SuPreME: A Supervised Pre-training Framework for Multimodal ECG Representation Learning

Mingsheng Cai[1,3], Jiuming Jiang[1,3], Wenhao Huang[2], Che Liu[3*], Rossella Arcucci[3]
[1]The University of Edinburgh, [2]Shenzhen Yinwang Intelligent Technology Co., Ltd,
[3]Imperial College London
{m.cai, jiuming.jiang}@ed.ac.uk[1], huangwenhao@yinwang.com[2],
{mingsheng.cai23, jiuming.jiang23, che.liu21, r.arcucci}@imperial.ac.uk[3]

Abstract

Cardiovascular diseases are a leading cause of death and disability worldwide. Electrocardiogram (ECG) is critical for diagnosing and monitoring cardiac health, but obtaining large-scale annotated ECG datasets is labor-intensive and time-consuming. Recent ECG Self-Supervised Learning (eSSL) methods mitigate this by learning features without extensive labels but fail to capture fine-grained clinical semantics and require extensive task-specific fine-tuning. To address these challenges, we propose **SuPreME**, a **Su**pervised **Pre**-training framework for **M**ultimodal **E**CG representation learning. SuPreME is pre-trained using structured diagnostic labels derived from ECG report entities through a one-time offline extraction with Large Language Models (LLMs), which help denoise, standardize cardiac concepts, and improve clinical representation learning. By fusing ECG signals with textual cardiac queries instead of fixed labels, SuPreME enables zero-shot classification of unseen conditions without further fine-tuning. We evaluate SuPreME on six downstream datasets covering 106 cardiac conditions, achieving superior zero-shot AUC performance of 77.20%, surpassing state-of-the-art eSSLs by 4.98%. Results demonstrate SuPreME's effectiveness in leveraging structured, clinically relevant knowledge for high-quality ECG representations.

1 Introduction

Supervised learning 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 and Yen, 2022). However, these methods typically rely on large-scale, high-quality annotated datasets, which are costly to create and difficult to scale.

To reduce dependence on annotations, recent advancements in ECG self-supervised learning (eSSL) have enabled the extraction of representative features from large-scale unannotated ECGs using contrastive or generative tasks (Eldele et al., 2021; Kiyasseh et al., 2021; Na et al., 2024). Despite their promise, these methods often rely on strong signal-level augmentations that may distort the semantic integrity of the signal and require complex pretext task designs (Kiyasseh et al., 2021). Multimodal learning approaches (Liu et al., 2024; Li et al., 2024) have also been proposed to learn ECG representations by leveraging free-text ECG reports. However, these methods face challenges due to noise in textual data and the complexities of language grammar, which can hinder learning efficiency (Wu et al., 2023).

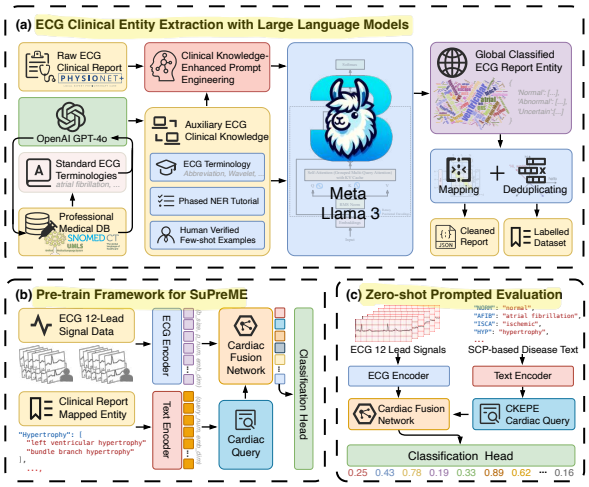


Figure 1: Overview including (a) ECG clinical entity extraction with LLMs, (b) supervised pre-training with SuPreME, and (c) zero-shot prompted evaluation.

To address these limitations and develop a scalable, simple, and effective ECG pre-training framework, we propose **SuPreME**, a **Su**pervised **Pre**-training framework for **M**ultimodal **E**CG representation learning. Our contributions are threefold: **(a)** We introduce an automated pipeline that extracts high-quality clinical entities from raw ECG reports using an instruction-tuned LLM enriched with domain-specific knowledge (Figure 1(a)). Ex-

*Correspondence: che.liu21@imperial.ac.uk

tracted entities are deduplicated and mapped to dataset-specific standardized diagnostic terms using clinician-validated resources (e.g., SNOMED CT, UMLS, SCP-ECG[1]), forming a dataset-specific global cardiac query list without manual annotation. This process enables scalable, consistent labeling and captures richer semantics than coarse-grained or free-text alternatives. **(b)** Leveraging these standardized cardiac queries, we propose SuPreME (Figure 1(b-c)), a multimodal framework that directly fuses ECG signals with cardiac queries through a lightweight Cardiac Fusion Network (CFN). Unlike prior methods (e.g., MERL) that rely on raw free-text inputs or handcrafted pretext tasks, SuPreME requires no signal-level augmentation or contrastive loss design, offering an efficient and interpretable multi-label supervision strategy grounded in standardized cardiac queries. **(c)** We pre-train SuPreME on 771,500 ECG signals paired with 295 global standardized cardiac queries from MIMIC-IV-ECG (Gow et al., 2023) (Appendix A.1.1). On six downstream datasets (e.g., PTB-XL, CPSC-2018, Chapman-Shaoxing-Ningbo; Appendix A.1.2), it achieves a new state-of-the-art zero-shot AUC of 77.20%, significantly outperforming existing eSSL and multimodal baselines, including those fine-tuned with 10–100% labeled data. SuPreME also shows strong data efficiency and generalization, with zero-shot performance under only 20% pre-training data surpassing fully fine-tuned eSSLs.

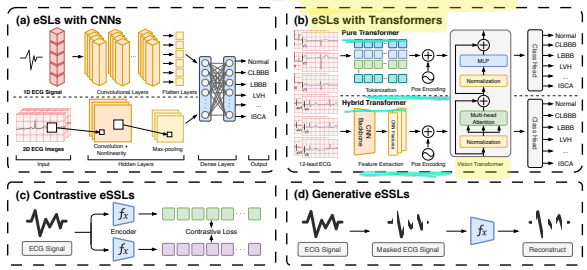


Figure 2: Current ECG representation learning methods, including (a) CNN-based supervised learning, (b) Transformer-based supervised learning, (c) contrastive learning, and (d) generative learning.

2 Related Work

ECG Supervised Learning. ECG supervised learning (eSL) methods, using CNNs or Transformers in Figure 2(a−b), 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 in Figure 2(c−d). 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-Text 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; Abbaspourazad 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

[1]SNOMED CT and UMLS are standardized clinical terminology databases; SCP-ECG refers to the Standard Communication Protocol for Computer-Assisted Electrocardiography.

representation learning and downstream flexibility.

3 Methodology

SuPreME extracts structured clinical entities from ECG reports via an instruction-tuned LLM (Section 3.1) to form cardiac queries, which are fused with ECG signals via Cardiac Fusion Network (CFN) in a shared latent space, enabling zero-shot classification of unseen cardiac conditions without fine-tuning (Section 3.3), thus yielding scalable, clinically meaningful representations.

3.1 LLM-based Clinical Entity Extraction

Enrich LLM with Domain Knowledge. ECG reports generated by 12-lead devices (Appendix A.2) contain diverse and nuanced descriptions of cardiac conditions. To contextualize these free-text reports for clinical entity extraction, we first construct a cardiac-specific vocabulary using GPT-4o-mini (Appendix A.3.1), by filtering, normalizing, and aggregating terminology from clinician-validated resources such as SNOMED CT, UMLS, and SCP-ECG (Bodenreider, 2004; Donnelly et al., 2006; Rubel et al., 2016). The vocabulary covers both complete diagnostic terms (e.g., `sinus rhythm`) and commonly used abbreviations (e.g., `LVH` for `Left Ventricular Hypertrophy`). This domain knowledge is then integrated into the LLM prompt design to guide the subsequent extraction process.

Knowledge-Guided Entity Extraction. We employ an instruction-tuned LLM to extract clinical entities from unstructured ECG reports. The model is prompted using structured instructions and few-shot examples (Appendix A.3.2), incorporating the curated cardiac vocabulary to enhance contextual understanding. It extracts diagnostic expressions (e.g., waveform patterns, cardiac abnormalities) along with their associated certainty. Extracted entities are categorized as `Normal`, `Abnormal`, or `Uncertain`, as illustrated in Figure 3(a). Entities marked with low certainty (e.g., containing "probably" or "cannot rule out") are discarded to improve diagnostic precision. Within each report, semantically similar expressions are merged to prepare for cross-report alignment and standardization.

Entity Deduplication and Mapping. Although the extraction is guided and structured, lexical variability due to differing physician writing styles and device-specific formats still leads to redundant or inconsistent entities. To resolve this, we apply the curated cardiac vocabulary for entity standardization. Both extracted entities and reference terms are encoded using MedCPT, a medical BERT model pre-trained for clinical semantic similarity (Jin et al., 2023). Cosine similarity between embeddings is computed to perform soft matching. Entities that exhibit high average similarity to a reference term are aligned and deduplicated with a threshold selected and clinically validated by experienced cardiologists with over ten years of practice. Case study is provided in Appendix A.3.3.

3.2 Multimodal ECG Supervised Learning

ECG Embedding with Vision Transformer. The Vision Transformer (ViT) (Dosovitskiy et al., 2020), designed for 2D image processing, reshapes images into sequences of flattened patches for Transformer-based analysis. Similarly, ECG signals exhibit temporal and structural patterns analogous to the spatial relationships in images. We then adapt its architecture by dividing ECG time series into fixed-size patches, as shown in Figure 3(b).

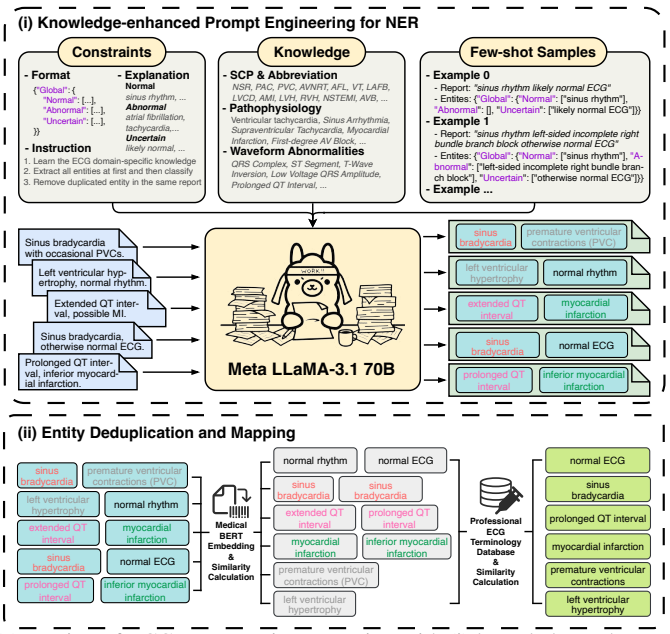
Multi-lead ECG signals are represented as $\mathbf{x} \in \mathbb{R}^{B\times L\times T}$, where B is the batch size, L the number of leads, and T the number of time steps. Each lead's signal is independently segmented into $N = T/P$ non-overlapping patches of length P, resulting in $\mathbf{x}_{i,j} \in \mathbb{R}^{B\times P}$ for lead i and patch index j. Each patch is flattened and passed through a shared linear projection layer $\mathbf{W}_p \in \mathbb{R}^{P\times D}$ to produce a token embedding $\mathbf{z}_{i,j} \in \mathbb{R}^{B\times D}$:

$$\begin{aligned} \mathbf{z}_{i,j} &= \mathbf{x}_{i,j}\mathbf{W}_p \\ \mathbf{z'}_{i,j} &= \mathbf{z}_{i,j} + \mathbf{e}_i + \mathbf{p}_j \end{aligned} \tag{1}$$

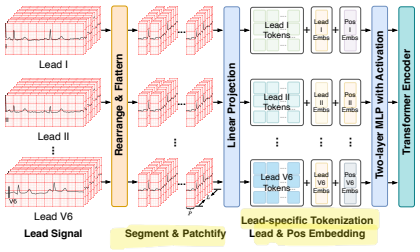
To preserve lead-specific features and spatial-temporal information, we introduce a unique learnable lead embedding $\mathbf{e}_i \in \mathbb{R}^D$ for each lead i, and a positional embedding $\mathbf{p}_j \in \mathbb{R}^D$ for each patch index j. These embeddings are element-wise added to the projected patch token $\mathbf{z}_{i,j}$, forming the enriched representation $\mathbf{z'}_{i,j}$. All enriched tokens from all leads and patches are then concatenated to form the Transformer encoder input sequence:

$$\begin{aligned} \mathbf{Z} &= [\mathbf{z'}_{1,1}, \ldots, \mathbf{z'}_{1,N}, \ldots, \mathbf{z'}_{L,N}] \\ \mathbf{F}_{\text{ECG}} &= \text{MLP}_{\text{ECG}}(\text{Dropout}(\mathbf{Z})) \end{aligned} \tag{2}$$

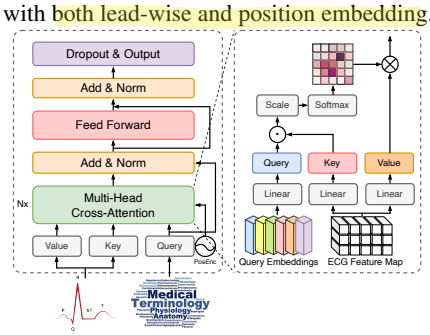
The final token sequence $\mathbf{Z} \in \mathbb{R}^{B\times(L\cdot N)\times D}$ is passed through Transformer encoders to extract high-level ECG representations. Each block consists of multi-head self-attention and feed-forward sublayers, with residual connections and layer normalization. To enhance generalization, we apply stochastic depth dropout to the residual paths.



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



(b) ECG 1D ViT encoder in the SuPreME, with both lead-wise and position embedding.



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

Figure 3: Implementation of the supervised ECG-Text multimodal pre-training framework including (a) ECG report entity extraction, (b) ECG 1D ViT encoder, and (c) architecture of the Cardiac Fusion Network.

Before multimodal fusion, the output features are passed through a modality-specific two-layer multilayer perceptron (MLP) projection head with an intermediate non-linear activation. This projection maps the ViT output from its internal width D to a shared latent dimension D' aligned with the textual modality, forming representation $\mathbf{F}_{\text{ECG}} \in \mathbb{R}^{B\times(L\cdot N)\times D'}$ used as the CFN input. Implementation details are provided in Appendix A.4.1, A.4.2.

Cardiac Query Embedding with MedCPT. Instead of relying on fixed categorical labels, our framework adopts a flexible and semantically meaningful approach based on textual cardiac queries. For the pre-training dataset, we construct fine-grained diagnostic queries by applying the LLM-based entity extraction pipeline described in Section 3.1. Specifically, we generate a dataset-specific global query list of size M, consisting of standardized cardiac terms derived from the deduplicated and mapped output of the extraction process.

Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_M\}$ denote the global query list. Each query q_i is encoded into a dense vector using the query encoder from MedCPT, which applies a Transformer (Trm) to the input sequence [CLS] q_i [SEP]. The final-layer [CLS] token embedding is used as the query representation:

$$\begin{aligned}\mathbf{E}[i,:] &= \text{Trm}([\text{CLS}]\ q_i\ [\text{SEP}]) \\ \mathbf{F}_{\text{Query}} &= \text{MLP}_{\text{Query}}(\text{Dropout}(\mathbf{E}))\end{aligned} \tag{3}$$

All M query embeddings $\mathbf{E} \in \mathbb{R}^{M\times 768}$ are then passed through a modality-specific two-layer MLP projection head with an intermediate activation function to obtain the final representations $\mathbf{F}_{\text{Query}} \in \mathbb{R}^{M\times D'}$ in the shared D'-dimensional latent space aligned with ECG token representations. Further implementations are provided in Appendix A.4.3.

Alignment by Cardiac Fusion Network. The Cardiac Fusion Network (CFN) fuses ECG signals with textual cardiac queries using a multi-layer Transformer decoder architecture, where query embeddings act as decoder inputs and ECG features serve as the encoder memory, following a standard cross-attention formulation (Figure 3(c)).

Given a batch of ECG features $\mathbf{F}_{\text{ECG}} \in \mathbb{R}^{B\times(L\cdot N)\times D'}$ and query embeddings $\mathbf{F}_{\text{Query}} \in \mathbb{R}^{M\times D'}$, CFN fuses the two modalities through cross-attention. During pre-training, each ECG sample is paired with the same M cardiac query embeddings, allowing CFN to learn a joint representation that captures query-conditioned patterns in the signal. The decoder attends to ECG patterns while grounding the prediction in the semantics of each diagnostic query, outputting $\mathbf{H} \in \mathbb{R}^{B\times M\times D'}$ which is passed through a single MLP classification head shared across all queries, producing M binary logits per ECG sample, where each logit indicates the relevance of a specific query to the signal input:

$$\textbf{Logits} = \text{MLP}_{\text{CFN}}(\mathbf{H}) \in \mathbb{R}^{B\times M} \tag{4}$$

In pre-training, we supervise CFN using weak binary labels derived from the entity extraction

pipeline. Each ECG report is matched against the global query list of M standardized diagnostic terms. A binary label of 1 is assigned if a mapped report entity matches a query; otherwise 0. This results in a sparse M-dimensional multi-label vector per ECG sample. To avoid data leakage and ensure modality separation, raw ECG reports are never used directly as input to the model. Instead, diagnostic query list serves as input with queries embedded independently and applied uniformly to every ECG sample.

This formulation enables SuPreME to perform open-set classification with a flexible, scalable query interface, supporting multi-label learning while maintaining clear supervision-query decoupling. CFN initialization is in Appendix A.4.4.

3.3 Zero-shot Prompted Classification

To enable zero-shot classification on unseen cardiac conditions without fine-tuning, we construct concise, clinically meaningful prompts (e.g., `left bundle branch block` for `LBBB`) derived from SCP-ECG codes in each downstream dataset. These prompts form a dataset-specific query list aligned with the pre-training query space and serve as inputs to the textual modality of SuPreME.

We follow a simplified version of Clinical Knowledge-Enhanced Prompt Engineering (CKEPE) (Liu et al., 2024), where SCP-ECG codes are translated into discriminative phrases validated by UMLS and SNOMED CT. Unlike full CKEPE pipelines that retrieve verbose descriptions (e.g., `a condition characterized by prolonged QRS complex...` for `LBBB`), our approach promotes clarity and cross-modal fusion by avoiding redundant or overly detailed textual artifacts. These prompts are used exclusively during inference and remain fixed for all ECG samples within a dataset.

During zero-shot evaluation, ECG signals and textual prompts are encoded via the pre-trained encoders into $\mathbf{F}^{\text{eval}}_{\text{ECG}}$ and $\mathbf{F}^{\text{eval}}_{\text{Query}}$, then passed into the Cardiac Fusion Network (CFN). The CFN performs cross-modal attention to align features and outputs one logit per query-ECG pair. The final prediction scores are computed as:

$$\mathbf{Pred} = \sigma(\text{CFN}(\mathbf{F}^{\text{eval}}_{\text{ECG}}, \mathbf{F}^{\text{eval}}_{\text{Query}})) \in \mathbb{R}^{B\times M'} \quad (5)$$

This setup decouples the prediction space from any fixed label vocabulary, allowing the model to generalize to arbitrary diagnostic queries. The query list can vary across downstream datasets, and the classifier is query-agnostic, meaning no structural change is required when adapting to new tasks. Evaluation is conducted without fine-tuning using AUROC (AUC) per class and mean AUC across all prompts. Details about simplified-CKEPE, evaluation metrics are in Appendix A.5.

4 Experiments

4.1 Configuration and Settings

Clinical Entity Extraction. Following Section 3.1, we extract and normalize clinical entities from MIMIC-IV-ECG using Llama3.1-70B-Instruct[2] with structured prompts to ensure high-quality annotations. Entities are deduplicated via MedCPT embeddings (cosine similarity > 0.8) and mapped to UMLS/SNOMED CT (average cosine similarity > 0.75)[3]. Experiments are run on 8 NVIDIA A100-SMX4-80GB GPUs using vLLM (Kwon et al., 2023). Statistics of extracted MIMIC-IV-ECG entities are in Appendix A.6.1.

Supervised ECG 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 16). Batch size is set to 256 on 4 NVIDIA A100-PCIE-40GB GPUs[4].

Downstream Classification Task. SuPreME is evaluated on six unseen datasets (e.g., PTB-XL, CPSC-2018, and Chapman-Shaoxing-Ningbo) using dataset-specific prompts (Section 3.3). Ablation studies assess the impact of different ECG/text encoders and the CFN module as well as other key procedures. Mainstream eSSLs are benchmarked with linear probing by freezing ECG encoders and fine-tuning a linear layer on 1%, 10%, and 100% of labeled data from the six datasets. All tasks are evaluated by average AUC across classes and datasets, following the data splits in Appendix A.7.1. Hyperparameters are provided in Appendix A.7.2 and overlap analysis in Appendix A.6.2.

[2] Used offline for one-time inference only; not required during deployment. A large LLM ensures high-quality labels. (Appendix A.10.1)

[3] Verified by cardiologists with 10+ years of experience; Appendix A.10.3.

[4] Compact pre-trained checkpoint (ViT-tiny + frozen MedCPT) runs on single GPU with $\geq$24GB memory, making it deployable in clinical or low-resource settings (Appendix A.10.2).

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 and 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
SuPreME (Ours)	*Z & L*	**77.20**	63.24	72.34	**84.48**

Table 1: Performance of SuPreME and eSSLs, with '*Z*' for zero-shot and '*L*' for linear probing. Best results are **bolded** and second best gray -flagged.

Dataset	*Linear Classification* ResNet	ViT	*Cardiac Fusion Network* ResNet	ViT
PTB-XL-Superclass	67.55	66.80	68.75	**78.20**
PTB-XL-Subclass	73.77	71.51	68.02	**77.52**
PTB-XL-Form	**64.34**	62.10	58.85	60.67
PTB-XL-Rhythm	75.68	75.34	68.69	**86.79**
CPSC-2018	**83.35**	79.13	60.38	79.83
CSN	72.61	72.32	65.07	**80.17**
Overall	72.88	71.23	64.96	**77.20**

Table 2: Performance of SuPreME and its variants on downstream datasets. Best results are **bolded**.

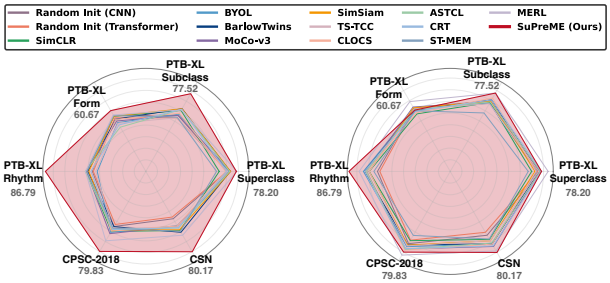


Figure 4: Comparison of SuPreME (zero-shot) and eSSLs (linear probing with 1% data on the left and 10% data on the right) across downstream datasets.

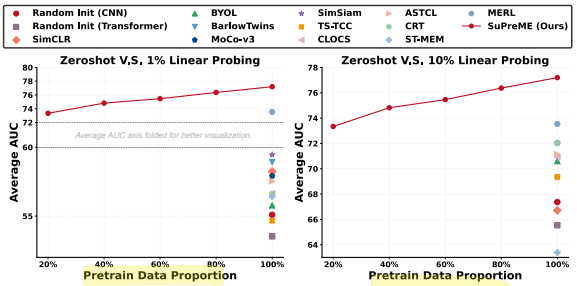


Figure 5: Data efficiency of SuPreME (zero-shot) and eSSLs (MERL in zero-shot, others in linear probing).

Framework	PTB-XL-Superclass 1%	10%	100%	PTB-XL-Subclass 1%	10%	100%	PTB-XL-Form 1%	10%	100%	PTB-XL-Rhythm 1%	10%	100%	CPSC-2018 1%	10%	100%	CSN 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
SuPreME (Ours)	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**

Table 3: Specific linear probing performance of SuPreME and eSSLs across six downstream datasets. Best results are **bolded** and second best gray -flagged.

4.2 Evaluation with Mainstream eSSLs

We evaluate SuPreME against mainstream eSSL frameworks across 106 classes in six downstream ECG datasets, conducting linear probing with eSSL ECG encoders across varying data proportions to facilitate performance comparison. Table 1 demonstrates AUC results of SuPreME and eSSLs under different evaluation approaches.

Our results demonstrate that SuPreME achieves superior performance compared to traditional eSSL frameworks. With an overall zero-shot AUC of 77.20% (Details in Appendix A.8), SuPreME outperforms all non-multimodal eSSLs, which require linear probing even with 1% (best: 59.46%) or 10% (best: 72.04%) of labeled data, showcasing its strong generalization capabilities and efficient utilization of pre-trained knowledge. Even without the CFN module, SuPreME remains highly competitive (Table 3) using only the pre-trained ECG encoder for linear probing. Its overall performance consistently surpasses non-multimodal eSSL models across 1%, 10%, and 100% labeled data (also outperforms in 15/18 tasks with different datasets and labelled data portion), and achieves comparable performance to multimodal contrastive learning frameworks like MERL (ViT backbone with explicit contrastive objectives, more comparisons in

Appendix A.10.4).

Since SuPreME's ECG backbone is optimized jointly with the CFN rather than via an explicit contrastive loss, part of the diagnostic knowledge is encoded within cross-modal interactions. Linear probing on the ECG encoder alone thus cannot fully utilize the rich alignment captured during pre-training. This explains why the full SuPreME pipeline including query prompts and CFN achieves stronger zero-shot performance, even compared to linear probing with more labeled data. We highlight zero-shot performance as our main evaluation objective, as it aligns with real-world clinical settings where labeled ECG data is scarce and fine-tuning is often impractical.

Figure 4 presents framework performance across individual datasets. SuPreME's advantage on the PTB-XL-Superclass dataset is minimal, likely due to the dataset's simplicity, as it includes only 5 broad cardiac condition labels (e.g., `NORM`, `STTC`, `MI`), making it difficult to differentiate model performance. All frameworks perform poorly on the PTB-XL-Form dataset, which focuses on 19 ECG waveform types that do not directly correspond to cardiac conditions, leading to ambiguous associations and reduced performance for all models.

To investigate SuPreME's sensitivity to pre-training scale, we evaluate its zero-shot performance under varying data proportions (Figure 5). SuPreME consistently improves with more data and maintains a clear advantage over non-multimodal eSSLs. Remarkably, with only 20% of pre-training data, SuPreME outperforms all non-multimodal eSSLs using 1% or 10% labeled data for linear probing, and matches the zero-shot performance of the multimodal baseline MERL trained on 100% of the pre-training data. Moreover, SuPreME's zero-shot performance with just 20% of data also exceeds MERL's linear probing result with 1% labels, highlighting its superior generalization and efficiency under limited supervision. Notably, SuPreME achieves these results with significantly fewer computational resources and shorter training time[5] (Appendix A.10.2).

4.3 Evaluation of SuPreME Architecture

Beyond comparisons with eSSLs, we assess the contribution of core components in SuPreME by varying its core modules, including the ECG backbone (ResNet vs. ViT) and (Linear vs. CFN) shown in Table 2 and Figure 6. Overall, SuPreME (ViT + CFN) achieves the highest average AUC of 77.20%, with strong results on PTB-XL-Rhythm (86.79%) and CSN (80.17%), demonstrating the effectiveness of cross-modal fusion for temporally and spatially complex signals.

Under linear classification, ResNet outperforms ViT across most datasets, reflecting its inductive bias toward local feature extraction. However, once CFN is introduced, ViT significantly benefits from its attention mechanisms and structured prompts, outperforming all other variants. This suggests that ViT's global receptive field aligns well with the query-driven fusion in CFN, while ResNet's local filters are less suited for attending over sparse textual queries.

Notably, the performance of ResNet + CFN is lower than ResNet + Linear across several datasets. We attribute this to a mismatch between ResNet's hierarchical, spatially localized features and CFN's attention-based fusion, which benefits more from globally contextualized inputs like ViT. CFN is designed to interpret semantically aligned queries over long-range dependencies, an area where ResNet lacks representational flexibility. This highlights the importance of matching the backbone's encoding characteristics with the fusion strategy. Details are in Appendix A.10.5.

Figure 7 further analyzes SuPreME's performance on individual cardiac conditions in PTB-XL-Subclass (Complete results in Appendix A.9). SuPreME consistently achieves high AUCs (many > 90), especially for nuanced conditions like LAFB/LPFB, CRBBB, CLBBB, and RVH, where both query semantics and signal patterns must be integrated. In contrast, ResNet + CFN underperforms in complex arrhythmias (e.g., ISCA, IRBBB), reinforcing our insight that strong multimodal fusion requires compatible encoders.

Unlike linear classifiers with fixed output dimensions, CFN enables flexible prompt-driven classification, aligning query semantics with signal patterns in a shared latent space as specified in Section 3.2, improving generalization to novel conditions without fine-tuning and demonstrates its strength especially when paired with ViT.

4.4 Ablation Analysis

Entity Extraction Model. Using Llama3.1-70B-Instruct for NER improves zero-shot AUC by 4.31% over the 8B variant (Table 4), reflecting bet-

[5] SuPreME: 4×A100-40GB GPUs for 16 epochs (∼90 minutes); MERL: 8×A100-40GB GPUs for 50 epochs (≥1 day).

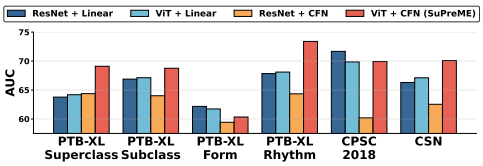


Figure 6: Specific zero-shot performance of SuPreME and its variants across downstream datasets.

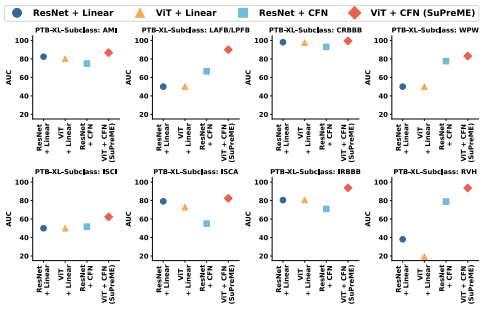


Figure 7: Specific zero-shot classification AUC performance of SuPreME and its variants on selected detailed categories in PTB-XL-Subclass.

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

Table 4: LLM for entity extraction.

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

Table 5: Entity deduplication.

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

Table 6: ECG backbone encoders.

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

Table 7: Language model encoders.

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

Table 8: Cardiac Fusion Network (CFN).

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**

Table 9: Zero-shot cardiac query prompts.

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

Table 10: Pre-training dropout ratios.

ter supervision quality. This step is offline and serves to build a high-quality labeled dataset, not for deployment.

Clinical Entity Mapping. Mapping to a standardized 295-term vocabulary improves zero-shot AUC from 65.94% to 77.20% (Table 5), likely by removing noise and resolving label redundancy to better represent distinct cardiac conditions.

ECG Encoder Backbone. Replacing ViT-tiny with ResNet18 drops zero-shot AUC by 12.24% (Table 6), suggesting ResNet is less effective at modeling long-range ECG dependencies than ViT (Appendix A.10.5).

Clinical Text Encoder. Among BioClinicalBERT, PubMedBERT, and MedCPT, the latter achieves the highest AUC, outperforming the others by over 14.25% (Table 7), likely due to its contrastive training objective, which better captures fine-grained clinical distinctions.

Cardiac Fusion Network Module. We compare CFN-based fusion with a simple linear projection (Table 8). CFN lifts the zero-shot AUC from 72.70% to 77.20%, highlighting the benefits of cross-attention in capturing multimodal synergies between ECG signals and text queries.

Customized Cardiac Prompts. Among three strategies (GPT-4o, detailed CKEPE, and simplified CKEPE), the simplified CKEPE achieves the best AUC (Table 9), with $\geq 8.04\%$ improvement, suggesting that concise, clinically focused prompts enhance alignment and reduce noise.

Dropout Ratio. In pre-training, we compare dropout rates of $\{0.05, 0.10, 0.15\}$. A rate of 0.10 yields the best AUC, striking a balance between regularization and signal retention (Table 10).

5 Conclusion

We present a novel LLM-based method for ECG clinical entity extraction and construct a high-quality labeled dataset from MIMIC-IV-ECG. Building on this, we propose SuPreME, a scalable supervised pre-training framework for multimodal ECG representation learning that fuses ECG signals with fine-grained, standardized medical terminologies rather than free-text reports. Its Cardiac Fusion Network (CFN) and simplified 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, SuPreME 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 multimodal learning, providing a strong basis for clinically oriented ECG representation learning.

Limitations

While SuPreME achieves strong zero-shot performance, several limitations remain. First, the clinical entity extraction relies on a large language model not fine-tuned for cardiology, which may miss rare or ambiguous terms and introduce noise. Second, SuPreME assumes generalization across clinical settings, but real-world data often involve device variability, demographic shifts, and class imbalance. Our experiments show lower performance on rare conditions, indicating sensitivity to distribution shift. Additionally, because the ECG encoder is trained jointly with CFN rather than via contrastive objectives, its features alone may not always outperform other baselines under linear probing. Lastly, most existing ECG baselines are single-modal and few of them support open zero-shot evaluation (e.g., MERL), underscoring the need for clinically motivated zero-shot benchmarks that better reflect practical deployment scenarios and support fairer comparison across methods.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. 2023. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*.

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

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

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

Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649.

Murside Degirmenci, Mehmet Akif Ozdemir, Elif Izci, and Aydin Akan. 2022. Arrhythmic heartbeat classification using 2d convolutional neural networks. *Irbm*, 43(5):422–433.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Benjamin Gow, Tom Pollard, Leslie A. Nathanson, Alistair Johnson, Benjamin Moody, Carla Fernandes, Natalie Greenbaum, Jonathan W. Waks, Parisa Eslami, Tiffany Carbonati, Anushka Chaudhari, Emily Herbst, Daniel Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. 2023. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0).

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. 2023. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78.

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

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

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

Mingfeng Jiang, Jiayan Gu, Yang Li, Bo Wei, Jucheng Zhang, Zhikang Wang, and Ling Xia. 2021. Hadln: hybrid attention-based deep learning network for automated arrhythmia classification. *Frontiers in Physiology*, 12:683025.

Jiarui Jin, Haoyu Wang, Jun Li, Sichao Huang, Jiahui Pan, and Shenda Hong. 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*.

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

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

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

Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2023a. 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*, pages 637–647. Springer.

Che Liu, Cheng Ouyang, Sibo Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2023b. G2d: From global to dense radiography representation learning via vision-language pre-training. *arXiv preprint arXiv:2312.01522*.

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

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. 2018. 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.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. 2023c. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, and Dahua Lin. 2023d. Improving pixel-based mim by reducing wasted modeling capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5361–5372.

Fazla Rabbi Mashrur, Amit Dutta Roy, and Dabasish Kumar Saha. 2019. Automatic identification of arrhythmia from ecg using alexnet convolutional neural network. In *2019 4th international conference on electrical information and communication technology (EICT)*, pages 1–5. IEEE.

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

Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Vij, and Jonathan Rubin. 2020. A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, pages 1–4. IEEE.

Paul Rubel, Danilo Pani, Alois Schloegl, Jocelyne Fayn, Fabio Badilini, Peter W Macfarlane, and Alpo Varri. 2016. Scp-ecg v3. 0: An enhanced standard communication protocol for computer-assisted electrocardiography. In *2016 Computing in Cardiology Conference (CinC)*, pages 309–312. IEEE.

Shinnosuke Sawano, Satoshi Kodera, Hirotoshi Takeuchi, Issei Sukeda, Susumu Katsushika, and Issei Komuro. 2022. 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*.

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

Huruy Tesfai, Hani Saleh, Mahmoud Al-Qutayri, Moath B Mohammad, Temesghen Tekeste, Ahsan Khandoker, and Baker Mohammad. 2022. Lightweight shufflenet based cnn for arrhythmia classification. *IEEE Access*, 10:111842–111854.

Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. 2024. Foundation model of ecg diagnosis: Diagnostics and explanations of any form and rhythm on ecg. *Cell Reports Medicine*, 5(12).

Ashish Vaswani. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2024. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36.

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

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383.

Kai Yang, Massimo Hong, Jiahuan Zhang, Yizhen Luo, Yuan Su, Ou Zhang, Xiaomao Yu, Jiawen Zhou, Liuqing Yang, Mu Qian, et al. Ecg-lm: Understanding electrocardiogram with large language model. *Health Data Science*.

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

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

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

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542.

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

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. 2020. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1):2898.

A Appendix

A.1 Dataset and Model Overview

A.1.1 Pre-training Dataset

MIMIC-IV-ECG. MIMIC-IV-ECG[6] 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.

A.1.2 Downstream Dataset

PTB-XL. PTB-XL[7] 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[8] 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[9] 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.

A.1.3 Llama3.1-70B-Instruct Model

Llama3.1-70B-Instruct[10] 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 4.4), we adopt Llama3.1-70B-Instruct to maximize annotation quality, particularly for downstream applications in clinical and low-resource settings that rely on precise structured supervision.

A.2 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

[6] MIMIC-IV-ECG is available at https://physionet.org/content/mimic-iv-ecg/1.0/.

[7] PTB-XL is available at https://physionet.org/content/ptb-xl/1.0.3/.

[8] CPSC-2018 is available at http://2018.icbeb.org/Challenge.html.

[9] Chapman-Shaoxing-Ningbo is available at https://physionet.org/content/ecg-arrhythmia/1.0.0/.

[10] Llama3.1-70B-Instruct is available at https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

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.

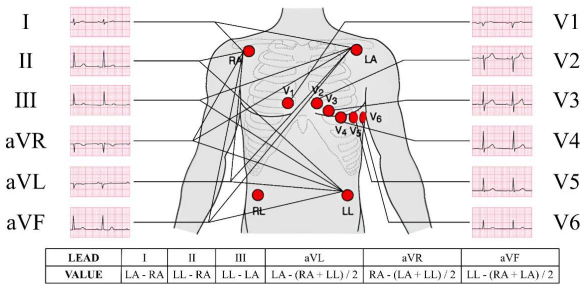


Figure 8: Standard 12-lead Electrocardiogram (ECG) showing 'sinus rhythm'.

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 8). 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.

A.3 Clinical NER Prompts, Statistics and Case Study

A.3.1 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"]}\"
```

A.3.2 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"]}\"
"""
```

A.3.3 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:

```
"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):

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)

Table 11: Example mapping from raw NER entities to standardized cardiac query labels.

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 11), these expressions are unified under concise, standardized queries aligned with SCP codes.

In Table 12 we show parts of the clustering and deduplication results on the pre-train dataset

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 (0.953), 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)

Table 12: Example clusters of raw NER entities mapped to standardized cardiac queries.

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.

A.4 Pre-training Framework Implementation

A.4.1 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 9.

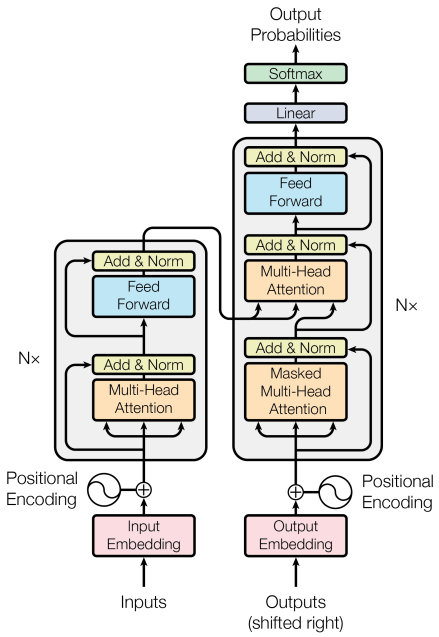


Figure 9: Encoder-decoder structure of Transformer, quoted from (Vaswani, 2017).

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:

$$\begin{aligned}\mathbf{Z}^{(k,1)} &= \mathbf{Z}^{(k-1)} + \text{Drop}(\mathbf{MHA}(\text{Norm}(\mathbf{Z}^{(k-1)})))\\ \mathbf{Z}^{(k,2)} &= \mathbf{Z}^{(k,1)} + \text{Drop}(\mathbf{FF}(\text{Norm}(\mathbf{Z}^{(k,1)})))\\ \mathbf{Z}^{\text{norm}} &= \text{Norm}(\mathbf{Z}^{(\text{final})})\end{aligned} \tag{6}$$

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 SuPreME.

A.4.2 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:

$$\begin{aligned}\mathbf{Emb}_{\text{ECG}}^{\text{hidden}} &= \mathbf{Z}_{\text{dropout}}\mathbf{W}_1 + \mathbf{b}_1\\ \mathbf{Emb}'^{\text{hidden}}_{\text{ECG}} &= \text{ReLU}(\mathbf{Emb}_{\text{ECG}}^{\text{hidden}})\\ \mathbf{F}_{\text{ECG}} &= \mathbf{Emb}'^{\text{hidden}}_{\text{ECG}}\mathbf{W}_2 + \mathbf{b}_2\end{aligned} \tag{7}$$

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 nonlinear 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).

A.4.3 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:

$$\begin{aligned}\mathbf{Emb}_{\text{Query}}^{\text{hidden}} &= \mathbf{E}_{\text{dropout}}\mathbf{W}_3 + \mathbf{b}_3\\ \mathbf{Emb}'^{\text{hidden}}_{\text{Query}} &= \text{GeLU}(\mathbf{Emb}_{\text{Query}}^{\text{hidden}})\\ \mathbf{F}_{\text{Query}} &= \mathbf{Emb}'^{\text{hidden}}_{\text{Query}}\mathbf{W}_4 + \mathbf{b}_4\end{aligned} \tag{8}$$

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.

A.4.4 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.

A.5 Zero-shot Evaluation Analysis

A.5.1 Specific 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.

A.5.2 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 3.3) 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.

A.5.3 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.:

$$\begin{aligned} \mathbf{TPR} &= \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \\ \mathbf{FPR} &= \frac{\mathbf{FP}}{\mathbf{FP} + \mathbf{TN}} \end{aligned} \tag{9}$$

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.

A.6 Statistics and Overlap Analysis

A.6.1 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 13). 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.

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 13: Statistics of unique cardiac Entities: Extraction, Filtering, Deduplication, and Mapping

Table 14 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 10. 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.

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%

Table 14: Clustering statistics of extracted cardiac entities on MedCPT embeddings

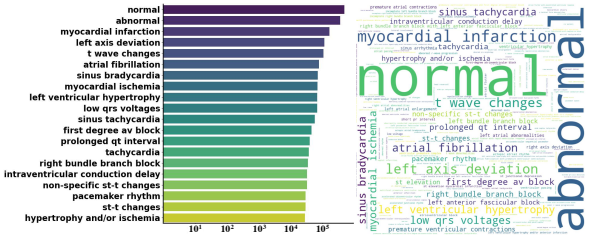


Figure 10: Frequency distribution of standardized ECG entities after deduplication and mapping in MIMIC-IV-ECG.

A.6.2 Dataset Overlap Analysis

We analyze the cardiac query overlap between the pre-training dataset and six downstream datasets specified in Section 4.1, as well as among the downstream datasets themselves, as illustrated in Figure 11. Specifically, we embed all entities from the pre-training 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.

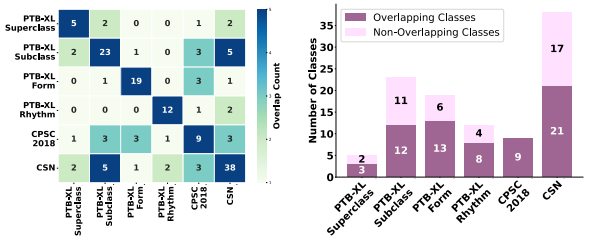


Figure 11: Overlap between ECG datasets, with left panel showing pairwise overlap counts between downstream datasets, and right panel showing the distribution of overlapping and non-overlapping classes between each downstream dataset and the pre-training dataset.

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-training dataset and the downstream datasets. Despite the pre-training dataset shares some similar queries, a substantial portion of queries remains unique to the downstream datasets, allowing the pre-training process to establish robust general-purpose representations while leaving room for downstream-specific adaptation.

Table 15 shows the overlap between entities from the pre-training dataset and cardiac queries from the downstream datasets, filtered using a cosine similarity threshold of 0.95.

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

Table 15: Overlap between pre-training dataset entities and downstream cardiac queries, filtered with similarity threshold of 0.95, sorted by similarity score.

A.7 Downstream Task Configuration

A.7.1 Downstream 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 16.

A.7.2 Downstream Experiment Configuration

The training configurations for downstream tasks, including optimizer, scheduler, and relevant hyperparameters, are detailed in Table 17.

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 16: Data splits and sample distribution for downstream datasets.

A.8 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 18 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.

A.9 Performance on Specific Cardiac Conditions

PTB-XL-Superclass. Figure 12 records the AUC performance of SuPreME on specific cardiac conditions in PTB-XL-Superclass dataset.

PTB-XL-Subclass. Figure 13 records the AUC performance of SuPreME on specific cardiac conditions in PTB-XL-Subclass dataset.

PTB-XL-Form. Figure 14 records the AUC performance of SuPreME on specific cardiac conditions in PTB-XL-Form dataset.

PTB-XL-Rhythm. Figure 15 records the AUC performance of SuPreME on specific cardiac conditions in PTB-XL-Rhythm dataset.

CPSC-2018. Figure 16 records the AUC performance of SuPreME on specific cardiac conditions in CPSC-2018 dataset.

CSN. Figure 17 records the AUC performance of SuPreME on specific cardiac conditions in CSN dataset.

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 Annealing	Cosine Annealing	Cosine Annealing	Cosine Annealing	Cosine Annealing	Cosine Annealing
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 17: Downstream dataset information and split proportions.

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

Table 18: Zero-shot AUC on downstream datasets using only non-overlapping (unseen) classes vs. using all classes.

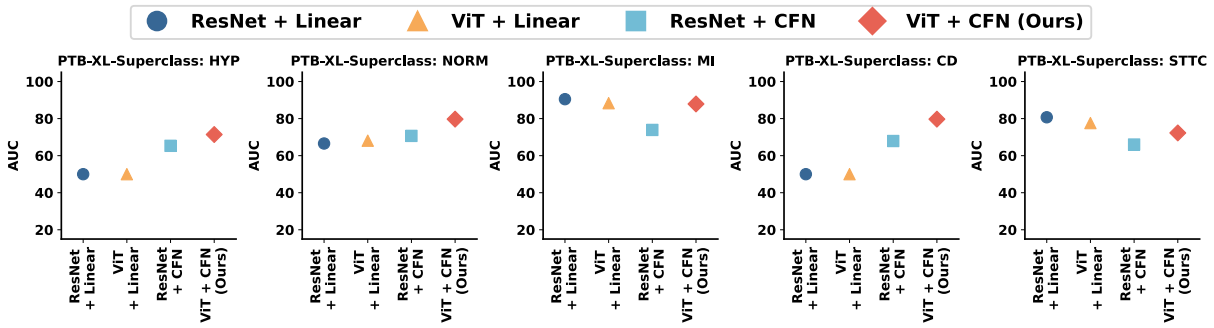
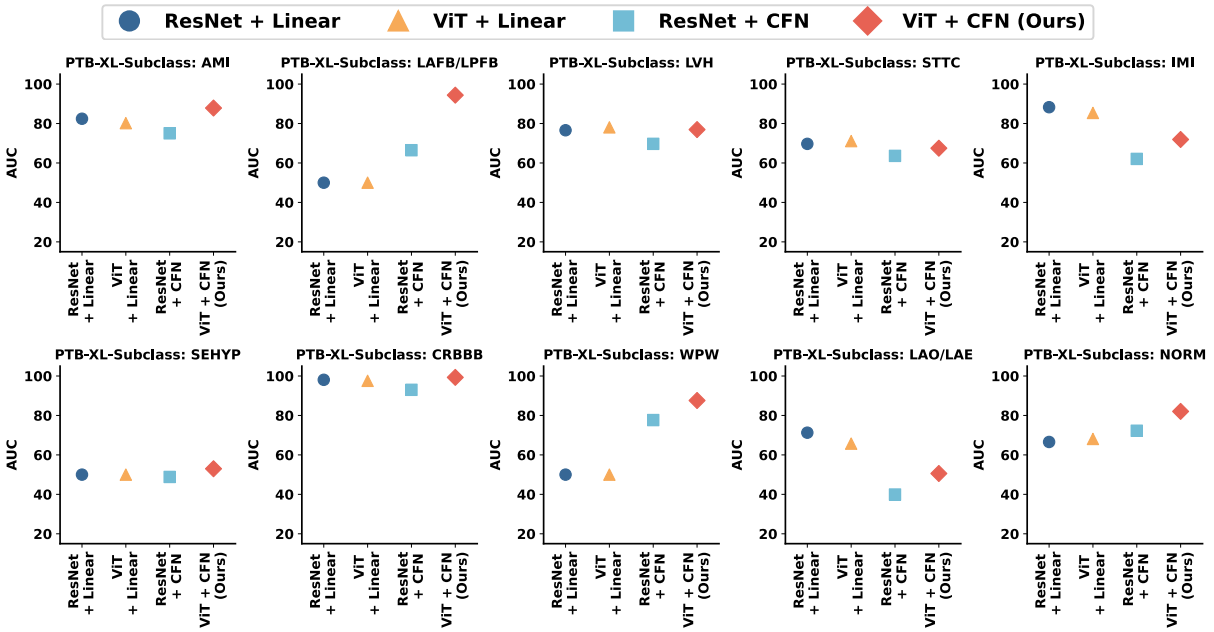


Figure 12: Zero-shot learning performance of SuPreME and its variants on specific categories in PTB-XL-Superclass.



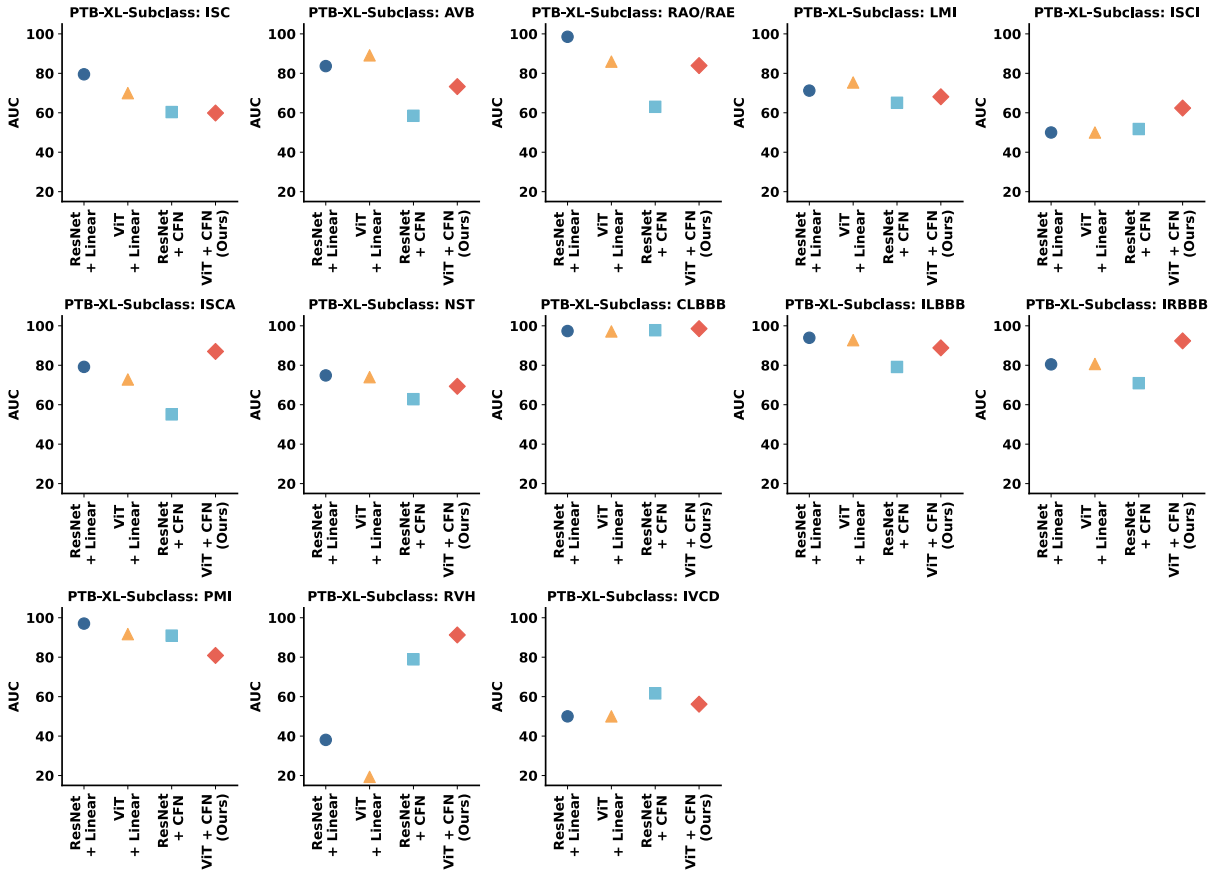
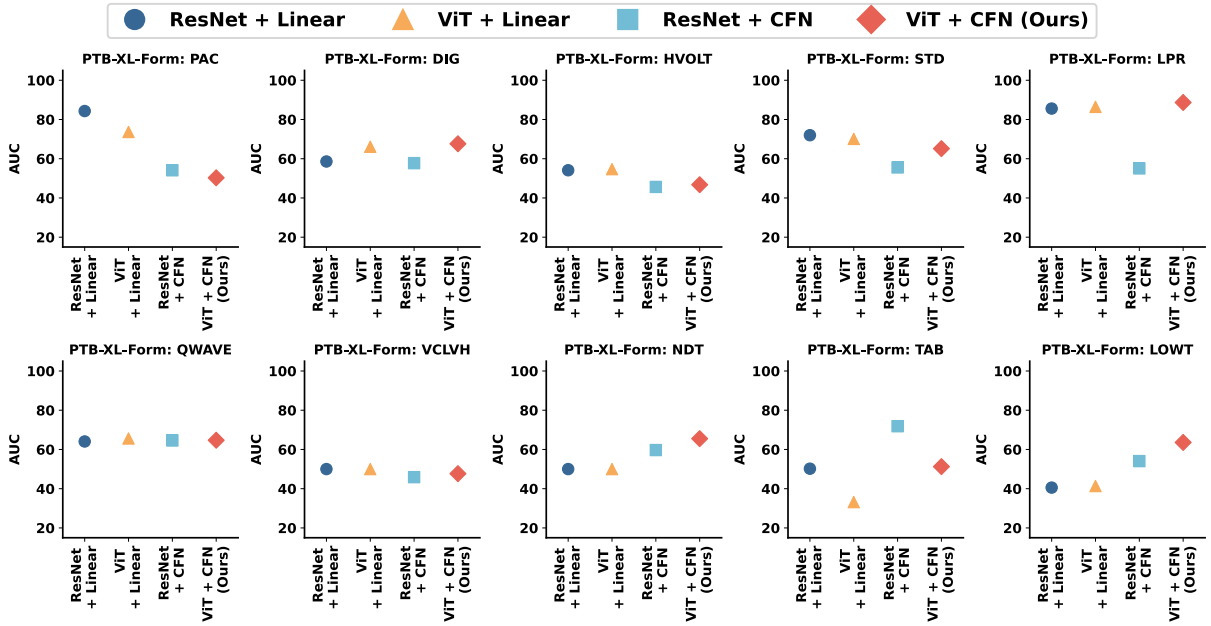


Figure 13: Zero-shot learning performance of SuPreME and its variants on specific categories in PTB-XL-Subclass.



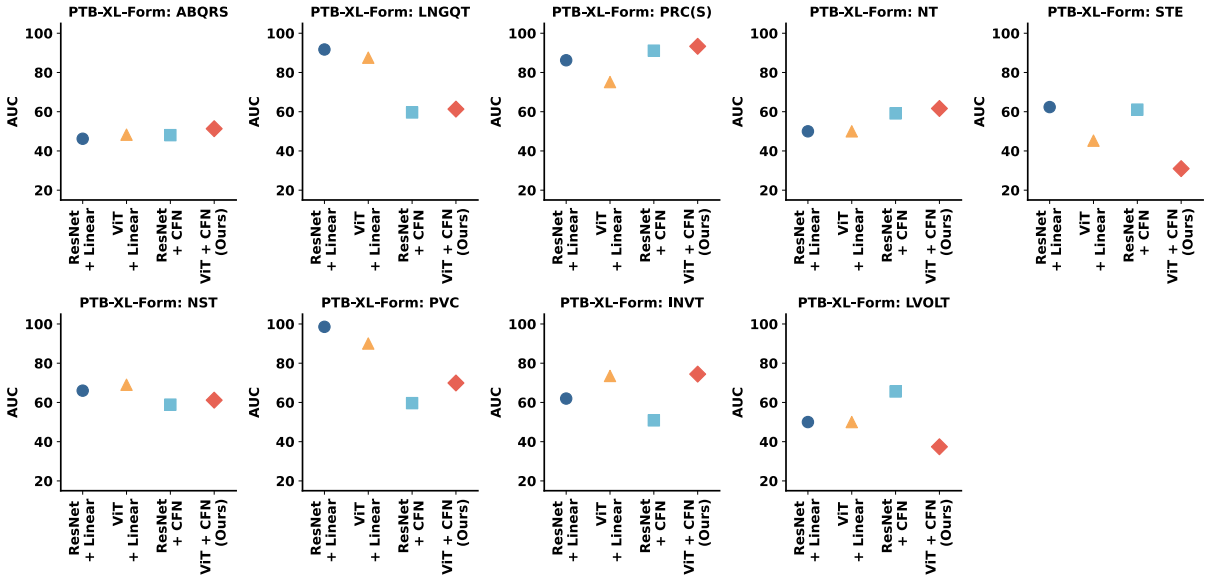


Figure 14: Zero-shot learning performance of SuPreME and its variants on specific categories in PTB-XL-Form.

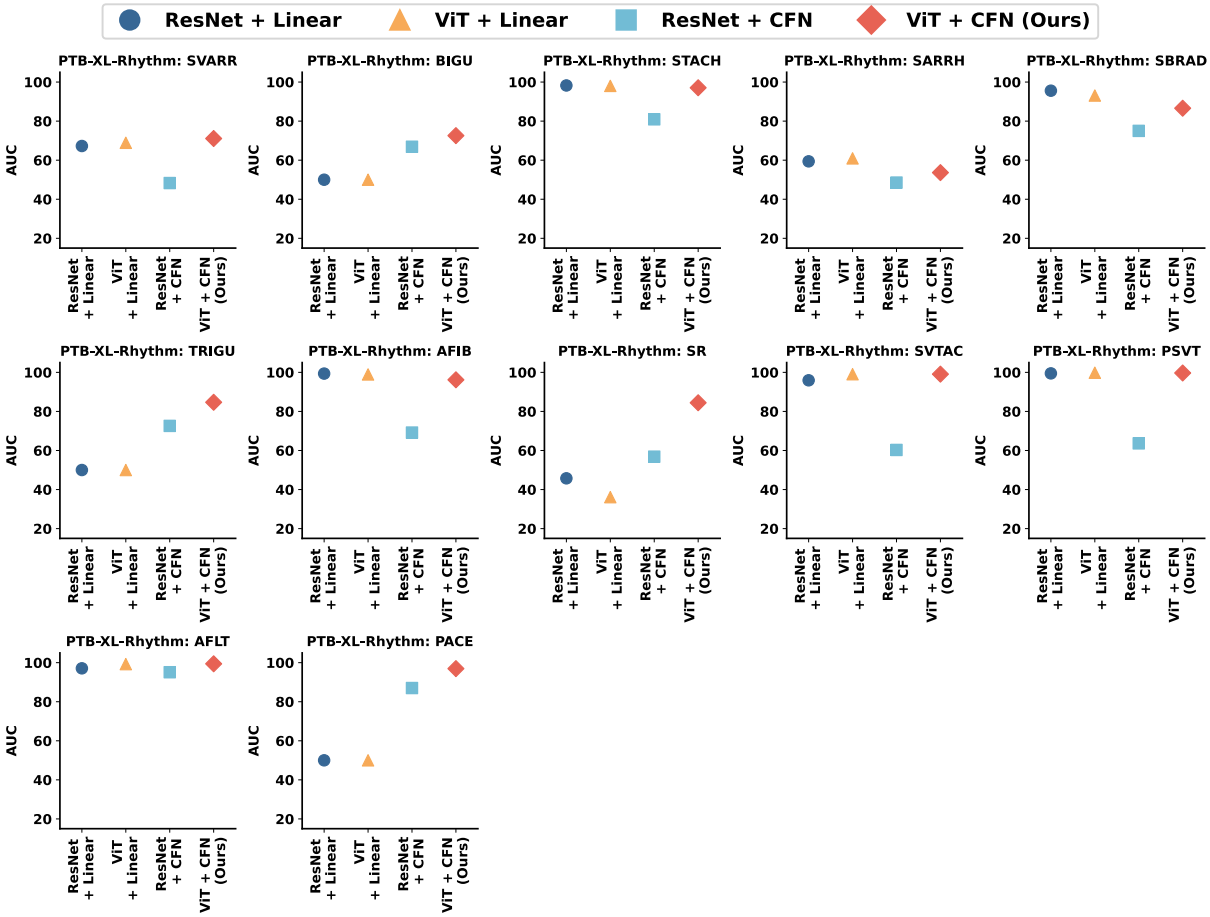


Figure 15: Zero-shot learning performance of SuPreME and its variants on specific categories in PTB-XL-Rhythm.

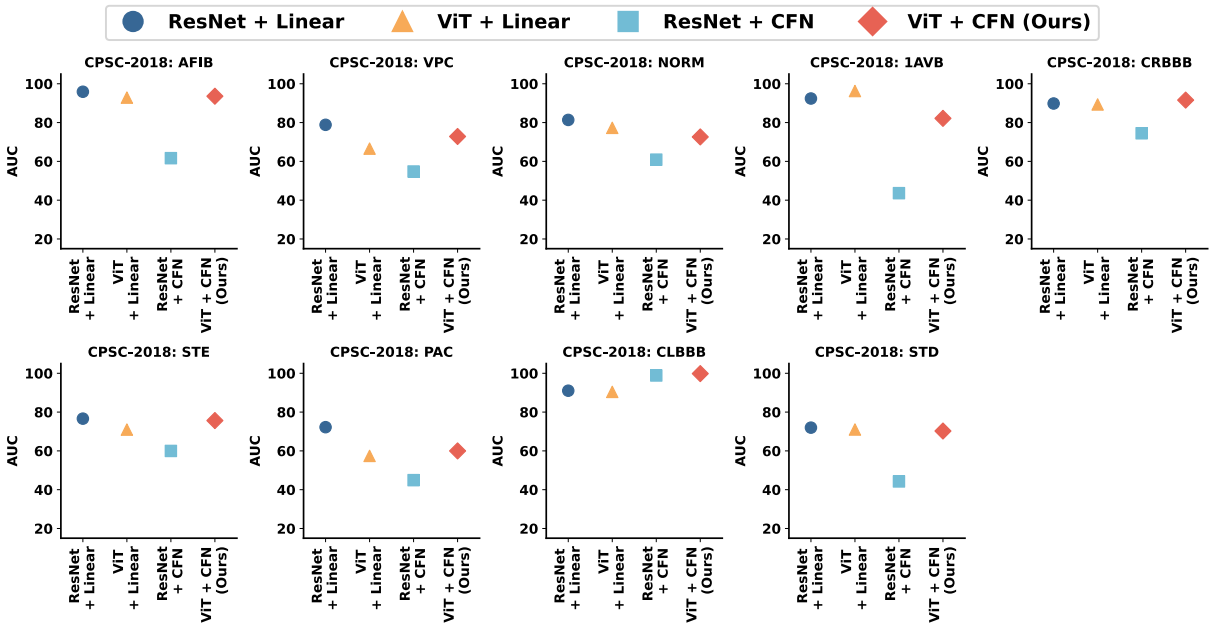
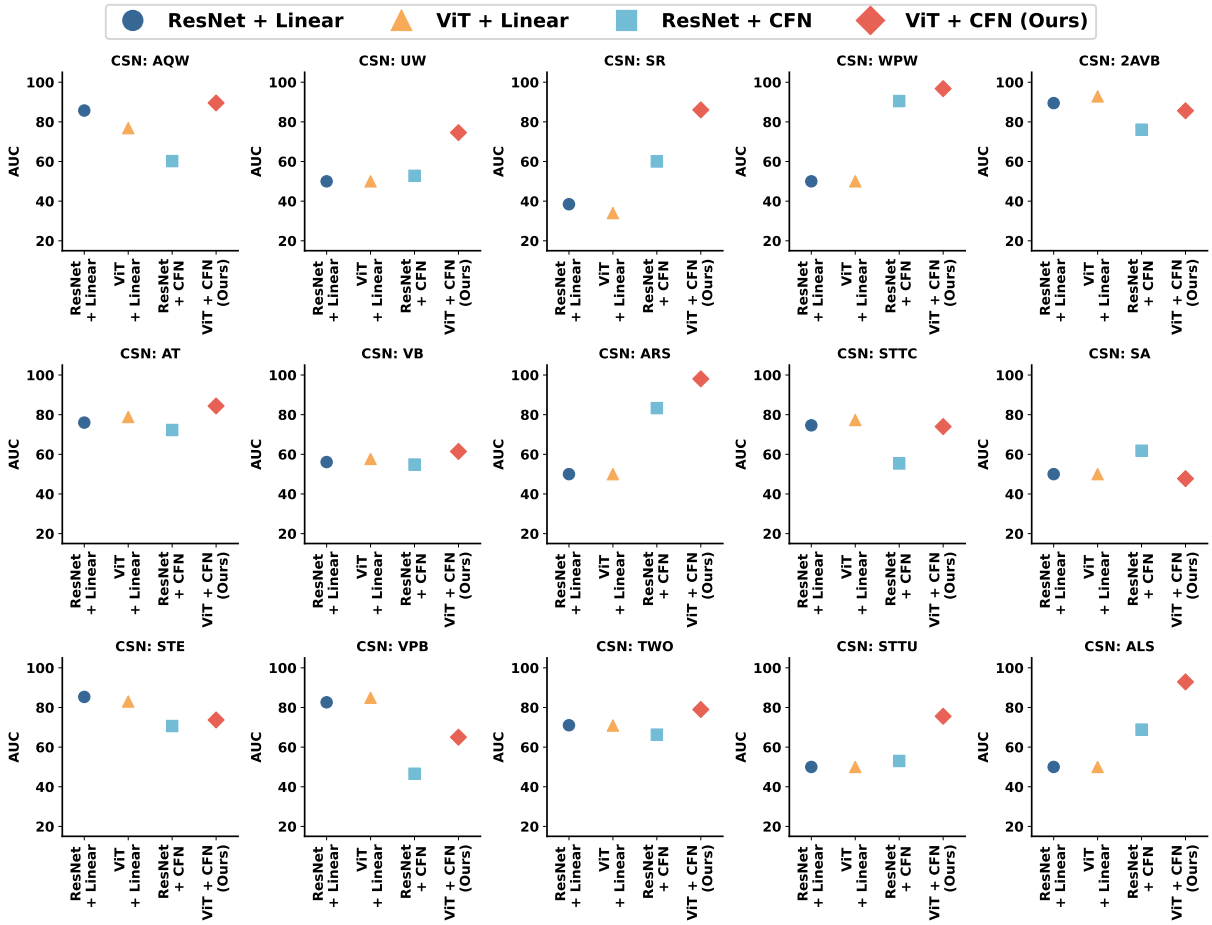


Figure 16: Zero-shot learning performance of SuPreME and its variants on specific categories in CPSC-2018.



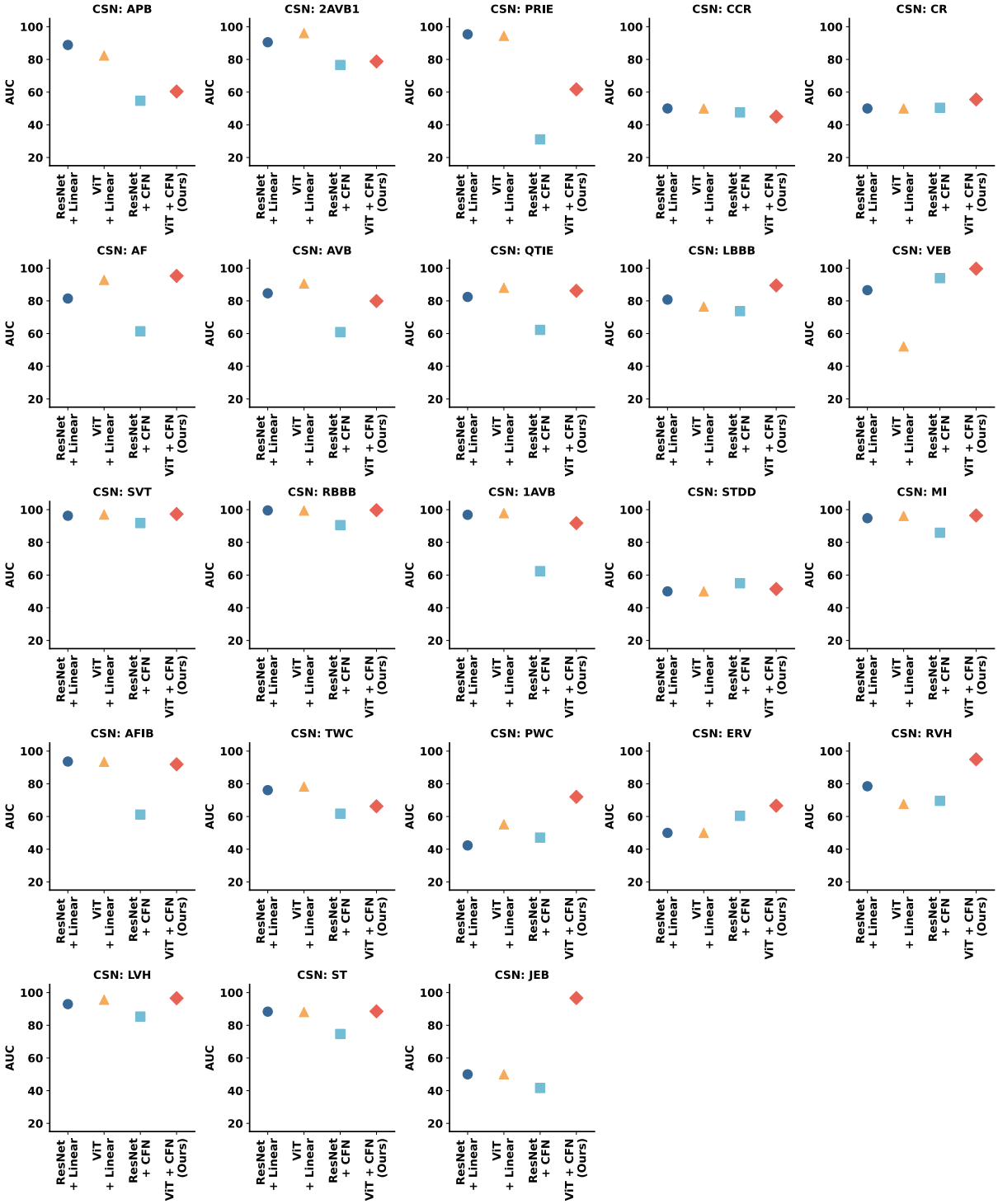


Figure 17: Zero-shot learning performance of SuPreME and its variants on specific categories in CSN.

A.10 Further Discussion

A.10.1 Offline Use of Large-sized LLM for Clinical NER

While we employ LLaMA3.3-70B-Instruct for clinical entity extraction, this step is performed offline only once during dataset construction and is not part of the SuPreME model's pre-training & 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 SuPreME model itself remains lightweight during inference.

A.10.2 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 LLaMA3.3–70B-Instruct with structured prompts. This step is performed only once as described above before pre-training SuPreME model. The output is a cleaned, deduplicated dataset of standardized diagnostic labels, which serves as supervision 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.

SuPreME Pre-training. SuPreME itself 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, SuPreME 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 SuPreME 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.

A.10.3 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 19).

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

Table 19: Incorrect matching examples between standard terminology and report entities.

"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.

A.10.4 Comparing SuPreME with MERL

To assess the relative effectiveness of our supervised multimodal framework, we compare SuPreME against MERL (Liu et al., 2024), a recent multimodal contrastive learning baseline that utilizes clinical reports and enhanced prompt engineering (Figure 18). While MERL employs contrastive objectives and handcrafted prompts, SuPreME leverages fine-grained diagnostic supervision through LLM-extracted entities and multimodal fusion via the Cardiac Fusion Network (CFN).

As shown in Table 20, SuPreME achieves consistently higher AUCs across all but one dataset. On average, SuPreME improves zero-shot performance by 3.66% absolute. Statistical testing confirms this improvement is significant: a paired t-test across the six datasets yields $t = 3.51$, $p = 0.0171$.

Moreover, SuPreME demonstrates significantly lower performance variance across datasets. While MERL exhibits a standard deviation of 2.30,

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
SuPreME	**78.20**	**77.52**	60.67	**86.79**	**79.83**	**80.17**	**77.20**

Table 20: Zero-shot AUC comparison between SuPreME and MERL.

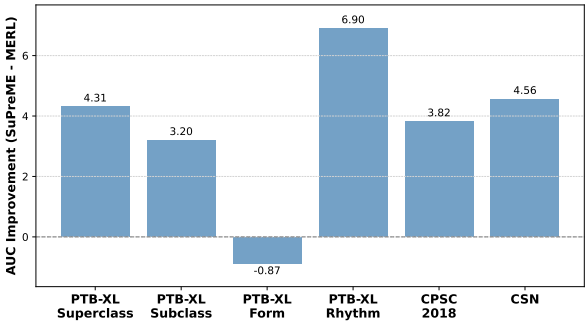


Figure 18: Per-dataset AUC difference between SuPreME and MERL.

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

Table 21: Average zero-shot AUC and standard deviation across six datasets.

SuPreME achieves a much smaller deviation of 0.21 (Table 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.

A.10.5 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 2), 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 19, 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.

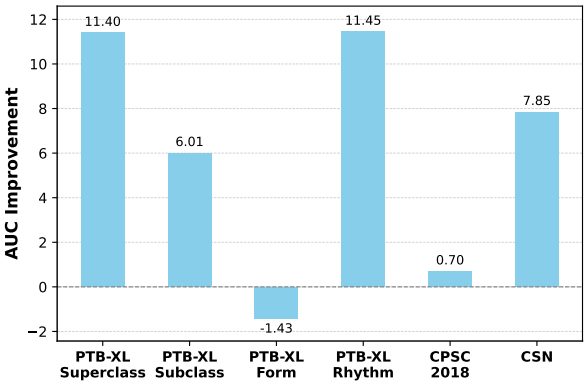


Figure 19: CFN vs. linear classification (Δ AUC) for ViT ECG backbone.

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 22, 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.

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 × Module	**290.65**	1	**6.60**	**0.018**
Residual	881.22	20	-	-

Table 22: Two-way ANOVA results on AUC with backbone and module as factors.

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.

A.10.6 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 SuPreME, 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.