Revisiting Masked Auto-Encoders for ECG-Language Representation Learning

Hung Manh Pham
Singapore Management University
hm.pham.2023@phdcs.smu.edu.sg

Aaqib Saeed
Eindhoven University of Technology
a.saeed@tue.nl

Dong Ma
Singapore Management University
dongma@smu.edu.sg

Abstract

We propose C-MELT , a novel framework for multimodal self-supervised learning of Electrocardiogram (ECG) and text encoders. C-MELT pre-trains a contrastive-enhanced masked auto-encoder architecture using ECG-text paired data. It exploits the generative strengths with improved discriminative capabilities to enable robust cross-modal alignment. This is accomplished through a carefully designed model, loss functions, and a novel negative sampling strategy. Our preliminary experiments demonstrate significant performance improvements with up to 12% in downstream cardiac arrhythmia classification and patient identification tasks. Our findings demonstrate C-MELT 's capacity to extract rich, clinically relevant features from ECG-text pairs, paving the way for more accurate and efficient cardiac diagnoses in real-world healthcare settings.

1 Introduction

Electrocardiograms (ECGs) provide critical insights into the heart's electrical activity through non-invasive electrodes, with the standard 12-lead ECG being key to diagnosing conditions like arrhythmias and myocardial infarction. While deep learning has revolutionized automated ECG interpretation, it often depends on large, labeled datasets, which are expensive to obtain. Self-supervised learning (SSL) has emerged as a promising alternative, allowing models to learn meaningful representations from vast unlabeled ECG data that can be fine-tuned or used for zero-shot learning on downstream tasks [22, 7, 19].

SSL methods in the ECG domain primarily follow two tracks: contrastive and generative. Contrastive approaches [2, 3, 10, 12, 16, 15] learn by distinguishing between positive and negative pairs, while generative approaches [11, 24, 25] aim to reconstruct missing segments of the ECG signal. Despite these advances, most SSL models overlook clinical text reports, which contain valuable diagnostic information [26, 4]. Recent efforts [14, 13] have begun integrating ECG signals and clinical reports through cross-modal contrastive learning, but joint ECG-text representation learning using generative methods remains underexplored. Furthermore, their contrastive methods often rely on randomly sampled negative pairs, which can be especially risky in the medical domain.

In this work, we introduce C-MELT , a hybrid framework combining contrastive and generative learning to capture ECG-text representations. Our model employs a masked multimodal autoencoder with carefully designed loss functions and a novel nearest-neighbor negative sampling strategy to enhance discriminative ability. We conduct extensive experiments by fine-tuning the pre-trained ECG encoder on popular downstream tasks, demonstrating that C-MELT significantly outperforms state-of-the-art baselines across all evaluations.

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

2 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 reconstruction tasks and contrastive learning objectives to capture intricate relationships between these modalities.

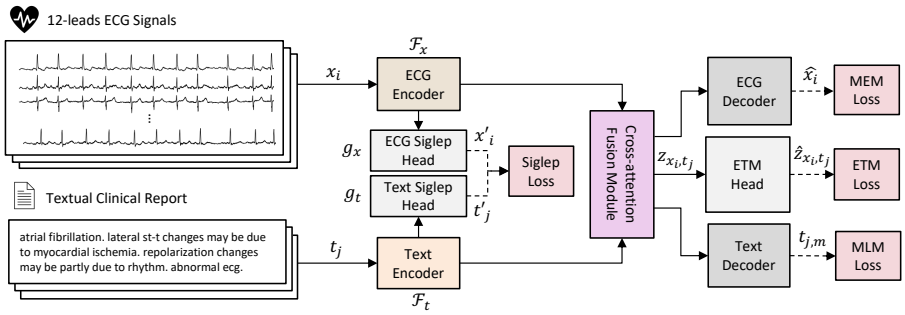

Figure 1: Illustration of our C-MELT framework for learning ECG-Text multimodal representations.

Figure 1 shows the architecture of C-MELT , comprising ECG and text encoders for cross-modal representation learning. The ECG encoder uses a transformer-based model [1] to process ECG signals into embeddings $\mathbf{H}_x$, while the text encoder employs the pre-trained Flan-T5 model [5] to extract embeddings $\mathbf{H}_t$ from clinical text. A fusion module with cross-attention integrates these representations into fused embeddings $\mathbf{H}_f$. The model includes decoders for reconstructing masked ECG signals and text, and a contrastive prediction head for ECG-text matching. We add projection heads g_x and g_t to facilitate discriminative representation learning with the Siglep loss. Our model is trained to optimize jointly four loss functions: masked language modeling ($\mathcal{L}_{MLM}$), masked ECG modeling ($\mathcal{L}_{MEM}$), ECG-text matching ($\mathcal{L}_{ETM}$), and the Siglep loss ($\mathcal{L}_{Siglep}$).

2.1 Multi-Modal masked auto-encoders.

ECG Encoder. We implement the ECG encoder (denoted as $\mathcal{F}_x$) using a transformer architecture [20] for efficient parallel processing of sequential data. Following [16], we apply a masking strategy to the ECG input $\mathbf{X} \in \mathbb{R}^{L \times C}$, where L is the signal length and C is the number of channels, to encourage robust feature learning. The masked input passes through convolutional layers with GELU activations and group normalization, projecting the features into a 768-dimensional space. We then employ eight transformer encoder layers with multi-head self-attention to capture complex dependencies in the ECG data. A feed-forward network further processes the features, and positional encoding is added to preserve the temporal order of the ECG sequence.

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 [17], which has been pre-trained on a massive and diverse text dataset covering numerous tasks, such as summarization and question answering.

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. Our model has three distinct network heads, 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. It 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:

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

where $\mathcal{B}$ is 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). MEM reconstructs masked ECG inputs, analogous to Masked Language Modeling. We embed the input sequence into a 384-dimensional space, incorporate learnable mask tokens and positional encodings to preserve the temporal structure and employ a multi-layer transformer decoder to capture sequence dependencies. A linear projection outputs the predicted ECG features, and we train MEM using the mean squared error loss (Equation 2):

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

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:

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

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.

2.2 Improving Contrastive Learning

Siglep Loss Function. To enhance the learning of discriminative features essential for downstream tasks, we address limitations of reconstruction-focused multi-modal masked autoencoders [4] and the ETM loss, which is not optimized for individual encoder discrimination. We adapt the Siglip method [23] to the ECG-text domain, introducing the Siglep loss function. Siglep operates independently on each ECG-text pair, eliminating the need for computationally expensive global normalization required by traditional softmax-based contrastive losses, thereby improving memory efficiency and scalability. We augment the ECG and text encoders with additional network heads, each comprising a pooling layer, a Tanh activation, and a dense layer to output 768-dimensional embeddings ($\mathbf{x'}_i, \mathbf{t'}_j \in \mathbb{R}^{768}$). The Siglep loss is defined as:

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

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

Nearest-neighbor-based negative sampling. In contrastive learning, effective negative sample selection is crucial [21]; random sampling often leads to false negatives in medical datasets due to report similarities, impeding learning. We propose a nearest-neighbor-based negative sampling strategy that enhances negative sample quality by selecting negatives dissimilar to positive samples in the Flan-T5 feature space. Specifically, we utilize a pre-trained Flan-T5 (small) to embed each text report $t \in \mathcal{D}_{train}$ as $\mathbf{v}_t \in \mathbb{R}^{512}$. During training, for each ECG and positive text pair (x_k, t_k^+) in half of the batch $\mathcal{B}$, we select the negative report t_k^- as one of the top 64 most dissimilar reports from $\mathbf{v}_{t_k^+}$ based on cosine distance. This approach ensures negatives are challenging yet distinct, promoting effective contrastive learning. We employ FAISS [6] for efficient nearest-neighbor search, enabling scalable application to large datasets.

Table 1: Test performances when fine-tuning on the five lead combinations. In fine-tuning, we fill unavailable leads with zero, which is denoted as P-N-lead (Padded-N-lead).

Methods	**Tasks**	**# Leads**				
		12-lead	P-6-lead	P-3-lead	P-2-lead	P-1-lead
W2V [1]	Dx.	71.4	64.3	67.6	61.1	52.5
	Id.	49.2	41.1	47.0	41.4	24.7
CMSC [12]	Dx.	62.5	52.2	57.5	50.7	40.6
	Id.	51.3	39.2	51.0	37.8	22.7
3KG [8]	Dx.	60.0	51.5	56.3	50.5	41.8
	Id.	40.7	32.0	36.7	31.0	19.8
SimCLR(RLM) [2]	Dx.	57.8	49.7	53.5	48.4	39.3
	Id.	35.3	28.9	36.8	30.4	19.2
W2V+CMSC [16]	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 [16]	Dx.	73.2	66.2	71.4	65.6	55.4
	Id.	57.7	45.9	54.8	45.7	31.3
Ours	**Dx.**	**85.7**	**81.1**	**84.2**	**81.9**	**76.5**
	Id.	**65.4**	**57.3**	**60.5**	**57.7**	**41.1**

3 Experiments

3.1 Implementation details.

We pre-trained our model on the MIMIC-IV-ECG v1.0 database [9], comprising 779,891 ECG-report pairs from 161,352 unique subjects after preprocessing. Each ECG is a 10-second, 500 Hz recording from Beth Israel Deaconess Medical Center, with corresponding text reports consolidated into a single diagnosis per recording. We removed invalid ECGs and cleaned text (lowercasing, stripping, punctuation removal) to prepare the dataset. Implementing our model with the fairseq-signals framework, we pre-trained it for 300,000 steps with a batch size of 128 on a single NVIDIA H100-80GB GPU. We optimized using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-6}$, weight decay 0.01) with a learning rate of 5×10^{-5}, adjusted via a tri-stage scheduler with ratios 0.1, 0.4, and 0.5.

We evaluate our pre-trained model on the PhysioNet 2021 dataset [18], focusing on subsets as described in [16]. Two downstream tasks are considered: 1) Cardiac Arrhythmia Classification (Dx.), a 26-multi-label task predicting cardiac abnormalities, and 2) Patient Identification (Id.), predicting patient ownership of ECG recordings. For evaluation, we add a single dense layer to the pre-trained ECG encoder and fine-tune the entire model. Performance is assessed using the CinC score for arrhythmia detection and accuracy for patient identification, across five lead combinations, as in [16].

3.2 Empirical Results.

Table 1 shows that our method consistently outperforms previous approaches in both tasks. In classification, our model achieves 76.5% accuracy with a single lead, surpassing the best baseline's 73.2% in all lead settings. The 3-lead combination provides nearly as good results, just 2% below using all leads, while the 2-lead and 6-lead combinations are comparable at around 81.5%. This suggests the selected leads (I, II, V2) provide effective information for the task. Similarly, for identification, our model reaches 41.1% accuracy with 1-lead, 60.5% with 3-lead, and 65.5% with all lead usage, outperforming the best baseline by 7%.

4 Conclusion

In this paper, we propose C-MELT a multimodal self-supervised learning technique for learning representations from ECG signals and corresponding texts, utilizing a novel masked transformer-based architecture. Our approach is a hybrid of generative and contrastive learning, enhanced with Siglep loss function, and nearest neighbor negative sampling to support contrastive aspects. The experimental results demonstrate that our method outperforms previous approaches in fully fine-tuned cardiac arrhythmia classification classification and patient identification tasks. C-MELT shows promise in advancing ECG-based diagnostic models.

References

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

[3] 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*, pages 9640–9649, 2021.

[4] 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*, pages 679–689. Springer, 2022.

[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.

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

[7] 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.

[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*, pages 156–167. PMLR, 2021.

[9] 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.

[10] 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.

[11] 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.

[12] 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*, pages 5606–5615. PMLR, 2021.

[13] 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.

[14] 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.

[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.

[16] 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*, pages 338–353. PMLR, 2022.

[17] 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.

[18] 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, pages 1–4. IEEE, 2021.

[19] 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.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[21] 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.

[22] 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)*, pages 898–905. IEEE, 2019.

[23] 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*, pages 11975–11986, 2023.

[24] 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, 2022.

[25] 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.

[26] 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*, pages 2–25. PMLR, 2022.