

FOCAL: Fine-Grained Optimal-Transport-Driven Contrastive Alignment of Language and ECGs with Waveform Enhancement

Haitao Li^{1,2} Che Liu³ Zhengyao Ding¹ Ziyi Liu⁴ Wenqi Shao⁵ Zhengxing Huang^{1*}

¹Zhejiang University ²Shanghai Innovation Institute ³Imperial College London

⁴Transtek Medical Electronics Co., Ltd ⁵Shanghai AI Lab

lihaitao@zju.edu.cn, zhengxinghuang@zju.edu.cn

Abstract

Electrocardiograms (ECGs) are essential non-invasive tools for diagnosing cardiovascular diseases. While recent multimodal ECG-Report contrastive learning methods have shown promise for zero-shot ECG interpretation, they predominantly rely on global representations, failing to capture the fine-grained relationship between localized waveform patches and specific pathological tags. This limitation is further exacerbated by the fact that nearly 55% of standard clinical reports (e.g., in MIMIC-ECG) lack explicit waveform descriptions. In this paper, we propose FOCAL, a novel framework that achieves precise, fine-grained alignment between localized ECG segments and individual report tags via Optimal Transport. Furthermore, because fine-grained alignment at the tag level exacerbates the false negative problem among reports sharing common diagnoses, we introduce a semantic similarity matrix to guide the contrastive objective and correct misalignments. To address the scarcity of detailed waveform text, we introduce a coarse-to-fine enrichment pipeline that leverages Large Language Models (LLMs) to recover missing semantics, utilizing a coarse model verification step to rigorously filter out hallucinations. Extensive experiments across six datasets demonstrate that FOCAL establishes new state-of-the-art performance in zero-shot prediction and linear probing. Our code and data are available at: <https://github.com/CuCl-2/FOCAL>.

1 Introduction

Electrocardiograms (ECGs) are essential non-invasive tools for detecting cardiac rhythm disorders in clinical practice (Sahoo et al., 2020; Rath et al., 2021; Ayano et al., 2022; Li et al., 2026; Ding et al., 2025). Recently, self-supervised learning (SSL) has emerged as a promising paradigm

for ECG representation learning, alleviating the reliance on large-scale annotated data and expert knowledge. Existing ECG SSL approaches can be broadly categorized into comparative methods (Chen et al., 2020, 2021; Wang et al., 2023; Edele et al., 2021) and generative methods (Zhang et al., 2022a; Hu et al., 2023; Na et al., 2024; Zhang et al., 2022b). However, most of these methods still struggle with unseen classes in zero-shot scenarios.

Recent studies leverage multimodal ECG-report alignment to enable zero-shot prediction. While foundational works (Li et al., 2024; Lalam et al., 2023) directly map paired ECG signals to clinical texts, subsequent methods introduce sophisticated enhancements: MERL (Liu et al., 2024b) incorporates uni-modal alignment alongside LLM descriptive prompts, and ESI (Yu et al., 2024) integrates RAG (Ni et al., 2025; Gao et al., 2023) pipelines. Furthermore, MELP (Wang et al., 2025) captures hierarchical ECG structures via multi-scale supervision, while D-BETA (Hung et al., 2025) combines generative masked auto-encoders with discriminative contrastive learning.

Despite these advances, current methods predominantly focus on aligning global ECG representations with entire reports. This macroscopic approach neglects the fine-grained, localized relationship between specific pathological tags and their exact waveform manifestations. Achieving this fine-grained alignment presents three critical challenges. First, existing architectures rely on global embeddings that fail to capture patch-level ECG features and tag-specific textual semantics. This limitation is further compounded by the lack of explicit annotations mapping individual report tags to their corresponding ECG segments, making direct supervision for such localized alignment impossible. Second, performing alignment at the individual tag level inevitably exacerbates the false negative problem during contrastive learning, as different reports frequently share common tags. Fi-

* Corresponding author.

nally, nearly 55% of ECG reports in the MIMIC-ECG dataset lack explicit waveform descriptions. Recovering these missing features using LLMs is complicated by the inherent hallucination problem and the complex, non-bijective relationship between waveform manifestations and diagnostic outcomes (Jin, 2018).

To address these challenges, we propose **FOCAL** (Fine-Grained Optimal-Transport-Driven Contrastive Alignment of Language and ECGs). First, we formulate the alignment between independent report tags and ECG patches as an Optimal Transport (OT) problem. This strict, parameter-free routing ensures that each pathological tag is forcefully grounded to its most distinct localized waveform features. Second, we introduce a semantic similarity matrix to dynamically mitigate the false negative collisions that occur when processing flattened, tag-level contrastive objectives. Finally, we design a coarse-to-fine LLM enrichment pipeline to recover missing waveform descriptions. Because LLM outputs are susceptible to hallucinations and complex non-bijective mappings, we employ a coarse FOCAL model to rigorously validate the generated waveform features before incorporating them into the pre-training corpus.

We validate our proposed FOCAL framework on six downstream ECG multi-label classification datasets. The results demonstrate that FOCAL establishes new state-of-the-art performance in zero-shot prediction, linear probing, and cross-domain transfer learning. Overall, our main contributions are as follows:

- We propose FOCAL, an Optimal Transport-driven architecture that achieves precise, fine-grained alignment between specific report tags and localized ECG patches.
- We present a semantic similarity matrix to guide contrastive learning, actively identifying and mitigating the false negative misalignments exacerbated by tag-level alignment.
- We introduce a coarse-to-fine training process that leverages LLMs to recover missing waveform features and validates them with a coarse model, effectively filtering out hallucinations.
- Extensive experiments across six datasets demonstrate that FOCAL achieves state-of-the-art performance.

2 Related Work

ECG-Report Contrastive Learning Recent advancements in multimodal contrastive learning (Radford et al., 2021) have spurred significant efforts in aligning ECG signals with clinical texts. Foundational approaches (Li et al., 2024; Liu et al., 2024a; Lalam et al., 2023) learn representations by directly pulling paired ECGs and reports together while pushing unpaired ones apart. Building on this, MERL (Liu et al., 2024b) introduces unimodal alignment and employs the CKEPE pipeline at inference to generate more descriptive prompts via LLMs. ESI (Yu et al., 2024) utilizes a retrieval-augmented generation (RAG) pipeline during training, integrating external medical knowledge to formulate more detailed descriptions. To capture the hierarchical nature of ECG signals, MELP (Wang et al., 2025) introduces multi-scale cross-modal supervision at the token, beat, and rhythm levels. Concurrently, D-BETA (Hung et al., 2025) proposes a hybrid framework integrating generative masked auto-encoders with boosted discriminative contrastive learning.

Despite these advances, most of the existing methods perform contrastive learning at the whole-ECG and report level, and overlook the absence of fine-grained waveform features in ECG reports.

Fine-Grained Contrastive Learning Recent studies in medical imaging have demonstrated the benefits of fine-grained alignment. Methods such as MedFILIP (Liang et al., 2025), fVLM (Shui et al., 2025) incorporate entity-level or region-level supervision by leveraging structured information extracted from reports or by segmenting images into anatomical regions. However, the application of fine-grained contrastive learning in the ECG domain remains largely unexplored.

False Negatives in Contrastive Learning Traditional contrastive learning (Radford et al., 2021) assumes all unpaired samples are strict negatives. This fails in the ECG domain, where frequent normal baselines and common diseases create inherent false negatives. Prior solutions involve adding regularization terms (Lavoie et al., 2024; Jiang et al., 2023; Li et al., 2023) or inter-report similarity matrices (Sun et al., 2023; Wang et al., 2022; Kim et al., 2025). In our framework, fine-grained alignment at the individual tag level significantly exacerbates this false negative problem since distinct reports frequently share identical diagnostic tags.

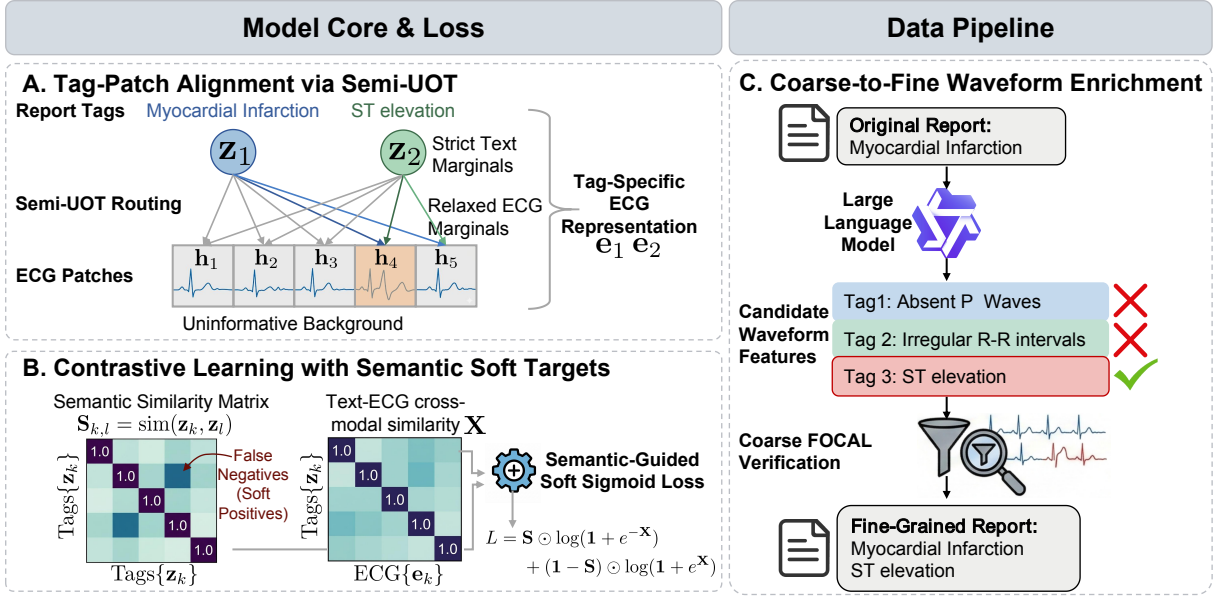


Figure 1: The overall framework of FOCAL. (A) **Tag-Patch Alignment via Semi-UOT**: Semi-Unbalanced Optimal Transport routes text tags to ECG patches, grounding all clinical semantics while dynamically ignoring uninformative background signals. (B) **Contrastive Learning with Semantic Soft Targets**: Integrating a text-derived semantic similarity matrix S into a soft sigmoid objective mitigates false negative bias by treating semantic identical tags across reports as soft positive targets. (C) **Coarse-to-Fine Waveform Enrichment**: To recover omitted clinical waveform features, an LLM generates candidates from the original report, which are rigorously verified by a Coarse FOCAL model to filter hallucinations and yield enriched reports.

3 Method

3.1 Fine-Grained Alignment Architecture

In this section, we describe how we achieve fine-grained alignment between specific ECG segments and tags in clinical reports. Given an ECG-report dataset $\mathcal{D} = \{(\mathbf{X}_i, \mathcal{R}_i)\}$, each report $\mathcal{R}_i$ consists of multiple clinically meaningful tags, such as *arrhythmia*, *myocardial infarction*, *atrial fibrillation*, etc. For each tag, our objective is to identify and align the corresponding ECG segment(s) that reflect this clinical finding.

Spatial-Temporal ECG Encoding. To capture both spatial relationships and temporal dynamics, we employ a Spatial-Temporal ViT encoder. Given a 12-lead ECG $\mathbf{X}_i \in \mathbb{R}^{L \times T}$ ($L = 12$, T is the signal length), we divide each lead into N_{time} non-overlapping patches. This yields a total of $N = L \times N_{time}$ patches, which are flattened into a 1D sequence. Each patch $\mathbf{x}_i^p$ ($p = 1, \dots, N$) is linearly projected to d dimensions. To preserve coordinate information, we add learnable lead and temporal embeddings to each projected patch:

$$\mathbf{u}_i^p = \text{Linear}(\mathbf{x}_i^p) + \mathbf{e}_{lead}^p + \mathbf{e}_{time}^p \quad (1)$$

where $\mathbf{e}_{lead}^p$ and $\mathbf{e}_{time}^p$ denote the specific lead and temporal embeddings corresponding to the p -th

patch. The token sequence $\mathbf{U}_i = [\mathbf{u}_i^1, \dots, \mathbf{u}_i^N]$ is then processed by the ECG encoder to yield the final fine-grained patch embeddings $\mathbf{H}_i = [\mathbf{h}_i^1, \dots, \mathbf{h}_i^N]$:

$$\mathbf{H}_i = \text{ECGEncoder}(\mathbf{U}_i) \in \mathbb{R}^{N \times d} \quad (2)$$

Report Tag-Level Encoding. Since ECG reports are highly structured, we can simply use commas to split the tags. We denote the set of tags in the i -th report as $\mathcal{T}_i = \{t_i^j\}_{j=1}^{m_i}$, where t_i^j represents the j -th tag in report i , and m_i is the number of tags in $\mathcal{R}_i$. Each tag t_i^j is then independently encoded via a text encoder to obtain its embedding representation. This yields the tag-level representation matrix for report i :

$$\mathbf{Z}_i = [\mathbf{z}_i^1, \dots, \mathbf{z}_i^{m_i}] \in \mathbb{R}^{m_i \times d} \quad (3)$$

$$\mathbf{z}_i^j = \text{TextEncoder}(t_i^j), \quad j = 1, \dots, m_i \quad (4)$$

Notably, both the ECG patch embeddings and tag embeddings are projected into the same latent space with dimension d . For clarity and conciseness, we omit explicit projection formulas in the equations.

Tag-Specific ECG Representation via Semi-Unbalanced Optimal Transport. With the fine-grained ECG patch embeddings and report tag

embeddings, our goal is to align them accurately. Standard Optimal Transport (OT) typically formulates such alignment by imposing strict, uniform marginal constraints on both modalities. However, this completely balanced formulation is ill-suited for the clinical reality of ECG-report pairs due to a structural asymmetry. On the textual side, every semantic tag extracted from the report contains critical diagnostic information and must be strictly grounded to the waveform. On the signal side, pathological manifestations in an ECG are highly localized and temporally sparse. Consequently, many background patches (e.g., normal baselines or regular heartbeats) carry no relevant diagnostic information. Forcing a strict uniform constraint on the ECG patches would erroneously compel the model to distribute the semantic mass of pathological tags onto these uninformative segments.

To resolve this asymmetry, we formulate the tag-patch alignment as a Semi-Unbalanced Optimal Transport (Semi-UOT) problem. Specifically, we treat the m_i tag embeddings $\mathbf{Z}_i$ and N ECG patch embeddings $\mathbf{H}_i$ as two discrete distributions. We first define a cost matrix $\mathbf{C} \in \mathbb{R}^{m_i \times N}$, where each element $C_{j,p}$ represents the cosine distance between the j -th tag and the p -th ECG patch:

$$C_{j,p} = 1 - \frac{(\mathbf{z}_i^j)^\top \mathbf{h}_i^p}{\|\mathbf{z}_i^j\|_2 \|\mathbf{h}_i^p\|_2} \quad (5)$$

Our objective is to find a transport plan $\mathbf{T} \in \mathbb{R}_+^{m_i \times N}$ that minimizes the overall transport cost. To accommodate the structural asymmetry, we strictly ensure every tag is grounded, but relax the constraint on ECG patches. To guarantee differentiability and computational efficiency, we introduce an entropic regularization term $\mathcal{H}(\mathbf{T})$ and a Kullback-Leibler (KL) divergence penalty, solving the following objective:

$$\mathbf{T}^* = \arg \min_{\mathbf{T} \mathbf{1}_N = \mathbf{a}} \left\{ \sum_{j=1}^{m_i} \sum_{p=1}^N T_{j,p} C_{j,p} - \epsilon \mathcal{H}(\mathbf{T}) + \tau \text{KL}(\mathbf{T}^\top \mathbf{1}_{m_i} \parallel \mathbf{b}) \right\} \quad (6)$$

where $\mathbf{a} = \frac{1}{m_i} \mathbf{1}_{m_i}$ and $\mathbf{b} = \frac{1}{N} \mathbf{1}_N$ represent the prior marginal distributions. The strict equality constraint $\mathbf{T} \mathbf{1}_N = \mathbf{a}$ ensures all clinical tags are fully assigned (the ‘‘balanced’’ aspect), while the KL divergence softens the marginal constraint on the ECG patches (the ‘‘unbalanced’’ aspect). The

hyperparameter τ controls this relaxation: a smaller τ allows the network to confidently assign near-zero mass to irrelevant ECG patches. This objective can be efficiently solved using the unbalanced Sinkhorn-Knopp algorithm.

Finally, to aggregate information from the most relevant ECG segments for each tag, we normalize the optimal transport plan such that each row sums to 1 (i.e., $\hat{T}_{j,p} = m_i \cdot T_{j,p}^*$). The Semi-UOT-routed, tag-specific ECG representation is then computed as a transport-weighted sum:

$$\mathbf{e}_i^j = \sum_{p=1}^N \hat{T}_{j,p} \cdot \mathbf{h}_i^p \quad (7)$$

By leveraging Semi-UOT, each tag forcefully captures its most distinct local ECG features while naturally filtering out noisy or irrelevant baselines. Importantly, because the iterative unbalanced Sinkhorn algorithm consists entirely of differentiable matrix operations, this module acts as a parameter-free layer. It can be seamlessly optimized end-to-end with the subsequent contrastive loss, allowing gradients to flow directly back to refine both the ECG and text encoders.

Fine-Grained Contrastive Learning with Semantic Similarity Matrix. To maximize data efficiency and construct a comprehensive negative sample pool, we leverage all available tags within a batch. For a batch of B reports, let $K = \sum_{i=1}^B m_i$ denote the total number of tags. We flatten these tags and their corresponding tag-specific ECG representations into two aligned sequences: $\{\mathbf{z}_k\}_{k=1}^K$ and $\{\mathbf{e}_k\}_{k=1}^K$.

Standard contrastive objectives, such as the pairwise Sigmoid loss (SigLIP) (Zhai et al., 2023), treat all non-paired samples as strict negatives. This hard-negative assumption is problematic in fine-grained medical alignment, as different reports frequently share identical pathological tags. Forcing the model to push apart these semantically equivalent but non-paired samples exacerbates the false negative problem.

To address this, we introduce a semantic similarity matrix $\mathbf{S} \in [0, 1]^{K \times K}$ computed from the text tag embeddings, where $\mathbf{S}_{k,l} = \text{sim}(\mathbf{z}_k, \mathbf{z}_l)$. Instead of using hard binary indicators, we replace them with this continuous semantic prior, formulating a

semantic-guided soft sigmoid Loss:

$$L = \frac{1}{K^2} \sum_{k=1}^K \sum_{l=1}^K \left[\mathbf{S}_{k,l} \log(1 + \exp(-x_{k,l})) + (1 - \mathbf{S}_{k,l}) \log(1 + \exp(x_{k,l})) \right] \quad (8)$$

where $x_{k,l} = t \cdot \text{sim}(\mathbf{e}_k, \mathbf{z}_l)$ is the temperature-scaled cosine similarity between the k -th ECG representation and the l -th text tag, and t is a learnable temperature parameter.

By integrating $\mathbf{S}$ directly into the binary cross-entropy formulation as soft targets, our model naturally mitigates the false negative bias. Semantically identical tags from different reports ($\mathbf{S}_{k,l} \approx 1$) act as positive soft-targets, elegantly eliminating the erroneous repulsive force at the gradient level. Through this unified process, pathological semantics are explicitly grounded to their most relevant ECG waveform manifestations.

3.2 Coarse-to-Fine Training Process

Fine-grained alignment relies on reports that include detailed waveform features. However, we observe that due to clinical habits, these intermediate diagnostic observations are frequently omitted from recorded reports. To recover this missing information, we query LLMs with the question, ‘*What waveform features are most likely to be present in electrocardiograms with these symptoms?*’ to identify potential overlooked characteristics. While this generates potential features, LLM outputs remain unreliable for two primary reasons. First, ECG waveform features and diagnoses are not in a one-to-one correspondence (Jin, 2018). A single disease may present different waveform characteristics across individuals, meaning doctors can infer a diagnosis from waveform features, but the reverse is not strictly possible. Second, the LLM’s output is inherently unstable due to hallucination issues (Huang et al., 2023; Günay et al., 2024).

To address these reliability challenges, we propose a novel coarse-to-fine training process, illustrated in Figure 1, consisting of three steps. We first train a coarse FOCAL model using contrastive learning on the original ECG-report pairs. Next, we validate the LLM-generated waveform features by computing their similarity to the actual ECG signal using this coarse model, selecting only high-confidence candidates for augmentation. Finally, these validated features are incorporated into the original report, and we continue training the coarse

model with this augmented data to obtain the final FOCAL model.

4 Experiments

4.1 Datasets

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

4.2 Implementation Details

Pre-training: We use a randomly initialized ViT (Dosovitskiy et al., 2020) with 120 non-overlapping patches for ECG encoding, and Bio-ClinicalBERT (Alsentzer et al., 2019) for text. Optimization is performed using AdamW (learning rate 2×10^{-5} , weight decay 1×10^{-4}) with a batch size of 100, a cosine annealing scheduler, and a 10% warmup. The temperature parameter t is initialized to $\log 10$. Coarse FOCAL is pre-trained for 10 epochs on original reports, followed by 3 epochs for fine FOCAL on fine-grained reports. Candidate waveform features are queried via Qwen3-8B (Yang et al., 2025) accelerated by vLLM (Kwon et al., 2023), and filtered using a 0.95 confidence threshold.

Downstream Tasks: For zero-shot classification, the entire model is frozen, and probabilities are computed via the similarity between ECG embeddings and category text prompts. For linear probing, the ECG encoder remains frozen while a newly initialized linear classifier is trained using 1%, 10%, and 100% data fractions. All task performances are measured by macro AUC.

4.3 Zero-Shot Ability

The zero-shot classification results across six datasets are presented in Table 1. Our proposed FOCAL framework consistently outperforms all state-of-the-art baselines, achieving the highest Macro AUC on every evaluated dataset and sub-

Table 1: Results of zero-shot classification.

Macro AUC	PTBXL-Super	PTBXL-Sub	PTBXL-Form	PTBXL-Rhythm	CPSC2018	CSN	Average
MERL (Liu et al., 2024b)	74.2	75.7	65.9	78.5	82.8	74.4	75.3
MELP (Wang et al., 2025)	76.2	81.2	69.1	85.4	84.2	77.6	79.0
D-BETA (Hung et al., 2025)	76.2	75.9	66.1	88.6	80.1	76.3	77.1
FOCAL(Ours)	77.7	83.5	70.3	90.8	85.0	79.2	81.1

Table 2: Results of Linear Evaluation. The best and second-best results are highlighted in bold and underlined.

Method	PTBXL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random Init	70.45	77.09	81.61	55.82	67.60	77.91	55.82	62.54	73.00	46.26	62.36	79.29	54.96	71.47	78.33	47.22	63.17	73.13
SimCLR	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
MERL	82.39	86.27	88.67	64.90	80.56	84.72	58.26	72.43	79.65	53.33	82.88	88.34	70.33	85.32	90.57	66.60	82.74	87.95
MELP	85.82	87.61	87.87	79.22	84.40	87.46	63.41	76.71	83.30	88.83	94.65	96.91	88.54	91.75	94.32	78.25	84.83	90.17
D-BETA	83.15	<u>88.36</u>	<u>90.11</u>	77.74	82.92	85.15	<u>70.10</u>	<u>78.91</u>	83.98	86.61	92.83	96.71	85.46	91.35	<u>94.92</u>	<u>80.04</u>	<u>87.36</u>	<u>90.71</u>
FOCAL	<u>85.50</u>	90.37	91.45	<u>78.24</u>	86.36	90.39	71.83	79.35	85.36	89.52	95.01	98.01	89.32	93.33	96.52	81.92	88.03	92.27

set. Notably, FOCAL achieves an average Macro AUC of 81.1%, surpassing the strongest baseline, MELP (Wang et al., 2025), by a significant margin of 2.1%.

A detailed examination reveals that FOCAL exhibits particularly strong performance on tasks requiring the detection of highly localized and complex waveform features, such as PTBXL-Rhythm and PTBXL-Form. This improvement directly validates our Optimal Transport-driven architecture, which accurately grounds specific pathological tags to precise ECG segments.

4.4 Linear Evaluation

We aim to evaluate the learned model transferability to downstream supervised tasks. We froze the ECG encoder and fine-tuned a randomly initialized linear classification head on the training data with binary cross-entropy loss. We compared a series of contrastive and generative self-supervised learning methods. Results in Table 2 show that FOCAL still achieves the best performances across all methods (Grill et al., 2020; Zbontar et al., 2021; Chen and He, 2021; Kiyasseh et al., 2021; Zhang et al., 2023) in most scenarios.

5 Analysis

In this section, we conduct a series of experiments to provide an in-depth analysis of FOCAL. All the reported metrics reflect the average zero-shot AUC across the six datasets described above.

5.1 Ablation Study

We conduct a comprehensive ablation study to validate FOCAL’s three core components, with results in Table 3.

Alignment Mechanisms. Removing fine-grained alignment entirely causes a severe performance drop. Substituting Semi-UOT with standard Cross-Attention also degrades performance, validating that OT enforces sparse, precise mapping over attention’s global smoothing. Furthermore, reverting to standard balanced OT (strict marginals) introduces severe noise by forcing uniform alignment mass onto uninformative background signals, highlighting the necessity of our semi-unbalanced relaxation for temporally sparse clinical data.

Loss Objective. Removing the semantic similarity matrix to revert to hard negative contrastive learning degrades performance. This confirms that strict binary pairings erroneously push apart semantically identical tags across different reports, a conflict effectively resolved by our semantic soft targets.

Coarse-to-Fine Pipeline. The Coarse FOCAL model trained solely on original reports underperforms, confirming the need for LLM waveform enrichment. Training the coarse model for 3 additional epochs without enrichment yielded no gains, indicating prior convergence and ruling out benefits from extended training alone. Additionally, training FOCAL from scratch on augmented reports achieves comparable performance but incurs

Table 3: Results of ablation study.

Model Setting	AUC
FOCAL (default)	81.1
<i>Alignment Mechanism Ablations</i>	
w/ Cross-Attention (instead of OT)	78.3
w/ Standard Balanced OT (Strict marginals)	79.2
w/o Fine-Grained Alignment (Global only)	76.0
<i>Loss Objective Ablations</i>	
w/o Semantic Similarity Matrix	77.6
<i>Training Pipeline Ablations</i>	
w/o LLM Waveform Enrichment (coarse FOCAL)	79.6
coarse FOCAL + 3 epochs (original reports)	79.5
FOCAL (trained from scratch)	81.1

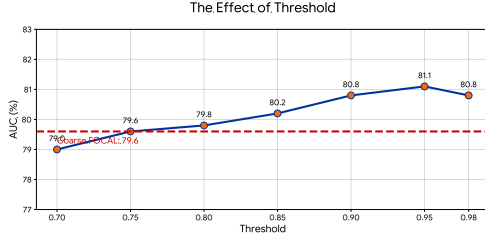


Figure 2: Effect of the filtering threshold

significantly higher computational costs, validating the efficiency of our curriculum.

Finally, Figure 2 demonstrates FOCAL’s robust performance across different LLM feature filtering thresholds. We retain a high default threshold of 0.95 to maximize precision and strictly filter out hallucinations.

5.2 Visual Activation Map

To further demonstrate the precise alignment between report tags and specific ECG segments, we present visual activation maps generated by FOCAL. We visualize the alignment weights for two highly localized pathological prompts in Figure 3. First, using the text prompt ‘RSR’ Pattern in V1-V3’, which is a key morphological feature for diagnosing right bundle branch block, FOCAL accurately localizes the corresponding waveform anomalies within the exact leads. Second, for the prompt ‘ST segment changes’, the model clearly highlights the specific temporal patches exhibiting abnormal elevations or depressions.

5.3 Validation and Transferability of Enriched Reports

To evaluate our waveform enrichment pipeline’s accuracy, we randomly sampled 200 MIMIC-ECG records originally lacking waveform descriptions. Three medical students independently annotated five key waveform features per ECG, establish-

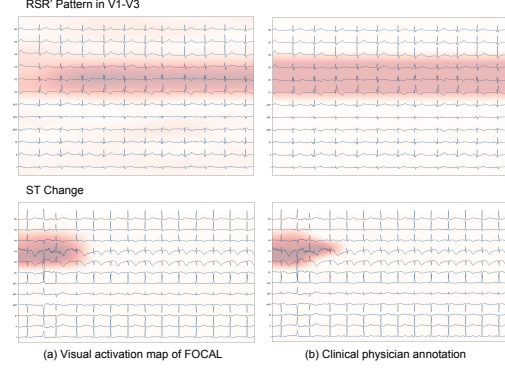


Figure 3: Visual activation maps generated by FOCAL.

Table 4: AUC of Generated Fine-Grained Reports (with/without verification).

Waveform Feature	w/o Veri.	w Veri.
Non-specific ST elevation	71.42	83.42
Long QT-interval	84.25	93.25
Abnormal QRS	76.68	91.68
Prolonged PR interval	75.13	86.13
Inverted T-waves	74.79	88.79

ing ground-truth labels via majority vote. To account for varying textual granularities, we calculated the AUC using the semantic similarity between the generated texts and target features. As Table 4 shows, unverified LLM-generated features yield low AUCs, confirming the risks of hallucinations and non-bijective clinical mappings. Conversely, verification by the Coarse FOCAL model substantially improves the AUC, demonstrating the pipeline’s reliability in filtering hallucinations to generate high-fidelity reports.

Beyond internal accuracy, our augmentation serves as a universally beneficial data engine. To prove model-agnostic transferability, we pre-trained the state-of-the-art MERL baseline (Liu et al., 2024b) using our enriched reports. Table 5 shows this verified corpus yields consistent zero-shot improvements across all six downstream datasets, achieving a +2.2% average gain over original texts. This plug-and-play capability confirms that recovering missing waveform semantics elevates overall pre-training corpus quality, extending its value beyond FOCAL to benefit the broader representation learning community.

5.4 Robustness to Data Distribution Shifts

To evaluate cross-domain generalizability, we tested models on unseen target domains following the experimental setting of MERL (Liu et al., 2024b) (Table 6). Unlike traditional uni-modal

Table 5: Impact of LLM-enriched reports on other baseline model (MERL).

Model Setting	PTBXL-Super	PTBXL-Sub	PTBXL-Form	PTBXL-Rhythm	CPSC2018	CSN	Average
MERL (Original Reports)	74.2	75.7	65.9	78.5	82.8	74.4	75.3
MERL (w/ Enriched Reports)	76.7	78.2	66.8	81.0	85.3	76.9	77.5
Gain (Δ)	+2.5	+2.5	+0.9	+2.5	+2.5	+2.5	+2.2

Table 6: Results under data distribution shift: ‘Source Domain’ denotes the dataset used for linear probing with the frozen pre-trained ECG encoder. ‘Target Domain’ refers to the corresponding test set. We include only those target domain samples that match categories from the source domain.

Source Domain	Zero-shot	Training Data	PTBXL-Super		CPSC2018		CSN	
Target Domain		Ratio	CPSC2018	CSN	PTBXL-Super	CSN	PTBXL-Super	CPSC2018
Random Init	$\times$	100%	68.62	75.31	55.74	68.92	56.57	61.16
SimCLR	$\times$	100%	69.62	73.05	56.65	66.36	59.74	62.11
BYOL	$\times$	100%	70.27	74.01	57.32	67.56	60.39	63.24
BarlowTwins	$\times$	100%	68.98	72.85	55.97	65.89	58.76	61.35
MoCo-v3	$\times$	100%	69.41	73.29	56.54	66.12	59.82	62.07
SimSiam	$\times$	100%	70.06	73.92	57.21	67.48	60.23	63.09
TS-TCC	$\times$	100%	71.32	75.16	58.47	68.34	61.55	64.48
CLOCS	$\times$	100%	68.79	72.64	55.86	65.73	58.69	61.27
ASTCL	$\times$	100%	69.23	73.18	56.61	66.27	59.74	62.12
CRT	$\times$	100%	70.15	74.08	57.39	67.62	60.48	63.33
ST-MEM	$\times$	100%	76.12	84.50	62.27	75.19	73.05	64.66
MERL	$\checkmark$	0%	88.21	78.01	76.77	76.56	74.15	<u>82.86</u>
MELP	$\checkmark$	0%	<u>87.75</u>	74.11	<u>77.89</u>	80.32	74.67	82.72
D-BETA	$\checkmark$	0%	72.09	79.11	77.12	<u>82.91</u>	<u>76.24</u>	80.10
FOCAL	$\checkmark$	0%	85.33	<u>82.98</u>	79.25	83.32	78.01	83.30

Table 7: Results of varying the LLM and Text Encoder components in our FOCAL framework.

Component Variant	Coarse FOCAL	FOCAL $\uparrow$
Different Large Language Models		
Qwen3-8B (Yang et al., 2025)	79.6	81.1
Qwen3-32B (Yang et al., 2025)	79.6	81.2
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	79.6	80.5
Gemma-3-12B-It (Team et al., 2024)	79.6	80.7
Meditron-7B (Chen et al., 2023)	79.6	81.0
medgemma-27b-text-it (Sjellergren et al., 2025)	79.6	81.4
Different Text Encoders		
BioClinicalBERT (Alsentzer et al., 2019)	79.6	81.1
PubMedBERT (Gu et al., 2021)	79.7	80.9
Med-CPT (Jin et al., 2023)	78.4	80.6
BioBERT (Lee et al., 2020)	78.6	80.5

SSL methods that require 100% of source domain labels for linear probing, FOCAL and other multi-modal baselines operate in a strict zero-shot setting. Despite this constraint, FOCAL achieves the best performance in four out of six cross-domain tasks and remains highly competitive in the others.

5.5 Robustness Across Different Components

We conducted experiments to evaluate the performance of our framework using various LLMs and text encoders. The results, presented in Table 7, demonstrate that FOCAL is highly robust across different configurations. Specifically, larger LLMs (Yang et al., 2025; Team et al., 2024; Chen et al., 2023; Sjellergren et al., 2025) yield only marginal gains, and domain-specific medical LLMs perform

comparably to general-purpose ones. This suggests that general LLMs already possess sufficient foundational medical knowledge for waveform recovery. More importantly, our coarse FOCAL verification acts as a rigorous quality filter, standardizing the enriched reports regardless of the LLM used. Furthermore, FOCAL consistently improves upon coarse baselines across all tested text encoders (Alsentzer et al., 2019; Gu et al., 2021; Jin et al., 2023; Lee et al., 2020), confirming the strong generalizability.

6 Conclusion

In this paper, we propose FOCAL, a novel framework for fine-grained ECG-text contrastive learning. By formulating cross-modal routing via Optimal Transport, FOCAL explicitly grounds pathological tags to localized ECG patches. To resolve false negative collisions inherent in tag-level alignment, we integrated a semantic similarity matrix via soft targets. Furthermore, our coarse-to-fine pipeline reliably recovers missing waveform semantics while filtering hallucinations. Extensive experiments demonstrate that FOCAL achieves state-of-the-art zero-shot performance, and our generated waveform-enriched reports provide a universally transferable data engine for the broader representation learning community.

Limitations

One primary limitation of our proposed framework is the computational and time overhead associated with the Large Language Model (LLM) waveform enrichment pipeline. Querying an LLM to generate potential waveform candidates for a large-scale dataset like MIMIC-ECG, which contains approximately 800,000 clinical reports, requires substantial inference compute. However, it is important to emphasize that this text augmentation process is strictly off-line and represents a one-time computational investment. Once the generated features are verified by the Coarse FOCAL model, the resulting fine-grained reports form a static, high-fidelity dataset. This enriched corpus not only facilitates the efficient training of our final model without adding any online training overhead, but it will also be open-sourced as a plug-and-play data asset. By providing this resource to the broader ECG representation learning community, the initial computational cost is effectively amortized over long-term, community-wide usage.

Acknowledgments

This work was partially supported by the Technical Innovation Key Project of Zhejiang Province (2025C01128), the National Natural Science Foundation of China (Grant Nos. 82272129, 82572380) and the Natural Science Foundation of Zhejiang Province (LHZSZ25H090002).

References

- Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.
- Yehualashet Megersa Ayano, Friedhelm Schwenker, Bisrat Derebssa Dufera, and Taye Girma Debelee. 2022. Interpretable machine learning techniques in ecg-based heart disease classification: a systematic review. *Diagnostics*, 13(1):111.
- 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.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Zhengyao Ding, Yujian Hu, Ziyu Li, Yiheng Mao, Haitao Li, Dongchen Zhou, Xuesen Chu, Long Yu, Ziyi Liu, Fei Wu, and 1 others. 2025. Ai modeling photoplethysmography to electrocardiography useful for predicting cardiovascular disease. *NPJ Digital Medicine*, 9(1):61.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 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*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).
- Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, and 1 others. Mimic-iv-ecg-diagnostic electrocardiogram matched subset.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, and 1 others. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

- Serkan Günay, Ahmet Öztürk, and Yavuz Yiğit. 2024. The accuracy of gemini, gpt-4, and gpt-4o in ecg analysis: A comparison with cardiologists and emergency medicine specialists. *The American journal of emergency medicine*, 84:68–73.
- Rui Hu, Jie Chen, and Li Zhou. 2023. Spatiotemporal self-supervised representation learning from multi-lead ecg signals. *Biomedical Signal Processing and Control*, 84:104772.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.
- Manh Pham Hung, Aaqib Saeed, and Dong Ma. 2025. Boosting masked ecg-text auto-encoders as discriminative learners. In *Forty-second International Conference on Machine Learning*.
- Chaoya Jiang, Wei Ye, Haiyang Xu, Shikun Zhang, Jie Zhang, Fei Huang, and 1 others. 2023. Vision language pre-training by contrastive learning with cross-modal similarity regulation. *arXiv preprint arXiv:2305.04474*.
- Jill Jin. 2018. Screening for cardiovascular disease risk with ecg. *Jama*, 319(22):2346–2346.
- 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.
- Myunsoo Kim, Seong-Woong Shim, and Byung-Jun Lee. 2025. Falcon: False-negative aware learning of contrastive negatives in vision-language pretraining. *arXiv preprint arXiv:2505.11192*.
- 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 E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, and 1 others. 2023. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*.
- Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wildon, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Haitao Li, Ziyu Li, Yiheng Mao, Ziyi Liu, Zhoujian Sun, and Zhengxing Huang. 2026. anyecg-chat: A generalist ecg-mlm for flexible ecg input and multi-task understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 597–605.
- Jun Li, Che Liu, Sibao 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.
- Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Zhongtian Du. 2023. Integrating language guidance into image-text matching for correcting false negatives. *IEEE Transactions on Multimedia*.
- Xinjie Liang, Xiangyu Li, Fanding Li, Jie Jiang, Qing Dong, Wei Wang, Kuanquan Wang, Suyu Dong, Gongning Luo, and Shuo Li. 2025. Medfilip: Medical fine-grained language-image pre-training. *IEEE Journal of Biomedical and Health Informatics*.
- Che Liu, Zhongwei Wan, Sibao Cheng, Mi Zhang, and Rossella Arcucci. 2024a. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8230–8234. IEEE.
- Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024b. 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, and 1 others. 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.
- 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*.
- Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, and 1 others. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Adyasha Rath, Debahuti Mishra, Ganapati Panda, and Suresh Chandra Satapathy. 2021. Heart disease detection using deep learning methods from imbalanced ecg samples. *Biomedical Signal Processing and Control*, 68:102820.
- S Sahoo, M Dash, S Behera, and S Sabut. 2020. Machine learning approach to detect cardiac arrhythmias in ecg signals: A survey. *Irbm*, 41(4):185–194.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and 1 others. 2025. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*.
- Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. 2023. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Patrick Wagner, Nils Strodthoff, Ralf-Dieter Boussejot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154.
- Fuying Wang, Jiacheng Xu, and Lequan Yu. 2025. From token to rhythm: A multi-scale approach for ecg-language pretraining. In *International Conference on Machine Learning*, pages 65059–65074. PMLR.
- 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*.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jiemeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Han Yu, Peikun Guo, and Akane Sano. 2024. Ecg semantic integrator (esi): A foundation ecg model pre-trained with llm-enhanced cardiological text. *arXiv preprint arXiv:2405.19366*.
- 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.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.
- Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. 2022a. Mae4e: 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. 2022b. Self-supervised time series representation learning via cross reconstruction transformer. *arXiv preprint arXiv:2205.09928*.
- Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. 2023. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*.
- J Zheng, H Guo, and H Chu. 2022. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022 Available online: <http://physionet.org/content/ecg-arrhythmia/1.0.0/>* (accessed on 23 November 2022).

A Dataset Analysis

We pre-train the FOCAL using the MIMIC-ECG dataset and test it on the PTB-XL, CPSC2018, and CSN datasets. All the ECGs in the datasets are 12-lead recordings. The PTB-XL dataset can be further divided into four subsets, and we follow the official train:validation:test split. For CPSC2018 and CSN, we split the dataset as 70%:10%:20% for the train:validation:test split. The statistics of the datasets used are presented in Table 8.

MIMIC-ECG The MIMIC-ECG dataset contains nearly 800,000 ECG-report pairs from approximately 160,000 unique patients. These diagnostic ECGs utilize 12 leads and are 10 seconds in duration, with a sampling rate of 500 Hz.

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

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

Chapman-Shaoxing-Ningbo (CSN) This dataset contains huge 12-lead ECGs with a 500 Hz sampling rate, which features multiple common rhythms and additional cardiovascular conditions, all labeled by professional experts.

Table 8: The statistics of used datasets.

Pretrain	# ECGs	# Reports			
MIMIC-ECG	773,268	773,268			
Evaluation	# Train	# Valid	# Test	# Classes	
PTBXL-Super	17,084	2,146	2,158	5	
PTBXL-Sub	17,084	2,146	2,158	23	
PTBXL-Form	7,197	901	880	19	
PTBXL-Rhythm	16,832	2,100	2,098	12	
CPSC2018	4,950	551	1,376	9	
CSN	16,546	1,860	4,620	38	

B Pseudo Code

The pseudo-code of our FOCAL training pipeline is summarized in Algorithm 1.

Algorithm 1 Coarse-to-Fine Training Pipeline of FOCAL

```

1: Input: Original dataset  $D = \{(x_{\text{ecg}_i}, x_{\text{txt}_i}) \mid i \in [0, n)\}$ , LLM prompt, threshold  $\tau$ 
2: Output: Optimized FOCAL model
3: # Step 1: Train Coarse Model
4: Train Coarse FOCAL model  $f_{\text{coarse}}$  using original dataset  $D$ 
5: # Step 2: LLM Waveform Enrichment & Verification
6: Initialize augmented dataset  $D_{\text{aug}} = \emptyset$ 
7: for  $i = 0$  to  $n - 1$  do
8:   features = LLM( $x_{\text{txt}_i}$ , prompt) # Generate candidate waveform features
9:    $x_{\text{txt}_i}^{\text{aug}} = x_{\text{txt}_i}$ 
10:  for each  $f_j \in \text{features}$  do
11:    sim =  $f_{\text{coarse}}(x_{\text{ecg}_i}, f_j)$  # Compute cross-modal similarity
12:    if sim >  $\tau$  then
13:       $x_{\text{txt}_i}^{\text{aug}} = x_{\text{txt}_i}^{\text{aug}} \cup \{f_j\}$  # Retain high-confidence feature
14:    end if
15:  end for
16:   $D_{\text{aug}} = D_{\text{aug}} \cup \{(x_{\text{ecg}_i}, x_{\text{txt}_i}^{\text{aug}})\}$ 
17: end for
18: # Step 3: Final FOCAL Training
19: Continue training  $f_{\text{coarse}}$  on augmented dataset  $D_{\text{aug}}$  to obtain final FOCAL

```

C Semantic Similarity Matrix

We visualize the semantic similarity matrix in Figure 4. The left side shows the semantic similarity matrix from a random batch. As illustrated, ECGs and tags from different records may share similarities to some extent. Ignoring these similarities would result in a diagonal matrix with ones on the diagonal and zeros elsewhere, which is obviously wrong. The right side displays a semantic similarity matrix where the first 16 entries are normal ECGs and the last 16 are abnormal ECGs. The matrix effectively captures the semantic similarities of the normal ECGs.

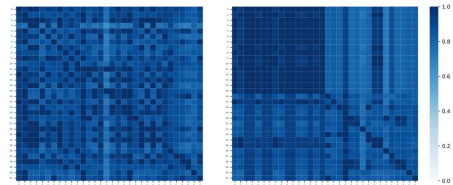


Figure 4: **The Heatmap of Semantic Similarity Matrix.** Left: a random batch; Right: with the first 16 as normal ECG and the last 16 as abnormal ECG.

D Stability of the Semantic Similarity Matrix

The semantic similarity matrix S is computed online from the text embeddings throughout the train-

Table 9: Relative Inference Time Comparison across baselines.

Method	LLM at Inference	Relative Time ↓	Notes
MERL (Liu et al., 2024b)	✓(CKEPE)	1.17×	LLM call required per sample at test time
MELP (Wang et al., 2025)	✗	1.08×	Overhead from multi-scale encoder
D-BETA (Hung et al., 2025)	✗	0.98×	MAE used only during training
FOCAL (Ours)	✗	1.00×	Sinkhorn-Knopp adds negligible overhead

Table 10: Full 2×2 Ablation of Fine-Grained Architecture and Waveform Enrichment (Zero-Shot Avg. AUC %).

Method	Fine-Grained Architecture	Waveform Enrichment	Avg. AUC (%)
Global only (w/o both)	✗	✗	74.8
w/o Fine-Grained Architecture [†]	✗	✓	76.0
Coarse FOCAL [†]	✓	✗	79.6
FOCAL[†]	✓	✓	81.1

ing process. A potential concern with online updates is whether the instability of the text encoder in the early stages of training could negatively affect the contrastive learning objective. In our framework, the text encoder is initialized from BioClinicalBERT (Alsentzer et al., 2019) weights, which are pre-trained on large-scale biomedical text. Consequently, these embeddings encode reliable semantic relationships among clinical terms from the very first training step. This robust initialization ensures that the online computation of $\mathbf{S}$ remains stable and consistently provides high-quality soft targets, preventing any training collapse or instability.

E Inference-Time Efficiency Analysis

To evaluate the computational overhead of our framework, we compare the relative inference time of FOCAL against key baselines in Table 9. We distinguish between training-time and inference-time costs. The LLM-based waveform enrichment in FOCAL is a strictly offline, one-time preprocessing step with zero online overhead. At inference time, the only addition is the unbalanced Sinkhorn-Knopp algorithm. Because this consists entirely of highly parallelizable matrix operations, its computational cost is negligible relative to standard neural network forward passes. As shown in Table 9, FOCAL is highly efficient, whereas methods like MERL (Liu et al., 2024b) require live LLM calls per sample via their CKEPE pipeline during inference.

F 2×2 Ablation of Fine-Grained Architecture and Enrichment

To further disentangle the performance gains contributed by our architectural design versus the data augmentation strategy, we provide a full 2×2 ablation study in Table 10. The results clearly demonstrate that the fine-grained Semi-UOT architecture is the dominant contributor to our framework’s success (yielding a 5.1% absolute improvement over the global baseline). Meanwhile, the LLM waveform enrichment provides a consistent additive gain (+1.5%) regardless of the underlying alignment configuration. This confirms that the improvements from the architecture and the enriched training text are orthogonal and complementary.

G Contribution of Alignment and Enrichment to Cross-Domain Transfer

To complement the evaluation and explicitly address how our core mechanisms affect generalization, we dissect the contributions of alignment and enrichment in the cross-domain transfer setting. Table 11 shows that the fine-grained OT-driven alignment is the primary driver of cross-domain transferability. Removing it causes a 5% $\sim$ 6% drop across all tasks. Text enrichment provides an additional 1% $\sim$ 2% gain by improving the quality of the pre-training corpus.

H Hyperparameter Sensitivity Analysis

We evaluate the robustness of our Semi-UOT formulation by varying its two core hyperparameters: the entropic regularization weight ϵ and the KL divergence penalty τ . As demonstrated in Table

Table 11: Ablation of Alignment and Enrichment on Cross-Domain Transfer (Zero-Shot AUC). "PTB" denotes PTBXL-Super.

Method	Alignment	Enrichment	PTB→CPSC	PTB→CSN	CPSC→PTB	CPSC→CSN	CSN→PTB	CSN→CPSC
w/o Alignment	✗	✓	79.51	76.83	73.21	77.61	72.53	77.42
Coarse FOCAL	✓	✗	83.52	81.24	77.82	81.91	76.44	81.83
FOCAL	✓	✓	85.33	82.98	79.25	83.32	78.01	83.30

12 and Table 13, FOCAL’s performance remains highly stable, fluctuating by less than 0.5% across a wide practical operating range. Extreme boundary cases predictably degrade performance: as $\epsilon \rightarrow 0$, the formulation approaches computationally unstable exact OT, and as $\tau \rightarrow \infty$, it reverts to standard balanced OT, which erroneously forces uniform alignment mass onto uninformative ECG baselines.

Table 12: Sensitivity to Entropic Regularization ϵ (with $\tau = 1.0$ fixed).

ϵ	0.01	0.05	0.1 (default)	0.2	0.5
Avg. AUC (%)	79.8	80.6	81.1	80.9	80.3

Table 13: Sensitivity to KL Divergence Penalty τ (with $\epsilon = 0.1$ fixed).

τ	0.1	0.5	1.0 (default)	2.0	5.0
Avg. AUC (%)	80.2	80.7	81.1	80.8	80.1

‘RSR’ Pattern, ‘Inverted T-waves,’ and ‘Low QRS voltages.’ Figure 8 shows the Top 3 retrieved ECGs with probabilities all greater than 0.99. Our model demonstrated strong capability in retrieving ECGs through waveform feature text, which can lead to two applications: (1) Helping doctors quickly find similar cases or specific ECG patterns, aiding in diagnosis and treatment decision-making; (2) In medical education, text-based retrieval can quickly find typical ECG cases, assisting in teaching and training, thereby improving educational effectiveness.

I Notation Table

Table 14 summarizes the key mathematical notations used in this paper.

J Running Cases for Generating Fine-Grained Reports

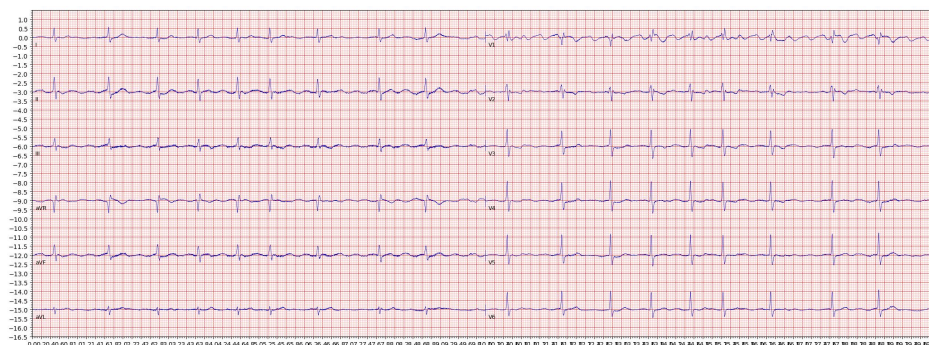
We present three case studies illustrating how the fine-grained reports with waveform features are generated step by step, as shown in Figure 5, 6, 7. The LLM is prompted with two explicit chain-of-thought instructions: ‘*What waveform features are most likely to be present in electrocardiograms with these symptoms?*’ followed by ‘*Organize these waveform features into a Python list, with each item representing a distinct waveform feature.*’

K ECG-Text Retrieval

We attempted to use FOCAL to retrieve electrocardiograms (ECGs) from the MIMIC-ECG dataset (Gow et al.) through text. To test our model’s ability to capture fine-grained waveform features, we tested a series of typical waveform features such as

Table 14: Summary of Key Notations in FOCAL

Symbol	Description
$\mathcal{D}$	The ECG-report multimodal dataset
$\mathbf{X}_i$	The i -th 12-lead ECG signal
$\mathcal{R}_i$	The clinical report paired with $\mathbf{X}_i$
L, T	Number of leads ($L = 12$) and temporal length of the ECG signal
N_{time}	Number of non-overlapping patches per lead
N	Total number of ECG patches ($L \times N_{time}$)
$\mathbf{x}_i^p$	The p -th flattened ECG patch
d	Dimension of the latent projection space
$\mathbf{e}_{lead}^p, \mathbf{e}_{time}^p$	Learnable lead and temporal embeddings for the p -th patch
$\mathbf{H}_i, \mathbf{h}_i^p$	Fine-grained patch embedding matrix and the p -th patch embedding
$\mathcal{T}_i$	Set of independent tags extracted from report $\mathcal{R}_i$
m_i	Number of clinical tags in report $\mathcal{R}_i$
$\mathbf{Z}_i, \mathbf{z}_i^j$	Tag-level representation matrix and the j -th tag embedding
$\mathbf{C}, C_{j,p}$	Cost matrix and the cosine distance between the j -th tag and p -th patch
$\mathbf{T}, \hat{T}_{j,p}$	Optimal transport plan and its normalized form
$\mathbf{a}, \mathbf{b}$	Prior marginal distributions for tags and ECG patches
ϵ, τ	Hyperparameters for entropic regularization and KL divergence penalty
$\mathbf{e}_i^j$	Tag-specific ECG representation routed via Semi-UOT
B, K	Batch size and the total number of tags in a batch
$\mathbf{S}$	Semantic similarity matrix computed from text tag embeddings
t	Learnable temperature parameter for contrastive learning



ID: p14578883/s43946433/43946433

Report: Atrial fibrillation., Right bundle branch block, Abnormal ECG

LLM Output:

Here is the list of waveform features organized into a Python list:

```
...
waveform_features = [
    "Irregularly irregular rhythm",
    "Absent or fragmented P-waves",
    "Wide QRS complexes (>120 ms)",
    "Variable RR intervals",
    "Fibrillatory waves (f-waves) or oscillations",
    "Wavy or sawtooth pattern",
    "Notched or slurred R wave in leads I, II, and V5-V6",
    "Inverted or absent R wave in lead V1",
    "Delta wave (slurred upstroke of the QRS complex) in leads I, II, and V5-V6",
    "RSR' pattern in leads V1-V3",
    "ST segment changes (elevation, depression, or inversion)",
    "T wave changes (inversion, flattening, or peaking)",
    "QT interval prolongation or shortening",
    "PR interval prolongation or shortening",
    "U waves or other abnormal waveforms"
]
...
```

Note that this list is not exhaustive, and there may be other waveform features that are not included here. Additionally, the presence of these features does not necessarily confirm a specific diagnosis, and further evaluation and testing may be necessary to establish a definitive diagnosis.

Potential Waveform features:

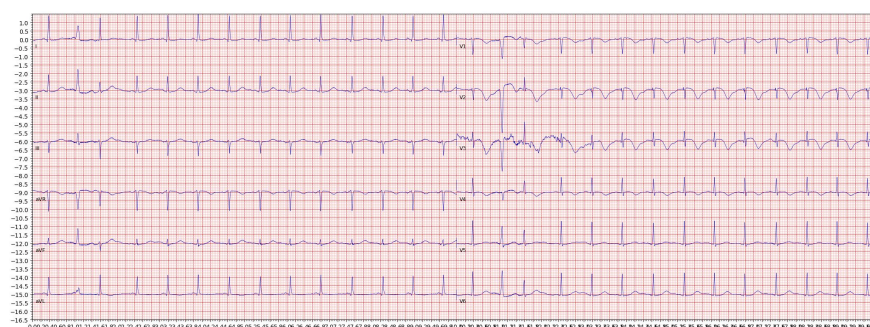
['Irregularly irregular rhythm', 'Absent or fragmented P-waves', 'Wide QRS complexes (>120 ms)', 'Variable RR intervals', 'Fibrillatory waves (f-waves) or oscillations', 'Wavy or sawtooth pattern', 'Notched or slurred R wave in leads I, II, and V5-V6', 'Inverted or absent R wave in lead V1', 'Delta wave (slurred upstroke of the QRS complex) in leads I, II, and V5-V6', 'RSR' pattern in leads V1-V3', 'ST segment changes (elevation, depression, or inversion)', 'T wave changes (inversion, flattening, or peaking)', 'QT interval prolongation or shortening', 'PR interval prolongation or shortening', 'U waves or other abnormal waveforms']

Probility:

[0.985297, 0.49500474, 0.95943975, 0.9952996, 0.96285915, 0.84171605, 0.88279337, 0.9455562, 0.98828167, 0.99366695, 0.62816966, 0.9148475, 0.99069655, 0.049276203, 0.9565827]

Fine-Grained Report: Atrial fibrillation., Right bundle branch block, Abnormal ECG, Irregularly irregular rhythm, Wide QRS complexes (>120 ms), Variable RR intervals, Fibrillatory waves (f-waves) or oscillations, Delta wave (slurred upstroke of the QRS complex) in leads I, II, and V5-V6, RSR' pattern in leads V1-V3, QT interval prolongation or shortening, U waves or other abnormal waveforms

Figure 5: Running Case1 with Atrial fibrillation and Right bundle branch block.



ID: p11799366/s45322712/45322712

Report: Probable atrial fibrillation, ventricular couplets, Long QTc interval, Ant/septal+lateral ST-T changes suggest myocardial infarction, Repolarization changes may be partly due to rhythm, Abnormal ECG

LLM Output:

Here is the list of waveform features organized in a Python list:

```
...
waveform_features = [
    "Irregularly irregular rhythm",
    "Absent or fragmented P waves",
    "Fibrillatory waves (f-waves) or oscillations in the atrial region",
    "Premature ventricular contractions (PVCs) with a different morphology from the normal QRS complex",
    "Ventricular couplets: two consecutive PVCs with a normal QRS complex in between",
    "Long QTc interval (> 440 ms for men, > 460 ms for women)",
    "Abnormal T-wave morphology (e.g., 'saddle-shaped' or 'inverted' T-wave)",
    "ST-segment elevation or depression in anteroseptal and lateral leads (e.g., V1-V3, I, aVL)",
    "T-wave inversion or abnormal T-wave morphology in anteroseptal and lateral leads",
    "Abnormal QRS complex morphology (e.g., wide or bizarre QRS complexes)",
    "Prolonged PR interval",
    "Abnormal QRS axis"
]
...
```

Note that I've kept the descriptions concise and focused on the key features, as they would be in a clinical setting.

Potential Waveform features:

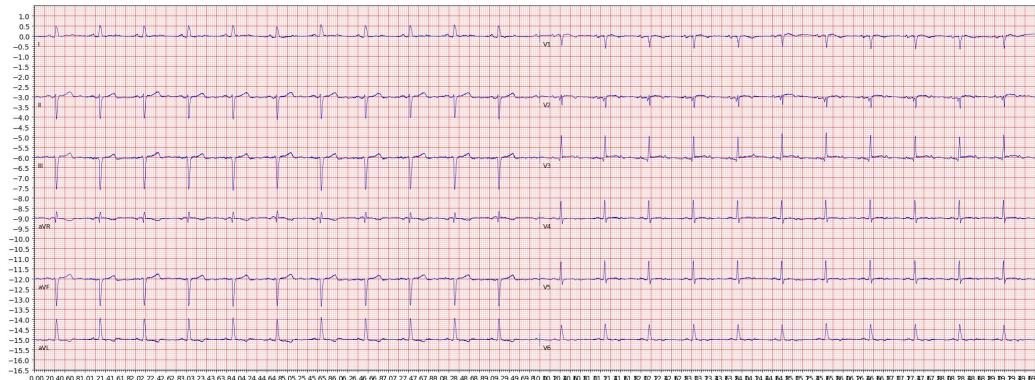
['Irregularly irregular rhythm', 'Absent or fragmented P waves', 'Fibrillatory waves (f-waves) or oscillations in the atrial region', 'Premature ventricular contractions (PVCs) with a different morphology from the normal QRS complex', 'Ventricular couplets: two consecutive PVCs with a normal QRS complex in between', 'Long QTc interval (> 440 ms for men, > 460 ms for women)', 'Abnormal T-wave morphology (e.g., 'saddle-shaped' or 'inverted' T-wave)', 'ST-segment elevation or depression in anteroseptal and lateral leads (e.g., V1-V3, I, aVL)', 'T-wave inversion or abnormal T-wave morphology in anteroseptal and lateral leads', 'Abnormal QRS complex morphology (e.g., wide or bizarre QRS complexes)', 'Prolonged PR interval', 'Abnormal QRS axis']

Probability:

[0.9616148, 0.9832206, 0.9936021, 0.9954572, 0.9854593, 0.99531144, 0.9919208, 0.9786107, 0.99453413, 0.99271095, 0.27789757, 0.51983744]

Fine-Grained Report: Probable atrial fibrillation, ventricular couplets, Long QTc interval, Ant/septal+lateral ST-T changes suggest myocardial infarction, Repolarization changes may be partly due to rhythm, Abnormal ECG, Irregularly irregular rhythm, Absent or fragmented P waves, Fibrillatory waves (f-waves) or oscillations in the atrial region, Premature ventricular contractions (PVCs) with a different morphology from the normal QRS complex, Ventricular couplets: two consecutive PVCs with a normal QRS complex in between, Long QTc interval (> 440 ms for men, > 460 ms for women), Abnormal T-wave morphology (e.g., 'saddle-shaped' or 'inverted' T-wave), ST-segment elevation or depression in anteroseptal and lateral leads (e.g., V1-V3, I, aVL), T-wave inversion or abnormal T-wave morphology in anteroseptal and lateral leads, Abnormal QRS complex morphology (e.g., wide or bizarre QRS complexes)

Figure 6: Running Case2 with Myocardial Infarction.



ID: p10474653/s42221714/42221714

Report: Sinus rhythm, Left anterior fascicular block, Possible septal infarct - age undetermined, Left ventricular hypertrophy, Lateral T wave changes are probably due to ventricular hypertrophy, Abnormal ECG

LLM Output:

Here is the list of waveform features organized in a Python list:

```
...
waveform_features = [
    "Sinus rhythm",
    "Left axis deviation (LAD) >30 ° ",
    "Q waves in leads I, II, or V5-V6",
    "ST segment depression in leads I, II, or V5-V6",
    "T wave inversion in leads I, II, or V5-V6",
    "Increased amplitude of the QRS complex (>3.5 mm) in leads I, II, or V5-V6",
    "Increased duration of the QRS complex (>120 ms)",
    "Deep S waves in leads I, II, or V5-V6",
    "T wave inversion or flattening in leads I, aVL, V5, and V6",
    "ST segment elevation or depression in other leads",
    "Prolonged QT interval",
    "Abnormal P wave morphology or axis"
]
...
```

Note that I've tried to be as specific as possible when describing each waveform feature, but some features may overlap or be related to each other.

Potential Waveform features:

['Sinus rhythm', 'Left axis deviation (LAD) >30 ° ', 'Q waves in leads I, II, or V5-V6', 'ST segment depression in leads I, II, or V5-V6', 'T wave inversion in leads I, II, or V5-V6', 'Increased amplitude of the QRS complex (>3.5 mm) in leads I, II, or V5-V6', 'Increased duration of the QRS complex (>120 ms)', 'Deep S waves in leads I, II, or V5-V6', 'T wave inversion or flattening in leads I, aVL, V5, and V6', 'ST segment elevation or depression in other leads', 'Prolonged QT interval', 'Abnormal P wave morphology or axis']

Probability:

[0.9861351, 0.5359271, 0.33153382, 0.71545005, 0.49900547, 0.42786044, 0.61200434, 0.18875751, 0.8920379, 0.78382677, 0.88176095]

Fine-Grained Report: Sinus rhythm, Left anterior fascicular block, Possible septal infarct - age undetermined, Left ventricular hypertrophy, Lateral T wave changes are probably due to ventricular hypertrophy, Abnormal ECG, Left axis deviation (LAD) >30 °

Figure 7: Running Case3 with Hypertrophy.

RSR' Pattern (rabbit ear pattern)



Inverted T-waves



Low QRS voltages

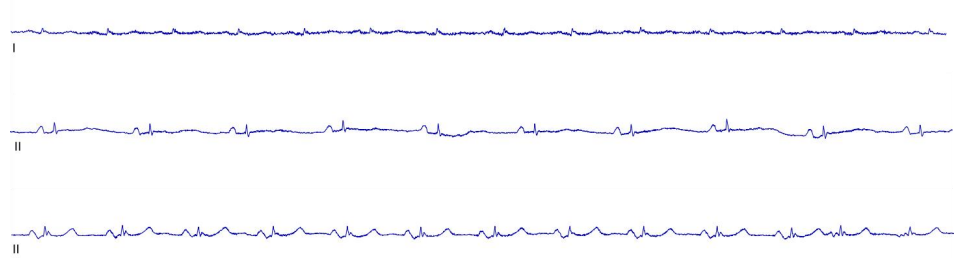


Figure 8: Top 3 retrieved ECG using FOCAL.