TAMER: A TRI-MODAL CONTRASTIVE ALIGNMENT AND MULTI-SCALE EMBEDDING REFINEMENT FRAMEWORK FOR ZERO-SHOT ECG DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cardiovascular disease (CVD) diagnosis relies heavily on electrocardiograms (ECGs). However, most existing self-supervised uni-modal methods suffer from limited representational capacity, while multi-modal frameworks are hindered by coarse-grained semantic alignment across modalities, thus restricting their generalizability in clinical settings. To address these limitations, we propose TAMER, a **T**ri-modal contrastive **A**lignment and **M**ulti-scale **E**mbedding **R**efinement framework that jointly models ECG recordings, spectrograms, and diagnostic reports. TAMER is composed of three key components: First, the tri-modal feature encoding and projection (TFEP) module employs modality-specific encoders to extract global and local features from ECG recordings, spectrograms, and diagnostic reports, and projects them into latent spaces. Then, the global-local temporal-spectral alignment (GLTSA) module captures complementary rhythm- and wave-level characteristics via contrastive alignment and attentive interaction between temporal and spectral modalities. Finally, the report-aware alignment and refinement (RAAR) module performs diagnostic-level alignment and wave-level refinement with clinical reports, enabling semantic enrichment of ECG representations. Extensive experiments on three public ECG datasets demonstrate that TAMER achieves state-of-the-art zero-shot classification performance (AUC: 81.2%) and strong cross-domain generalization (AUC: 83.1%), outperforming existing uni-modal and multi-modal baselines methods. The source code is available at `https://anonymous.4open.science/r/TAMER-FB58`.

1 INTRODUCTION

Early detection of cardiovascular diseases (CVDs) is critical for improving patient outcomes and reducing healthcare costs Tripathi et al. (2022); Elvas et al. (2025). Subtle irregularities in ECG recordings can indicate early signs of arrhythmia, ischemia, and other cardiac abnormalities, often before the onset of severe symptoms Acharya et al. (2016); Hong et al. (2020); Yagi et al. (2024). Consequently, ECG serves as a vital tool for identifying at-risk individuals, enabling timely interventions that help prevent disease progression and reduce mortality.

In recent years, the increasing availability of clinical data and advances in deep learning have led to significant progress in automated ECG diagnostic models Ribeiro et al. (2020); Huang et al. (2022); Liu et al. (2023); Al-Zaiti et al. (2023); Ameen et al. (2024). However, several challenges continue to hinder their widespread clinical adoption. First, the scarcity of labeled data for specific or rare clinical conditions poses a major obstacle, making it difficult to train reliable models that generalize well across diverse patient populations. Second, uni-modal ECG signals, which primarily reflect electrical activity, not only fail to capture the complex structural and functional abnormalities associated with cardiovascular disease, but also suffer from inherent noise and variability that hinder effective multi-modal fusion Zhang et al. (2023); Tripathi et al. (2022); Ameen et al. (2024).

To mitigate the scarcity of annotated data, self-supervised learning (SSL) has emerged as a powerful paradigm for ECG representation learning Wang et al. (2023); Zhang et al. (2022). Existing ECG SSL approaches generally fall into two categories: contrastive learning Mehari & Strodthoff (2022); Li et al. (2022); Oh et al. (2022) and generative learning Zhang et al. (2023); Na et al. (2024).

Contrastive methods learn discriminative representations by constructing positive and negative pairs in the embedding space, whereas generative methods rely on masked reconstruction tasks to model latent structural patterns. However, these methods predominantly focus on uni-modal time-series data, limiting their capacity to capture complex pathological features.

To effectively leverage other modalities, recent research has focused on multi-modal ECG modeling (Zhang et al., 2023; Lalam et al., 2023). One promising approach involves extracting time-frequency joint representations Bui et al. (2024); Yang et al. (2024), which improve sensitivity to local perturbations and non-stationary rhythms. Another emerging trend incorporates clinical diagnostic reports Liu et al. (2024); PHAM et al. (2024), providing high-level semantic supervision. However, several key challenges remain unresolved: (1) Temporal and spectral modalities emphasize distinct feature types and are subject to modality-specific noise Singh & Krishnan (2023), leading to semantic misalignment and fusion instability. (2) Most ECG-report alignment methods focus primarily on global coarse matching, neglecting local correspondences between waveform anomalies and diagnostic phrases Liu et al. (2024); PHAM et al. (2024), which limits the detection of subtle abnormalities.

To address these challenges, we propose TAMER, a tri-modal contrastive alignment and multi-scale embedding refinement framework for zero-shot ECG diagnosis. TAMER is composed of three key components: (1) The tri-modal feature encoding and projection (TFEP) module employs modality-specific encoders and projections to extract global and local features from ECG recordings, spectrograms, and clinical reports, projecting each into latent spaces. (2) The global-local temporal-spectral alignment (GLTSA) module performs rhythm-level contrastive alignment and wave-level attentive interaction between temporal and spectral ECG features, producing a unified ECG representation that captures multi-scale diagnostic patterns. (3) The report-aware alignment and refinement (RAAR) module integrates report-anchored diagnostic-level alignment and report-guided wave-level refinement to enable semantic awareness, yielding robust ECG representations.The main contributions are summarized as follows:

- We propose a tri-modal self-supervised ECG framework that jointly models ECG recordings, spectrograms, and clinical reports, extracting complementary diagnostic information from underexplored modalities.
- We introduce the GLTSA and RAAR modules to improve tri-modal ECG representations by enforcing temporal-spectral consistency and enabling cross-modal, global-local semantic alignment between ECG signals and clinical reports.
- Extensive experiments conducted on three public ECG datasets demonstrate that TAMER outperforms state-of-the-art methods in zero-shot classification and cross-domain generalization, highlighting its strong transferability and clinical relevance.

2 Related Work

2.1 Self-Supervised Learning in ECG Analysis

In recent years, SSL, broadly categorized into contrastive and generative methods, has made notable progress in intelligent ECG diagnosis. For contrastive learning, Mehari & Strodthoff (2022) adapted classical visual contrastive techniques to ECG data, demonstrating their feasibility and effectiveness in modeling medical time-series signals. Meanwhile, Wang et al. (2023) proposed ASTCL, which enhances the robustness and spatiotemporal representation of ECG signals via adversarial contrastive learning. However, contrastive methods often rely on augmentations that can introduce non-physiological features, reducing sensitivity to critical patterns. In contrast, generative approaches such as ST-MEM Na et al. (2024) and MAFE Zhang et al. (2022) utilize Vision Transformers and masking strategies to reconstruct occluded segments, capturing local morphological patterns and temporal dependencies. Moreover, CRT Zhang et al. (2023) and MassMIB Yang et al. (2024) performed cross-domain reconstruction of time- and frequency-domain representations, exploiting their complementary characteristics to enhance robustness of ECG representations. However, these methods typically demand significant computational resources and large-scale datasets.

Overall, current ECG SSL methods limited in capturing complex or subtle abnormal patterns due to underutilization of diverse information inherent in clinical data, highlighting the need for multi-modal frameworks integrating time-domain signals, frequency-domain features, and clinical context.

2.2 Multi-Modal Learning for ECG Analysis

The multi-modal nature of ECG data is increasingly recognized as a key factor in enhancing diagnostic performance. Integrating ECG signals, spectrograms, and clinical reports yields more comprehensive representations, motivating current multi-modal methods. On the one hand, frequency-domain features complement time-domain signals by capturing rhythmic and instantaneous frequency variations (Yang & Hong, 2022; Duan et al., 2024; Zhou et al., 2024). This has motivated ECG-specific time-frequency model methods such as CRT Zhang et al. (2023) and MassMIB Yang et al. (2024), which employ masked reconstruction across time and frequency views to capture global context and enhance cross-view robustness. On the other hand, inspired by CLIP Radford et al. (2021), vision-language contrastive learning has been widely adopted in medical image analysis, where clinical reports are increasingly regarded as a key modality for providing high-level diagnostic supervision (Wang et al., 2022; Cheng et al., 2023). More recently, MERL Liu et al. (2024) extended this paradigm to ECG data, introducing a multimodal contrastive learning framework that aligns ECG signals and clinical reports in a shared embedding space for zero-shot diagnosis.

In clinical practice, diagnostic decisions are based on specific waveform patterns, which are reflected in reports through textual descriptions. However, existing multi-modal approaches often overlook the alignment between localized waveform features and diagnostic semantics. To the best of our knowledge, no existing method jointly models time-domain signals, frequency-domain features, and clinical reports within a unified framework. To address this gap, we propose a tri-modal framework that enhances both modality diversity and alignment granularity, particularly over MERL Liu et al. (2024), by incorporating spectrogram features and introducing dual-level semantic alignment for more precise cross-modal understanding.

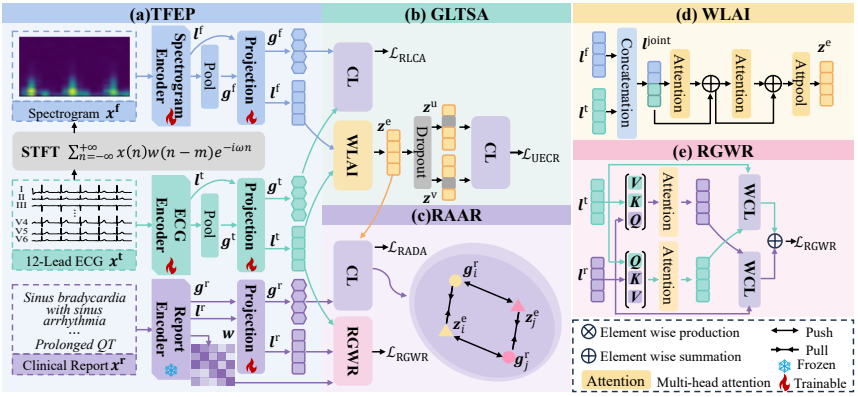

Figure 1: Overview of the TAMER framework. (a) Tri-modal Feature Encoding and Projection (TFEP). (b) Global-Local Temporal-Spectral Alignment (GLTSA). (c) Report-Aware Alignment and Refinement (RAAR). Subfigures (d) and (e) provide detailed illustrations of the internal mechanisms of the Wave-Level Attentive Interaction (WLAI) and Report-Guided Wave-Level Refinement (RGWR) respectively. **CL** and **WCL** denote the contrastive loss and the weighted contrastive loss.

3 Methods

Figure 1 illustrates the proposed TAMER framework. Given an unlabeled tri-modal dataset, $\mathcal{D} = \{(x_i^{\mathrm{t}}, x_i^{\mathrm{f}}, x_i^{\mathrm{r}})\}_{i=1}^{N}$, each training sample consists of a 12-lead ECG signal x_i^{t}, its corresponding spectrogram x_i^{f}, and the associated clinical report x_i^{r}. The TFEP module encodes each modality to extract both global features $(\boldsymbol{g}_i^{\mathrm{t}}, \boldsymbol{g}_i^{\mathrm{f}}, \boldsymbol{g}_i^{\mathrm{r}})$ and local features $(\boldsymbol{l}_i^{\mathrm{t}}, \boldsymbol{l}_i^{\mathrm{f}}, \boldsymbol{l}_i^{\mathrm{r}})$. Subsequently, the GLTSA performs rhythm-level contrastive alignment on $(\boldsymbol{g}_i^{\mathrm{t}}, \boldsymbol{g}_i^{\mathrm{f}})$ to ensure cross-modal consistency, and wave-level attentive interaction on $(\boldsymbol{l}_i^{\mathrm{t}}, \boldsymbol{l}_i^{\mathrm{f}})$ to produce a unified ECG representation $\boldsymbol{z}_i^{\mathrm{e}}$. $\boldsymbol{z}_i^{\mathrm{e}}$ is further regularized through dropout-based perturbation to produce augmented views for consistency learning. Simultaneously, both $\boldsymbol{z}_i^{\mathrm{e}}$ and $\boldsymbol{l}_i^{\mathrm{t}}$ are fed into the RAAR module, which performs dual-level

contrastive learning with $\boldsymbol{g}_i^{\mathrm{r}}$ and $\boldsymbol{l}_i^{\mathrm{r}}$, promoting deep semantic alignment between ECG signals and clinical reports.

3.1 Tri-Modal Feature Encoding and Projection

Multi-modal ECG data comprises time-domain signals that capture rhythm and waveform characteristics, spectrograms that reveal frequency-domain anomalies (e.g., transient bursts, non-stationary spectral patterns), and clinical reports that contain diagnostic knowledge. We hypothesize that modality-specific information is essential for learning high-quality representations.

We propose the tri-modal feature encoding and projection (TFEP) module, as illustrated in Figure 1(a). First, we transform the raw ECG waveform $x_i^{\mathrm{t}} \in \mathbb{R}^{L\times T}$, where L denotes the number of leads (typically 12) and T is the temporal length, into a spectrogram $x_i^{\mathrm{f}} \in \mathbb{R}^{L\times F\times M}$ using the short-time Fourier transform (STFT) Bui et al. (2024). F and M are the numbers of frequency bins and temporal frames, respectively. We then extract local features $\boldsymbol{l}_i^{\mathrm{t}}$ and $\boldsymbol{l}_i^{\mathrm{f}}$ from x_i^{t} and x_i^{f} via the hidden layers of the ECG and spectrogram encoders. We also extract global features $\boldsymbol{g}_i^{\mathrm{t}}$ and $\boldsymbol{g}_i^{\mathrm{f}}$ by applying average pooling over $\boldsymbol{l}_i^{\mathrm{t}}$ and $\boldsymbol{l}_i^{\mathrm{f}}$. For the clinical report x_i^{r}, we apply a frozen report encoder to obtain the global representation $\boldsymbol{g}_i^{\mathrm{r}}$ and local representations $\boldsymbol{l}_i^{\mathrm{r}}$, along with the attention weights of the [CLS] token, represented as a vector $\boldsymbol{w}$, which is used to estimate the importance of each report token based on its contribution to the [CLS] token. All extracted features are then projected into modality-specific latent spaces to facilitate downstream contrastive alignment and fusion.

3.2 Global-Local Temporal-Spectral Alignment

ECG signals and spectrograms exhibit periodicity and semantic complementarity in the time-frequency domain. To fully leverage this internal consistency, we introduce three modules: (1) rhythm-level contrastive alignment (RLCA) for enforcing global temporal-spectral consistency. (2) wave-level attentive interaction (WLAI) module for enhancing local feature interaction and derive a unified ECG representation, and (3) uni-ECG consistency regularization (UECR) for improveing the robustness of the unified representation, as shown in Figure 1(b).

3.2.1 Rhythm-Level Contrastive Alignment.

Although temporal and spectral features originate from the same physiological signal, the STFT transformation introduces differences in temporal resolution. Coupled with modality-specific noise, this often lead to semantic misalignment between global cardiac rhythms across modalities, hindering the model's ability to detect periodic abnormalities. To address this, we propose the RLCA, which enforces global alignment between temporal and spectral modalities via contrastive learning. Specifically, inspired by the contrastive learning Radford et al. (2021), given a pair of features $(\boldsymbol{\eta}^{\mathrm{a}}, \boldsymbol{\eta}^{\mathrm{b}})$ from modalities a and b, we minimize the distance between positive pairs $(\boldsymbol{\eta}_i^{\mathrm{a}}, \boldsymbol{\eta}_i^{\mathrm{b}})$, while maximizing that between negative pairs $(\boldsymbol{\eta}_i^{\mathrm{a}}, \boldsymbol{\eta}_j^{\mathrm{b}})$. The contrastive loss (CL) $\mathcal{L}_{\mathrm{CL}}(\cdot)$ is defined as:

$$\begin{gathered}\mathcal{L}_{i,j}^{a2b} = -\log\left(\frac{\exp(\mathrm{sim}(\boldsymbol{\eta}_i^{\mathrm{a}}, \boldsymbol{\eta}_i^{\mathrm{b}})/\tau)}{\sum_{j=1}^{N}\mathbf{1}_{[j\neq i]}\exp(\mathrm{sim}(\boldsymbol{\eta}_i^{\mathrm{a}}, \boldsymbol{\eta}_j^{\mathrm{b}})/\tau)}\right), \quad \mathcal{L}_{i,j}^{b2a} = -\log\left(\frac{\exp(\mathrm{sim}(\boldsymbol{\eta}_i^{\mathrm{b}}, \boldsymbol{\eta}_i^{\mathrm{a}})/\tau)}{\sum_{j=1}^{N}\mathbf{1}_{[j\neq i]}\exp(\mathrm{sim}(\boldsymbol{\eta}_i^{\mathrm{b}}, \boldsymbol{\eta}_j^{\mathrm{b}})/\tau)}\right), \\ \mathcal{L}_{\mathrm{CL}}(\boldsymbol{\eta}^{\mathrm{a}}, \boldsymbol{\eta}^{\mathrm{b}}) = \frac{1}{2B}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathcal{L}_{i,j}^{\mathrm{a2b}} + \mathcal{L}_{i,j}^{\mathrm{b2a}}\right),\end{gathered} \tag{1}$$

where τ is the temperature parameter, B is the batch size, $\mathbf{1}(\cdot)$ is the indicator function, and $\mathrm{sim}(\cdot)$ denotes the cosine similarity.

For temporal-spectral pairs $(\boldsymbol{g}^{\mathrm{t}}, \boldsymbol{g}^{\mathrm{f}})$, the loss $\mathcal{L}_{\mathrm{RLCA}}$ is defined as $\mathcal{L}_{\mathrm{RLCA}} = \mathcal{L}_{\mathrm{CL}}(\boldsymbol{g}^{\mathrm{t}}, \boldsymbol{g}^{\mathrm{f}})$. By enforcing this global rhythm-level alignment, RLCA enhances the temporal-spectral consistency and improves the model's ability to recognize rhythmic patterns and periodic abnormalities.

3.2.2 Wave-Level Attentive Interaction.

While RLCA captures global rhythm consistency, it lacks fine-grained modeling of diagnostic waves, such as the QRS complex or ST segment, which reflect critical pathological patterns across cycles. To this end, we introduce the WLAI module, which enhances wave-level interactions and

constructs a unified ECG representation. Specifically, local temporal and spectral features, $\boldsymbol{l}_i^{\mathrm{t}}$ and $\boldsymbol{l}_i^{\mathrm{f}}$ are concatenated to preserve modality-specific characteristics. A two-stage residual attention mechanism is then applied to adaptively reweigh salient features and align complementary semantics. Finally, a learnable class token and attention-based pooling are incorporated to aggregate diagnostic-sensitive waves into a compact embedding $\boldsymbol{z}_i^{\mathrm{e}}$, which captures compound or co-existing pathological patterns. The overall process is:

$$\boldsymbol{l}_i^{\text{joint}} = \text{concatenation}(\boldsymbol{l}_i^{\mathrm{t}}, \boldsymbol{l}_i^{\mathrm{f}}), \quad \boldsymbol{z}_i^{(1)} = \boldsymbol{l}_i^{\text{joint}} + \text{att}(\boldsymbol{l}_i^{\text{joint}}), \quad \boldsymbol{z}_i^{(2)} = \boldsymbol{z}_i^{(1)} + \text{att}(\boldsymbol{z}_i^{(1)}), \quad \boldsymbol{z}_i^{\mathrm{e}} = \text{attpool}(\boldsymbol{z}_i^{(2)}), \tag{2}$$

where att denotes the multi-head attention mechanism, and attpool$(\cdot)$ represents attention-based aggregation (Vaswani et al., 2017).

Unlike conventional fusion methods that directly sum or concatenate modalities, often leading to information redundancy or loss, WLAI selectively integrates clinically relevant features, resulting in a coherent and semantically enriched representation.

3.2.3 Uni-ECG Consistency Regularization.

Although the WLAI produces a unified representation, it may still be instability due to modality-specific noise, motion artifacts, or modality discrepancies. To enhance robustness and representation consistency, we introduce the UECR module. UECR aims to enforce view-invariant representations by perturbing the fused embedding $\boldsymbol{z}_i^{\mathrm{e}}$. Specifically, we apply dropout to to generate two stochastic views $\boldsymbol{z}_i^{\mathrm{u}}$ and $\boldsymbol{z}_i^{\mathrm{v}}$. We then apply a contrastive loss, $\mathcal{L}_{\text{UECR}} = \mathcal{L}_{\text{CL}}(\boldsymbol{z}^{\mathrm{u}}, \boldsymbol{z}^{\mathrm{v}})$, using the contrastive function defined in Eq. equation 1 to encourage their alignment in the embedding space.

3.3 Report-Aware Alignment and Refinement

In clinical practice, diagnostic reports interpret ECG signals from a medical perspective, providing high-level complementary information. Exploiting the complementarity of these two modalities enhances the model's ability to capture disease-related features. To this end, the RAAR module leverages a frozen text encoder to provide stable diagnostic semantics and enhances ECG representations through two sub-modules: the report-anchored diagnostic-level alignment (RADA) enhances the model's awareness of global diagnostic semantics, while the report-guided wave-level refinement (RGWR) strengthens attention to key diagnostic waves and improves the identification of fine-grained abnormalities, as illustrated in Figure 1(c).

3.3.1 Report-Anchored Diagnostic-Level Alignment.

The representation $\boldsymbol{z}_i^{\mathrm{e}}$ produced by the WLAI module integrates key waveform segments to form a comprehensive diagnostic embedding, while the global report embedding $\boldsymbol{g}_i^{\mathrm{r}}$ captures diagnostic semantics from textual descriptions. These two modalities respectively provide the physiological and clinical perspectives necessary for cardiovascular disease diagnosis. To align them, the RADA module enforces global semantic consistency across modalities via contrastive learning. The objective is defined as $\mathcal{L}_{\text{RADA}} = \mathcal{L}_{\text{CL}}(\boldsymbol{z}^{\mathrm{e}}, \boldsymbol{g}^{\mathrm{r}})$ (Eq. equation 1).

3.3.2 Report-Guided Wave-Level Refinement.

While RADA captures global semantics, it lacks fine-grained alignment between ECG and diagnostic report. Since WLAI already fuses temporal and spectral features, RGWR focuses on local interactions between temporal ECG waves and diagnostic reports, enabling wave-level semantic refinement and improving abnormality localization.

Let $T_i = \{\boldsymbol{t}_i^k\}_{k=1}^K, \boldsymbol{R}_i = \{\boldsymbol{r}_i^m\}_{m=1}^M, \boldsymbol{t}_i^k, \boldsymbol{r}_i^m \in \mathbb{R}^D$ denote the local feature sets from ECG recordings and report respectively. A dual cross-attention mechanism computes contextual representations:

$$\boldsymbol{c}_i^k = \sum_{m=1}^{M} \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{t}_i^k \cdot \boldsymbol{K}\boldsymbol{p}_i^m}{\sqrt{D}}\right) \cdot \boldsymbol{V}\boldsymbol{r}_i^m, \quad \boldsymbol{c}_i^m = \sum_{k=1}^{K} \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{r}_i^m \cdot \boldsymbol{K}\boldsymbol{t}_i^k}{\sqrt{D}}\right) \cdot \boldsymbol{V}\boldsymbol{t}_i^k, \tag{3}$$

with $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{D\times D}$ as learnable matrices.

To dynamically weight the diagnostic importance of different tokens and ECG waves, we utilize token-level attention weights $\boldsymbol{w}_i^j$ generated by the report encoder. The weighted contrastive loss

Table 1: Comparison of different methods on the PTBXL-Super dataset (best in **bold**).

Method	Training Ratio	Zero-Shot	Domain Shift	
		PTBXL-Super	CPSC2018	CSN
SimCLR	100%	×	69.62	73.05
BYOL	100%	×	70.27	74.01
BarlowTwins	100%	×	68.98	72.85
MoCo-v3	100%	×	69.41	73.29
SimSiam	100%	×	70.06	73.92
TS-TCC	100%	×	71.32	75.16
CLOCS	100%	×	68.79	72.64
ASTCL	100%	×	69.23	73.18
CRT	100%	×	70.15	74.08
ST-MEM	100%	×	**76.12**	**84.50**
MERL	0%	74.2	**88.21**	78.01
C-MET	0%	76.2	72.09	79.11
TAMER	0%	**76.5**	84.71	80.95

(WCL) for local ECG-report refinement denoted as $\mathcal{L}_{\text{RGWR}}$ and computed as follows:

$$\mathcal{L}_{\text{ECG}} = -\frac{1}{2NK}\sum_{i=1}^{N}\sum_{j=1}^{K} \boldsymbol{w}_i^j \log\left(\frac{\exp(\text{sim}(\boldsymbol{t}_i^j, \boldsymbol{c}_i^j)/\lambda)}{\sum_{k=1}^{K}\exp(\text{sim}(\boldsymbol{t}_i^j, \boldsymbol{c}_i^k)/\lambda)}\right), \quad \mathcal{L}_{\text{report}} = -\frac{1}{2NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \boldsymbol{w}_i^j \log\left(\frac{\exp(\text{sim}(\boldsymbol{r}_i^j, \boldsymbol{c}_i^j)/\lambda)}{\sum_{m=1}^{M}\exp(\text{sim}(\boldsymbol{r}_i^j, \boldsymbol{c}_i^m)/\lambda)}\right),$$
$$\mathcal{L}_{\text{RGWR}} = \mathcal{L}_{\text{ECG}} + \mathcal{L}_{\text{report}}, \tag{4}$$

where λ is the temperature parameter. The RGWR highlights waveform segments that are most relevant to the diagnostic report, enhancing interpretability and fine-grained disease recognition. Attention weights $\boldsymbol{w}_i^j$ dynamically adjust focus based on the diagnostic importance of each token, improving sensitivity to critical abnormalities.

By integrating global and local contrastive losses, the RAAR alignment loss is defined as:

$$\mathcal{L}_{\text{RAAR}} = \mathcal{L}_{\text{RADA}} + \mathcal{L}_{\text{RGWR}}. \tag{5}$$

3.4 Overall Loss Function

Finally, TAMER jointly optimizes the key loss functions as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RLCA}} + \mathcal{L}_{\text{UECR}} + \mathcal{L}_{\text{RAAR}}. \tag{6}$$

By integrating these components, TAMER significantly enhances cross-modal consistency and improves representation robustness.

4 Experiments and Results

4.1 Pre-Training

4.1.1 Pre-Training Dataset.

We pre-train our model on the MIMIC-ECG dataset Gow et al. (2023), which contains 800,035 ECG-report pairs collected from 161,252 patients. Each sample consists of a 12-lead ECG signal recorded at 500 Hz for 10 seconds, accompanied by a structured or free-text diagnostic report. Data processing follows the standard pipeline proposed in MERL Liu et al. (2024), including signal normalization, clinical report cleaning, and semantic filtering. After processing, a total of 771,693 high-quality triplets are retained for unsupervised tri-modal training.

4.1.2 Pre-Training Implementation Details.

Our model is implemented in PyTorch and trained on a single NVIDIA A100-PCIE-40GB GPU. The ECG encoder adopts a randomly initialized 1D ResNet-34 He et al. (2016), the spectrogram encoder is a 2D CNN, and the report encoder employs a frozen Med-CPT Query Encoder Jin et al. (2023) for semantic stability.

We use the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-7}. To dynamically adjust the learning rate during training, we adopt a cosine annealing warm restart scheduler with an initial restart period of $T_0 = 40,000$. The temperature factor λ is set to be 0.04. The model is trained for 50 epochs with a batch size of 256.

Table 2: Comparison of different methods on the CPSC2018 dataset (best in **bold**).

Method	Training Ratio	Zero-Shot	Domain Shift	
		CPSC2018	PTBXL-Super	CSN
SimCLR	100%	×	56.65	66.36
BYOL	100%	×	57.32	67.56
BarlowTwins	100%	×	55.97	65.89
MoCo-v3	100%	×	56.54	66.12
SimSiam	100%	×	57.21	67.48
TS-TCC	100%	×	58.47	68.34
CLOCS	100%	×	55.86	65.73
ASTCL	100%	×	56.61	66.27
CRT	100%	×	57.39	67.62
ST-MEM	100%	×	62.27	75.19
MERL	0%	82.8	76.77	76.56
C-MET	0%	80.1	77.12	82.91
TAMER	0%	**88.3**	**82.00**	**86.73**

Table 3: Comparison of different methods on the CSN dataset (best in **bold**).

Method	Training Ratio	Zero-Shot	Domain Shift	
		CSN	PTBXL-Super	CPSC2018
SimCLR	100%	×	59.74	62.11
BYOL	100%	×	60.39	63.24
BarlowTwins	100%	×	58.76	61.35
MoCo-v3	100%	×	59.82	62.07
SimSiam	100%	×	60.23	63.09
TS-TCC	100%	×	61.55	64.48
CLOCS	100%	×	58.69	61.27
ASTCL	100%	×	59.74	61.12
CRT	100%	×	60.48	63.33
ST-MEM	100%	×	73.05	64.66
MERL	0%	74.4	74.15	82.86
C-MET	0%	76.3	76.24	80.10
TAMER	0%	**78.7**	**76.49**	**87.62**

4.2 Downstream Tasks

4.2.1 Downstream Task Datasets.

To evaluate the generalization of our pre-trained TAMER across various clinical scenarios, we conduct downstream experiments on three public ECG datasets, each providing 12-lead signals (500 Hz, 10 s). Data splitting and preprocessing follow the protocol in MERL Liu et al. (2024).

PTBXL-Super Dataset. A PTBXL subset Wagner et al. (2020) with 21,837 ECG recordings from 18,885 patients across five major CVD categories is used for evaluation.

CSN Dataset. The CSN dataset Zheng et al. (2020; 2022) includes 23,026 ECG recordings annotated with 38 diagnostic labels.

CPSC2018 Dataset. The CPSC2018 dataset Liu et al. (2018) comprises 6,877 12-lead ECG recordings, with durations ranging from 6 to 60 seconds, and includes 9 diagnostic labels. We retain recordings with durations $\geq$ 10 seconds and truncate all signals to 10 seconds, resulting in 6,867 records used for evaluation.

4.2.2 Downstream Task Implementation Details.

To comprehensively assess the generalization ability of the pre-trained model under real-world clinical constraints, we design two zero-shot evaluation scenarios:

Zero-Shot Classification Across Unseen Labels. We evaluate the model's ability to recognize previously unseen disease categories while keeping all pre-trained parameters frozen. We adopt the CKEPE prompt dictionary Liu et al. (2024) to generate class-level textual descriptions. The similarity between each ECG representation and the semantic prompt is computed and used as the prediction score on the downstream test set.

Zero-Shot Classification Under Domain Shift. To simulate domain shifts frequently encountered in clinical environments, such as varying patient populations or acquisition protocols, we conduct cross-dataset evaluations where the source and target domains share semantically aligned diagnostic labels but differ in data distribution. Label mapping and merging are performed following the pro-

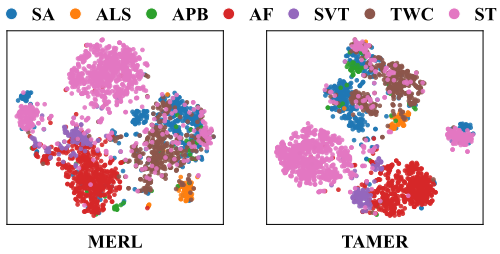

Figure 2: t-SNE Visualization of ECG Features (MERL vs. TAMER) on CSN Dataset

tocol in MERL Liu et al. (2024). The model is directly evaluated on the target domain without any additional tuning.

4.3 Evaluation Metrics

All evaluations are based on macro-AUC, ensuring a fair comparison in the presence of class imbalance and providing a robust measure of model reliability across domains.

4.4 Comparison with State-of-the-art Methods

To comprehensively evaluate the effectiveness and generalization capability of the proposed TAMER framework, we conduct systematic comparisons with state-of-the-art SSL methods under two evaluation settings: zero-shot classification and zero-shot classification under domain shift. Zero-shot classification is performed for multi-modal methods to assess their capacity for semantic understanding and classification. For zero-shot classification under domain shift, multi-modal approaches such as MERL Liu et al. (2024) and C-MET PHAM et al. (2024) are evaluated on the target domain without any fine-tuning, better reflecting their ability to generalize across domains through cross-modal semantic alignment. All uni-modal SSL baselines: SimCLR Chen et al. (2020), BYOL Grill et al. (2020), BarlowTwins Zbontar et al. (2021), MoCo-v3 Chen et al. (2021), SimSiam Chen & He (2021), TS-TCC Eldele et al. (2021), CLOCS Kiyasseh et al. (2021), ASTCL Wang et al. (2023), CRT Zhang et al. (2023), ST-MEM Na et al. (2024) are fine-tuned on 100% of the labeled source domain data and then evaluated on the target domain.

4.4.1 Zero-Shot Classification.

As illustrated in Tables 1, 2, and 3,TAMER achieves the best overall performance, with AUCs of 76.5%/88.3%/78.7% on PTBXL-Super, CPSC2018, and CSN, respectively. This notably surpasses MERL (74.2%/82.8%/74.4%) and C-MET (76.2%/80.1%/76.3%), especially on CPSC2018 where TAMER an AUC of 88.3%, demonstrating its exceptional capability in handling complex, multi-label ECG classification under zero-shot settings.

4.4.2 Zero-Shot Classification under Domain Shift.

Under domain shift conditions, TAMER continues to outperform all compared SSL, achieving an average AUC of 81.2% on three cross-domain evaluation settings, without any fine-tuning, as shown in Tables 1, 2, and 3. This highlights its robustness to distribution shifts and the advantage of leveraging tri-modal semantics.

Several insights can be drawn from the results. First, under domain shift scenarios, multi-modal methods generally outperform uni-modal approaches that rely on source-domain fine-tuning, emphasizing the effectiveness of incorporating clinical text as semantic priors to mitigate the impact of distributional shifts. Second, TAMER achieves average AUCs of 81.2% and 83.1% across three datasets on the two downstream tasks, consistently surpassing all uni-modal and multi-modal self-supervised baselines. These results demonstrate its superior generalization capability and validate the effectiveness of our proposed tri-modal architecture and report-aware alignment and refinement module in improving cross-domain modeling.

4.5 Analysis of TAMER

4.5.1 Ablation Study.

We conduct ablation experiments by individually removing the RLCA, WLAI, and RGWR modules to evaluate their contributions to model performance. As shown in Table 4, removing RLCA, WLAI, and RGWR results in average AUC drops of 0.57%, 5.03%, and 2.46%, respectively, on the zero-shot task, and 1.59%, 4.14%, and 3.64% on the domain shift task. These results clearly demonstrate the essential role of all three modules in enhancing the model's generalization and diagnostic performance. Specifically, RLCA and WLAI enhance semantic consistency between temporal and spectral modalities at the global and local levels, respectively. RGWR performs fine-grained semantic alignment between ECG signals and diagnostic phrases, facilitating the precise identification of critical abnormalities. Together, these three modules promote deep interaction and semantic enrichment across ECG tri-modal features, significantly improving the model's discriminative capability and clinical adaptability in automated ECG diagnosis. These findings validate the effectiveness of the proposed method in multi-modal medical scenarios.

Table 4: Results of ablation experiments on key modules.

	WLAI	RGWR	Zero-Shot	Domain Shift
×	✓	✓	80.62	81.49
✓	×	✓	76.16	78.94
✓	✓	×	78.73	79.44
✓	✓	✓	**81.19**	**83.08**

4.5.2 Effects of Temperature Parameter λ.

The temperature parameter λ controls the concentration level of the similarity distributions. We assess the impact of λ using values of 0.03, 0.04, and 0.05. As shown in Table 5, $\lambda = 0.04$ yields the best performance in both zero-shot and domain shift settings, indicating a balanced trade-off between training stability and discriminative learning.

Table 5: Effects of temperature parameter λ.

λ	Zero-Shot	Domain Shift
0.03	80.42	81.66
0.04	**81.19**	**83.08**
0.05	79.91	81.62

4.5.3 Visualization of Feature Representations.

We employ t-SNE to visualize the ECG embeddings extracted from the CSN test set, as shown in Figure 2. To enhance clustering clarity, multi-label samples and classes with fewer than 50 instances are excluded. The visualization shows that TAMER produces more distinct and compact clusters across various diagnostic categories. Compared to MERL, TAMER exhibits clearer group boundaries for easily confusable classes such as ALS, APB, and TWC, indicating improved separation and reduced overlap in the feature space. For well-separated categories like AF and ST, both models perform similarly, confirming that TAMER retains the original discriminative capacity while also producing more refined embeddings for ambiguous cases.

5 Conclusion

In this work, we propose TAMER, a unified tri-modal pre-training framework for robust and generalizable ECG representation learning. TAMER integrates three key components: a tri-modal feature encoding and projection module, a global-local temporal-spectral alignment module, and a report-aware alignment and refinement module. By jointly modeling ECG signals, spectrograms, and clinical diagnostic reports, TAMER effectively captures heterogeneous and localized semantic information. The fusion and contrastive alignment modules promote consistent, interpretable, and discriminative representations. Extensive evaluations on three public datasets demonstrate that TAMER consistently outperforms state-of-the-art uni-modal and multi-modal baselines in both zero-shot classification and cross-domain transfer tasks.

REFERENCES

U Rajendra Acharya, Hamido Fujita, Vidya K Sudarshan, Shu Lih Oh, Muhammad Adam, Joel EW Koh, Jen Hong Tan, Dhanjoo N Ghista, Roshan Joy Martis, Chua K Chua, et al. Automated detection and localization of myocardial infarction using electrocardiogram: a comparative study of different leads. *Knowledge-Based Systems*, 99:146–156, 2016.

Salah S Al-Zaiti, Christian Martin-Gill, Jessica K Zègre-Hemsey, Zeineb Bouzid, Ziad Faramand, Mohammad O Alrawashdeh, Richard E Gregg, Stephanie Helman, Nathan T Riek, Karina Kraevsky-Phillips, et al. Machine learning for ecg diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 29(7):1804–1813, 2023.

Asmaa Ameen, Ibrahim Eldesouky Fattoh, Tarek Abd El-Hafeez, and Kareem Ahmed. Advances in ecg and pcg-based cardiovascular disease classification: a review of deep learning and machine learning methods. *Journal of Big Data*, 11(1):159, 2024.

Nhat-Tan Bui, Dinh-Hieu Hoang, Thinh Phan, Minh-Triet Tran, Brijesh Patel, Donald Adjeroh, and Ngan Le. Tsrnet: Simple framework for real-time ecg anomaly detection with multimodal time and spectrogram restoration network. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9640–9649, 2021.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21361–21371, 2023.

Jufang Duan, Wei Zheng, Yangzhou Du, Wenfa Wu, Haipeng Jiang, and Hongsheng Qi. Mf-clr: multi-frequency contrastive learning representation for time series. In *Forty-first International Conference on Machine Learning*, 2024.

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

Luis B Elvas, Ana Almeida, and Joao C Ferreira. The role of ai in cardiovascular event monitoring and early detection: Scoping literature review. *JMIR Medical Informatics*, 13(1):e64349, 2025.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. *Type: dataset*, 6:13–14, 2023.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122:103801, 2020.

Yu Huang, Gary G Yen, and Vincent S Tseng. Snippet policy network v2: Knee-guided neuroevolution for multi-lead ecg early classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2167–2181, 2022.

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

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

Sravan Kumar Lalam, Hari Krishna Kunderu, Shayan Ghosh, Harish Kumar, Samir Awasthi, Ashim Prasad, Francisco Lopez-Jimenez, Zachi I Attia, Samuel Asirvatham, Paul Friedman, et al. Ecg representation learning with multi-modal ehr data. *Transactions on Machine Learning Research*, 2023.

Simon A Lee, Anthony Wu, and Jeffrey N Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text, 2025.

Fangyu Li, Hui Chang, Min Jiang, and Yihuan Su. A contrastive learning framework for ecg anomaly detection. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 673–677. IEEE, 2022.

Che Liu, Sibo Cheng, Weiping Ding, and Rossella Arcucci. Spectral cross-domain neural network with soft-adaptive threshold spectral enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 31949–31963, 2024.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141:105114, 2022.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*, 2024.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pp. 338–353. PMLR, 2022.

Hung Manh PHAM, Aaqib SAEED, and Dong MA. Revisiting masked auto-encoders for ecg-language representation learning. In *The first NeurIPS workshop on Time Series in the Age of Large Models, Vancouver*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature Communications*, 11(1):1760, 2020.

Anupreet Kaur Singh and Sridhar Krishnan. Ecg signal feature extraction trends in methods and applications. *BioMedical Engineering OnLine*, 22(1):22, 2023.

Prashant Mani Tripathi, Ashish Kumar, Rama Komaragiri, and Manjeet Kumar. A review on computational methods for denoising and detecting ecg signals to detect cardiovascular diseases. *Archives of Computational Methods in Engineering*, 29(3):1875–1914, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022.

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

Ryuichiro Yagi, Yuichiro Mori, Shinichi Goto, Taku Iwami, and Kosuke Inoue. Routine electrocardiogram screening and cardiovascular disease events in adults. *JAMA Internal Medicine*, 184(9): 1035–1044, 2024.

Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International Conference on Machine Learning*, pp. 25038–25054. PMLR, 2022.

Shunxiang Yang, Cheng Lian, Zhigang Zeng, Bingrong Xu, Yixin Su, and Chenyang Xue. Masked self-supervised ecg representation learning via multiview information bottleneck. *Neural Comput. Appl.*, 36:7625–7637, 2024.

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

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

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

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

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study. *PhysioNet*, 2022.

Ziyu Zhou, Gengyu Lyu, Yiming Huang, Zihao Wang, Ziyu Jia, and Zhen Yang. Sdformer: transformer with spectral filter and dynamic attention for multivariate time series long-term forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Jeju, Republic of Korea*, pp. 3–9, 2024.

A Appendix

A.1 LLM Assistance Statement

During the preparation of this manuscript, we utilized AI-based assistance tools (OpenAI's ChatGPT) to support the writing and editing process. The AI was used primarily to:

- Refine and polish the language of certain paragraphs to improve clarity, readability, and conciseness.
- Suggest alternative wording or phrasing for specific terms to enhance precision and academic tone.
- Provide guidance on restructuring sentences or paragraphs for better logical flow.

All scientific content, including experimental design, methodology, results, analysis, and conclusions, was authored and verified solely by the human authors. The AI did not generate any original scientific claims or analyses; it assisted only with language expression and clarity.

A.2 Choices of Text Encoders.

We evaluate the effectiveness of different report encoders derived from three representative medical language models: PubMedBERT Gu et al. (2021), Clinical ModernBERT Lee et al. (2025), and Med-CPT Jin et al. (2023). Each text encoder is assessed under identical pre-training and downstream evaluation settings. As shown in Table A1, Med-CPT consistently achieves the best performance across both zero-shot classification and domain shift tasks, significantly outperforming the other encoders. This advantage is attributed to Med-CPT's contrastive pre-training strategy, which is more effective at modeling semantic consistency and capturing fine-grained features in medical reports, thereby improving cross-modal alignment performance.

Table A1: Performance comparison of different text encoders

Text encoder	Zero-Shot	Domain Shift
PubMedBERT	72.23	74.36
Clinical ModernBERT	76.92	79.61
Med-CPT	**81.19**	**83.08**

A.3 Comparison of different methods under Zero-shot.

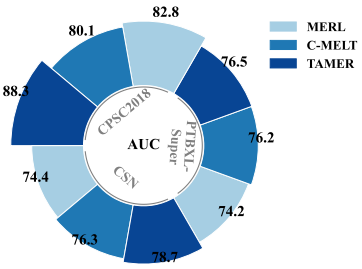

Figure A1: Zero-shot AUC (%) on three ECG datasets: MERL vs. C-MET vs. TAMER. AUC performance (%) of MERL, C-MET, and TAMER across three ECG datasets in the zero-shot setting.