELT	**76.2**	**75.9**	**66.1**	**88.6**	80.1	**76.3**	**77.1**

ECG encoder in two downstream tasks: 1) 26-multi-label cardiac arrhythmia classification (Dx.); 2) patient identification (Id.), predicting patient ownership of ECG recordings.

PTB-XL. The PTB-XL dataset includes 21,837 ECG signals collected from 18,885 patients. Each sample has a 12-lead ECG recording sampled at 500 Hz over 10 seconds and corresponding cardiac labels. We follow (Liu et al., 2024) to split the dataset, including four sub-groups (super, sub, form, and rhythm). We consider them as the four separated datasets and prepare each of them with the same train, val, and test set split as in the original paper (Wagner et al., 2020).

CSN. This dataset consists of 23,026 ECG recordings sampled at 500 Hz for 10 seconds with 38 distinct labels. Therefore, it also supports the evaluation in a classification task. We use 70%:10%:20% data split as processed in (Liu et al., 2024).

CPSC2018. The dataset contains 6,877 standard 12-lead ECG recordings (500 Hz), which cover 9 distinct categories. Similarly, we also use the same data configuration following (Liu et al., 2024).

CODE-test: This contains 827 12-lead ECG samples (400 Hz) at varying lengths covering 6 abnormalities, annotated by several experienced residents and medical students. We resample the ECG signals to 500 Hz and adjust the lengths to 10 seconds.

Experimental Configurations. To evaluate our model's performance on downstream tasks, we conduct three experiments: 1) First, we integrate a linear layer on top of the pre-trained ECG encoder and fine-tune the entire model to test its efficacy in two tasks within the Physionet 2021 dataset: Dx. (by CinC score) and Id. (by % accuracy). We report the results with five cases of lead combinations, as presented in (Oh et al., 2022); 2) Second, we also implement a linear classifier but keep the ECG encoder frozen. This linear probing approach is applied at different training set sizes (1%, 10%, and 100%) to assess the macro AUC score (%) on the PTB-XL, CSN, and CPSC2018 test datasets, facilitating a comparison with our baseline (Liu et al., 2024); 3) Finally, we investigate zero-shot classification (AUC) on PTB-XL, CSN, CPSC2018 and CODE-test datasets. Here, the texts used are obtained by passing the category names through GPT-4o for capturing better medical context. The detailed configuration on each experiment is mentioned in *Appendix* A.1.

4.2 Quantitative Results

Full Fine-tuning Classifier. As shown in Table 1, our method consistently outperforms previous approaches Oh et al. (2022) in both examined tasks. In the classification task, our model achieves 85.7% accuracy with all 12 leads, significantly higher than the best baseline (W2V+CMSC+RLM), which is 73.2%. This number is even lower than our setting with only 1 lead usage (76.5%). Interestingly, the 3-lead combination yields the second-highest result, only 1.5% lower than using all leads,

Table 4: Zero-shot performance (AUC in %) under data distribution shift.

Source Domain	Zero-shot	Training Ratio	PTBXL-Super		CPSC2018		CSN	
Target Domain			**CPSC2018**	**CSN**	**PTBXL-Super**	**CSN**	**PTBXL-Super**	**CPSC2018**
SimCLR (Chen et al., 2020)	✗	100%	69.62	73.05	56.65	66.36	59.74	62.11
BYOL (Grill et al., 2020)	✗	100%	70.27	74.01	57.32	67.56	60.39	63.24
BarlowTwins (Zbontar et al., 2021)	✗	100%	68.98	72.85	55.97	65.89	58.76	61.35
MoCo-v3 (Chen et al., 2021)	✗	100%	69.41	73.29	56.54	66.12	59.82	62.07
SimSiam (Chen & He, 2021)	✗	100%	70.06	73.92	57.21	67.48	60.23	63.09
TS-TCC (Eldele et al., 2021)	✗	100%	71.32	75.16	58.47	68.34	61.55	64.48
CLOCS (Kiyasseh et al., 2021)	✗	100%	68.79	72.64	55.86	65.73	58.69	61.27
ASTCL (Wang et al., 2023)	✗	100%	69.23	73.18	56.61	66.27	59.74	62.12
CRT (Zhang et al., 2023)	✗	100%	70.15	74.08	57.39	67.62	60.48	63.33
ST-MEM (Na et al., 2024)	✗	100%	76.12	84.50	62.27	75.19	73.05	64.66
MERL (Liu et al., 2024)	✓	0%	**88.21**	78.01	76.77	76.56	74.15	**82.86**
C-MELT	✓	**0%**	72.09	**79.11**	**77.12**	**82.91**	**76.24**	80.10

Table 5: ECG interpretation comparison (AUC in %): Human experts vs. DNN (Ribeiro et al., 2020) vs. C-MELT.

Cardio Resident	Emergency Resident	Medical Student	DNN	C-MELT (Zero-shot)
92.07	90.52	93.61	96.59	**96.79**

while the 2-lead and 6-lead combinations produce comparable results, both around 81.5%. This suggests that the selected leads (I, II, V2) capture sufficient information for accurate performance. A similar pattern emerges in the identification task, where our model achieves 41.1% accuracy with a single lead, 60.5% with 3 leads, and 65.4% with 12 leads, surpassing the best baseline by 7%.

Linear Probing Classifier. Table 2 presents the linear probing results, where our method demonstrates a clear advantage over the baseline approaches Liu et al. (2024). Notably, with only 1% of the training data, our method shows a substantial improvement over MERL, especially in CSN (14% enhancement) and PTBXL-Rhythm (33%) datasets. Similarly, impressive results are observed at 10% and 100% of the data. For example, on the PTBXL-Rhythm dataset, our method achieves approximately a 10% improvement at the 10% configuration. On the CPSC2018 dataset, we also observe a considerable increase from 90.57% to 94.92% when using 100% of the training data.

Zero-shot Classifier. We first compare our method with MERL in conventional zero-shot settings across six datasets, as shown in Table 3. On average, our method achieves 77%, outperforming MERL by 2%. Notably, MERL performs particularly impressive on the CPSC2018 dataset, while its results on the other five datasets are consistently lower than ours. Next, we extend the comparison of our method with MERL and other SSL baselines Liu et al. (2024) under data distribution shifts. Specifically, we compare linear probing (100% training size) of SSL methods with MERL's and our zero-shot approach. In this setup, *source domain* and *target domain* share some common categories. Details on this implementation can be found in *Appendix* A.1. As shown in Table 4, our results surpass MERL and other methods, except when CPSC2018 is the target domain, which aligns with our previous observations. Finally, Table 5 shows that our zero-shot model outperforms three experienced cardiologists (over 3%) and also the in-domain model (Ribeiro et al., 2020), i.e., trained with millions of annotated ECG examples. We will discuss more on zero-shot settings in the *Appendix* A.3.

4.3 Ablation Studies

We evaluate the impact of the key model components, the choice of language encoders, and varying the number of transformer layers in the ECG encoder for ablation studies. Here, we focus on three downstream tasks, including full fine-tuned diagnosis classification (results across five lead combinations), linear probing at 1% training size, and zero-shot classification using category names (results across PTB-XL, CSN, and CPSC2018 datasets).

Effects of Key Components. To assess the contribution of different model components, including Flan-T5, Siglep, and N3S, we systematically remove one component at a time from the default proposed model. Specifically, we start by eliminating the N3S and train the model with randomly selected negative samples. Subsequently, we take the Siglep loss away to assess its effectiveness in capturing rich representative embeddings in both encoders. Lastly, by replacing the Flan-T5 language encoder with a standard Bert-base architecture (Devlin, 2018), we consider this as the baseline model. Table 6 demonstrates the results of this experiment. It can be seen that Siglep significantly enhances performance, showing an improvement of approximately 15% in both full fine-tuning and linear probing settings over the baseline model. Meanwhile, adding N3S improves zero-shot classification by 2%, and introducing Flan-T5 enhances performance in linear probing by 4% compared to the baseline. These results underscore the effectiveness of each component in optimizing the model's performance.

To better understand how our method improves downstream performance, we visualize and compare the t-SNE embeddings generated by our ECG encoder on the CSN test set with those from MERL. For clearer visualization, we include only samples from unique categories and exclude categories with fewer than 50 samples. Figure 3 reveals that our embeddings show more well-defined and distinct clusters representing different ECG diagnoses, which aligns with expectations.

Table 6: Effects of model components: ① FlanT5, ② Siglep, ③ N3S.

①	②	③	Full fine-tune	Linear probing	Zero-shot
✓	✓	✓	**81.88 ± 3.52**	**80.52 ± 6.08**	**72.50± 9.01**
✓	✓		80.93 ± 3.74	78.29 ± 6.19	70.61 ± 8.10
✓			78.29 ± 3.87	67.19 ± 6.14	–
			76.81 ± 3.96	63.50 ± 6.95	–

Table 7: Effects of different language encoders.

Lang encoder	Full fine-tune	Linear probing	Zero-shot
Flan-T5	**81.88 ± 3.52**	**80.52 ± 6.08**	**72.50± 9.01**
Med-CPT	81.02 ± 3.61	79.57 ± 6.32	71.81 ± 9.14
Deberta	79.23 ± 3.65	78.24 ± 6.21	70.67 ± 9.88
Bert	78.08 ± 3.91	77.58 ± 6.49	69.14 ± 9.97

Table 8: Effects of the number of transformer layers from ECG encoder. By default, our model contains 8 transformer layers.

# Layers	Full fine-tune	Linear probing	Zero-shot
8	**81.88 ± 3.52**	**80.52 ± 6.08**	**72.50± 9.01**
4	77.63 ± 4.14	70.17 ± 7.60	70.64 ± 8.63
1	69.40 ± 4.55	66.83 ± 7.52	69.43 ± 9.51

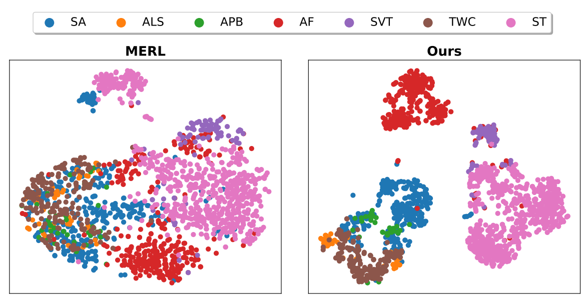

Figure 3: T-SNE visualization on the CSN test set.

Choice of Language Encoders. In this ablation study, we evaluate the performance of four pretrained language models, namely Bert (Devlin, 2018), Deberta (He et al., 2020), Med-CPT (Jin et al., 2023), and Flan-T5 (Chung et al., 2024) to determine the most suitable language encoder for our model. Here, only the base versions were tested. As shown in Table 7, Flan-T5 outperforms the others across multiple metrics, highlighting the importance of choosing a model that excels not only in general text processing but also in capturing domain-specific nuances, such as ECG reports.

Choice of Number ECG Transformer Layers. As part of our ablation study, we explore the impact of varying the number of transformer layers (1, 4, 8) in the ECG encoder. As shown in Table 8, increasing the number of layers significantly improves performance. Specifically, the 1-layer model performs 11% worse than the 8-layer model in full fine-tuning and 13% worse in linear probing. For zero-shot, the 8-layer model still delivers superior results, with 2% and 3% higher performance than the 4-layer and 1-layer models, respectively. Although these differences are smaller than in full fine-tuning, they highlight the language encoder's impact in improving performance.

5 Conclusion

We propose C-MELT to pre-train a model on ECG signals and corresponding texts, utilizing a novel contrastive masked transformer-based architecture. Our approach is generative self-supervised learning, enhanced with Siglep loss, and nearest-neighbor negative sampling to support contrastive aspects. Experimental results demonstrate that our method outperforms previous approaches across multiple datasets and on a range of downstream tasks, including under full fine-tuning, linear probing, and zero-shot classification. C-MELT shows promise in advancing ECG-based diagnostic models, paving the way for more accurate, efficient, and personalized cardiac care.

Reproducibility Statement

We provide detailed information, including dataset descriptions, experiment configurations, and other discussions in the Appendix. The code and pre-trained model will be made publicly available once the paper is accepted.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020. 2, 6

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 1, 2, 6, 7, 8

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. 1, 2, 7, 8

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. 1, 2, 7, 8

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 679–689. Springer, 2022. 2, 5

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 2, 3, 9

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 9

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024. 6

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning, 2023. URL https://arxiv.org/abs/2305.01233. 3

Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020. 1

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021. 7, 8

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pp. 156–167. PMLR, 2021. 2, 6

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. *Type: dataset*, 2023. 6, 13, 14

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020. 1, 2, 7, 8

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning, 2021. URL https://arxiv.org/abs/2010.09709. 2

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 2, 9

Rui Hu, Jie Chen, and Li Zhou. Spatiotemporal self-supervised representation learning from multi-lead ecg signals. *Biomedical Signal Processing and Control*, 84:104772, 2023. 1, 2

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. 9

Saehoon Kim, Sungwoong Kim, and Juho Lee. Hybrid generative-contrastive representation learning, 2021. URL https://arxiv.org/abs/2106.06162. 1

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pp. 5606–5615. PMLR, 2021. 1, 2, 6, 7, 8

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023. 2

Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Self-knowledge distillation based self-supervised learning for covid-19 detection from chest x-ray images. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022a. doi: 10.1109/icassp43922.2022.9746540. URL http://dx.doi.org/10.1109/ICASSP43922.2022.9746540. 2

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pp. 402–415. PMLR, 2024. 2

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022b. 1

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models, 2024. URL https://arxiv.org/abs/2301.06267. 3

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024. 2, 3, 7, 8, 13, 16

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018. 6, 14, 15

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024. 1

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024. 2, 7, 8

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pp. 338–353. PMLR, 2022. 1, 2, 4, 6, 7

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 3

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683. 4

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners, 2023. URL https://arxiv.org/abs/2212.03640. 3

Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *2021 Computing in Cardiology (CinC)*, volume 48, pp. 1–4. IEEE, 2021. 6, 14, 15

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 2020. doi: https://doi.org/10.1038/s41467-020-15432-4. 6, 8, 13, 14

Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019. 2

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. 1

Junho Song, Jong-Hwan Jang, Byeong Tak Lee, DongGyun Hong, Joon-myoung Kwon, and Yong-Yeon Jo. Foundation models for electrocardiograms. *arXiv e-prints*, pp. arXiv–2407, 2024. 1

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding, 2021. URL `https://arxiv.org/abs/2106.00750`. 2

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`. 3

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1): 1–15, 2020. 6, 7, 14, 15

Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 7, 8

Lanling Xu, Jianxun Lian, Wayne Xin Zhao, Ming Gong, Linjun Shou, Daxin Jiang, Xing Xie, and Ji-Rong Wen. Negative sampling for contrastive representation learning: A review. *arXiv preprint arXiv:2206.00212*, 2022. 5

Genshen Yan, Shen Liang, Yanchun Zhang, and Fan Liu. Fusing transformer model with temporal features for ecg heartbeat classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 898–905. IEEE, 2019. 1

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021. 7, 8

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. 3, 5

Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2022a. 1, 2

Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 2, 7, 8

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022b. 2

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022c. 2

J Zheng, H Guo, and H Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022Available online httpphysionet orgcontentecg arrhythmia10 0accessed on*, 23, 2022. 6, 14, 15

A Appendix

A.1 Data and training details.

In this section, we first visualize representative examples of ECG-text pairs from the MIMIC IV ECG dataset (Gow et al., 2023), as shown in Figure 4. We also indicate the top 30 common unique reports (before merging) in Figure 5. Prominent terms such as ”abnormal ecg”, ”normal ecg”, ”atrial fibrillation”, and ”sinus tachycardia” indicate common diagnoses, which suggests prevalent cardiovascular conditions and typical annotations within this dataset.

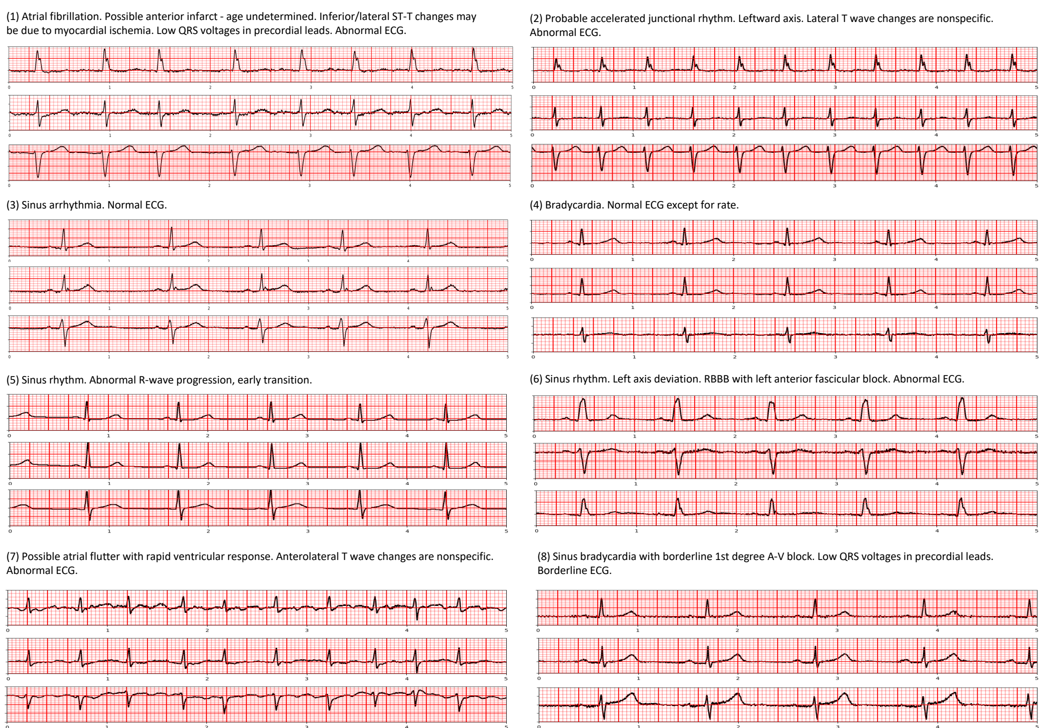

Figure 4: Examples of ECG-text pairs in MIMIC IV ECG dataset (Gow et al., 2023). We visualize three leads (I, II, V2) out of twelve.

Next, we provide more details on data configurations in Table 9, including data split, number of classes, metrics, and the corresponding tasks with the given downstream dataset.

CODE-test: Particularly, this data is from the work of Ribeiro et al. (2020), which is the test set used for evaluating their trained model's performance compared with cardiology resident medical doctors. It is worth noting that their training set consists of over 2 million ECG records from 1,676,384 different patients in 811 counties. We evaluate the performance of our method on the same released test set of 827 samples in a zero-shot manner. These samples are originally sampled at 400 Hz, with durations of either 10 seconds or 7 seconds. Therefore, we resampled to 500 Hz and adjusted by truncating or padding with zeros as needed to get 10-second samples. For the gold standard (ground truth), two expert cardiologists provided their diagnoses. If they agree with each other, their consensus becomes the gold standard. In cases of disagreement, a third specialist reviews their diagnoses and determines the final decision.

We also indicate important hyper-parameters during the fine-tuning process in Table 10. We keep training 200 epochs, batch size at 128, and learning rate at 0.001 for the first three datasets. When conducting full fine-tuning experiments, we only need to train 100 epochs and specifically lower the learning rates with 0.00005 and 0.0001 for Dx. and Id. tasks, respectively.

For the distribution shift experiment, we follow the SCP-codes (classes) matching settings in (Liu et al., 2024), which can be seen in Table 11. This is to support three dataset matches (PTBXL-Super

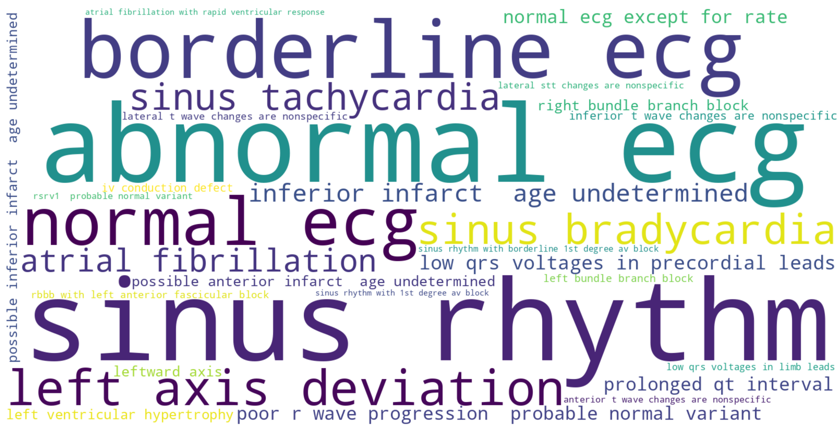

Figure 5: WordCloud visualization on the top 30 common unique reports from MIMIC IV ECG dataset.

Table 9: Details on data configurations on five evaluated datasets. Here, LP, ZS are linear probing and zero-shot respectively, while FFT means full fine-tuning.

Dataset	Tasks	Metric	# Classes	Train	Valid	Test
PTBXL-Super (Wagner et al., 2020)	LP, ZS	AUC	5	17,084	2,146	2,158
PTBXL-Sub (Wagner et al., 2020)	LP, ZS	AUC	23	17,084	2,146	2,158
PTBXL-Form (Wagner et al., 2020)	LP, ZS	AUC	19	7,197	901	880
PTBXL-Rhythm (Wagner et al., 2020)	LP, ZS	AUC	12	16,832	2,100	2,098
CPSC2018 (Liu et al., 2018)	LP, ZS	AUC	9	4,950	551	1,376
CSN (Zheng et al., 2022)	LP, ZS	AUC	38	16,546	1,860	4,620
Physionet2021-Dx. (Reyna et al., 2021)	FFT	CinC	26	32,640	4,079	4,079
Physionet2021-Id. (Reyna et al., 2021)	FFT	Accuracy	2,127	147,444	17,670	2,127
CODE-test (Ribeiro et al., 2020)	ZS	AUC	6	–	–	827

and CPSC2018), (PTBXL-Super and CSN), and (CPSC2018 and CSN). It is worth noting that the None value indicates the target dataset does not have a matching label for given labels in the source dataset.

A.2 Contrastive learning discussion.

Why Using ETM Only Is Not A True Way To Zero-shot Learning. As mentioned in the Method section, ETM functions as a contrastive learning technique in the masked auto-encoder architecture. However, it heavily relies on binary classification tasks with explicit ECG-text pairs to learn cross-modal correspondences. It is not designed for zero-shot learning which strongly requires the model to generalize to unseen tasks or classes without the need for such supervised pairings or fused information during training. This motivates us to use Siglep, boosting the model's zero-shot ability.

Why N3S Can Enhance The Performance. In medical datasets, particularly the MIMIC-IV ECG dataset (Gow et al., 2023), we observe a significant amount of duplicate or highly similar text samples: among nearly 800,000 records, only approximately 180,000 are unique. For instance, over 100,000 samples share an identical text report, which is ”sinus rhythm normal ecg”. Randomly selecting negative samples for contrastive loss training is not a suitable approach in this scenario. Therefore, we propose using the N3S technique to more effectively differentiate between similar and dissimilar samples, improving contrastive learning by selecting more meaningful negatives. Notably, we observe that during training, the ETM accuracy without N3S stagnates around 75% while with N3S, it exceeds 96%, demonstrating the significant impact of this approach.

Table 10: Details on training configurations on the fine-tuned datasets. For optimizer, we keep using Adam in all experiments.

Dataset	**# Epoch**	**Batch size**	**Learning rate**
PTBXL-Super (Wagner et al., 2020)	200	128	0.001
PTBXL-Sub (Wagner et al., 2020)	200	128	0.001
PTBXL-Form (Wagner et al., 2020)	200	128	0.001
PTBXL-Rhythm (Wagner et al., 2020)	200	128	0.001
CPSC2018 (Liu et al., 2018)	200	128	0.001
CSN (Zheng et al., 2022)	200	128	0.001
Physionet2021-Dx. (Reyna et al., 2021)	100	256	0.00005
Physionet2021-Id. (Reyna et al., 2021)	100	256	0.0001

Table 11: Domain transfer category matching.

PTBXL-Super	**CPSC2018**
HYP	None
NORM	NORM
CD	1AVB, CRBBB, CLBBB
MI	None
STTC	STE, STD
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
CPSC2018	**CSN**
AFIB	AFIB
VPC	VPB
NORM	SR
1AVB	1AVB
CRBBB	RBBB
STE	STE
PAC	APB
CLBBB	LBBB
STD	STE, STTC, STTU, STDD

A.3 ENHANCING ZERO-SHOT PERFORMANCE WITH LLM.

(1) Response with merging subtypes reducing capability on new tasks

"**AFIB**":"**Atrial Fibrillation**, **Paroxysmal** Atrial Fibrillation, **Persistent** Atrial Fibrillation, **Long-standing Persistent** Atrial Fibrillation, **Permanent** Atrial Fibrillation."

"**SEHYP**": "septal hypertrophy, **left** ventricular septal hypertrophy, **right** ventricular septal hypertrophy, apical septal hypertrophy, **mid**-septal hypertrophy."

(2) Response showing limitations on LLM's searching and hallucination

"**AF**": "Atrial Flutter**, Atrial Fibrillation**, Paroxysmal Atrial Flutter, Persistent Atrial Flutter, Long-standing Persistent Atrial Flutter."

"**BIGU**": "**Based on the input, I generated** the following subtypes and attributes for Bigeminal pattern …**Let me know if this meets your requirements!**."

Figure 6: Limitations on MERL's enhanced texts.

Why Using LLMs But Not As MERL. In zero-shot learning, models typically rely on category names alone to make predictions. However, by incorporating Large Language Models (LLMs), we can enhance the context by generating richer, clinically relevant descriptions of the categories, as discussed in MERL (Liu et al., 2024). However, we observe two main drawbacks in their enhanced text reports, as shown in Figure 6: 1) MERL's performance heavily depends on their sub-types and attributes searching prompt and additional database. This leads to a limitation when testing detailed analysis with labels that are different sub-types themselves. Moreover, this also raises suspicion about the performance when new tasks require labels that are not able to search sub-types and attributes in the database; 2) Following that point, MERL's enhanced texts might be uncontrollable to the outputs where the LLMs provide wrong sub-types or unnecessary context. For example, "Atrial Fibrillation" is already in "AFIB" type but shown to misleadingly be in "AF"-"Atrial Flutter".

How Our Work Leverages LLM's Strength. We address these points using a straightforward prompt strategy with explicit instructions. Specifically, we employ a prompt: *"You are an experienced cardiologist. For a given clinical term such as 'normal ECG', your job is to describe each term clinically and apply your medical domain knowledge to include other relevant explanations that will help a text encoder like Flan-T5 fully understand medical concepts. Do not include any recommendations in the description."* This makes the LLM generate clinically accurate and more focused explainable descriptions, enhancing the text encoding without introducing irrelevant or redundant information. For example, with the code "AFIB", our prompt on GPT-4o can output: "Atrial Fibrillation (AFIB). Irregular and often rapid heart rate due to uncoordinated atrial activity.".

Additional Experiments Here, we present additional experiments to highlight the effectiveness of ETM loss and masking modeling techniques (e.g., MLM, MEM). Specifically, we perform zero-shot classification with GPT-4o support (reported in AUC (%)) on four datasets: PTBXL-Super, PTBXL-Form, CSN, and CODE-Test.

Table 12: Impact of ETM. Results report zero-shot classification in AUC (%).

	PTBXL-Super	PTBXL-Form	CSN	CODE-Test
w/o ETM	73.2	65.8	**76.6**	96.2
w ETM	**76.2**	**66.1**	76.3	**96.8**

As indicated in Table 12, the impact of ETM is demonstrated where, removing ETM slightly decreases performance across most datasets, particularly in PTBXL-Super (76.2 to 73.2). However, the effect on CSN is minimal, suggesting dataset-specific sensitivity to ETM.

Table 13: Impact of MLM and MEM. Results report zero-shot classification in AUC (%).

	PTBXL-Super	PTBXL-Form	CSN	CODE-Test
w/o MLM + MEM	70.3	**67.4**	74.5	94.6
w MLM + MEM	**76.2**	66.1	**76.3**	**96.8**

Next, we can see that incorporating MLM and MEM noticeably improves performance across all evaluated datasets in Table 13. Especially, gains are observed in PTBXL-Super (+5.9%), and CODE-Test (+2.2%), demonstrating that the reconstruction tasks play an important role in enhancing the model's ability for better performance, aligned with our motivation.