Two-Stage Contrastive Language Electrocardiogram Pre-training for Fine-Grained Waveform Features

Haitao Li [1] **Che Liu** [2] **Zhengyao Ding** [1] **Ziyi Liu** [3] **Zhengxing Huang** [1]

Abstract

Electrocardiograms (ECGs) play a vital role in diagnosing cardiovascular diseases. While recent ECG-text contrastive learning methods have shown promise, they often overlook the incomplete nature of clinical reports. Typically, a report is generated by identifying key waveform features and then deriving a diagnosis, yet these intermediate features are rarely documented. This gap limits the model's ability to capture waveform patterns and understand the underlying diagnostic reasoning. To address this, we propose FG-CLEP (Fine-Grained Contrastive Language ECG Pre-training), which leverages large language models (LLMs) to recover the missing waveform features from incomplete reports. To further improve performance, we introduce a semantic similarity matrix to reduce false negatives caused by the prevalence of common diagnoses and adopt a sigmoid-based loss function to handle the multi-label nature of ECG tasks. Experiments on six datasets show that FG-CLEP consistently outperforms state-of-the-art methods in both zero-shot prediction and linear probing.

1. Introduction

Electrocardiograms (ECGs) are widely used, non-invasive tools for detecting heart rhythm disorders (Sahoo et al., 2020; Rath et al., 2021; Ayano et al., 2022). ECG self-supervised learning leverages unlabeled data to reduce reliance on expert-labeled annotations. Existing methods fall into two categories: comparative self-supervision (Chen et al., 2020; 2021; Wang et al., 2023; Eldele et al., 2021), which distinguishes positive and negative samples, and generative self-supervision (Zhang et al., 2022a; Hu et al., 2023; Na et al., 2024; Zhang et al., 2022b), which reconstructs masked signals. However, both types often require labeled samples for fine-tuning and struggle with unseen classes during inference.

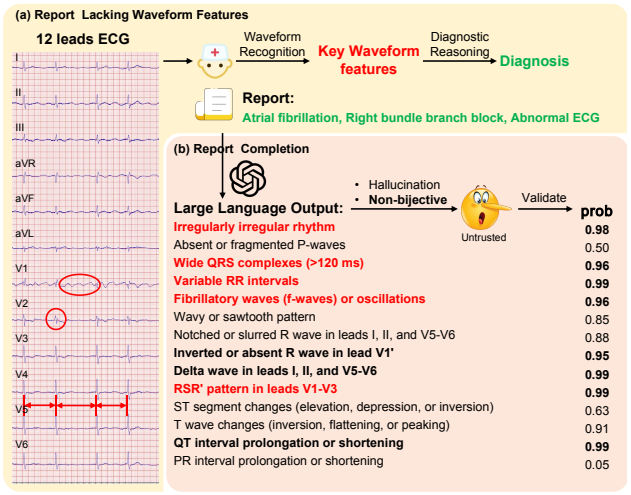

Figure 1. (a) Doctors typically make diagnoses based on waveform features. (b) Using LLMs to recover waveform features from diagnosis faces two challenges: hallucinations and the non-bijective relationship

To tackle the zero-shot challenge, recent works draw inspiration from multimodal models like CLIP (Radford et al., 2021). Methods such as (Li et al., 2024; Liu et al., 2024a; Lalam et al., 2023) align ECGs with corresponding reports, enabling zero-shot predictions. MERL (Liu et al., 2024b) adds uni-modal alignment and uses LLMs during inference to generate richer prompts. ESI (Yu et al., 2024) improves training-time reports using retrieval-augmented generation. However, these methods overlook a key issue: the absence of waveform features in ECG reports. As Figure 1(a) shows, doctors base diagnoses on waveform features, which are often omitted in reports. Simply using LLMs to augment reports (Yu et al., 2024), as in Figure 1(b), is unreliable due to hallucinations (Huang et al., 2023; Günay et al., 2024) and the non-bijective relation between features and diagnoses (Jin, 2018).

[1]Zhejiang University [2]Imperial College London [3]Transtek Medical Electronics Co., Ltd. Correspondence to: Zhengxing Huang <zhengxinghuang@zju.edu.cn>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

We propose FG-CLEP, a novel training approach to recover these missing waveform features. It includes three steps: (1) training a CLEP model on original ECG-report pairs via contrastive learning, (2) prompting LLMs to generate potential waveform features from reports and validating them using CLEP, and (3) continuing training with augmented reports containing validated features. The key is to use CLEP to filter LLM-generated features based on similarity to the ECG signal, keeping only high-confidence features. This addresses both the feature-diagnosis ambiguity and LLM hallucinations.

In addition, ECG data differ significantly from images, making CLIP-style models less effective without adaptation. Most ECGs are normal (Thai et al., 2017; Yogarajan et al., 2021), leading to similar report semantics. Prior work treats only matched ECG-report pairs as positives, which can cause false negatives. We introduce a semantic similarity matrix to identify and downweight such cases. Furthermore, ECG tasks are often multi-label, while prior methods use softmax-based InfoNCE loss (Oord et al., 2018), which assumes single-label classification. We replace it with a sigmoid-based loss to better support multi-label settings.

We validate FG-CLEP on six multi-label ECG datasets under zero-shot and linear probing settings. Our model achieves competitive or superior performance compared to state-of-the-art baselines in both evaluations.

2. Method

We introduce **FG-CLEP**, a contrastive learning framework for ECG-report representation learning. It includes a novel pre-training strategy to recover missing waveform features and a model architecture designed to handle multi-label classification and reduce false negatives. Figure 2 illustrates the full pipeline.

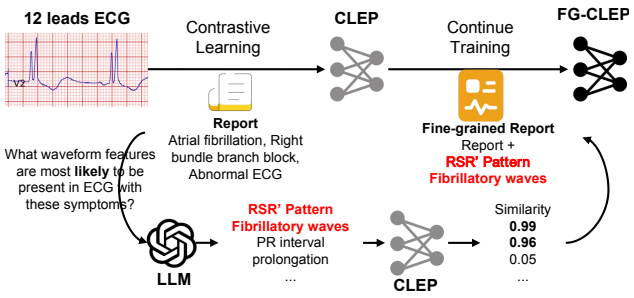

Figure 2. **Training Porcess of FG-CLEP.** (1) training the CLEP model using contrastive learning on original ECG-report pairs, (2) generating potential waveform features based on the original report using LLMs and validating them with CLEP, and (3) continuing to train the CLEP model with this augmented report containing waveform features to obtain the final FG-CLEP model.

2.1. Model Architecture

FG-CLEP includes an ECG encoder and a text encoder.

ECG Encoder. The ECG encoder E_{ecg} maps ECG signals to embeddings, followed by a projection head f_e:

$$\mathbf{e}_p = f_e(E_{\text{ecg}}(x_{\text{ecg}})) \tag{1}$$

Text Encoder. Similarly, the text encoder E_{txt} processes clinical reports, followed by a projection head f_t:

$$\mathbf{t}_p = f_t(E_{\text{txt}}(x_{\text{txt}})) \tag{2}$$

Both $\mathbf{e}_p$ and $\mathbf{t}_p$ are projected to the same dimension P for contrastive learning.

Semantic Similarity Matrix False negatives in the pre-training phase arise from the assumption that ECGs and reports from different patients are unmatched. However, due to the prevalence of common diagnoses, many ECGs exhibit similar symptoms, leading to reports with similar semantics. To address this, we introduce a Semantic Similarity Matrix similar to (Sun et al., 2023; Wang et al., 2022) to measure the similarity of reports from different patients during the pre-training phase.

We denote an ECG-report dataset as $D = \{(x_{\text{ecg}_i}, x_{\text{txt}_i}) \mid i \in [0, n)\}$, where $(x_{\text{ecg}_i}, x_{\text{txt}_i})$ represents a sample with paired ECG-report content. The ECG and text signals are encoded into $(\mathbf{e}_p, \mathbf{t}_p)$ as disscussed above, and the semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ is defined as follows:

$$\mathbf{S}_{ij} = \text{sim}(\mathbf{t}_{pi}, \mathbf{t}_{pj}) = \frac{\mathbf{t}_{pi} \cdot \mathbf{t}_{pj}}{\|\mathbf{t}_{pi}\| \|\mathbf{t}_{pj}\|} \tag{3}$$

which measures cosine similarity between reports.

Loss Function

The loss function of our FG-CLEP framework consists of two parts: the sigmoid-based contrastive loss L_{sig} and the false negative mitigation loss L_{fnm}. These components together enhance the alignment between ECG signals and their corresponding textual reports.

Sigmoid-based Contrastive Loss

$$L_{\text{sig}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} \log \left(\frac{1}{1 + e^{-z_{ij}(-t \cdot \text{sim}(\mathbf{e}_{pj}, \mathbf{t}_{pj}) + b)}} \right) \tag{5}$$

This loss replaces the traditional softmax with a sigmoid function, making it more suitable for multi-label ECG classification. The temperature parameter t controls the smoothness of similarity scores, and the learnable bias b helps

Table 1. Results of zero-shot classification. ENS: ensemble inference

macro AUC	PTB-XL-Super	PTBXL-Sub	PTBXL-Form	PTBXL-Rhythm	CPSC2018	CSN
MERL	74.20	75.70	65.90	78.50	82.80	74.40
CLEP	77.50	81.85	66.29	88.60	85.15	80.10
CLEP_{ENS}	75.64	82.55	64.74	88.67	83.91	81.24
FG-CLEP	79.28	83.57	67.77	**92.31**	**88.24**	82.46
FG-CLEP $_{\text{ENS}}$	**79.68**	**83.65**	**70.79**	91.52	87.08	**84.60**

mitigate the imbalance between positive and negative pairs during early training.

False Negative Mitigation Loss

$$L_{\text{fnm}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}|\text{sim}(\mathbf{e}_{pi}, \mathbf{t}_{pj}) - \mathbf{S}_{ij}| \tag{6}$$

This loss uses a semantic similarity matrix $\mathbf{S}$ to reduce the impact of false negatives, which occur when non-paired but semantically similar samples are treated as negatives. It encourages embedding similarity to match semantic similarity.

Combined Loss

$$L = L_{\text{sig}} + \lambda L_{\text{fnm}} \tag{7}$$

The final loss combines both terms, where λ is a balancing hyperparameter.

2.2. Training Process

As shown in Figure 2, our training process includes three steps: (1) training the CLEP model on ECG-report pairs via contrastive learning, (2) using LLMs to generate potential waveform features from reports and validating them with CLEP, and (3) continuing training with reports augmented by validated features to obtain the final FG-CLEP model. Since waveform features—key to diagnostic reasoning—are often missing from reports, we prompt the LLM with: "What waveform features are most likely to be present in ECGs with these symptoms?" and format the output as a Python list. Due to the non-bijective nature of ECG features and diagnoses (Jin, 2018) and LLM hallucinations (Huang et al., 2023; Günay et al., 2024), we validate LLM-generated features by measuring their similarity to the ECG using coarse CLEP, keeping only high-confidence results for augmentation. See Appendix D for pseudo code.

3. Experiments

3.1. Datasets and Implementation

We pre-train the FG-CLEP framework using the MIMIC-ECG (Gow et al.) dataset and test it on the PTB-XL (Wagner et al., 2020), CPSC2018 (Liu et al., 2018), and CSN (Zheng et al., 2022) datasets, following the benchmark proposed by (Liu et al., 2024b). All the ECGs in the datasets are 12-lead recordings. The MIMIC-ECG dataset contains nearly 800,000 ECG-report pairs. To improve data quality, we excluded samples with an empty report or reports containing fewer than three words, removed reports without useful information, and discarded ECGs with unexpected situations. Details regarding the train:validation:test split and other dataset-specific information are provided in Appendix B. The detailed implementation is provided in Appendix C.

3.2. Zero-Shot Ability

We conducted a zero-shot ECG classification evaluation on four PTB-XL subsets, CPSC2018, and CSN. The results are illustrated in Table 1. Both CLEP and FG-CLEP performed well. A detailed examination of the data reveals that FG-CLEP significantly outperforms CLEP on PTBXL-Form, demonstrating that continue training using fine-grained reports substantially enhanced the model's ability to capture local ECG waveform features. This improvement is particularly evident when using the ensemble method, which extends the label text to 12 leads ('label in lead x', where x represents any of the 12 leads). This further indicates FG-CLEP's fine-grained waveform feature capture capability. However, the ensemble inference method often proves detrimental to CLEP, as seen in PTBXL-Super, PTBXL-Form, and CPSC2018.

3.3. Linear Evaluation

We evaluate the transferability of the learned model to downstream supervised tasks by freezing the ECG encoder and training a randomly initialized linear classification head using binary cross-entropy loss. Several contrastive and generative self-supervised methods are compared. As shown in Table 2, FG-CLEP consistently outperforms other methods in most cases.

Moreover, comparing the linear probe results (Table 2) with the zero-shot results (Table 1), we find that FG-CLEP's zero-shot performance is comparable to linear probing with 10% of the training data on PTBXL-Sub, PTBXL-Form, CPSC2018, and CSN. Notably, in PTBXL-Form, zero-shot performance is even close to the result using the full dataset.

Table 2. Results of Linear Evaluation.

Method	PTB-XL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random Init	70.45	77.09	81.61	55.82	67.60	77.91	55.82	62.54	73.00	46.26	62.36	79.29	54.96	71.47	78.33	47.22	63.17	73.13
SimCLR(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	57.96	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., 2024b)	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
CLEP	84.04	88.79	89.82	69.09	**86.08**	92.50	67.89	72.35	82.59	61.79	**91.86**	90.18	83.12	93.42	96.56	63.00	80.03	93.35
FG-CLEP	**84.89**	**89.51**	**90.77**	**69.96**	85.75	**92.62**	**68.91**	**74.80**	**85.42**	**68.99**	91.35	**94.08**	**83.35**	**93.60**	**96.65**	62.59	79.35	**93.46**

These findings further demonstrate the robustness and generalizability of our framework.

3.4. Semantic Similarity Matrix

We visualize the semantic similarity matrix in Figure 3. The left side shows the semantic similarity matrix from a random batch. As illustrated, ECGs and reports from different records may share similarities to some extent. Ignoring these similarities would result in a diagonal matrix with ones on the diagonal and zeros elsewhere, which is obviously wrong. The right side displays a semantic similarity matrix where the first 16 entries are normal ECGs and the last 16 are abnormal ECGs. The matrix effectively captures the semantic similarities of the normal ECGs.

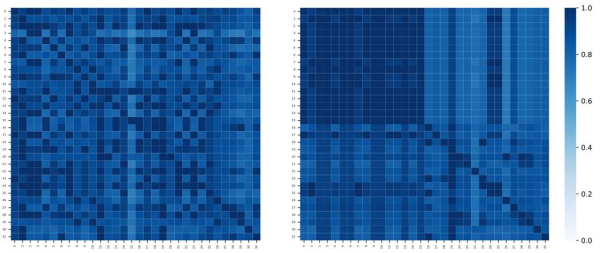

Figure 3. **The Heatmap of Semantic Similarity Matrix.** Left: from a random batch; Right: with the first 16 as normal ECG and the last 16 as abnormal ECG.

3.5. Different LLMs

In the fine-grained pre-training stage, we used LLMs to extract potential waveform features associated with ECG reports. To evaluate the feasibility of our framework, we conducted experiments using a variety of LLMs, including both general-purpose models and domain-specific medical LLMs. The results are presented in Table 3. Notably, while larger LLMs offered some performance improvements, the gains were modest. Although medical LLMs possess more specialized knowledge, they proved less effective at formatting the waveform feature outputs required by our method. As a result, their performance did not surpass that of the general-purpose models.

Table 3. Results on different LLMs.

LLM	AUC
Phi-3-mini-4k-instruct (Abdin et al., 2024)	81.51
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023a)	81.89
Llama3-8B-Instruct (AI@Meta, 2024)	82.27
Llama3-70B-Instruct (AI@Meta, 2024)	82.80
BioMistral-7B (Labrak et al., 2024)	81.67
Llama3-OpenBioLLM-8B (Ankit Pal, 2024)	82.36

4. Conclusion

In this paper, we introduced FG-CLEP, a novel approach for fine-grained contrastive learning in ECG-text tasks, addressing the critical issue of incomplete ECG reports, particularly the absence of key waveform features. By incorporating LLMs to generate potential waveform features and validating them with our pipeline, we were able to enhance the quality of ECG reports and better capture the diagnostic reasoning process. Additionally, we tackled the challenges associated with multi-label classification and frequent false negatives by introducing a sigmoid-based loss function and a semantic similarity matrix to guide contrastive learning respectively. Experimental results across six ECG datasets, including PTB-XL, CPSC2018, and CSN, demonstrate that FG-CLEP outperforms existing state-of-the-art methods in both zero-shot prediction and linear probing, highlighting its effectiveness in improving ECG classification and facilitating more accurate diagnostic insights.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

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

Ankit Pal, M. S. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

Ayano, Y. M., Schwenker, F., Dufera, B. D., and Debelee, T. G. Interpretable machine learning techniques in ecg-based heart disease classification: a systematic review. *Diagnostics*, 13(1):111, 2022.

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

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

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

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

Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Günay, S., Öztürk, A., and Yiğit, Y. The accuracy of gemini, gpt-4, and gpt-4o in ecg analysis: A comparison with cardiologists and emergency medicine specialists. *The American journal of emergency medicine*, 84:68–73, 2024.

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

Hu, R., Chen, J., and Zhou, L. Spatiotemporal self-supervised representation learning from multi-lead ecg signals. *Biomedical Signal Processing and Control*, 84: 104772, 2023.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Jiang, C., Ye, W., Xu, H., Zhang, S., Zhang, J., Huang, F., et al. Vision lanauge pre-training by contrastive learning with cross-modal similarity regulation. *arXiv preprint arXiv:2305.04474*, 2023b.

Jin, J. Screening for cardiovascular disease risk with ecg. *Jama*, 319(22):2346–2346, 2018.

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

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

Lalam, S. K., Kunderu, H. K., Ghosh, S., Kumar, H., Awasthi, S., Prasad, A., Lopez-Jimenez, F., Attia, Z. I., Asirvatham, S., Friedman, P., et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023.

Lavoie, S., Kirichenko, P., Ibrahim, M., Assran, M., Wildon, A. G., Courville, A., and Ballas, N. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*, 2024.

Li, J., Liu, C., Cheng, S., Arcucci, R., and Hong, S. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pp. 402–415. PMLR, 2024.

Li, Z., Guo, C., Feng, Z., Hwang, J.-N., and Du, Z. Integrating language guidance into image-text matching for correcting false negatives. *IEEE Transactions on Multimedia*, 2023.

Liu, C., Wan, Z., Cheng, S., Zhang, M., and Arcucci, R. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8230–8234. IEEE, 2024a.

Liu, C., Wan, Z., Ouyang, C., Shah, A., Bai, W., and Arcucci, R. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *arXiv preprint arXiv:2403.06659*, 2024b.

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

Na, Y., Park, M., Tae, Y., and Joo, S. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Rath, A., Mishra, D., Panda, G., and Satapathy, S. C. Heart disease detection using deep learning methods from imbalanced ecg samples. *Biomedical Signal Processing and Control*, 68:102820, 2021.

Sahoo, S., Dash, M., Behera, S., and Sabut, S. Machine learning approach to detect cardiac arrhythmias in ecg signals: A survey. *Irbm*, 41(4):185–194, 2020.

Sun, W., Zhang, J., Wang, J., Liu, Z., Zhong, Y., Feng, T., Guo, Y., Zhang, Y., and Barnes, N. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6420–6429, 2023.

Thai, N., Nghia, N., Binh, D., Hai, N., and Hung, N. Long-tail effect on ecg classification. In *2017 International Conference on System Science and Engineering (ICSSE)*, pp. 34–38. IEEE, 2017.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

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

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Yogarajan, V., Pfahringer, B., Smith, T., and Montiel, J. Improving predictions of tail-end labels using concatenated biomed-transformers for long medical documents. *arXiv preprint arXiv:2112.01718*, 2021.

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

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

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

Zhang, W., Yang, L., Geng, S., and Hong, S. Self-supervised time series representation learning via cross reconstruction transformer. *arXiv preprint arXiv:2205.09928*, 2022b.

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

Zheng, J., Guo, H., and Chu, H. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022Available online: http://physionet. org/content/ecg-arrhythmia/1.0. 0/(accessed on 23 November 2022)*, 2022.

A. Related Work

ECG Self-Supervised Learning Self-supervised learning in ECG analysis has primarily been explored through two paradigms: Contrastive self-supervision (Chen et al., 2020; 2021; Wang et al., 2023; Eldele et al., 2021), which typically involves augmenting the same ECG signal into two different views as positive samples, while different ECG signals serve as negative samples; Generative self-supervision (Zhang et al., 2022b; Hu et al., 2023; Zhang et al., 2022a; Na et al., 2024), which first masks a portion of the ECG signal and then attempts to recover the masked part using the unmasked portion. Unlike previous ECG self-supervised methods that rely on annotations and struggle with unseen classes during fine-tuning, our FG-CLEP enables direct zero-shot prediction on downstream tasks.

ECG-Report Contrastive Learning Recently, inspired by the strong zero-shot ability of image-caption multimodal contrastive learning methods like CLIP (Radford et al., 2021), significant efforts have been made in ECG-Report contrastive learning(Li et al., 2024; Liu et al., 2024b;a; Yu et al., 2024; Lalam et al., 2023). Similar to CLIP (Radford et al., 2021), (Li et al., 2024; Liu et al., 2024a; Lalam et al., 2023) learns ECG representations by pulling ECGs with their paired reports while pushing them from unpaired reports. MERL (Liu et al., 2024b) further introduces uni-modal alignment and employs the CKEPE pipeline at inference to generate more descriptive prompts via LLMs. However, enhancing textual prompts only during inference creates a distribution mismatch between training and testing text. In contrast, ESI (Yu et al., 2024) enhances ECG reports during training using a retrieval-augmented generation (RAG) pipeline, integrating LLMs and external medical knowledge for more detailed descriptions.

Despite these advances, existing methods overlook the absence of fine-grained waveform features in ECG reports. Additionally, due to medical LLM hallucinations (Huang et al., 2023; Günay et al., 2024) and the variability of waveform features for the same disease across patients (Jin, 2018), relying solely on LLMs (Yu et al., 2024) for report augmentation is unreliable. To address these challenges, we propose the FG-CLEP training process.

False Negatives in Contrastive Learning Traditional multi-modal contrastive learning (Radford et al., 2021) assumes that only images and captions from the same record are positive pairs. However, this assumption often fails in the ECG domain, where most ECGs are normal, and abnormalities typically involve common diseases, leading to frequent false negatives. There have been several attempts to address this issue (Lavoie et al., 2024; Jiang et al., 2023b; Sun et al., 2023; Li et al., 2023; Wang et al., 2022). Some approaches (Jiang et al., 2023b; Li et al., 2023) attempt to add a regularization term to mitigate false negatives. Others (Sun et al., 2023; Wang et al., 2022) introduce a matrix to measure the similarity between different reports, guiding contrastive learning to identify and address false negatives. In this paper, we explore the application of the latter approach in the ECG multi-modal contrastive learning domain.

B. Dataset Analysis

We pre-train the FG-CLEP using the MIMIC-ECG dataset and test it on the PTB-XL, CPSC2018, and CSN datasets. All the ECGs in the datasets are 12-lead recordings. The PTB-XL dataset can be further divided into four subsets, and we follow the official train:validation:test split. For CPSC2018 and CSN, we split the dataset as 70%:10%:20% for the train:validation:test split. The statistics of the datasets used are presented in Table 4.

MIMIC-ECG The MIMIC-ECG dataset contains nearly 800,000 ECG-report pairs from approximately 160,000 unique patients. These diagnostic ECGs utilize 12 leads and are 10 seconds in duration, with a sampling rate of 500 Hz. https://physionet.org/content/mimic-iv-ecg/1.0/.

PTB-XL The PTB-XL ECG dataset is a large dataset of 21,799 clinical 12-lead ECGs from 18,869 patients of 10-second length. There are four subsets with multi-label classification tasks: Superclass (5 categories), Subclass (23 categories), Form (19 categories), and Rhythm (12 categories). Notably, these four subsets have different numbers of samples. https://physionet.org/content/ptb-xl/1.0.3/.

CPSC2018 This publicly accessible dataset comprises 6,877 standard 12-lead ECG records, each sampled at a rate of 500 Hz, with durations ranging from 6 to 60 seconds. The dataset is annotated with 9 distinct labels. http://2018.icbeb.org/Challenge.html.

Chapman-Shaoxing-Ningbo (CSN) This dataset contains 12-lead ECGs of 45,152 patients with a 500 Hz sampling rate. It features multiple common rhythms and additional cardiovascular conditions, all labeled by professional experts. https://physionet.org/content/ecg-arrhythmia/1.0.0/.

Table 4. The statistics of used datasets.

Pretrain	**# ECGs**	**# Reports**		
MIMIC-ECG	773,268	773,268		
Evaluation	**# Train**	**# Valid**	**# Test**	**# Classes**
PTB-XL Super	17,084	2,146	2,158	5
PTB-XL Sub	17,084	2,146	2,158	23
PTB-XL Form	7,197	901	880	19
PTB-XL Rhythm	16,832	2,100	2,098	12
CPSC2018	4,800	684	1,383	9
CSN	31,606	4,515	9,031	51

C. Implementation Details

Pre-training Implementation: In the pre-training stage, we utilize a randomly initialized 1D-ResNet50 model (He et al., 2016) as the ECG encoder and BioClinicalBERT (Alsentzer et al., 2019) for text encoding. The AdamW optimizer is selected with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-4}. CLEP is pre-trained for 10 epochs with original reports and FG-CLEP is trained for another 3 epochs with fine-grained reports, using a cosine annealing scheduler for learning rate adjustments and a warmup phase for the first 10% of training steps. A batch size of 100 is maintained. The temperature parameters t and b are initialized to $\log 10$ and -10, respectively. The default hyperparameter λ is set to 0.5 and the default threshold for selecting high-confidence waveform features is set to 0.95. We use LLaMA3-8B (AI@Meta, 2024) as our LLM to query potential waveform features and use vLLM (Kwon et al., 2023) to speed up inference. All experiments used two NVIDIA A800 80GB GPUs, except LLaMA3-70B ablation, which used four.

Downstream Task Implementation: We evaluated the downstream tasks using both zero-shot and linear probe settings. For the zero-shot setting, we froze the entire model and used the text of each category as the prompt. We computed the similarity between the ECG embedding and the category text embedding as the classification probability. Additionally, we employed an ensemble method to enhance zero-shot performance. Specifically, in addition to using the category as text, we also added 'category in lead x' (x represents any of the 12 leads) as text to compute the probability and used the highest probability as the final probability for that category. For linear probing, we kept the ECG encoder frozen and updated only the parameters of a newly initialized linear classifier. We conducted linear probing for each task using 1%, 10%, and 100% of the training data. For all downstream tasks, we used macro AUC as the metric.

D. Pseudo Code

The pseudo-code of our FG-CLEP training process is shown in algorithm 1

Algorithm 1 FG-CLEP Training Process

Input: $D = \{(x_{\text{ecg}_i}, x_{\text{txt}_i}) \mid i \in [0, n)\}$
Output: FG-CLEP
Perform contrastive training on CLEP using D
Generate fine-grained reports
for $i = 0$ to $n - 1$ **do**
  $f_{\text{features}} = \text{LLM}(x_{\text{txt}_i}, \text{prompt})$
  for $j = 1$ to m **do**
    where m is the number of waveform features generated
    $\text{sim} = \text{CLEP}(x_{\text{ecg}_i}, f_j)$
    if sim $>$ threshold **then**
      $x_{\text{txt}_i} = x_{\text{txt}_i} + f_j$
    end if
  end for
end for
Continue training CLEP on $\{(x_{\text{ecg}_i}, x_{\text{txt}_i})\}$ to obtain FG-CLEP

E. Embedding Visualization

We also demonstrate the effectiveness of our representation learning framework by plotting t-SNE visualizations of ECG embeddings produced for PTB-XL ECGs in five classical waveform features. As shown in Fig. 4, our model produces well-clustered representations. Furthermore, as expected, FG-CLEP learns more fine-grained local waveform features of ECGs. Specifically, FG-CLEP clusters 'prolonged PR interval' much better than CLEP.

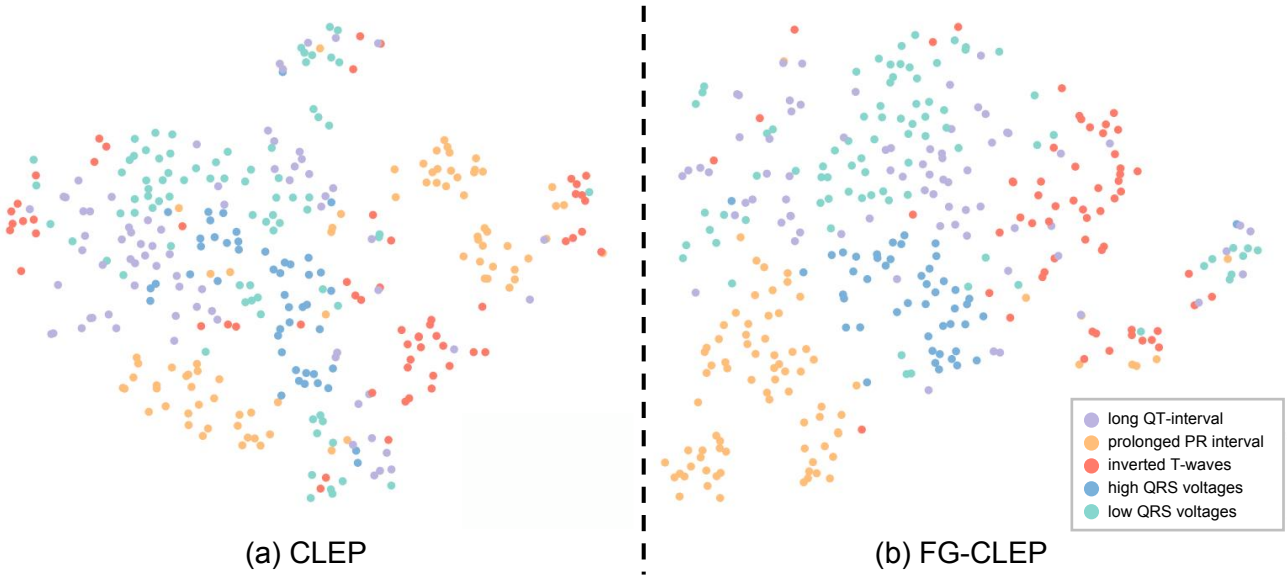

Figure 4. Embeddings visualization of PTB-XL ECGs in 5 waveform features by (a)CLEP and (b)FG-CLEP. Dimension reduced by t-SNE.

F. ECG-Text Retrieval

We attempted to use FG-CLEP to retrieve electrocardiograms (ECGs) from the MIMIC-ECG dataset through text. To test our model's ability to capture fine-grained waveform features, we tested a series of typical waveform features such as 'RSR' Pattern,' 'Inverted T-waves,' and 'Low QRS voltages.' Figure 5 shows the Top 3 retrieved ECGs with probabilities all greater than 0.99. Our model demonstrated strong capability in retrieving ECGs through waveform feature text, which can lead to two applications: (1) Helping doctors quickly find similar cases or specific ECG patterns, aiding in diagnosis and treatment decision-making; (2) In medical education, text-based retrieval can quickly find typical ECG cases, assisting in teaching and training, thereby improving educational effectiveness.

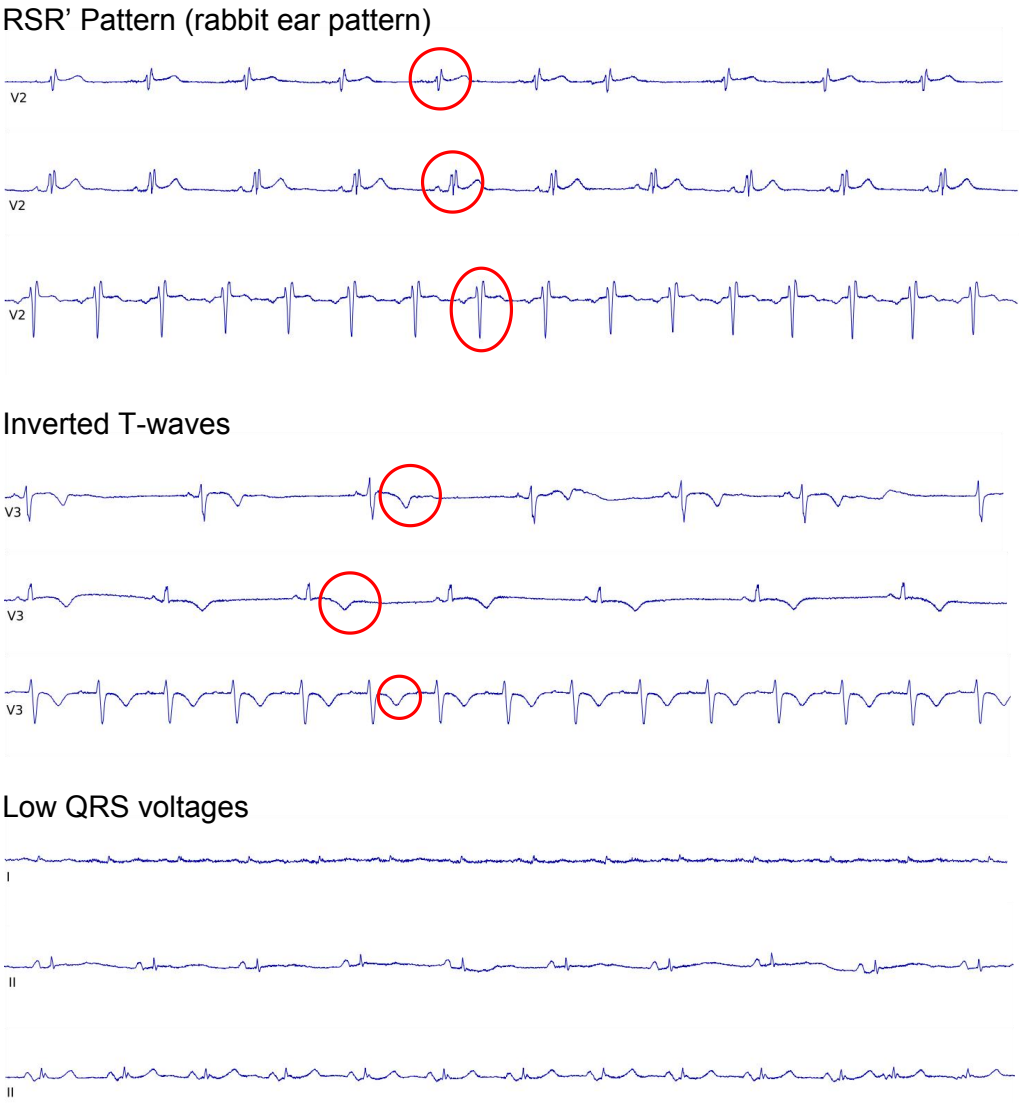

Figure 5. Top 3 retrieved ECG using FG-CLEP.