Towards Generalizable Multimodal ECG Representation Learning with LLM-extracted Clinical Entities

Mingsheng Cai [1 2] **Jiuming Jiang** [1] **Wenhao Huang** [3] **Che Liu** [2] **Rossella Arcucci** [2]

Abstract

Electrocardiogram (ECG) recordings are essential for cardiac diagnostics but require large-scale annotation for supervised learning. In this work, we propose a supervised pre-training framework for multimodal ECG representation learning that leverages Large Language Model (LLM) based clinical entity extraction from ECG reports to build structured cardiac queries. By fusing ECG signals with standardized queries rather than categorical labels, our model enables zero-shot classification of unseen conditions. Experiments on six downstream datasets demonstrate competitive zero-shot AUC of 77.20%, outperforming state-of-the-art self-supervised and multimodal baselines by 4.98%. Our findings suggest that incorporating structured clinical knowledge via LLM-extracted entities leads to more semantically aligned and generalizable ECG representations than typical contrastive or generative objectives.[1]

1. Introduction

Supervised learning (eSL) methods have proven effective in classifying cardiac conditions using Electrocardiogram (ECG), a widely utilized clinical tool for monitoring the heart's electrical activity (Huang et al., 2023; Huang & Yen, 2022). However, eSLs typically rely on large-scale, expert-annotated datasets, which are costly and difficult to scale.

To reduce annotation dependence, recent advances in ECG self-supervised learning (eSSL) use contrastive or generative pretext tasks to learn signal-level features from unlabelled data (Eldele et al., 2021; Kiyasseh et al., 2021; Na et al., 2024). While promising, these methods often rely on handcrafted augmentations that distort physiological semantics and require non-trivial task engineering, thus may lack clinical interpretability and generalizability (Liu et al., 2024; Kiyasseh et al., 2021). ECG-Text Multimodal approaches attempt to incorporate the rich context of free-text ECG reports to improve representation learning. However, the unstructured nature of clinical narratives introduces linguistic noise, inconsistencies, and limited diagnostic coverage (Liu et al., 2024; Li et al., 2024; Wu et al., 2023). ECG introduction and related works are detailed in Appendix A, B.

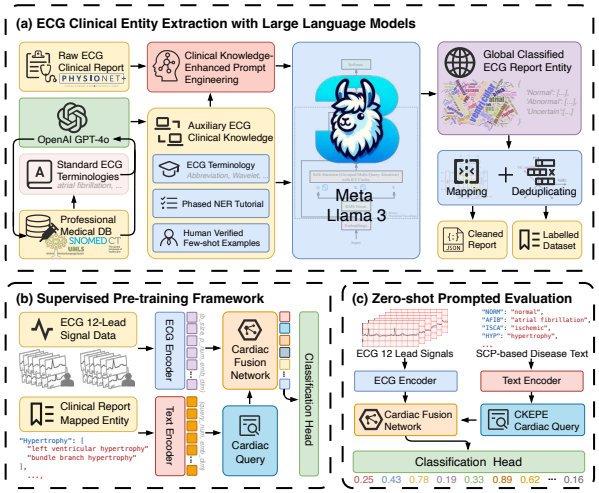


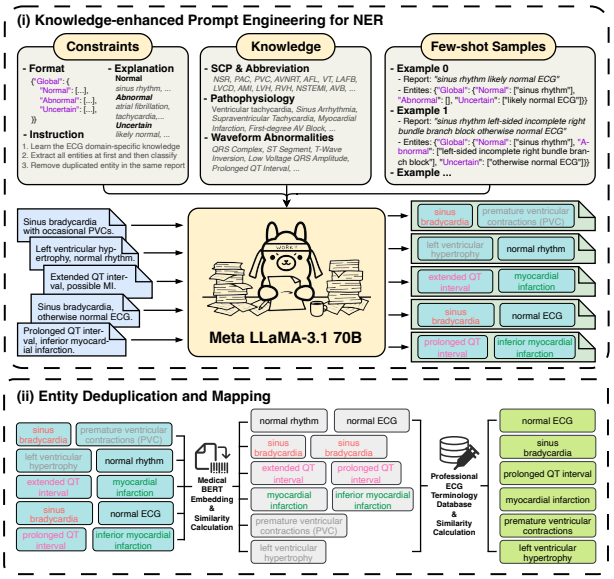
Figure 1. Overall framework including (a) ECG clinical entity extraction with LLMs, (b) supervised ECG-Text multimodal pre-training, and (c) zero-shot prompted evaluation.

We address these limitations by introducing a supervised multimodal pre-training framework that leverages structured clinical entities extracted by LLMs in Figure 1. Instead of relying on noisy free-text or proxy tasks, we align ECG signals with standardized diagnostic concepts, further enabling zero-shot classification of unseen cardiac conditions. Our contributions are threefold: **(a)** A scalable pipeline that extracts structured entities from noisy free-text ECG reports using an instruction-tuned LLM enriched with domain knowledge. Entities are then mapped to a curated set of standardized diagnostic terms using medical databases (e.g., SNOMED CT, UMLS), producing fine-grained, clinically meaningful supervision cardiac query labels without hu-

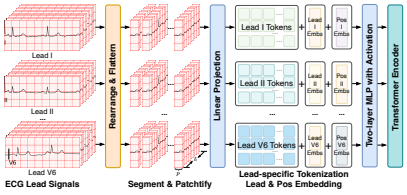
[1]School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom [2]Data Science Institute, Imperial College London, London, United Kingdom [3]Shenzhen Yinwang Intelligent Technology Co., Ltd, Shenzhen, China. Correspondence to: Che Liu <che.liu21@imperial.ac.uk>.

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

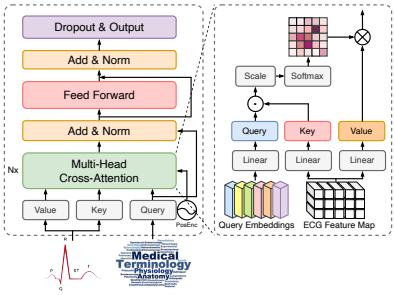
[1]Code, datasets, and model checkpoints are available at https://github.com/mingscai/SuPreME.



(a) ECG report entity extraction with (i) knowledge-enhanced prompt engineering, and (ii) candidate entity deduplication and mapping.



(b) ECG 1D ViT encoder with lead-wise and position embedding.



(c) Architecture of the Cardiac Fusion Network (CFN) in the framework.

Figure 2. Implementation of the clinical NER and supervised ECG-Text multimodal pre-training framework: (a) ECG report entity extraction, (b) ECG 1D ViT encoder, and (c) architecture of the Cardiac Fusion Network.

man annotation; **(b)** A supervised multimodal pre-training framework that fuses ECG signals with clinical queries using a Cardiac Fusion Network (CFN) without the need for signal-level augmentations or handcrafted contrastive losses, thus directly learning noiseless semantics from structured supervision; **(c)** Extensive pre-training on 771,500 ECGs paired with 295 global cardiac queries from MIMIC-IV-ECG (Gow et al., 2023), and evaluation on six downstream datasets (e.g., PTB-XL, CPSC-2018, Chapman-Shaoxing-Ningbo; Appendix C). Our approach achieves a state-of-the-art zero-shot AUC (77.20%), outperforming leading eSSL and multimodal baselines, even those fine-tuned with 10–100% labeled data. Our model also shows strong data efficiency and generalization, with performance under only 20% pre-training data surpassing fully fine-tuned eSSLs.

2. Methodology

Our framework extracts clinical entities using LLMs to form cardiac queries, which are fused with ECG signals via Cardiac Fusion Network (CFN) in a shared latent space for zero-shot diagnosis.

Clinical Entity Extraction with LLMs. We employ an instruction-tuned LLM to extract entities from unstructured ECG reports (Figure 2(a)). To enhance accuracy, we use structured prompts and few-shot examples with medical terminologies from clinician-validated resources (e.g., SNOMED CT, UMLS, SCP-ECG) (Bodenreider, 2004; Donnelly et al., 2006; Rubel et al., 2016). The model identifies diagnostic phrases (e.g., "sinus rhythm") and their certainty (e.g., "abnormal"). Uncertain Entities are discarded. To standardize entities with variations, we compile a cardiac vocabulary from SNOMED CT and UMLS via LLM filtering ECG terminologies. Both extracted and reference terms are encoded using MedCPT, a BERT model initialized from PubMedBERT, and clustered using cosine similarity (Jin et al., 2023; Gu et al., 2021). Entities with high intra-cluster similarity are grouped, and then mapped to standard terms if the average similarity to the reference exceeds threshold set with clinician verification. The resulting deduplicated terms form a global query set for supervised pre-training. Prompts, statistics and case study are included in the Appendix D.

Supervised ECG Multimodal Learning. ECG signals exhibit temporal and structural patterns analogous to the spatial relationships in images. We thus adapt the architecture of Vision Transformer (ViT) by dividing ECG time series into fixed-size patches, as shown in Figure 2(b).

Given ECG signals $\mathbf{x} \in \mathbb{R}^{B\times L\times T}$ with B batches, L leads, and T time steps, we segment each lead independently into $N = T/P$ non-overlapping patches of length P. This yields $\mathbf{x}_{i,j} \in \mathbb{R}^{B\times P}$ for lead i and patch j. Each patch is linearly projected by $\mathbf{W}_p \in \mathbb{R}^{P\times D}$ into a D-dim embedding:

$$\mathbf{z}_{i,j} = \mathbf{x}_{i,j}\mathbf{W}_p \quad \mathbf{z}'_{i,j} = \mathbf{z}_{i,j} + \mathbf{e}_i + \mathbf{p}_j$$
$$\mathbf{Z} = [\mathbf{z}'_{1,1}, \ldots, \mathbf{z}'_{1,N}, \ldots, \mathbf{z}'_{L,N}] \in \mathbb{R}^{B\times(L\cdot N)\times D} \quad (1)$$

Here, $\mathbf{e}_i \in \mathbb{R}^D$ and $\mathbf{p}_j \in \mathbb{R}^D$ are learnable lead and positional embeddings. The enriched tokens $\mathbf{z}'_{i,j}$ are concatenated into a full sequence $\mathbf{Z}$, which is processed by stacked Transformer encoders with residual connections, layer normalization, and stochastic depth dropout. The output is then passed through a modality-specific two-layer MLP with intermediate activation, yielding $\mathbf{F}_{\text{ECG}} \in \mathbb{R}^{B\times(L\cdot N)\times D'}$ for

multimodal fusion in shared D'-dim.

Rather than relying on fixed per-sample labels, we construct a global query list of M standardized diagnostic terms derived from the entity extraction pipeline. Each query q_i is encoded independently using the MedCPT query encoder. All M query embeddings are also passed through a modality-specific MLP to obtain projected embeddings $\mathbf{F}_{\text{Query}} \in \mathbb{R}^{M \times D'}$ in the shared D'-dim.

$$\mathbf{E}[i,:] = \text{QEnc}(q_i) = \text{Trm}([\text{CLS}]\ q_i\ [\text{SEP}]) \in \mathbb{R}^{768} \quad (2)$$

To align ECG features with cardiac queries, we employ a Cardiac Fusion Network (CFN) composed of Transformer decoder layers. Given ECG features $\mathbf{F}_{\text{ECG}} \in \mathbb{R}^{B \times (L \cdot N) \times D'}$ and cardiac queries $\mathbf{F}_{\text{Query}} \in \mathbb{R}^{M \times D'}$, the CFN performs cross-attention between queries (input) and ECG features (memory), producing a fused representation $\mathbf{H} \in \mathbb{R}^{B \times M \times D'}$. This output is passed through a shared classification head to yield M binary logits per ECG sample:

$$\mathbf{Logits} = \text{MLP}_{\text{CFN}}(\mathbf{H}) \in \mathbb{R}^{B \times M} \quad (3)$$

During pre-training, each sample is weakly supervised by a binary vector indicating which global queries match its extracted diagnostic entities. A label of 1 is assigned if a query aligns with a mapped report entity, 0 otherwise. To prevent data leakage, raw reports are excluded from model input, and the global query list is reused across all samples, supporting scalable and decoupled multi-label learning. For further details on the ECG backbone, projection layers, and CFN setup, please refer to Appendix E.

Zero-shot Prompted Classification. To enable generalization to unseen cardiac conditions without fine-tuning, we convert SCP-ECG codes from downstream datasets into concise, clinically meaningful textual prompts (e.g., "left bundle branch block" for LBBB), forming a dataset-specific prompt set aligned with the query space used during pre-training. Following Clinical Knowledge-Enhanced Prompt Engineering (CKEPE) (Liu et al., 2024), we adopt a simplified variant that avoids verbose LLM-generated descriptions (e.g., "a condition characterized by prolonged QRS complex..." for LBBB), instead using compact expressions. This reduces redundancy and allows CFN to focus on cross-modal fusion rather than memorizing textual artifacts.

At inference, ECG signals and cardiac prompts are independently encoded by the ECG encoder and frozen MedCPT text encoder. Let $\mathbf{Emb}_{\text{ECGs}} \in \mathbb{R}^{B \times N \times D}$ and $\mathbf{Emb}_{\text{Queries}} \in \mathbb{R}^{M \times D}$ denote the embeddings. These are fused via CFN to produce logits over cardiac conditions:

$$\mathbf{Pred} = \sigma(\text{CFN}(\mathbf{F}^{\text{eval}}_{\text{ECG}}, \mathbf{F}^{\text{eval}}_{\text{Query}})) \quad (4)$$

We apply sigmoid activation for multi-label classification and report AUROC (AUC) per condition, followed by macro-averaging. Evaluation details are in Appendix F.

3. Experiments

3.1. Configurations

Clinical Entity Extraction. Following Section 2, we extract and normalize clinical entities from MIMIC-IV-ECG using Llama3.1-70B-Instruct with structured prompts to ensure high-quality annotations. Entities are deduplicated via MedCPT embeddings (cosine similarity > 0.8) and mapped to UMLS/SNOMED CT if average similarity exceeds 0.75[2]. Experiments are run on 8 NVIDIA A100-SMX4-80GB GPUs using vLLM (Kwon et al., 2023).

Supervised Pre-training. We use a 1D ViT-tiny encoder (patch size = 125, i.e., 0.25s) and a frozen MedCPT text encoder. Training employs AdamW (LR=1×10^{-3}, weight decay=1×10^{-8}) with cosine annealing (T_0=5000, T_{mult}=1, min LR=1×10^{-8}), for up to 50 epochs with early stopping (patience=10, best AUC at epoch 16). We train with batch size 256 on 4 NVIDIA A100-PCIE-40GB GPUs[3].

Downstream Classification. We evaluate on six unseen datasets (e.g., PTB-XL, CPSC-2018, Chapman; Appendix C) in zero-shot settings. Ablations test the effect of ECG/text backbones and CFN. We also benchmark mainstream eSSLs via linear probing (1%, 10%, 100% label use), freezing the ECG encoder. All tasks are evaluated by average AUC across classes and datasets (splits in Appendix G).

3.2. Evaluation Results

The performance evaluation is carried out against 11 ECG-only or multimodal eSSLs on six downstream ECG datasets covering 106 unique cardiac conditions. Comparisons are made under two settings: (i) zero-shot inference, and (ii) linear probing with varying data proportions.

Table 1(a) shows the performance of our framework and eSSLs. Our framework achieves the highest zero-shot AUC of 77.20%, outperforming most eSSLs even when those are fine-tuned with 100% labeled data, showcasing its strong generalization capabilities. Without the CFN module in linear probing (detailed in Appendix H)), it also outperforms all ECG-only eSSLs and achieves competitive performance with MERL[4] (with explicit contrastive objectives), while requiring significantly fewer training epochs (16 vs. 50).

Figure 3(a) presents dataset-wise comparisons. our framework excels on challenging benchmarks such as PTB-XL-Rhythm, CPSC-2018, and CSN. The gap is narrower on PTB-XL-Superclass, likely due to its limited label granularity (5 broad classes). All methods perform poorly on PTB-

[2]Verified by cardiologists with 10+ years of experience.

[3]Compact checkpoint (ViT-tiny + frozen MedCPT) runs on single GPU with ≥24GB memory, making it deployable in clinical or low-resource settings without pre-training.

[4]Former SOTA model, ViT backbone for fair comparision.

Table 1. Performance evaluation and ablation study of our framework across model components and training configurations. Reported values are zero-shot AUC (%).

(a) Overall performance with 'Z' for zero-shot and 'L' for linear probing. Best results are **bolded** and second best gray -flagged.

Framework	Evaluation Approach	Zero-shot 0%	Linear Probing 1%	10%	100%
From Scratch					
Random Init (CNN)	L	-	55.09	67.37	77.21
Random Init (Transformer)	L	-	53.53	65.54	75.52
ECG Only					
SimCLR (Chen et al., 2020)	L	-	58.24	66.71	72.82
BYOL (Grill et al., 2020)	L	-	55.78	70.61	74.92
BarlowTwins (Zbontar et al., 2021)	L	-	58.92	70.85	75.39
MoCo-v3 (Chen et al., 2021)	L	-	57.92	72.04	75.59
SimSiam (Chen & He, 2021)	L	-	59.46	69.32	75.33
TS-TCC (Eldele et al., 2021)	L	-	54.66	69.37	76.95
CLOCS (Kiyasseh et al., 2021)	L	-	56.67	70.91	75.86
ASTCL (Wang et al., 2023)	L	-	57.53	71.15	75.98
CRT (Zhang et al., 2023a)	L	-	56.62	72.03	76.65
ST-MEM (Na et al., 2024)	L	-	56.42	63.39	69.60
Multimodal Learning					
MERL (Liu et al., 2024)	Z & L	73.54	**63.57**	**78.35**	83.68
Our Framework	Z & L	**77.20**	63.24	72.34	**84.48**

(b) LLM for entity extraction.

LLM Size	Zero-shot AUC
Llama3.1-8B-Instruct	72.89 ± 0.49
Llama3.1-70B-Instruct (Ours)	**77.20 ± 0.21**

(c) Entity deduplication.

Deduplication	Zero-shot AUC
Not Deduplicated	65.94 ± 0.49
Deduplicated (Ours)	**77.20 ± 0.21**

(d) ECG backbone encoders.

Backbone	Zero-shot AUC
ResNet	64.96 ± 0.20
ViT (Ours)	**77.20 ± 0.21**

(e) Language model encoders.

Language Model	Zero-shot AUC
BioClinicalBERT	62.95 ± 0.53
PubMedBERT	62.51 ± 2.21
MedCPT (Ours)	**77.20 ± 0.21**

(f) Cardiac Fusion Network.

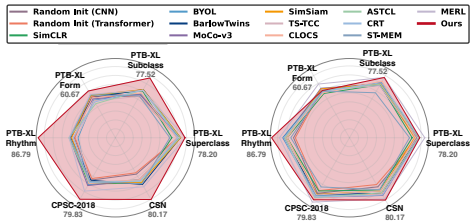
Module	Zero-shot AUC
w/o CFN (Linear)	72.70 ± 0.42
CFN (Ours)	**77.20 ± 0.21**

(g) Zero-shot cardiac query prompts.

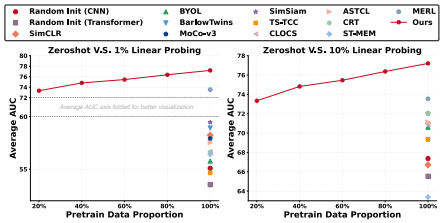
Prompt Strategy	Zero-shot AUC
GPT-4o Generated	60.83 ± 0.26
CKEPE Detailed	69.16 ± 1.94
CKEPE Simplified (Ours)	**77.20 ± 0.21**

(h) Pre-train dropout ratios.

Dropout Ratio	Zero-shot AUC
0.05	75.98 ± 0.56
0.10 (Ours)	**77.20 ± 0.21**
0.15	75.63 ± 0.63



(a) Comparison of our framework (zero-shot) and eSSLs (linear probing with 1% data on the left and 10% data on the right).



(b) Data efficiency of our framework (zero-shot) and eSSLs (MERL in zero-shot, others in linear probing).

Figure 3. Visualization of comparison of our framework with mainstream eSSLs, including both unimodal and multimodal baselines.

XL-Form, which defines waveforms that lack strong clinical specificity, reducing semantic alignment with queries.

We also evaluate the data efficiency of our framework under varying pre-training set sizes (Figure 3(b)). Even with 20% of the pre-train data, it exceeds eSSLs fine-tuned with 10% labeled data; it also achieves performance on par with multimodal baseline MERL, demonstrating strong data efficiency and robustness under limited resource conditions. More evaluations of our framework including unseen-query metrics, different ECG encoders (ResNet vs. ViT), and the effect of CFN are reported in Appendix I, J.

3.3. Ablation Studies

We conduct ablation studies on key components of our framework, summarized in Table 1(b–h). Using Llama3.1-70B-Instruct for entity extraction yields the highest zero-shot AUC (77.20%), outperforming its 8B variant (72.89%), suggesting large LLMs produce higher-quality annotations, while smaller models remain viable lightweight alternatives. Replacing our 295 standardized queries with the duplicated 1,095 terms drops AUC to 65.94%, indicating the effectiveness of semantic deduplication. Substituting ViT with ResNet18 degrades performance by 12.24%, highlighting the benefit of self-attention for temporal modeling. Among text encoders, MedCPT achieves the best AUC, surpassing BioClinicalBERT and PubMedBERT by up to 14.69%, likely due to its contrastive training. Removing the CFN and using a linear head reduces AUC to 72.70%, underscoring the importance of multimodal attention. Simplified CKEPE prompts outperform GPT-generated or verbose versions by at least 8.04%, demonstrating the value of concise, clinically focused queries. Finally, a dropout rate of 0.10 offers the best generalization, with both lower and higher values leading to minor performance declines.

4. Conclusion

We present a novel LLM-based method for ECG clinical entity extraction and introduce a scalable supervised pre-training framework for multimodal ECG representation learning that fuses ECG signals with fine-grained, standardized cardiac queries rather than free-text reports. Its Cardiac Fusion Network (CFN) and Clinical Knowledge-Enhanced Prompt Engineering (CKEPE) eliminate the need for further fine-tuning, enabling robust zero-shot classification with concise cardiac queries. Benchmarked on six downstream datasets, our framework achieves superior zero-shot performance against 11 eSSLs, underscoring both data efficiency and diagnostic precision. Our results highlight the value of explicit entity-level supervision over raw text alignment in ECG-Text multimodal learning, providing a strong basis for clinically oriented ECG representation learning. Discussions and limitations are presented in Appendix K, L.

Acknowledgements

We extend our gratitude to the Edinburgh Compute and Data Facility (ECDF, http://www.ecdf.ed.ac.uk/) and the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232) for providing essential computational resources for this research.

Impact Statement

This work proposes a zero-shot ECG-classification framework intended to lower the barrier to automated ECG analysis in settings where labelled cardiology data are scarce. By converting public, de-identified MIMIC-IV-ECG reports into high-quality supervisory signals, we hope to accelerate open research and ultimately improve access to expert-level screening in under-served hospitals. At the same time, models trained on a single-centre dataset may encode demographic or device-specific biases; misclassification could lead to delayed or inappropriate care if the system is deployed without rigorous external validation and clinician oversight. All code and checkpoints will therefore be released under a research-only licence, and we strongly advise prospective users to conduct site-specific audits, fairness analyses and human-in-the-loop evaluations before any clinical integration. No additional privacy concerns arise, as all data are fully de-identified, and the large-language-model component operates only on those public texts.

References

Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

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.

Degirmenci, M., Ozdemir, M. A., Izci, E., and Akan, A. Arrhythmic heartbeat classification using 2d convolutional neural networks. *Irbm*, 43(5):422–433, 2022.

Donnelly, K. et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.

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., Waks, J. W., Eslami, P., Carbonati, T., Chaudhari, A., Herbst, E., Moukheiber, D., Berkowitz, S., Mark, R., and Horng, S. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0), 2023. URL https://doi.org/10.13026/4nqg-sb35.

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

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

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

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., and Shen, D. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, 2023.

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

Huang, Y., Yen, G. G., and Tseng, V. S. Snippet policy network for multi-class varied-length ecg early classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6349–6361, 2022.

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

Jiang, M., Gu, J., Li, Y., Wei, B., Zhang, J., Wang, Z., and Xia, L. Hadln: hybrid attention-based deep learning network for automated arrhythmia classification. *Frontiers in Physiology*, 12:683025, 2021.

Jin, J., Wang, H., Li, J., Huang, S., Pan, J., and Hong, S. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

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

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

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

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.

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

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

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*, 2024.

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

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

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

Mashrur, F. R., Roy, A. D., and Saha, D. K. Automatic identification of arrhythmia from ecg using alexnet convolutional neural network. In *2019 4th international conference on electrical information and communication technology (EICT)*, pp. 1–5. IEEE, 2019.

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.

Natarajan, A., Chang, Y., Mariani, S., Rahman, A., Boverman, G., Vij, S., and Rubin, J. A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, pp. 1–4. IEEE, 2020.

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

Sawano, S., Kodera, S., Takeuchi, H., Sukeda, I., Katsushika, S., and Komuro, I. Masked autoencoder-based self-supervised learning for electrocardiograms to detect left ventricular systolic dysfunction. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022.

Strodthoff, N., Wagner, P., Schaeffter, T., and Samek, W. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Tesfai, H., Saleh, H., Al-Qutayri, M., Mohammad, M. B., Tekeste, T., Khandoker, A., and Mohammad, B. Lightweight shufflenet based cnn for arrhythmia classification. *IEEE Access*, 10:111842–111854, 2022.

Tian, Y., Li, Z., Jin, Y., Wang, M., Wei, X., Zhao, L., Liu, Y., Liu, J., and Liu, C. Foundation model of ecg diagnosis: Diagnostics and explanations of any form and rhythm on ecg. *Cell Reports Medicine*, 5(12), 2024.

Vaswani, A. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 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):1–15, 2020.

Wan, Z., Liu, C., Zhang, M., Fu, J., Wang, B., Cheng, S., Ma, L., Quilodrán-Casas, C., and Arcucci, R. Medunic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21372–21383, 2023.

Yang, K., Hong, M., Zhang, J., Luo, Y., Su, Y., Zhang, O., Yu, X., Zhou, J., Yang, L., Qian, M., et al. Ecg-lm: Understanding electrocardiogram with large language model. *Health Data Science*.

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

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

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

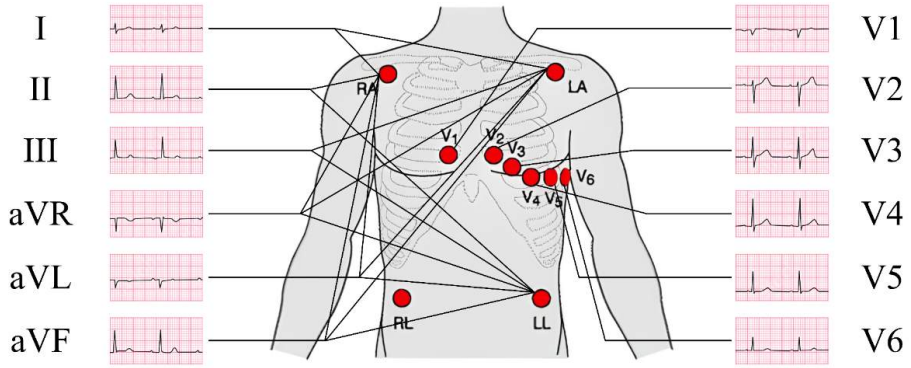
Zhang, X., Wu, C., Zhang, Y., Xie, W., and Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14 (1):4542, 2023b.

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

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

A. Electrocardiogram (ECG)

In the medical field, electrocardiogram (ECG) is an important tool for recording and analyzing patients' cardiac activities, which helps healthcare professionals identify various kinds of cardiac problems by detecting the electrical changes in different leads. The standard 12-lead ECG is the most common method of recording ECGs, and it can capture relatively comprehensive range of cardiac signals through placing electrodes at different locations on the body, providing information of the heart's health conditions.



LEAD	I	II	III	aVL	aVR	aVF
VALUE	LA - RA	LL - RA	LL - LA	LA - (RA + LL) / 2	RA - (LA + LL) / 2	LL - (RA + LA) / 2

Figure 4. Standard 12-lead Electrocardiogram (ECG) showing 'sinus rhythm'.

The basic components of the 12-lead ECG include the limb leads and the precordial leads. The limb leads contain I, II, III, aVR, aVL, and aVF, each of them consists of a combination of electrodes located primarily in the right arm, left arm, left leg, and right leg (as shown in Figure 4). The precordial leads contain V1, V2, V3, V4, V5, and V6, which all correspond to specific single electrodes at different locations on the chest, and are used to observe in detail the electrical activity of the anterior, lateral, and posterior walls of the heart.

B. Related Work

ECG Supervised Learning. ECG supervised learning (eSL) methods, using CNNs or Transformers, achieve high accuracy in cardiovascular disease diagnosis. CNNs excel at capturing spatial and temporal patterns in 1D ECG signals or 2D ECG images (Tesfai et al., 2022; Degirmenci et al., 2022; Mashrur et al., 2019; Huang et al., 2022), while Transformers use attention mechanisms to model global dependencies (Natarajan et al., 2020; Jiang et al., 2021; He et al., 2023). Despite their strengths, eSLs rely heavily on large-scale datasets with expert-verified annotations, making them costly and impractical for pre-training tasks (Strodthoff et al., 2020). This dependence limits their scalability and generalizability, particularly when addressing diverse datasets or unseen cardiac conditions.

ECG Self-supervised Learning. To overcome the annotation bottleneck, ECG self-supervised learning (eSSL) methods have been introduced, enabling representation learning from unannotated ECG signals. Contrastive learning frameworks, such as CLOCS and ASTCL (Kiyasseh et al., 2021; Wang et al., 2023), explore temporal and spatial invariance in ECG data (Eldele et al., 2021; Chen et al., 2020; 2021). Generative eSSL techniques reconstruct masked segments to capture signal-level features (Zhang et al., 2022; Sawano et al., 2022; Na et al., 2024; Jin et al.). Despite their successes, eSSLs fail to incorporate clinical semantics from associated medical reports and require fine-tuning for downstream tasks (Liu et al., 2023d;c; He et al., 2022), limiting their utility in zero-shot scenarios.

ECG Multimodal Learning. Multimodal learning has advanced significantly in biomedical applications, especially in vision-language pre-training (VLP) frameworks for radiology (Liu et al., 2023b;a; Wan et al., 2024; Zhang et al., 2023b; Wu et al., 2023), which align radiology images with structured knowledge from reports to reduce noise and improve robustness. However, ECG-Text multimodal learning holds substantial potential for further development. Methods like MERL (Liu et al., 2024) and ECG-LM (Yang et al.) integrate ECG signals and raw text reports but struggle with noise and inconsistencies in unstructured reports. Others, such as KED (Tian et al., 2024), use structured labels and contrastive learning strategies but face challenges from label noise and LLM-generated knowledge hallucinations. Our approach addresses these issues

by structuring reports into meaningful entities, reducing noise, and aligning them with ECG signals without reliance on LLM-augmented content, minimizing hallucination risks while enabling efficient representation learning and downstream flexibility.

C. Datasets and Models

C.1. Pre-train Dataset

MIMIC-IV-ECG. MIMIC-IV-ECG[5] is a comprehensive database containing 800,035 diagnostic ECG samples from 161,352 unique patients, with 12-lead recordings in 10 second length and sampled at 500 Hz (Gow et al., 2023). These data have been matched with patient records in the MIMIC-IV clinical database, allowing for the association of waveforms with reports when a cardiologist's report is available through provided linking information. To enhance the usability of the data, we exclude empty reports as well as reports containing fewer than 3 words, and replace 'NaN' and 'Inf' values in the ECG records with the average of 6 neighboring points. Ultimately, the dataset used for clinical entity extraction tasks includes 771,500 samples, each comprising 18 machine-generated ECG reports based on rules and the corresponding ECG data. After clinical NER and deduplication on the 18 ECG reports of each sample, the dataset holds 295 labels of professional medical terminologies.

C.2. Downstream Dataset

PTB-XL. PTB-XL[6] is a large open-source ECG dataset, comprising 21,799 clinical ECG records from 18,869 patients, with each lead sampled at a rate of 500 Hz and a duration of 10 seconds (Wagner et al., 2020). A total of 71 different ECG reports are SCP-ECG compliant, covering diagnostic, form and rhythm reports. PTB-XL also provides a recommended train-test split and includes multi-level ECG annotations, covering Superclass (5 categories), Subclass (23 categories), Form (19 categories), and Rhythm (12 categories). Notably the 4 subsets have different sample sizes.

CPSC-2018. The CPSC-2018[7] dataset originates from the China Physiological Signal Challenge (CPSC) 2018, including 6,877 records from 9,458 patients, with durations ranging from 6 to 60 seconds (Liu et al., 2018). The standard 12-lead ECG data is sampled at a rate of 500 Hz, collected from 11 hospitals and categorized into 9 different labels: 1 normal type and 8 abnormal types.

Chapman-Shaoxing-Ningbo (CSN). The CSN[8] 12-lead ECG dataset is created with the support of Chapman University, Shaoxing People's Hospital and Ningbo First Hospital, which includes 12-lead ECGs from 45,152 patients, with a sampling rate of 500 Hz and a duration of 10 seconds (Zheng et al., 2020; 2022). It contains expert annotated features that cover variety of common heart rhythms and other cardiovascular conditions. We exclude ECG records with "unknown" annotations and get 23,026 ECG records with 38 different labels.

C.3. Llama3.1-70B-Instruct Model

Llama3.1-70B-Instruct[9] is a 70-billion parameter large language model released by Meta AI as part of the Llama 3 family. Built on a transformer decoder architecture, it is optimized for instruction following and few-shot generalization through extensive supervised fine-tuning and reinforcement learning from human feedback (RLHF). Compared to its predecessors, Llama3.1-70B-Instruct demonstrates substantial improvements in reasoning, factuality, and alignment with user intent across a wide range of NLP tasks.

In our framework, we leverage Llama3.1-70B-Instruct to extract fine-grained diagnostic entities from free-text ECG reports in the MIMIC-IV-ECG dataset. The scale and instruction-tuning of this model make it well suited for domain-specific named entity recognition (NER) in noisy clinical narratives. Our objective is to construct a high-quality, large-scale set of cardiac entities and their mapped terminologies, enabling robust supervision for ECG-text multimodal learning and promoting reproducibility in future research.

Although smaller models can provide acceptable results (see Section 3.3), we adopt Llama3.1-70B-Instruct to maximize

[5]MIMIC-IV-ECG is available at https://physionet.org/content/mimic-iv-ecg/1.0/.
[6]PTB-XL is available at https://physionet.org/content/ptb-xl/1.0.3/.
[7]CPSC-2018 is available at http://2018.icbeb.org/Challenge.html.
[8]Chapman-Shaoxing-Ningbo is available at https://physionet.org/content/ecg-arrhythmia/1.0.0/.
[9]Llama3.1-70B-Instruct is available at https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

annotation quality, particularly for downstream applications in clinical and low-resource settings that rely on precise structured supervision.

D. LLM-based Clinical Entity Extraction

D.1. Prompt for Medical Database Terminology Filtering

```
system_message = """\
You are a clinical NLP assistant specializing in identifying ECG related terminologies from medical databases.

Your primary task is to serve as a strict terminology filter that judges whether the provided terminology is related to ECG or not, and output your judgement in a **strictly formatted JSON object** that conforms exactly to the following schema:

{
  "IS_ECG_TERM": true/false,
}

**Strict constraints**:
- Return **only** the JSON object. Do not include any natural language explanation or commentary.
- Do not hallucinate or invent fields not specified above.

Your output will be used in real-life clinical settings. Any deviation from this format may cause serious issues in downstream applications. Be precise and compliant.
"""

def get_prompt(row):
    return f"""\
Please read and give your judgement on the following terminology.

Terminology:
\"{row["ENG_TERM"]}\"
"""
```

D.2. Prompt for Report Entity Extraction

```
system_message = """\
You are a clinical NLP assistant specializing in information extraction from medical ECG (electrocardiogram) reports. Your role is to serve as a strict, schema-aware entity extractor that produces structured annotations for downstream machine learning and clinical data analysis tasks.

Please learn the knowledge including common ECG terminologies and abbreviations first:

**Common ECG terminologies**:
Normal: "normal sinus rhythm", "normal ecg", "sinus rhythm", "within normal limits", "no abnormalities detected", ...
Abnormal: "atrial fibrillation", "ventricular tachycardia", "left ventricular hypertrophy", "right bundle branch block", "ST elevation, "T wave inversion", "prolonged QT interval", "first degree AV block", "pacemaker rhythm", ...
Uncertain: "possible infarction", "borderline ecg", "nonspecific ST-T changes", "probable left ventricular hypertrophy", "cannot rule out ischemia", ...

**Demo Abbreviations**:
NSR: "Normal Sinus Rhythm",
AFIB: "Atrial Fibrillation",
AFL: "Atrial Flutter",
VT": "Ventricular Tachycardia",
```

```
PVC: "Premature Ventricular Contraction",
PAC: "Premature Atrial Contraction",
LVH: "Left Ventricular Hypertrophy",
RVH: "Right Ventricular Hypertrophy",
RBBB: "Right Bundle Branch Block",
LBBB: "Left Bundle Branch Block",
AVB1: "First Degree AV Block",
AVB2: "Second Degree AV Block",
AVB3: "Third Degree AV Block",
STEMI: "ST-Elevation Myocardial Infarction",
NSTEMI: "Non-ST-Elevation Myocardial Infarction",
TW": "T Wave Inversion",
QTc: "Corrected QT Interval",
BBB: "Bundle Branch Block",
LAD: "Left Axis Deviation",
RAD: "Right Axis Deviation",
SA: "Sinoatrial",
PVCs: "Premature Ventricular Contractions",
PACs: "Premature Atrial Contractions"

Your primary task is to identify all relevant entities in an ECG report and then classify based on diagnosis certainty, afterwards output them in a **strictly formatted JSON object** that conforms exactly to the following schema:

```json
{
    "global": [...],    # All ECG entities from the provided report
    "classification": {
        "normal": [...],     # Entities confidently labeled as clinically "normal" (e.g., "normal ECG", "sinus rhythm")
        "abnormal": [...],   # Entities labeled as clinically "abnormal" (e.g., "atrial fibrillation", "ST elevation")
        "uncertain": [...]   # Entities with uncertainty or ambiguity in the report context (e.g., "possible LVH", "undetermined".)
    }
}
```

Strict constraints:

- Return **only** the JSON object. Do not include any natural language explanation or commentary.
- Do not hallucinate or invent fields not specified above.
- Do not extract adjectives or modifiers (e.g., "nonspecific", "mild", "marked", "possibly", "likely") as standalone entities. If a descriptive modifier qualifies an entity (e.g., "nonspecific ST-T changes", "likely normal ecg"), include it in the full entity string.
- Do not extract entire sentences or diagnostic phrases as a single entity. If a sentence contains multiple medical concepts, extract each as a separate entity.
- If an entity contains conjunctions (e.g., "and", "or", "and/or"), causal phrases (e.g., "due to", "with"), or multiple anatomical locations (e.g., "inferior/lateral"), you must split it into separate entities.
- If there are entities with clinically same meanings in the given report, only retain one with better expression.

Some examples:

- [Modifier + Entity]:
  Input: "lateral st-t changes are probably due to ventricular hypertrophy"
  Output: {"global": ["lateral st-t changes", "ventricular hypertrophy"], "classification": {"normal": [], "abnormal": ["lateral st-t changes", "ventricular hypertrophy"], "uncertain": []}}

- [Entity A with/and/or/'/' Entity B]:
```

```
  Input: "sinus rhythm with pacs. hypertrophy and/or ischemia. inferior/lateral st-t changes."
  Output: {"global": ["sinus rhythm", "pacs", "hypertrophy", "ischemia", "inferior st-t changes", "lateral st-t changes"], "classification": {"normal": ["sinus rhythm"], "abnormal": ["pacs", "hypertrophy", "ischemia", "inferior st-t changes", "lateral st-t changes"], "uncertain": []}}

- [Entity + Further Description]:
  Input: "inferior infarct - age undetermined. pacemaker rhythm - no further analysis. poor r wave progression - probable normal variant."
  Output: {"global": ["inferior infarct", "age undetermined", "pacemaker rhythm", "poor r wave progression", "probable normal variant"], "classification": {"normal": [], "abnormal": ["inferior infarct", "pacemaker rhythm", "poor r wave progression"], "uncertain": ["age undetermined", "probable normal variant"]}}  # "no further analysis" is not a medical entity

Your output will be used in real-life clinical settings. Any deviation from this format may cause serious issues in downstream applications. Be precise and compliant.
"""


def get_prompt(row):
    return f"""\
Please extract all relevant clinical entities from the following ECG report.

Return the output strictly in the JSON format described in the system prompt.
Do not include any explanation or additional text.

ECG report text:
\"{row["total_report"]}\"
"""
```

D.3. Statistics of Extracted MIMIC-IV-ECG Entities

We extract over 3.4 million clinical entities from free-text ECG reports in the MIMIC-IV-ECG dataset using an instruction-tuned LLM. At the term level, this results in 1,168 unique raw entities (Table 2). Among these, 93.75% remain after filtering out uncertain or ambiguous expressions. To resolve redundancy and lexical variation, we apply embedding-based clustering using MedCPT representations, reducing the vocabulary to 341 cluster representatives. Further manual verification and mapping to UMLS/SNOMED CT terminologies yield a final set of 295 standardized cardiac entities used as global queries during supervised pre-training.

Table 2. Statistics of unique cardiac entities: Extraction, Filtering, Deduplication, and Mapping.

Entity Type	Count	Proportion
Raw extracted entities	3,419,064	100% (sample-level)
Unique raw extracted entities	1,168	100% (term-level)
Terms after uncertainty filtering	1,095	93.75% (vs. 1,168)
Entity cluster representatives (post-deduplication)	341	29.20% (vs. 1,168)
Final unique standardized entities (post-mapping)	295	25.26% (vs. 1,168)

Table 3 provides additional statistics on the clustering process. The average cluster contains 3.39 entities, with some clusters merging up to 29 semantically similar terms. In total, 86.51% of clusters are successfully mapped to standardized terms. The distribution of standardized entity frequencies is illustrated in Figure 5. The left panel shows a log-scaled histogram of the most common cardiac terms, with ”normal”, ”abnormal”, and ”myocardial infarction” being the most frequent. The right panel presents a word cloud that qualitatively reflects term prevalence and semantic variety. Together, these visualizations confirm that while a few diagnostic terms dominate the corpus, a long tail of clinically significant but less frequent entities is preserved, supporting robust coverage in downstream classification.

Table 3. Clustering statistics of extracted cardiac entities on MedCPT embeddings.

Clustering Metric	Value
Number of clusters formed	341
Average number of entities per cluster	3.39
Maximum / Minimum cluster size	29 / 1
Proportion of clusters mapped to standard terms	86.51%

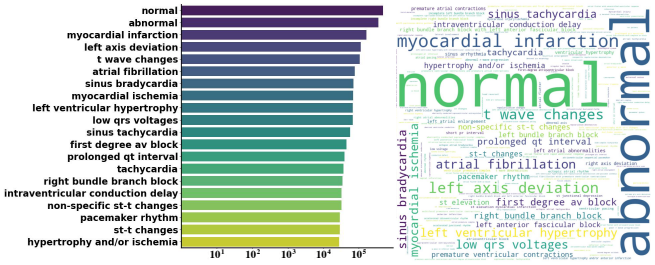


Figure 5. Frequency distribution of standardized ECG entities after deduplication and mapping in MIMIC-IV-ECG.

D.4. Case Study of Deduplication and Mapping

To address concerns about how descriptive cardiac queries are constructed and how they reduce noise compared to raw NER outputs, we present a representative case study from the MIMIC-IV-ECG dataset.

Original Clinical Report:

> "Sinus rhythm w/ PACs, QTc prolonged, Left axis deviation, RBBB with left anterior fascicular block, Inferior/lateral T changes may be due to myocardial ischemia, Low QRS voltages in precordial leads."

Extracted Raw Entities (via LLM-based NER):

```
 "sinus rhythm", "PACs", "QTc prolonged", "Left axis deviation",
"RBBB", "Left anterior fascicular block", "Inferior/lateral T changes",
"Myocardial ischemia", "Low QRS voltages in precordial leads"
```

Mapped and Standardized Queries (after Deduplication and Mapping):

Table 4. Example mapping from raw NER entities to standardized cardiac query labels.

SCP Code	Standardized Query	Matched Raw Entities (Cosine Similarity)
SR	sinus rhythm	sinus rhythm (1.0000)
PAC	premature atrial complex	PACs (0.8976)
LNGQT	prolonged QT interval	QTc prolonged (0.9434)
ALS	axis left shift	Left axis deviation (0.8723)
RBBB	right bundle branch block	RBBB (1.0000)
LAFB	left anterior fascicular block	Left anterior fascicular block (1.0000)
NT	non-specific T wave changes	Inferior/lateral T changes (0.7751)
MI	myocardial infarction	Myocardial ischemia (0.9231)
LVOLT	low QRS voltages	Low QRS voltages in precordial leads (0.8919)

This example illustrates how the same clinical concept may be expressed in different lexical forms (e.g., "PACs" vs. "premature atrial complex") or contain verbose phrasing (e.g., "Low QRS voltages in precordial leads"), leading to noisy or redundant supervision if used directly. By clustering and mapping using MedCPT embeddings and similarity thresholds (Table 4), these expressions are unified under concise, standardized queries aligned with SCP codes.

Table 5. Example Clusters of Raw NER Entities Mapped to Standardized Cardiac Queries.

SCP Code	Standard Cardiac Query	Mapped Raw NER Entities (Cosine Similarity)
NORM	normal	Normal (1.000), of normal (0.995), Normal result (0.979), Normal interest (0.953), percent of normal (0.942)
IMI	inferior myocardial infarction	Inferior myocardial ischemia (0.956), Inferior MI on ECG (0.935), ECG shows inferior MI (0.928), Myocardial infarction (0.923), Old inferior MI (0.907)
LVH	left ventricle hypertrophy	Left ventricular hypertrophy (0.992), Severe LVH (0.967), Hypertensive LVH (0.948), Acquired LVH (0.940), Congenital LVH (0.904)

In Table 5 we show parts of the clustering and deduplication results on the pre-train dataset MIMIC-IV-ECG. This process prevents redundant terms from introducing duplicate supervision, normalizes entities with modifiers (e.g., "in precordial leads"), and enforces semantic consistency across ECG samples. These standardized queries form the global label set used for training, enabling clean multimodal supervision and robust generalization in zero-shot settings.

E. Pre-training Framework Implementation

E.1. Transformer Block Structure.

The Transformer architecture (Vaswani, 2017) is widely used for seq2seq modeling, learning global dependencies via self-attention instead of recurrent or convolutional structures. It consists of an encoder-decoder design, where both the encoder and decoder utilize stacked self-attention and feed-forward layers, as shown in Figure 6.

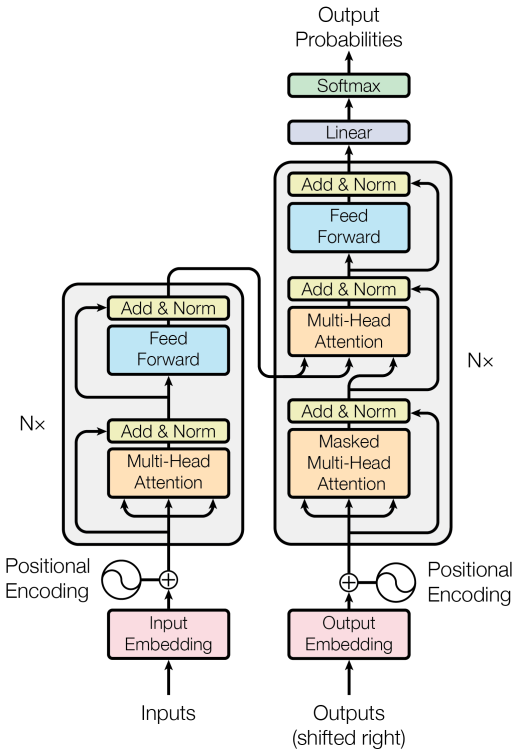


Figure 6. Encoder-decoder structure of Transformer, quoted from (Vaswani, 2017).

Each encoder block applies a residual connection around its multi-head self-attention (**MHA**) and position-wise feed-forward (**FF**) sublayers, followed by layer normalization:

$$\begin{aligned}\mathbf{Z}^{(k,1)} &= \mathbf{Z}^{(k-1)} + \text{Drop}(\mathbf{MHA}(\text{Norm}(\mathbf{Z}^{(k-1)}))) \\ \mathbf{Z}^{(k,2)} &= \mathbf{Z}^{(k,1)} + \text{Drop}(\mathbf{FF}(\text{Norm}(\mathbf{Z}^{(k,1)}))) \\ \mathbf{Z}^{\text{norm}} &= \text{Norm}(\mathbf{Z}^{(\text{final})})\end{aligned} \tag{5}$$

where $\mathbf{Z}^{(k-1)}$ is the input to the k-th Transformer block, $\mathbf{Z}^{(k,1)}$ represents the intermediate state after multi-head attention and residual connection, and $\mathbf{Z}^{(k,2)}$ is the output after the feed-forward network. The final normalized representation $\mathbf{Z}^{\text{norm}}$ is used for downstream ECG classification.

The decoder extends the encoder structure by introducing an additional multi-head attention sublayer that attends to encoder outputs, while also incorporating masked self-attention to ensure autoregressive sequence modeling. These layers collectively enable flexible cardiac feature extraction in our framework.

E.2. Projection of ECG Embeddings

Following the Transformer encoder stack in the Vision Transformer (ViT) backbone, the resulting ECG token sequence $\mathbf{Z} \in \mathbb{R}^{B\times(L\cdot N)\times D}$ is passed through a modality-specific projection head to align its dimensionality with the shared multimodal latent space used in fusion.

The projection head is implemented as a two-layer multilayer perceptron (MLP_{ECG}), consisting of:

- A linear transformation from the ViT output width D to an intermediate hidden size D_h;
- A non-linear activation function (ReLU);
- A linear transformation from D_h to the final projected dimension D', shared with the text modality.

Formally, the projection can be written as:

$$\begin{aligned}\mathbf{Emb}_{\text{ECG}}^{\text{hidden}} &= \mathbf{Z}_{\text{dropout}}\mathbf{W}_1 + \mathbf{b}_1 \\ \mathbf{Emb}'^{\text{hidden}}_{\text{ECG}} &= \text{ReLU}(\mathbf{Emb}_{\text{ECG}}^{\text{hidden}}) \\ \mathbf{F}_{\text{ECG}} &= \mathbf{Emb}'^{\text{hidden}}_{\text{ECG}}\mathbf{W}_2 + \mathbf{b}_2\end{aligned} \tag{6}$$

where ReLU is the activation function, $\mathbf{b}_1$ and $\mathbf{b}_2$ bias terms, and $\mathbf{W}_1 \in \mathbb{R}^{D\times D_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{D_h\times D'}$ are learnable parameters.

This projection layer serves to improve non-linear representational capacity before multimodal alignment, and to map ViT-specific features to a dimensionally consistent space with text query embeddings, enabling efficient cross-modal attention in the Cardiac Fusion Network (CFN).

E.3. Projection of Text Query Embeddings

To align cardiac query embeddings with ECG features in the multimodal latent space, we apply a modality-specific projection head to the output of the MedCPT query encoder (QEnc). Given M queries encoded into a matrix $\mathbf{E} \in \mathbb{R}^{M\times 768}$, the projection head transforms each 768-dimensional embedding into a D'-dimensional representation compatible with ECG tokens.

The projection is implemented as a two-layer multilayer perceptron ($\text{MLP}_{\text{Query}}$) as well, consisting of:

- A linear transformation from 768 to a hidden dimension D_h;
- A non-linear activation function (GELU);
- A linear transformation from D_h to the target fusion dimension D'.

Formally, the operation is defined as:

$$\begin{aligned}\mathbf{Emb}_{\text{Query}}^{\text{hidden}} &= \mathbf{E}_{\text{dropout}}\mathbf{W}_3 + \mathbf{b}_3 \\ \mathbf{Emb}'^{\text{hidden}}_{\text{Query}} &= \text{GeLU}(\mathbf{Emb}_{\text{Query}}^{\text{hidden}}) \\ \mathbf{F}_{\text{Query}} &= \mathbf{Emb}'^{\text{hidden}}_{\text{Query}}\mathbf{W}_4 + \mathbf{b}_4\end{aligned} \tag{7}$$

where GeLU is the activation function, $\mathbf{b}_3$ and $\mathbf{b}_4$ bias terms, and $\mathbf{W}_3 \in \mathbb{R}^{768\times D_h}$ and $\mathbf{W}_4 \in \mathbb{R}^{D_h\times D'}$ are learnable parameters.

This projection head enables cross-modal alignment by transforming domain-specific textual semantics into a shared feature space used by the Cardiac Fusion Network (CFN). The structure mirrors the ECG-side projection to maintain architectural symmetry and training stability.

E.4. Initialization of Cardiac Fusion Network

All weights in linear layers and attention modules are initialized with a normal distribution, $\mathbf{W} \sim \mathcal{N}(0, 0.02)$. To support batch processing, the text embeddings $\mathbf{F}_{\text{text}}$ are expanded to match the batch size B. Both ECG and text embeddings undergo layer normalization to improve training stability and convergence.

F. Zero-shot Evaluation Analysis

F.1. Classification Mechanism

During zero-shot evaluation, the class set (i.e., diagnostic query set $\mathcal{Q}$) is dynamically specified per downstream dataset but remains fixed for all samples within that dataset. The model computes one score per query in $\mathcal{Q}$ for a given ECG sample. These scores are produced via a sigmoid-activated MLP head following the Cardiac Fusion Network (CFN) output, where each query representation attends over the ECG feature sequence. Importantly, this design supports variable-sized query sets across datasets, and prediction is always performed over the currently defined $\mathcal{Q}$. The classifier weights are not pre-defined or fixed, but learned representations aligned to query embeddings through cross-modal attention, ensuring full flexibility across unseen classes.

F.2. Simplified Clinical Knowledge-Enhanced Prompt Engineering

In our implementation of simplified CKEPE query construction, we follow the general design principle introduced in MERL (Liu et al., 2024). The original CKEPE pipeline in MERL employs GPT-4 with web browsing to retrieve attributes and subtypes of each cardiac condition from clinical knowledge sources such as SNOMED CT and SCP-ECG. The prompt typically used is:

```
"Which attributes and subtypes does <cardiac condition> have?"
```

The responses are then validated against the external databases to avoid hallucination and finally organized into detailed clinical descriptions used as prompts for downstream evaluation (see MERL Section 3.4 and Figure 3).

In contrast, we adopt a simplified version of this process (Section 2) aimed at reducing verbosity while preserving clinical specificity. Specifically, we use GPT-4o with the following style of prompt:

```
"Provide the standard clinical definition of <SCP diagnostic code> based
on the SCP-ECG protocol."
```

The generated responses are then automatically validated by external databases as well to reduce hallucinated content. Rather than expanding into all potential attributes or phenotypes (as done in MERL), we retain only the concise, high-precision diagnostic phrase for each class, enabling cleaner alignment with the downstream label space.

Take a simple case study as example, for the diagnostic class LBBB (Left Bundle Branch Block), MERL would produce a long-form prompt such as:

> "A conduction abnormality characterized by delayed depolarization of the left ventricle, typically resulting in a widened QRS complex (>120 ms), often associated with underlying structural heart disease or ischemia."

In contrast, our simplified prompt (after GPT-4o generation and medical verification) becomes:

> "left bundle branch block"

This compact form reduces potential noise in query encoding while retaining diagnostic specificity. It aligns with our hypothesis that multimodal fusion benefits more from semantically discriminative labels than verbose natural language definitions.

F.3. Dataset Overlap Analysis

We analyze the cardiac query overlap between the pre-train dataset and six downstream datasets specified in Section 3.1, as well as among the downstream datasets themselves, as illustrated in Figure 7. Specifically, we embed all entities from the pre-train dataset and cardiac queries from the downstream datasets, compute their cosine similarity, and apply a threshold of 0.95 verified by cardiologists with 10+ years of experience as well to filter overlapping queries.

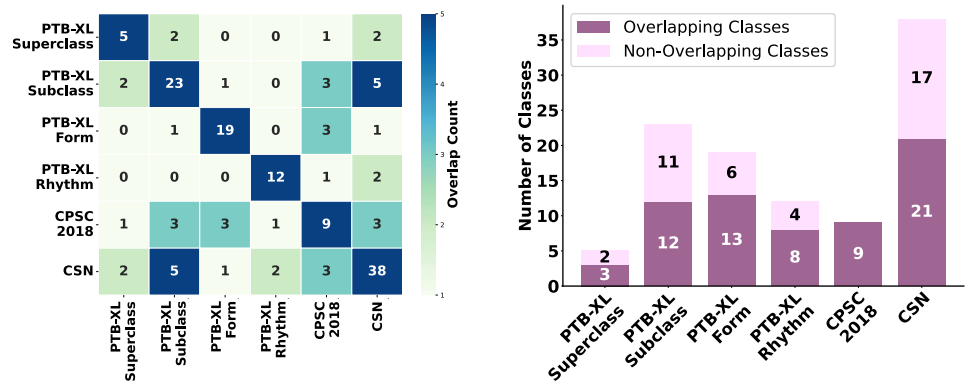


Figure 7. Overlap between ECG datasets, with left panel showing pairwise overlap counts between downstream datasets, and right panel showing the distribution of overlapping and non-overlapping classes between each downstream dataset and the pre-training dataset.

The heatmap on the left shows that pairwise overlaps among downstream datasets are generally limited, reflecting the diversity of cardiac query prompts. The bar chart on the right reveals that 57 cardiac queries overlap between the pre-train dataset and the downstream datasets. Despite the pre-train dataset shares some similar queries, a substantial portion of queries remains unique to the downstream datasets, allowing the pre-train process to establish robust general-purpose representations while leaving room for downstream-specific adaptation.

Table 6 shows the overlap between entities from the pre-train dataset and cardiac queries from the downstream datasets, filtered using a cosine similarity threshold of 0.95.

F.4. Evaluation Metrics

We use zero-shot learning and linear probing to evaluate the performance of our framework and mainstream eSSL frameworks. The primary metric is Area Under the Receiver Operating Characteristic (AUROC, also referred to as AUC). AUROC is widely used to evaluate the performance of binary classification models. The ROC curve plots the True Positive Rate (TPR) on the vertical axis against the False Positive Rate (FPR) on the horizontal axis. By varying the classifier's threshold, $\mathbf{TPR}$ and $\mathbf{FPR}$ are calculated and then plotted to form the curve, where $\mathbf{TP}$ refers to True Positive, $\mathbf{FN}$ refers to False Negative, $\mathbf{FP}$ refers to False Positive, and $\mathbf{TN}$ refers to True Negative.:

$$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$$
$$\mathbf{FPR} = \frac{\mathbf{FP}}{\mathbf{FP} + \mathbf{TN}} \tag{8}$$

Table 6. Overlap between pre-train dataset queries and downstream dataset queries (similarity ≥ 0.95), shown side-by-side.

Pre-train Entity	Downstream Query	Sim.	Pre-train Entity	Downstream Query	Sim.
atrial fibrillation	atrial fibrillation	1.0000	atrial flutter	atrial flutter	1.0000
supraventricular tachycardia	supraventricular tachycardia	1.0000	Sinus Tachycardia	sinus tachycardia	1.0000
ventricular preexcitation	ventricular preexcitation	1.0000	Sinus Bradycardia	sinus bradycardia	1.0000
right bundle branch block	right bundle branch block	1.0000	first degree AV block	first degree av block	1.0000
myocardial infarction	myocardial infarction	1.0000	premature complex	premature complex	1.0000
atrial premature complex	atrial premature complex	1.0000	ST-T change	st-t changes	0.9968
Prolonged QT interval	prolonged qt interval	1.0000	premature atrial complex	atrial premature complex	0.9961
T wave abnormalities	t wave abnormalities	1.0000	left ventricle hypertrophy	left ventricular hypertrophy	0.9924
ST depression	st depression	1.0000	right ventricle hypertrophy	right ventricular hypertrophy	0.9920
AV block	av block	1.0000	Q wave present	q wave	0.9903
T wave Changes	t wave changes	1.0000	complete right bundle branch block	right bundle branch block	0.9891
sinus bradycardia	sinus bradycardia	1.0000	high QRS voltage	high qrs voltages	0.9878
left anterior fascicular block	left anterior fascicular block	1.0000	complete left bundle branch block	left bundle branch block	0.9861
sinus arrhythmia	sinus arrhythmia	1.0000	second degree AV block(Type one)	second degree av block	0.9817
left bundle branch block	left bundle branch block	1.0000	anteroseptal myocardial infarction	anteroseptal infarction	0.9809
sinus tachycardia	sinus tachycardia	1.0000	ischemic	ischemia	0.9804
abnormal Q wave	abnormal q wave	1.0000	second degree AV block(Type two)	second degree av block	0.9795
ventricular premature complex	ventricular premature complex	1.0000	third degree av block	second degree av block	0.9795
Prolonged PR interval	prolonged pr interval	1.0000	low amplitude T wave	high t wave amplitude	0.9741
Atrial Tachycardia	atrial tachycardia	1.0000	abnormal QRS	abnormal qrs morphology	0.9737
Supraventricular Tachycardia	supraventricular tachycardia	1.0000	suggests digitalis-effect	digitalis effect	0.9726
left posterior fascicular block	left posterior fascicular block	1.0000	supraventricular arrhythmia	supraventricular tachycardia	0.9684
normal	normal	1.0000	anterolateral myocardial infarction	anterolateral infarction	0.9667
second degree AV block	second degree av block	1.0000	paroxysmal supraventricular tachycardia	supraventricular tachycardia	0.9611
anterior myocardial infarction	anterior myocardial infarction	1.0000	left front bundle branch block	left bundle branch block	0.9537
incomplete left bundle branch block	incomplete left bundle branch block	1.0000	inferior myocardial infarction	inferior infarction	0.9512
incomplete right bundle branch block	incomplete right bundle branch block	1.0000	right atrial hypertrophy	right atrial enlargement	0.9570
ST elevation	st elevation	1.0000			

AUROC is the area under the ROC curve, with values ranging from 0 to 1, reflecting the overall classification ability of the model. **AUROC = 0.5** indicates that the model's classification ability is equivalent to random guessing, while **AUROC > 0.5** and values closer to **1** indicate that the model is able to classify with greater accuracy.

G. Downstream Task Configuration

G.1. Data Split

For PTB-XL, we adopt the official train-test split recommended by the dataset authors (Wagner et al., 2020), ensuring consistency with prior works and a balanced distribution of ECG categories. This split is directly applied across the Superclass, Subclass, Form, and Rhythm subsets of PTB-XL. For CPSC-2018 and CSN, we follow the data splitting approach used by (Liu et al., 2024), which randomly divides the datasets into training, validation, and testing subsets in a 70%:10%:20% ratio.

Details of the splits, including the specific number of samples allocated to each subset, are summarized in Table 7.

Table 7. Data splits and sample distribution for downstream datasets.

Dataset	Category Number	Train Set	Validation Set	Test Set
PTB-XL				
Superclass	5	17,084	2,146	2,158
Subclass	23	17,084	2,146	2,158
Form	19	7,197	901	880
Rhythm	12	16,832	2,100	2,098
Others				
CPSC-2018	9	4,950	551	1,376
CSN	38	16,546	1,860	4,620

Table 8. Downstream dataset information and split proportions.

Hyperparameter	PTB-XL-Superclass	PTB-XL-Subclass	PTB-XL-Form	PTB-XL-Rhythm	CPSC-2018	CSN
Optimizer						
Type	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Learning Rate	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
Weight Decay	1e-8	1e-8	1e-8	1e-8	1e-8	1e-8
Scheduler						
Type	Cosine Anealing	Cosine Anealing	Cosine Anealing	Cosine Anealing	Cosine Anealing	Cosine Anealing
Warmup Steps	5	5	5	5	5	5
General						
Batch Size	16	16	16	16	16	16
Epochs	100	100	100	100	100	100

Table 9. Specific linear probing performance of our framework and eSSLs across six downstream datasets. Best results are **bolded** and second best gray -flagged.

	PTB-XL-Superclass			PTB-XL-Subclass			PTB-XL-Form			PTB-XL-Rhythm			CPSC-2018			CSN		
Framework	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
From Scratch																		
Random Init (CNN)	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
Random Init (Transformer)	70.31	75.27	77.54	53.56	67.56	77.43	53.47	61.84	72.08	45.36	60.33	77.26	52.93	68.00	77.44	45.55	60.23	71.37
ECG Only																		
SimCLR	63.41	69.77	73.53	60.84	68.27	73.39	54.98	56.97	62.52	51.41	69.44	77.73	59.78	68.52	76.54	59.02	67.26	73.20
BYOL	71.70	73.83	76.45	57.16	67.44	71.64	48.73	61.63	70.82	41.99	74.40	77.17	60.88	74.42	78.75	54.20	71.92	74.69
BarlowTwins	72.87	75.96	78.41	62.57	70.84	74.34	52.12	60.39	66.14	50.12	73.54	77.62	55.12	72.75	78.39	60.72	71.64	77.43
MoCo-v3	73.19	76.65	78.26	55.88	69.21	76.69	50.32	63.71	71.31	51.38	71.66	74.33	62.13	76.74	75.29	54.61	74.26	77.68
SimSiam	73.15	72.70	75.63	62.52	69.31	76.38	55.16	62.91	71.31	49.30	69.47	75.92	58.35	72.89	75.31	58.25	68.61	77.41
TS-TCC	70.73	75.88	78.91	53.54	66.98	77.87	48.04	61.79	71.18	43.34	69.48	78.23	57.07	73.62	78.72	55.26	68.48	76.79
CLOCS	68.94	73.36	76.31	57.94	72.55	76.24	51.97	57.96	72.65	47.19	71.88	76.31	59.59	77.78	77.49	54.38	71.93	76.13
ASTCL	72.51	77.31	81.02	61.86	68.77	76.51	44.14	60.93	66.99	52.38	71.98	76.05	57.90	77.01	79.51	56.40	70.87	75.79
CRT	69.68	78.24	77.24	61.98	70.82	78.67	46.41	59.49	68.73	47.44	73.52	74.41	58.01	76.43	82.03	56.21	73.70	78.80
ST-MEM	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
Multimodal Learning																		
MERL	**78.64**	**83.90**	85.27	61.41	**77.55**	82.98	56.32	**69.11**	**77.66**	52.16	**78.07**	81.83	**69.25**	**82.82**	**89.44**	63.66	**78.67**	84.87
Our framework	73.58	79.07	**87.67**	**66.30**	74.20	**84.84**	**58.94**	58.93	74.06	**56.92**	76.27	**84.42**	58.28	70.51	86.74	**65.42**	75.08	**89.16**

G.2. Experiment Configuration

The training configurations for downstream tasks, including optimizer, scheduler, and relevant hyperparameters, are detailed in Table 8.

H. Performance on Linear probing.

Table 9 shows the linear probing AUC performance of our framework's and other eSSLs' ECG encoders on specific six downstream datasets.

I. Performance of Non-overlapped Cardiac Conditions

While evaluating performance exclusively on non-overlapping (i.e., unseen) downstream classes is not a standard practice in existing ECG literature, including MERL and other multimodal or self-supervised frameworks, we acknowledge its value in assessing true generalization. To address this, we conduct an additional analysis where we evaluate zero-shot AUC only on downstream classes that do not appear in the pre-training dataset.

Table 10 presents the comparison between AUC scores on all downstream classes versus only non-overlapping ones. As expected, performance on unseen classes is moderately lower, yet remains strong across datasets, confirming our framework's ability to generalize beyond pre-trained diagnostic categories. This analysis complements our main results and provides deeper insights into model robustness.

Table 10. Zero-shot AUC on downstream datasets using only non-overlapping (unseen) classes vs. using all classes.

Setting	PTB-XL-Super	PTB-XL-Sub	PTB-XL-Form	PTB-XL-Rhythm	CPSC-2018	CSN	Overall
Non-overlapping Classes	75.97	69.30	61.36	83.83	–	75.73	73.24
All Classes	78.20	77.52	60.67	86.79	79.83	80.17	77.20

J. Performance of Architecture Varients

Beyond comparisons with eSSLs, we directly assess our framework's zero-shot classification performance by varying its core modules, including the ECG backbone and CFN. Table 11 reports the results for our framework and its variants, where the ViT + CFN architecture achieves the highest average AUC of 77.20%, with particularly strong performance on PTB-XL-Rhythm and CSN.

Table 11. Zero-shot classification AUC performance of our framework and its variants on six downstream ECG datasets, with best results **bolded**.

	Linear Classification		*Cardiac Fusion Network*	
Dataset	ResNet	ViT	ResNet	ViT
PTB-XL-Superclass	67.55	68.37	68.75	**78.20**
PTB-XL-Subclass	73.77	74.25	68.02	**77.52**
PTB-XL-Form	**64.34**	63.46	58.85	60.67
PTB-XL-Rhythm	75.68	76.20	68.69	**86.79**
CPSC-2018	**83.35**	79.71	60.38	79.83
CSN	72.61	74.22	65.07	**80.17**
Overall	72.88	72.70	64.96	**77.20**

Under the linear classification setup, ResNet slightly outperforms ViT across few datasets (e.g. PTB-XL-Form, CPSC-2018), demonstrating its effectiveness in extracting essential features without contextual mechanisms. However, the CFN significantly improves performance, with ViT + CFN achieving a notable boost in AUC, particularly on PTB-XL-Superclass (78.20%), PTB-XL-Rhythm (86.79%), and CSN (80.17%). Figure 8 highlights the strength of ViT's attention mechanisms combined with CFN, which excels at capturing complex temporal and spatial dependencies in ECG signals.

By employing prompts to generate meaningful class queries, the CFN enables a more flexible and adaptable approach to classification, allowing flexible adaptation to diverse downstream class structures. Unlike linear classification, our framework dynamically aligns pre-trained knowledge with new cardiac conditions, eliminating the need for explicit class mapping between pre-training and downstream datasets and improving generalization.

Notably, the ResNet + CFN combination performs worse than ViT + Linear in several cases. The performance difference is primarily due to the intrinsic architectural mismatch between ResNet-based encoders and the design of our Cardiac Fusion Network (CFN). Specifically, the CFN is designed to leverage fine-grained, structured token representations to facilitate flexible cross-modal query fusion. This design aligns naturally with ViT-based ECG encoders, which output a sequence of patch-level embeddings that can be effectively interacted with text queries at the token level. In contrast, ResNet produces globally pooled feature vectors without fine-grained temporal tokenization. Such representations inherently lack the structural granularity required for effective token-level cross-modal fusion, limiting the potential benefit of the CFN. In this case, a simple linear classifier can more directly exploit the coarse-grained ResNet features without introducing additional complexity. We would like to emphasize that this observation reflects the architectural characteristics of ResNet rather than a limitation of the CFN itself. As demonstrated in Table 11, when paired with a ViT encoder that naturally provides tokenized representations, the CFN significantly improves zero-shot performance across all datasets (e.g., Overall AUC: 77.20 vs. 72.70).

We further analyze performance on specific cardiac conditions in PTB-XL-Subclass, shown in Figure 9. Our framework consistently achieves high AUC scores, particularly for critical conditions like LAFB/LPFB, CRBBB, CLBBB, IRBBB, and RVH (AUC > 90). In contrast, other variants show lower performance, especially for conditions requiring nuanced spatial and temporal patterns. ResNet + Linear performs competitively on simpler cases but struggles with complex conditions. ResNet + CFN exhibits significant drops, particularly for IMI, AVB, RAO/RAE, ISCA, and IRBBB, highlighting its limitations in effectively capturing intricate dependencies.

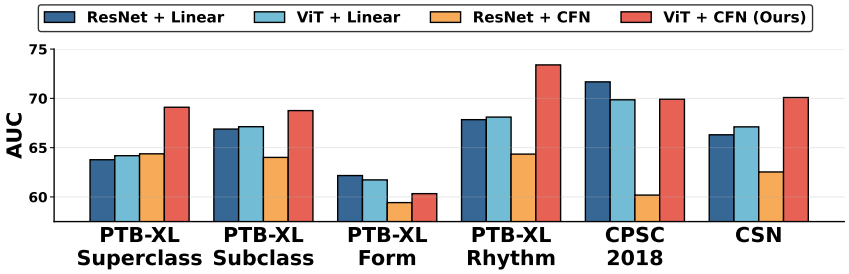


Figure 8. Specific zero-shot performance of our framework and its variants across downstream datasets.

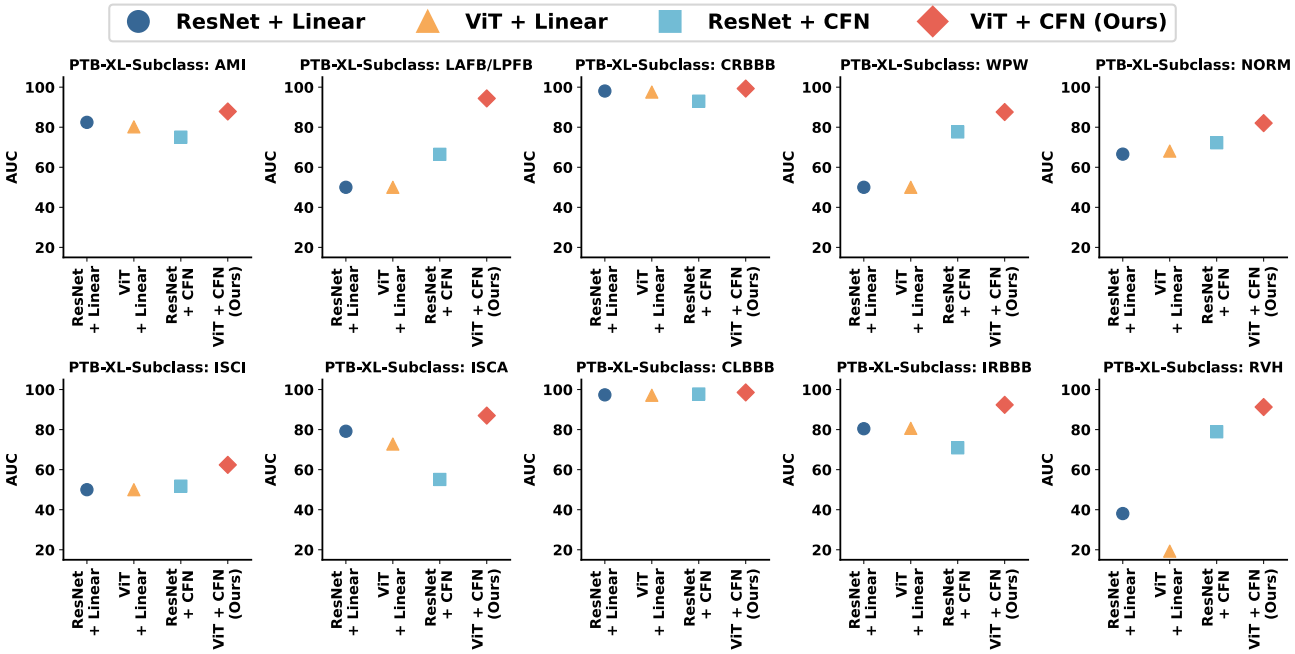


Figure 9. Specific zero-shot classification AUC performance of our framework and its variants on selected detailed categories in PTB-XL-Subclass.

Other condition-wise performance is shown below.

PTB-XL-Superclass. Figure 10 records the AUC performance of our framework on specific cardiac conditions in PTB-XL-Superclass dataset.

PTB-XL-Subclass. Figure 11 records the AUC performance of our framework on specific cardiac conditions in PTB-XL-Subclass dataset.

PTB-XL-Form. Figure 12 records the AUC performance of our framework on specific cardiac conditions in PTB-XL-Form dataset.

PTB-XL-Rhythm. Figure 13 records the AUC performance of our framework on specific cardiac conditions in PTB-XL-Rhythm dataset.

CPSC-2018. Figure 14 records the AUC performance of our framework on specific cardiac conditions in CPSC-2018 dataset.

CSN. Figure 15 records the AUC performance of our framework on specific cardiac conditions in CSN dataset.

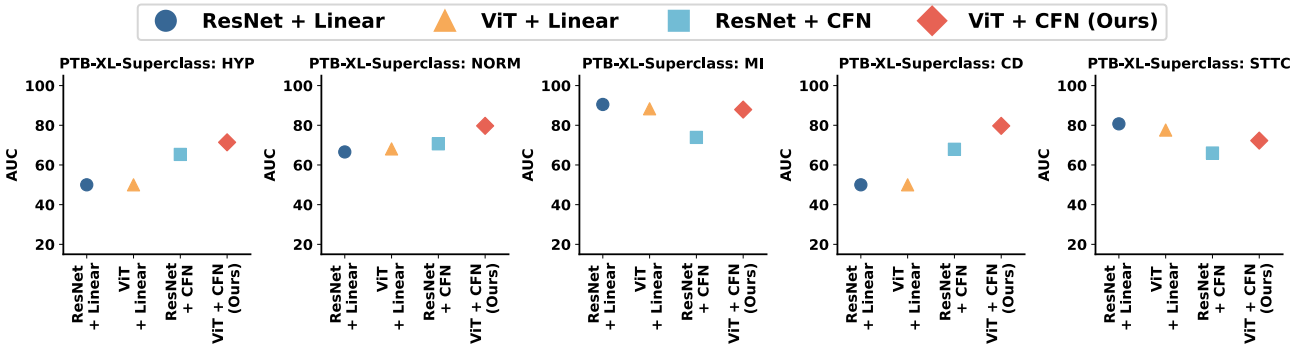
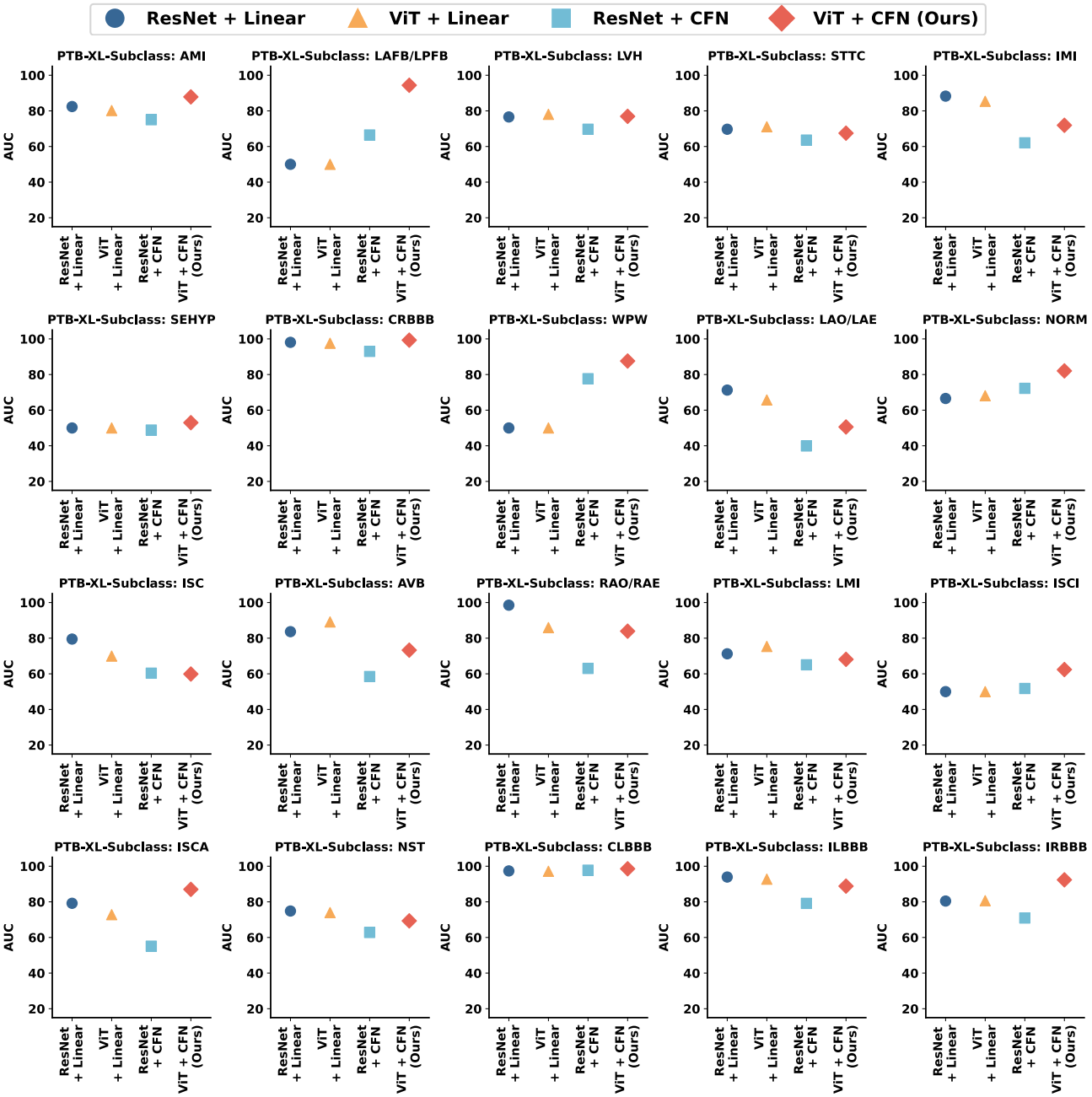


Figure 10. Zero-shot learning performance of our framework and its variants on specific categories in PTB-XL-Superclass.



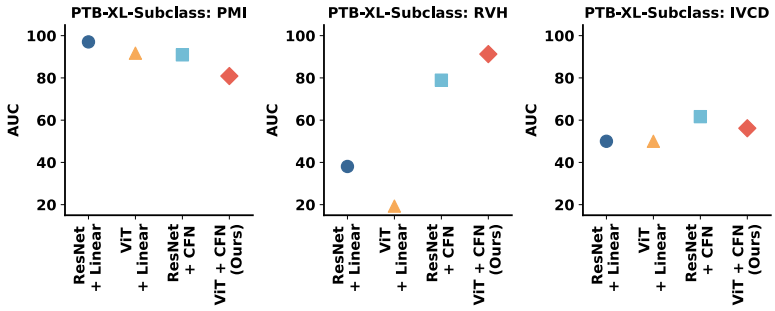


Figure 11. Zero-shot learning performance of our framework and its variants on specific categories in PTB-XL-Subclass.

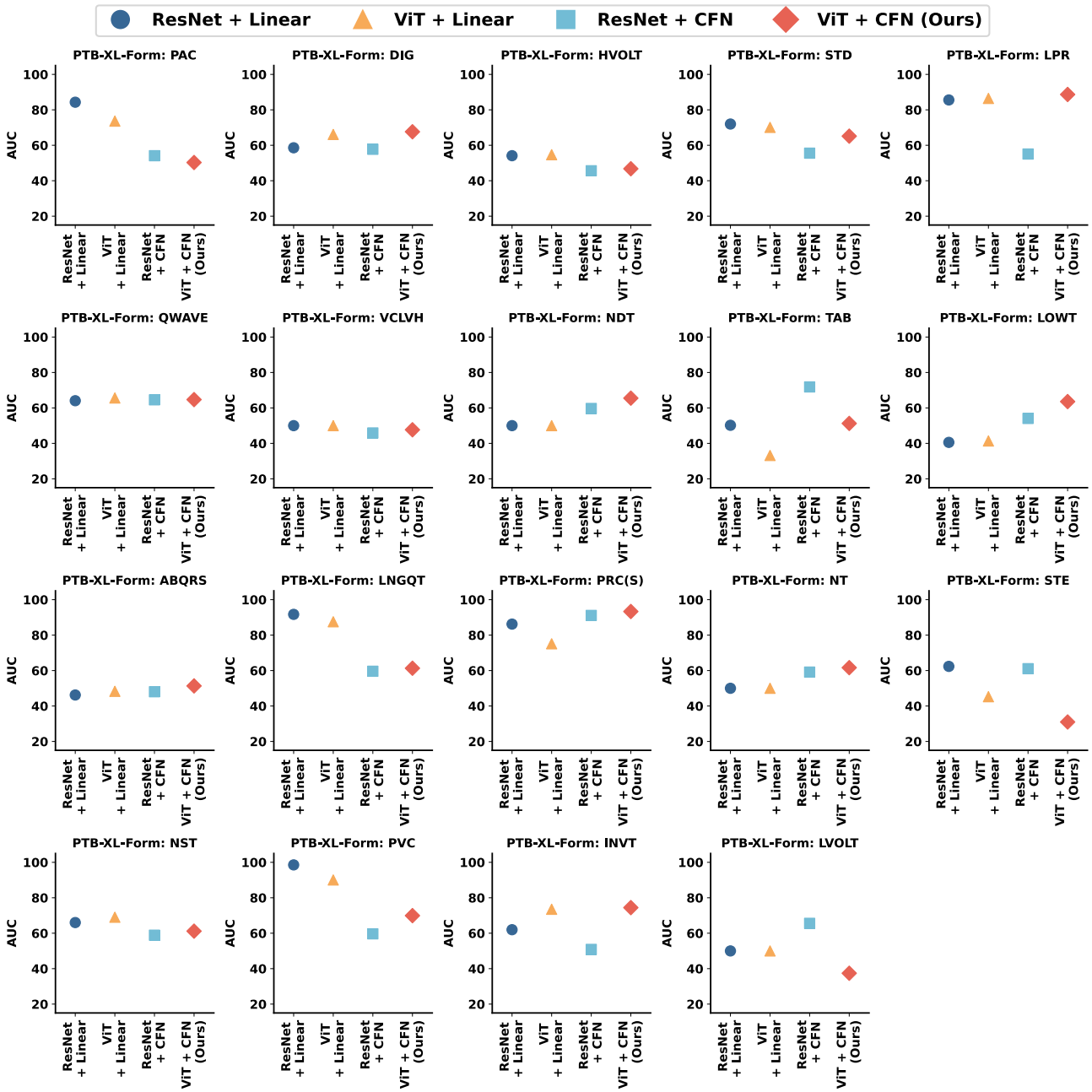


Figure 12. Zero-shot learning performance of our framework and its variants on specific categories in PTB-XL-Form.

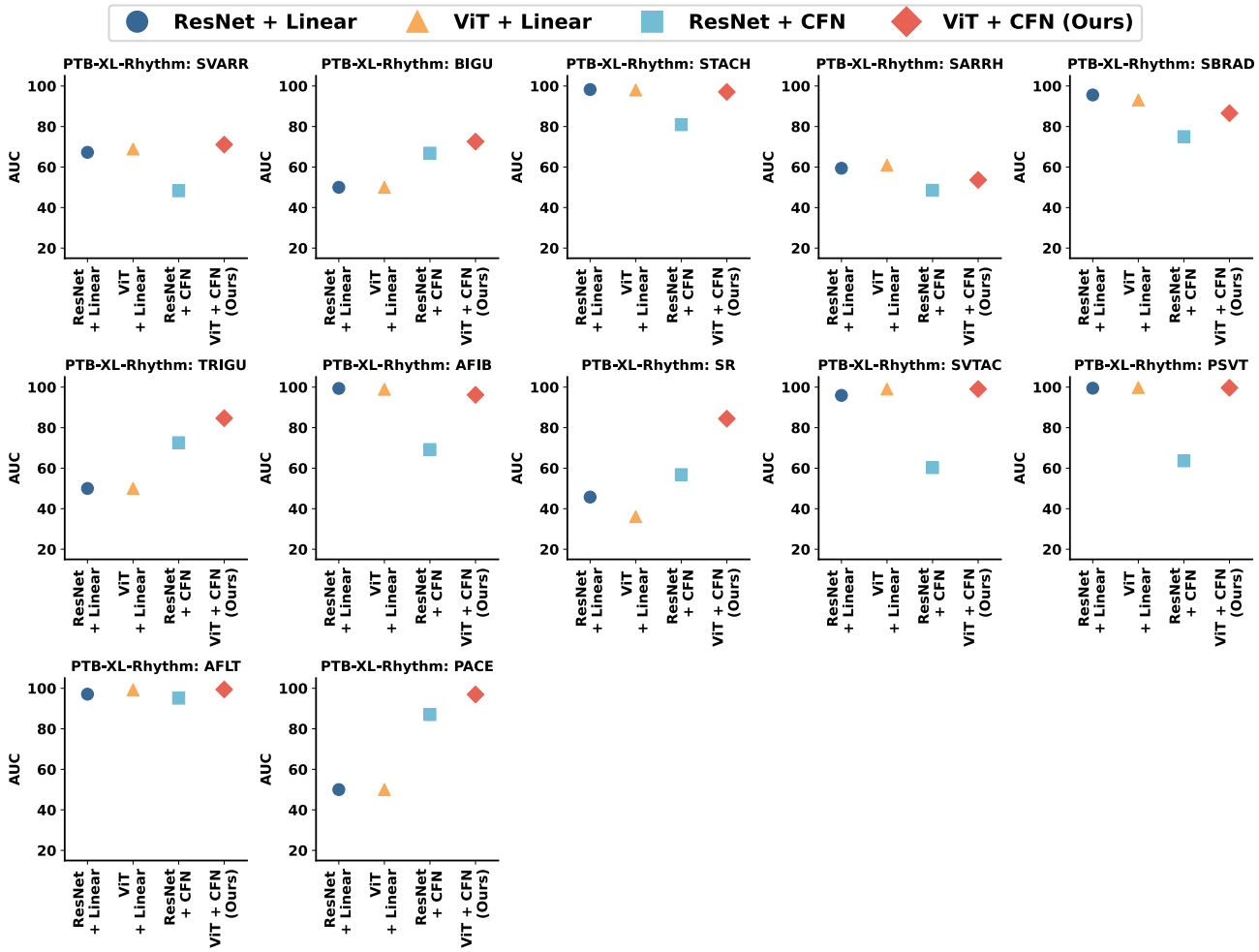


Figure 13. Zero-shot learning performance of our framework and its variants on specific categories in PTB-XL-Rhythm.

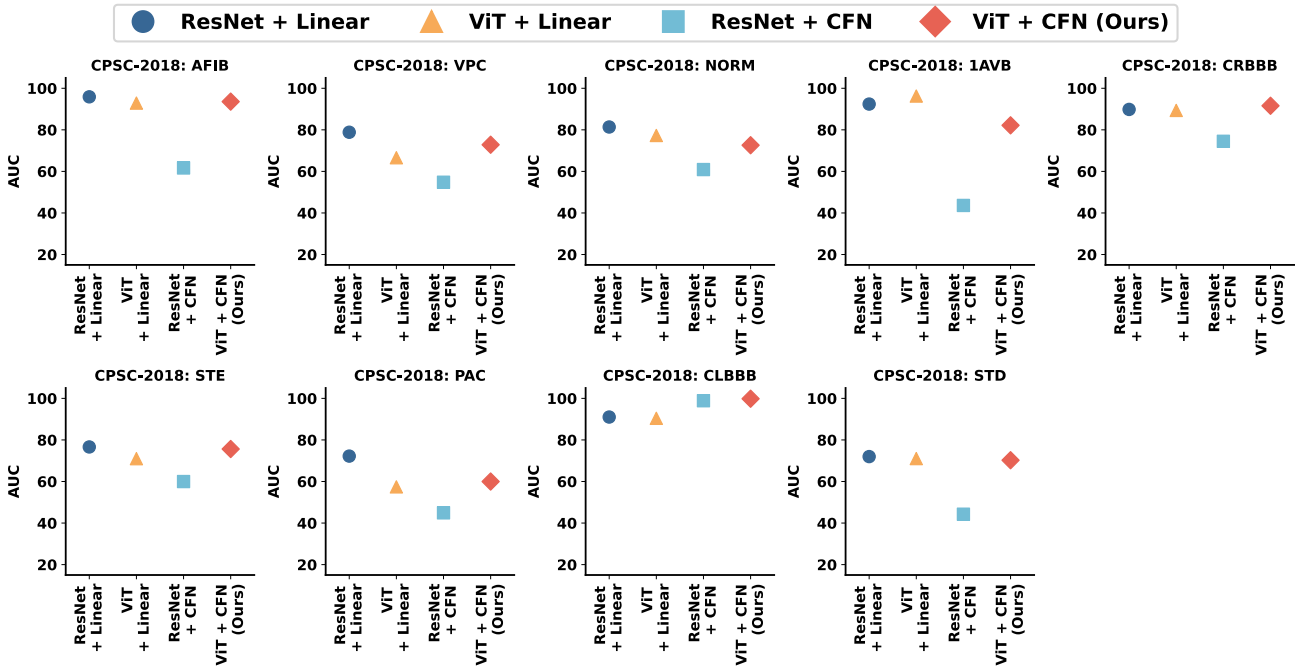
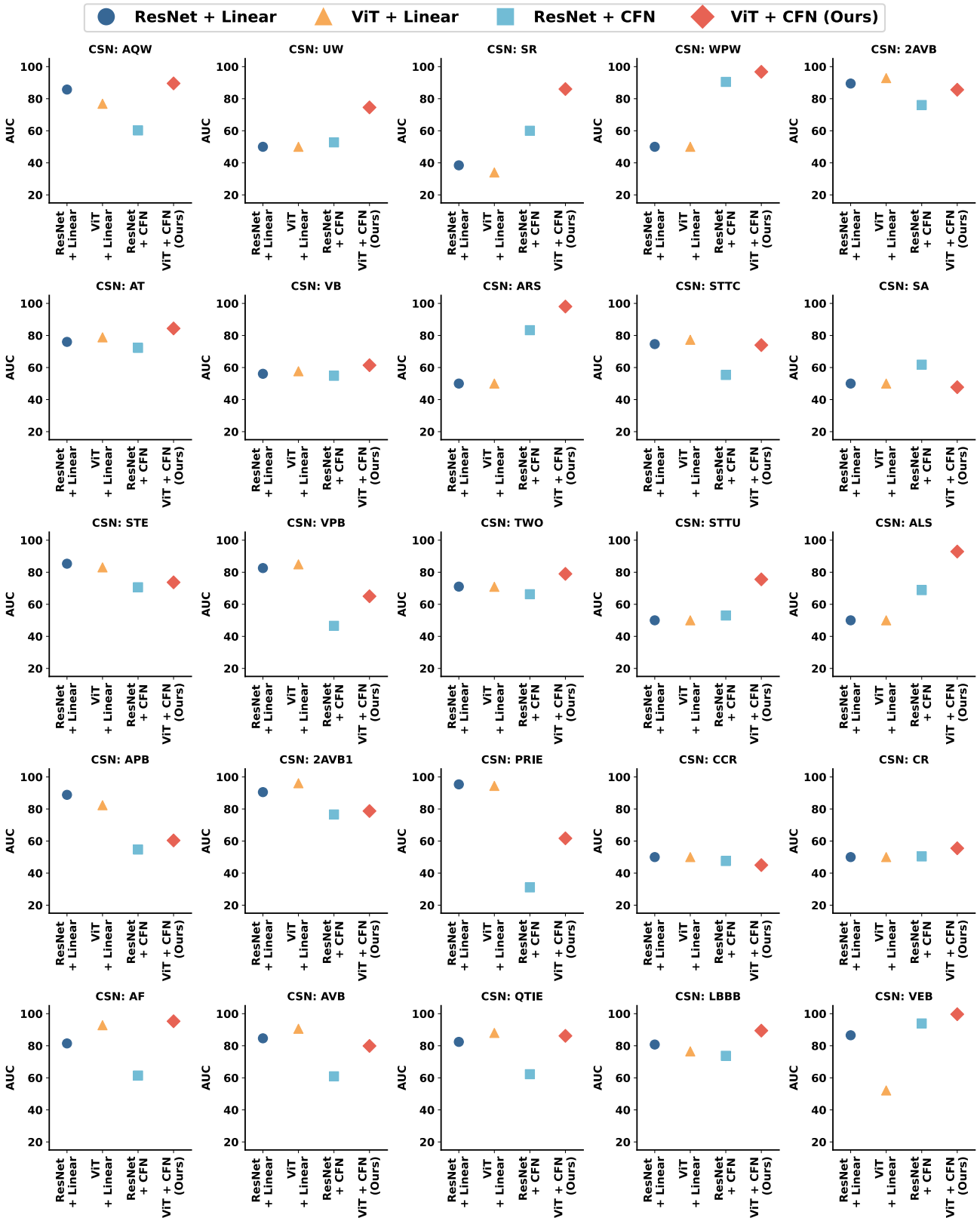


Figure 14. Zero-shot learning performance of our framework and its variants on specific categories in CPSC-2018.


ResNet + Linear
ViT + Linear
ResNet + CFN
ViT + CFN (Ours)
CSN: AQW
CSN: UW
CSN: SR
CSN: WPW
CSN: 2AVB
CSN: AT
CSN: VB
CSN: ARS
CSN: STTC
CSN: SA
CSN: STE
CSN: VPB
CSN: TWO
CSN: STTU
CSN: ALS
CSN: APB
CSN: 2AVB1
CSN: PRIE
CSN: CCR
CSN: CR
CSN: AF
CSN: AVB
CSN: QTIE
CSN: LBBB
CSN: VEB
AUC
100
80
60
40
20

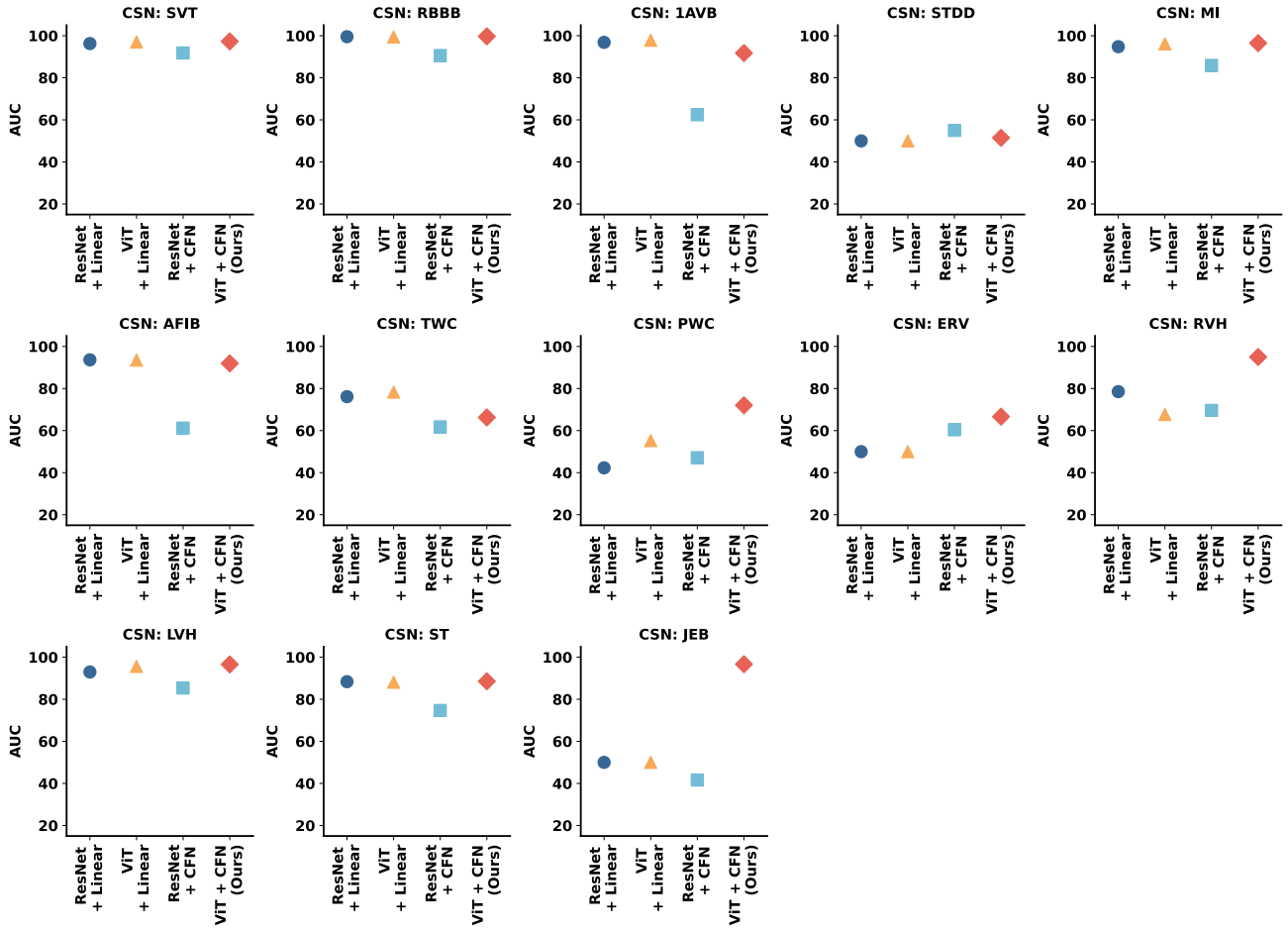


Figure 15. Zero-shot learning performance of our framework and its variants on specific categories in CSN.

K. Further Discussion

K.1. Offline Use of Large-sized LLM for Clinical NER

While we employ LLaMA3.1-70B-Instruct for clinical entity extraction, this step is performed offline only once during dataset construction and is not part of our framework's pre-train & inference pipeline. The motivation for using a larger model is to ensure high annotation quality for the pre-training dataset. Once entities are extracted and mapped, they form a standardized query list used throughout training and evaluation. Therefore, clinical deployments do not require access to large LLMs, and the our framework itself remains simple and lightweight during pre-training and inference.

K.2. Computation Cost and Practical Deployment

Data Processing (Offline NER). To obtain high-quality supervision labels, we extract and normalize diagnostic entities from MIMIC-IV-ECG reports using LLaMA3.1–70B-Instruct with structured prompts. This step is performed only once as described above before pre-training our framework. The output is a cleaned, deduplicated dataset of standardized diagnostic labels, which serves as global cardiac queries for training. The annotation process takes approximately 6 hours on 8 NVIDIA A100-SMX4-80GB GPUs, and the tested minimum reproducing resources are 4 NVIDIA A100-PCIE-40GB GPUs without parallelized LLM inference. The resulting standardized dataset is reused across training and downstream evaluation.

Model Pre-training. Our proposed ECG-Text framework consists only of a ViT-based ECG encoder, query-based supervision, and a lightweight Cardiac Fusion Network (CFN). Training is efficient, around 1.5 hours on 4 NVIDIA A100-PCIE-40GB GPUs achieving best AUC performance (16 epochs), and does not require contrastive sampling or further fine-tuning in deployment.

Deployment and Inference Once pre-trained, our framework supports zero-shot ECG classification via a set of concise cardiac query prompts. Inference only involves a forward pass through the ECG encoder and CFN, taking milliseconds per ECG sample. No LLMs or textual reports are needed at test time, making our framework highly practical for deployment in real-world clinical settings. We empirically verify that inference can be efficiently performed on a single NVIDIA A5000-PCIE-24GB GPU or NVIDIA RTX4090-PCIE-24GB GPU.

K.3. Similarity Threshold Determination

The similarity thresholds in our entity deduplication and mapping pipeline were determined in consultation with experienced cardiologists (over 10 years of clinical practice), based on joint analysis of the results under various threshold settings in each phase.

Through this process, we observed that setting the thresholds too high (e.g., above 0.9 in entity mapping) would exclude valid clinical variants due to minor wording differences, while setting them too low (e.g., below 0.7 in entity mapping) could introduce semantic ambiguity by incorrectly matching unrelated conditions (Table 12).

Table 12. Incorrect matching examples between standard terminology and report entities.

Standard Terminology	Report Entity	Similarity Score
left ventricle hypertrophy	Ventricular fibrillation	0.6400
non-specific ST changes	ST elevation	0.6343
inferior myocardial infarction	anterior wall abnormality	0.5717

”left ventricle hypertrophy” and ”ventricular fibrillation” shows a similarity score of 0.64, but are entirely unrelated - one refers to structural enlargement of the left ventricle, while the other refers to a life-threatening arrhythmia. The selected thresholds reflect a balance between preserving clinically meaningful variants and minimizing noise.

K.4. On the Effectiveness of CFN with Different Backbones

To address concerns regarding the effectiveness of the Cardiac Fusion Network (CFN), particularly its relatively lower performance when paired with a ResNet backbone (cf. Table 11), we conduct statistical analyses to better understand the interaction between backbone architecture and the CFN module.

Δ AUC Comparison. We compare the AUC improvement brought by CFN over linear classification for both ViT and ResNet backbones across six downstream datasets. As shown in Figure 16, CFN brings consistent performance gains when combined with ViT, with average improvement of +5.97 AUC. In contrast, CFN shows little to negative improvement with ResNet, indicating that the quality of the underlying feature representations plays a critical role in effective cross-modal fusion.

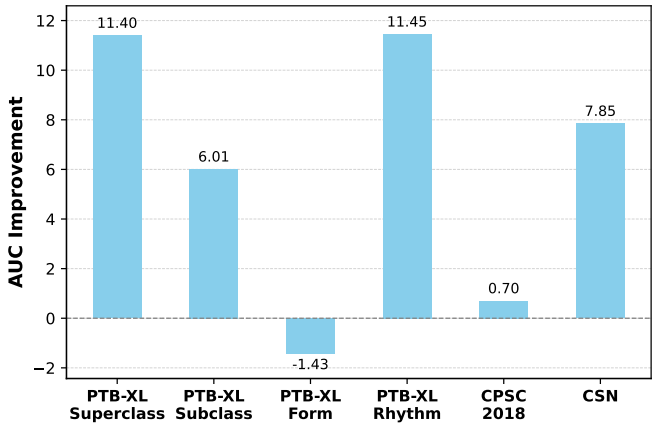


Figure 16. CFN vs. linear classification (Δ AUC) for ViT ECG backbone.

Statistical Significance. To verify this trend, we perform a paired t-test and Wilcoxon signed-rank test on the Δ AUC values. Both tests confirm that CFN yields significantly greater improvements on ViT than ResNet:

- **Paired t-test**: $t = 4.99$, $p = 0.0021$
- **Wilcoxon test**: $W = 21.0$, $p = 0.0156$

These results provide strong statistical evidence that ViT synergizes better with CFN compared to ResNet, likely due to ViT's superior capacity in capturing global temporal dependencies in ECG signals.

CFN is designed to align high-level ECG features with cardiac queries via cross-attention. However, ResNet provides only local, convolutional features with limited contextual depth, especially compared to ViT's global receptive field. As a result, the decoder lacks sufficient global representations to effectively condition on query semantics. This bottleneck explains the performance drop observed in ResNet + CFN.

Two-Way ANOVA. We further conduct a two-way ANOVA with Backbone (ResNet vs. ViT) and Module (Linear vs. CFN) as factors. As shown in Table 13, the interaction term is statistically significant ($F = 6.60$, $p = 0.018$), confirming that the effect of CFN depends on the choice of backbone. Notably, neither factor alone is significant, suggesting that their combination determines performance.

Table 13. Two-way ANOVA results on AUC with backbone and module as factors.

Source	Sum of Squares	df	F-value	p-value
Backbone	167.06	1	3.79	0.066
Module	5.57	1	0.13	0.726
Backbone $\times$ Module	**290.65**	1	**6.60**	**0.018**
Residual	881.22	20	-	-

Takeaway. These findings reinforce CFN's role as a powerful fusion mechanism when paired with a backbone (like ViT) that produces expressive feature sequences. The drop in performance with ResNet may stem from its less structured output, which lacks the sequential token-style organization needed for effective query-based attention. Thus, the CFN is not inherently ineffective, but its utility hinges on a compatible encoder design.

K.5. Comparing Our framework with MERL

To assess the relative effectiveness of our supervised multimodal framework, we compare our framework against MERL (Liu et al., 2024), a recent multimodal contrastive learning baseline that utilizes clinical reports and enhanced prompt engineering (17). While MERL employs contrastive objectives and handcrafted prompts, our framework leverages fine-grained diagnostic supervision through LLM-extracted entities and multimodal fusion via the Cardiac Fusion Network (CFN).

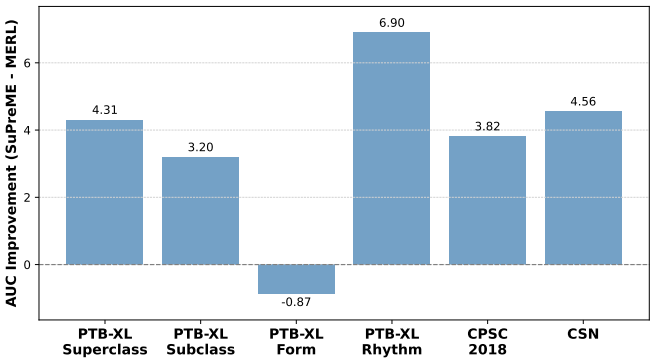


Figure 17. Per-dataset AUC difference between our framework and MERL.

As shown in Table 14, our framework achieves consistently higher AUCs across all but one dataset. On average, it improves

Table 14. Zero-shot AUC comparison between our framework and MERL.

Framework	PTB-XL-Superclass	PTB-XL-Subclass	PTB-XL-Form	PTB-XL-Rhythm	CPSC-2018	CSN	Avg
MERL	73.89	74.32	61.54	79.89	76.01	75.61	73.54
Our framework	**78.20**	**77.52**	60.67	**86.79**	**79.83**	**80.17**	**77.20**

zero-shot performance by 3.66% absolute. Statistical testing confirms this improvement is significant: a paired t-test across the six datasets yields $t = 3.51$, $p = 0.0171$.

Table 15. Average zero-shot AUC and standard deviation across six datasets.

Framework	**Zero-shot AUC (%)**
MERL	73.54 ± 2.30
Our framework	$\mathbf{77.20 \pm 0.21}$

Moreover, our framework demonstrates significantly lower performance variance across datasets. While MERL exhibits a standard deviation of 2.30, our framework achieves a much smaller deviation of 0.21, indicating greater robustness and stability across diverse cardiac classification tasks. These results collectively support the effectiveness of our proposed supervised pre-training framework and its entity-level modality fusion strategy, even when compared to a strong multimodal baseline.

K.6. Domain Scope and Generalization Potential

While our framework is evaluated on 12-lead ECG data, we believe that this modality represents a highly impactful and widely applicable domain in clinical practice. ECG is routinely used across diverse medical contexts—including emergency rooms, intensive care units (ICUs), outpatient cardiology clinics, and even home-based healthcare monitoring—due to its low cost, non-invasiveness, and real-time ability to reflect cardiac electrical activity. As such, improving automated ECG interpretation has direct clinical relevance across resource settings and specialties.

Moreover, although this work focuses on ECG, the core methodology of our framework, namely multimodal learning between biomedical signals and clinically meaningful queries, can be generalized to other physiological signal domains such as EEG (electroencephalogram) or PPG (photoplethysmography). These modalities are similarly structured (multi-channel, time-series signals) and increasingly available in clinical and wearable settings. However, to the best of our knowledge, there is currently a lack of large-scale, publicly accessible datasets that pair these signals with detailed, free-text clinical reports suitable for training our entity extraction module.

We hope our work can inspire future efforts toward building such paired datasets for other biomedical signals, enabling the broader application of query-based multimodal learning frameworks beyond ECG.

L. Limitations

Although our framework achieves state-of-the-art zero-shot performance, several aspects merit closer scrutiny. First, the clinical entity extraction pipeline relies on a large language model that has not been fine-tuned on cardiology-specific corpora; consequently, highly specialised or context-dependent terms, for example, vernacular descriptions of rare channelopathies, may be missed or hallucinated, potentially propagating noise into subsequent stages. Second, our framework implicitly assumes that the pre-trained model will generalise well across diverse clinical contexts like different institutions, devices and patient demographics. Real-world data often contain higher artefact levels and markedly imbalanced class distributions; preliminary experiments already indicate that performance on infrequent rhythms can lag behind that on common diagnoses, suggesting room for improvement under distribution shift.

Moreover, because the ECG backbone is optimized jointly with CFN rather than via an explicit contrastive objective, its features along do not always yield superior linear-probe accuracy; in our own experiments, it only achieves comparative linear probing (use ECG backbone only) performance with MERL. Despite that we centre our study on the zero-shot setting, which reflects many anticipated deployment scenarios where high-quality ECG annotations are scarce or prohibitively

expensive, exploring end-to-end or contrastive-hybrid objectives remains a promising direction for future work, as such strategies could further enhance performance when a modest amount of supervision becomes available. Finally, open zero-shot baselines for ECG classification remain limited: apart from MERL, most existing methods are single-modal and were not designed for zero-shot evaluation, underscoring the need for broader public benchmarks to facilitate systematic comparison.

From a representation learning perspective, our evaluation is predominantly quantitative. As future work, incorporating qualitative analyses could provide additional insights into the structure and separability of the learned features. Furthermore, presenting realistic inference examples (e.g., how the system might behave in deployment or how predictions appear to clinicians, researchers, or engineers) could help illustrate practical impact.