Tracing the Heart's Pathways: ECG Representation Learning from a Cardiac Conduction Perspective

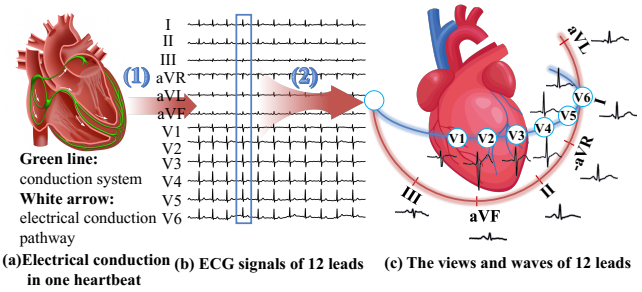
Tan Pan[1, 2*], **Yixuan Sun**[1, 2*], **Chen Jiang**[1, 2†], **Qiong Gao**[6], **Rui Sun**[1, 2], **Xingmeng Zhang**[1, 2], **Zhenqi Yang**[2], **Limei Han**[1, 2], **Yixiu Liang**[4, 5, 2], **Yuan Cheng**[1, 2†], **Kaiyu Guo**[2, 3]

[1] Fudan University
[2] Shanghai Academy of Artificial Intelligence for Science
[3] The University of Queensland
[4] Department of Cardiology, Zhongshan Hospital of Fudan University
[5] National Heart and Lung Institute, Imperial College London
[6] Department of Electrocardiology, Zhongshan Hospital of Fudan University
{pant23,sunyx23}@m.fudan.edu.cn, jiangchen@sais.org.cn, cheng_yuan@fudan.edu.cn

Abstract

The multi-lead electrocardiogram (ECG) stands as a cornerstone of cardiac diagnosis. Recent strides in electrocardiogram self-supervised learning (eSSL) have brightened prospects for enhancing representation learning without relying on high-quality annotations. Yet earlier eSSL methods suffer a key limitation: they focus on consistent patterns across leads and beats, overlooking the inherent differences in heartbeats rooted in cardiac conduction processes, while subtle but significant variations carry unique physiological signatures. Moreover, representation learning for ECG analysis should align with ECG diagnostic guidelines, which progress from individual heartbeats to single leads and ultimately to lead combinations. This sequential logic, however, is often neglected when applying pre-trained models to downstream tasks. To address these gaps, we propose CLEAR-HUG, a two-stage framework designed to capture subtle variations in cardiac conduction across leads while adhering to ECG diagnostic guidelines. In the first stage, we introduce an eSSL model termed **C**onduction-**LEA**d **R**econstructor (CLEAR), which captures both specific variations and general commonalities across heartbeats. Treating each heartbeat as a distinct entity, CLEAR employs a simple yet effective sparse attention mechanism to reconstruct signals without interference from other heartbeats. In the second stage, we implement a **H**ierarchical lead-**U**nified **G**roup head (HUG) for disease diagnosis, mirroring clinical workflow. Experimental results across six tasks show a 6.84% improvement, validating the effectiveness of CLEAR-HUG. This highlights its ability to enhance representations of cardiac conduction and align patterns with expert diagnostic guidelines.

Code — https://github.com/Ashespt/CLEAR-HUG



(a)Electrical conduction in one heartbeat (b) ECG signals of 12 leads (c) The views and waves of 12 leads

Figure 1: Cardiac conduction and the relationship between the 12 ECG leads: (1) Electrical activity propagates through the heart (a) and reflects on the 12 leads in the same time window (b). (2) The 12 leads capture the heart's electrical activity from different views (c).

Introduction

ECG plays a crucial role in the diagnosis of various cardiac conditions (Berkaya et al. 2018; Auer et al. 2012). In recent years, AI-assisted ECG analysis has demonstrated significant potential in enhancing diagnostic accuracy (Zhou et al. 2024; Poterucha et al. 2025) and enabling real-time ECG monitoring (Xia, Asif, and Zhao 2013; Alimbayeva et al. 2024). To leverage large-scale unannotated data while addressing the dependency on expert knowledge and high-quality annotations, recent advances in ECG self-supervised learning (eSSL) (Na et al. 2024a; Jin et al. 2025) have enhanced ECG signal analysis.

Self-supervised learning (SSL) methods for medical data can be divided into two approaches: contrastive learning (Pan et al. 2025; Wu, Zhuang, and Chen 2024) and generative learning (Wang et al. 2023b; Hatamizadeh et al. 2021). Advanced eSSL methods build on these techniques by incorporating spatial and temporal priors to improve representation power (Na et al. 2024a; Zhang et al. 2022). For instance, ST-MEM uses a lead-wise shared decoder for better lead differentiation (Na et al. 2024a), while (Zhang

[*] Equal contribution. This work was conducted during an internship at the Shanghai Academy of Artificial Intelligence for Science.
[†] Corresponding authors.

et al. 2023) applies cross-reconstruction across time and frequency domains. These methods typically partition signals uniformly, whereas HeartLang (Jin et al. 2025) partitions signals by heartbeats, treating them like words in language models.

While existing methods demonstrate promising performance, they primarily focus on consistent patterns across leads and beats, which leads to the following two failure cases: **(1) Occasional anomalies.** For some abnormal cardiac conditions, the corresponding ECG signals lack consistently periodic or cyclic patterns. For instance, ventricular premature beats are characterized by abnormal contractions that arise prematurely within the cardiac cycle. These beats do not repeat consistently in every heartbeat and appear sporadically. **(2) Various focuses of leads across different diseases.** For example, premature ventricular contractions (PVCs) are characterized by premature, wide QRS complexes without a preceding P wave, typically seen in leads V1 and V2. Right bundle branch block (RBBB), on the other hand, shows an RSR' pattern in V1 and a broad R wave in V6, highlighting the focus on different lead combinations for diagnosis.

The aforementioned cases can be summarized as two problems: 1) the heartbeat- and lead-specific variations; 2) diagnosis-driven lead combination. To address these challenges, we reformulate the ECG recording process from the perspective of cardiac conduction and propose a two-stage framework for learning representations progressively, first from heartbeats to leads at the pretraining stage and then from leads to their combinations at the finetuning stage.

During the pretraining stage, the relation between heartbeats and leads serves as auxiliary information. Specifically, as shown in Fig. 1 (1) and (2), this information primarily exhibits two characteristics: (1) the 12 leads corresponding to the same heartbeat share the same time window and the same electrical activity, which enables them to provide complementary views of the heart's electrical activity to each other; (2) it can be observed that a single heartbeat consists of signals in two dimensions: heartbeat-specific and lead-specific, which leads to our assumption of conduction-guided and view-guided information. Based on the observation, we propose a reconstruction framework, termed **C**onduction-**LEA**d **R**econstructor (CLEAR), which aims to learn meaningful representations through recovering the signal of a heartbeat by utilizing both conduction-guided and view-guided information. Particularly, we propose a sparse attention mechanism with a tailored attention mask to highlight the conduction-guided and view-guided information through the reconstruction.

In the finetuning stage, the design of our method is guided by the clinical diagnosis workflow, enhancing the explainability of our model. As per the established ECG analysis guidelines (Kligfield et al. 2007), cardiac conditions are diagnosed at multiple levels: individual heartbeats, single-lead signals, and combinations of leads, which motivates us to design a multi-level linear model for specific diagnosis tasks. Specifically, we propose to integrate the representations learned from different lead combinations through various groups of linear models, which is termed as **H**ierarchical lead-**U**nified **G**roup (HUG) head. By simulating the clinical diagnosis, this design demonstrates more promising experimental results compared with conventional methods.

As mentioned above, each stage in our framework is motivated by a novel and reasonable insight. Therefore, the main contributions of this paper are threefold:

(1) Pretraining stage: We introduce the concept of conduction-guided and view-guided information during pretraining, observed on the relationship between heartbeats and leads. Based on the insight, we propose a novel eSSL framework, CLEAR, to learn meaningful representations with conduction-guided and view-guided information.

(2) Finetuning stage: We introduce a novel design inspired by the clinical diagnostic workflow to better align model finetuning with ECG analysis guidelines. Based on the design, we propose a linear model HUG to learn integral representations from different lead combinations.

(3) Experimental improvements: Our method outperforms state-of-the-art (SOTA) techniques across 6 datasets, with an average improvement of 6.8%. Notably, CLEAR-HUG achieved an 8.25% improvement under 1% of the training data, highlighting its strong potential for few-shot transfer. The code will be available upon publication.

Related Work

Self-supervised Learning. Current self-supervised learning (SSL) methods can be classified into contrastive-based and reconstruction-based methods. For the contrastive-learning paradigm, SimCLR (Chen et al. 2020) focuses on learning representations by maximizing agreement between augmented views of the same data instance and minimizing agreement between different instances. On the other hand, BYOL (Grill et al. 2020) enhances contrastive learning by eliminating the need for negative samples, leveraging a bootstrap mechanism. Similarly, SimSiam (Chen and He 2021) avoids the use of negative pairs entirely and instead focuses on a simple predictor network to learn useful representations. In contrast, MoCo-v3 (Chen, Xie, and He 2021) incorporates momentum encoders and dynamic dictionaries to further refine the learning of representations. While contrastive-based methods heavily rely on data augmentations, masked autoencoder (MAE) (He et al. 2022) follows a different approach, learning meaningful image representations by reconstructing missing or masked parts of the input data.

Self-supervised Learning For ECG Signals. ESSL has made significant strides in recent years, with methods primarily falling into contrastive learning and generative approaches. CLOCS (Kiyasseh, Zhu, and Clifton 2021) utilizes multi-lead ECG temporal alignment to create positive pairs, while TS-TCC (Eldele et al. 2021) introduces temporal contrastive strategies, excelling in ECG classification tasks. Unlike contrastive learning, generative methods learn representations through reconstruction tasks. ST-MEM (Na et al. 2024a) treats ECG as a spatiotemporal 2D signal for joint masking, while CRT (Zhang et al. 2023) performs cross-domain reconstruction in both time and frequency domains. Recently, HeartLang (Jin et al. 2025) proposes that

partitioning ECG signals following the heartbeat and learning it with heartbeat vocabulary. These approaches directly model data distributions and tend to learn global representations across all leads. In contrast, our work focuses on modeling the cardiac conduction, which carries unique physiological signatures.

Insights from Cardiac Conduction

This section outlines ECG pattern insights and clinical guidelines for ECG diagnosis based on cardiac conduction features, which form our method's foundation. We define notations for a 12-lead ECG, though our approach is applicable to ECGs with any number of leads.

Insight from ECG Patterns. Based on the cyclic pattern of heartbeats and the principles of ECG signal acquisition (Mirvis and Goldberger 2001), ECG signals exhibit inter-lead heartbeat conduction synchronization (Fig. 1 (1)) and intra-lead heterogeneity (Fig. 1 (2)). Specifically, during a single heartbeat, the ECG signals recorded by the 12 leads $\{L_i\}_{i=1}^{12}$ are derived from the same physiological process (one cardiac conduction), capturing the electrical activities of the heart at different views. Within the j-th heartbeat signal b_i^j of the i-th lead, the formulations of the signals from the information of two perspectives are presented as follows:

(1) Conduction-guided perspective: The information I_c originates from the same heartbeat conduction, reflecting the temporal synchronization shared across all leads during the j-th heartbeat. Thus, the set of signals across all leads within the j-th heartbeat can be denoted as $B^j = \{b_i^j\}_{i=1}^{12}$.

(2) View-guided perspective: The information I_v captures the spatial heterogeneity among leads, where each lead L_i provides a distinct view of cardiac activity across all sampled heartbeats. The set of signals within the lead L_i can be denoted as $B_i = \{b_i^j\}_{j=1}^{N}$, where N represents the total number of sampled heartbeats.

The lead combination guidelines of ECG diagnosis. Considering ECG diagnosis tasks T, some tasks can be addressed using single-lead signals, while others require the integration of information across multiple leads. According to the recommendations for the standardization and interpretation of ECG (Kligfield et al. 2007), the 12 leads can be categorized into three groups: the first group $G_1 = \{I, II, III\}$ consists of bipolar limb leads; the second group $G_2 = \{\mathrm{aVR}, \mathrm{aVL}, \mathrm{aVF}\}$ comprises unipolar augmented limb leads; and the third group $G_3 = \{\mathrm{V1}, \mathrm{V2}, \mathrm{V3}, \mathrm{V4}, \mathrm{V5}, \mathrm{V6}\}$ includes the precordial leads. More details can be found in the Appendix. To emulate the diagnostic reasoning of clinicians, we propose a hierarchical and explainable model structure in the following section to enhance diagnostic performance during the fine-tuning stage.

Method

Similar to previous SSL methods (Tsai et al. 2020; Wang et al. 2022), we divide representation learning into two stages: (1) pretraining to learn individual lead representations, and (2) fine-tuning to learn task-driven representations of lead combinations, simulating the diagnosis workflow.

Preliminary

Masked Autoencoder. We present the formulation of masked autoencoder (MAE) (He et al. 2022) tailored for sequence modeling, which constitutes a core component and the baseline of our approach. MAE is a widely used self-supervised learning framework that learns meaningful representations by reconstructing missing parts of the input. Following MAE, we formulate the framework as follows. Given an input sequence $\mathcal{B} = (b_1, b_2, \ldots, b_T)$ and a random masked position k, the modified sequence can be formulated as $\mathcal{B}_{masked} = (b_1, b_2, \ldots, \beta_k, \ldots, b_T)$, where β_k denotes the masked element filled with zero values at the k-th position in the sequence. Note that there is more than one masked position within the input sequence in practice. Given an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, the reconstructed data can be represented as $\hat{\mathcal{B}} = \mathcal{D}(\mathcal{E}(\mathcal{B}_{masked}))$. Then, the learning object of MAE is:

$$\mathcal{L} = \mu(||\hat{\mathcal{B}} - \mathcal{B}||_2^2), \tag{1}$$

where $\mu(\cdot)$ is the average operator.

Attention Mask. Attention mask (Vaswani et al. 2017) is a technique suitable for next-token prediction tasks to avoid data leak, which is widely used in attention layers of transformer models. Given an input feature X, the procedure of the self-attention mechanism can be formulated as follows:

$$ATN(X) = softmax(\frac{XW_Q W_K^\top X^\top}{\sqrt{d_k}})XW_V,$$

where W_Q, W_K, and W_V are the weights for query, key, and value in the attention. d_k is the dimension of the key weight. To capture the important information and accelerate the computation, the mask technique is applied in the attention, which can be formulated as follows:

$$ATN(X, \mathcal{M}) = softmax(\frac{XW_Q W_K^\top X^\top}{\sqrt{d_k}} + \mathcal{M})XW_V, \tag{2}$$

where $\mathcal{M}$ is a $n_t \times n_t$ matrix, where n_t is the number of the tokens in X. Specifically, $\mathcal{M}_{[i,j]} = -\infty$ represents ignoring position (i, j) and $\mathcal{M}_{[i,j]} = 0$ represents keep the position (i, j) during attention computation.

Pretraining: Conduction-Lead Reconstructor

To capture individual heartbeats' details, we first segment a lead L_i into a set of beats (tokens) $B_i = \{b_i^j\}_{j=1}^{N}$, where $i \in \{1, \ldots, 12\}$. This process resembles tokenization, similar to the strategy in prior work (Jin et al. 2025).

Motivated by the conduction-view guided perspectives, we assume that a masked token can be recovered by conduction-guided information I_c and view-guided information I_v. We define (L_i, j) as j-th beat of i-th lead L_i and the relationship can be formulated as $b_i^j = f(I_c^{(\cdot,j)}, I_v^{(L_i,\cdot)})$, where $I_c^{(\cdot,j)}$ represents the conduction-guided information of the j-th beat and $I_v^{(L_i,\cdot)}$ represents the view-guided information of the L_i, f is a mapping to reconstructed beat from corresponding information. Specifically, for the conduction-guided $I_c^{(\cdot,j)}$, we assume that B^j can reflect the characteristics of j-th heartbeat conduction; for view-guided information $I_v^{(L_i,\cdot)}$, we introduce cls tokens for 12 leads $C =$

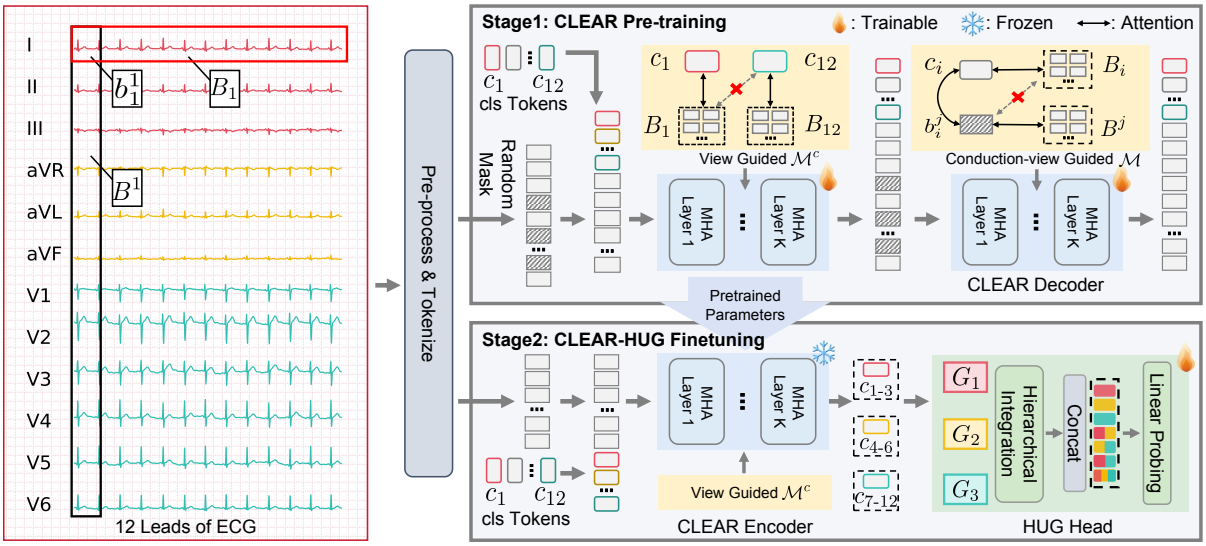


Figure 2: Illustration of proposed CLEAR-HUG framework. CLEAR-HUG is composed of two stages: (1) CLEAR Pre-training stage to learn specified representations of 12 leads, and (2) CLEAR-HUG Finetuning to integrate the lead feature from the pretrained encoder and simulate the clinical diagnosis procedure to provide predictions for downstream tasks. In which MHA layer stands for multi-head attention layer (Vaswani et al. 2017).

$\{c_i\}_{i\in\{1,2,\ldots,12\}}$ to capture the global context of each view of cardiac activity. Thus, the reconstructed beat $\hat{b}_i^j$ can be:

$$\hat{b}_i^j = f(B^j, c_i). \tag{3}$$

This formulation represents that a heartbeat token can be reconstructed by its same heartbeats but different viewed (lead) tokens and the specific global context of the lead. Therefore, the problem becomes how to learn the lead-specific representation c_i and the synchronized heartbeat representation shared by B^j.

Based on the above discussion, we propose a **C**onduction-**LEA**d **R**econstructor (**CLEAR**) to build our pretraining framework, learning representations from heartbeats to leads by the relationship between b_i^j, B^j, and c_i. We first randomly initialize 12 *cls* tokens $C = \{c_i\}_{i\in\{1,2,\ldots,12\}}$. If we mask the token b_i^j, we hope it can be reconstructed only by j-th heartbeat information and the view-guided information c_i, like Equation 3. Thus, **The reconstruction process of b_i^j should be restricted to interactions with the set B^j, c_i. For c_i, it should "see" only the tokens B_i from the same lead L_i, ensuring that it maintains both lead-specific and global representations.** To achieve this sparse relationship and interactions, CLEAR utilizes a sparse attention mechanism as shown in Stage 1 of Fig. 2

To be specific, the input $X \in \mathcal{R}^{12(N+1)\times d_t}$, where d_t is the dimension of the tokens, consists of the tokens from 12 leads and the initialized 12 *cls* tokens. Specifically, $X = [C, B_1, B_2, \ldots, B_{12}]$. Given the input X and the masked tokens $\{\beta_k, k \in K\}$, where $K \in [1, 12N]$ is the position set of the masked tokens, we acquire masked input as X_{masked}. Then, we propose a new sparse attention mechanism tailored for the ECG data in the transformer. Specifically, we separate the attention mask $\mathcal{M} = [m_{p,q}]_{(12N+12)\times(12N+12)}$ into two parts: $\mathcal{M}^c = [m_{p,q}]_{12\times(12N+12)}$ and $\mathcal{M}^b = [m_{p,q}]_{12N\times(12N+12)}$. The attention mask for *cls* tokens $\mathcal{M}^c$ is defined as follows:

$$\mathcal{M}^c_{[p,q]} \triangleq \begin{cases} 0, & if\ q \in Q_p \\ & or\ p = q \\ -\infty, & else. \end{cases} \tag{4}$$

where $Q_p = [(p-1)N+13, pN+12], Q_p \subset Z^+$. And, the attention mask for heartbeat tokens $\mathcal{M}^b$ is defined as:

$$\mathcal{M}^b_{[p,q]} \triangleq \begin{cases} 0, & if\ p \in K, q \in R_p \\ & or\ p \notin K, q \in C_p \\ & or\ q = \left\lfloor \frac{p-1}{N} \right\rfloor + 1 \\ -\infty, & else. \end{cases} \tag{5}$$

where $R_p = \{o+12 | o \bmod 12 = p \bmod 12, o \le 12N\, o \in Z^+\}$, i.e., the position indicates the same heartbeat, $C_p = \{o+12 | o \bmod N = o \bmod N, o \le 12N, o \in Z^+\}$, i.e., the position indicates the same lead. The final attention mask $\mathcal{M}$ is defined as $\begin{bmatrix} \mathcal{M}^c \\ \mathcal{M}^b \end{bmatrix}$.

The view-guided and conduction-guided attention focuses on specific tokens as per the equation. $\mathcal{M}^c$ is applied to both $\mathcal{E}$ and $\mathcal{D}$, while $\mathcal{M}$ is applied only to $\mathcal{D}$, as $\mathcal{E}$'s input doesn't contain masked tokens. Finally, the reconstruction loss of the masked input X_{masked} can be formulated as:

$$\mathcal{L} = \|\mathcal{D}(\mathcal{E}(X_{masked})) - X\|_2^2, \tag{6}$$

Finetuning: Hierarchical Lead-unified Group

According to the insights of lead combination guidelines in ECG diagnosis discussed above, we propose a lead grouping

strategy, **H**ierarchical lead-**U**nified **G**roup (**HUG**), as shown in Fig. 3, to adapt the pretrained eSSL model CLEAR to downstream tasks. As shown in Fig. 1, HUG has 7 linear heads $\{\phi_i\}_{i=1}^{7}$ to simply and effectively learn information of the lead combinations, which aims to mimic the workflow of clinical diagnosis. Given 12 *cls* tokens $\{c_i\}_{i\in\{1,2,...,12\}}$ from the output of CLEAR Encoder, the finetuning pipeline can be summarized as 3 steps:

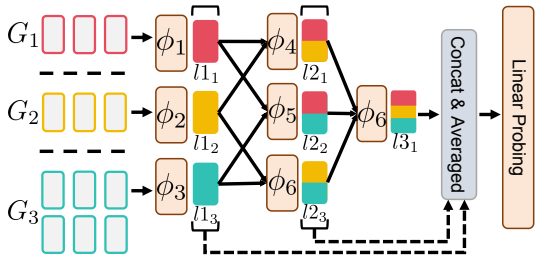


Figure 3: Illustration of proposed HUG head. The HUG head integrates the three ECG lead groups via a three-stage hierarchical framework.

Step 1: Following principles of ECG signal acquisition (Kligfield et al. 2007), we divide *cls* tokens of 12 leads into 3 groups by the guidance aforementioned, i.e., $G_1 = \{I, II, III\}$, $G_2 = \{aVR, aVL, aVF\}$, and $G3 = \{V1, V2, V3, V4, V5, V6\}$. Each group is processed by a separate linear layer ϕ_i, and the outputs are averaged:

$$l1 = \{\mu(\phi_1(G_1)), \mu(\phi_2(G_2)), \mu(\phi_3(G_3))\}$$

Step 2: The outputs of Step 1 are reallocated in pairwise combinations. The recombined token pairs are denoted as $\mathcal{P} = \{(l1_1, l1_2), (l1_1, l1_3), (l1_2, l1_3)\}$, and each pair is processed by a separate head and then averaged:

$$l2 = \{\mu(\phi_4(\mathcal{P}_1)), \mu(\phi_5(\mathcal{P}_2)), \mu(\phi_6(\mathcal{P}_3))\},$$

Step 3: Aggregate the outputs from Step 2:

$$l3 = \mu(\phi_7(l2)).$$

At last, the outputs of the 3 steps are concatenated to form the input of the final linear model.

$$f_g = \mu(l1 \cup l2 \cup \{l3\}).$$

The lead combination of the diagnostic process is emulated through this hierarchical group head, where the seven groups correspond to distinct lead combinations. The hierarchical structure enables progressive integration of information, with each level building upon the representations learned at the previous level.

Experimental Setup

We organized experiments according to the configuration in MERL (Liu et al. 2024). Following MERL's configuration, the CLEAR-HUG is first pre-trained on MIMIC-IV-ECG (Gow et al. 2023) and evaluated on downstream tasks. And we select the SSL methods, SimCLR (Chen et al. 2020), BYOL (Grill et al. 2020), MoCo-v3 (Chen, Xie, and He 2021), and SimSiam (Chen and He 2021), time-series SSL method, TS-TCC (Eldele et al. 2021), and eSSL methods, CLOCS (Kiyasseh, Zhu, and Clifton 2021), ASTCL (Wang et al. 2023a), CRT (Zhang et al. 2023), ST-MEM (Na et al. 2024b), and HeartLang (Jin et al. 2025) for comparison.

Setup of CLEAR Pre-training

Datasets. Following previous works (Zhang et al. 2023; Na et al. 2024a; Jin et al. 2025), we use MIMIC-IV-ECG (Gow et al. 2023) to pre-train CLEAR-HUG. This dataset comprises 161,352 subjects with 800,035 12-lead ECG recordings. Each sample lasts for 10 seconds at 500 Hz. We further process the dataset with (1) splitting the signal according to heartbeat detection; (2) replacing the "NaN" and "Inf" values in the ECG recordings with the average of six neighboring points, similar to HeartLang (Jin et al. 2025).

Implementation Detail. We downsample all the ECG records of MIMIC to 100 Hz and tokenize each lead as 15 heartbeat tokens with a $[cls]$ token using the designed tokenizer. We set the learning rate to 5×10^{-4} and trained for 100 epochs. We set the mask ratio of input tokens as 80%. Analysis of the hyperparameter can be found in the appendix. The AdamW optimizer and cosine annealing scheduler are applied for learning rate adjusting. All experiments were conducted on 6 NVIDIA A100 GPUs, and the batch size is set to 256 for each GPU.

Setup of Downstream Funetuning

Datasets. We conduct experiments on downstream tasks using three publicly accessible datasets: PTB-XL (Wagner et al. 2020), CPSC2018 (Liu et al. 2018), and Chapman-Shaoxing-Ningbo (CSN) (Zheng et al. 2020).

PTB-XL: PTB-XL includes 21,837 12-lead ECG recordings from 18,885 patients, each lasting 10 seconds with 500 Hz sampling rate. The dataset is divided into four subsets based on the SCP-ECG protocol: Form (19 classes), Rhythm (12 classes), Superclass (5 classes), and Subclass (23 classes). We followed the official configuration for splitting the training, validation, and test sets.

CPSC2018: CPSC2018 contains 6,877 12-lead ECG recordings a 500 Hz sampling rate. Unlike PTB-XL, the ECG duration ranges from 6 to 60 seconds, with 9 labels. We split the data into training(70%), validation (10%), and testing (20%) sets, following the HeartLang configuration.

CSN: CSN includes 45,152 12-lead ECG recordings sampled at 500 Hz, each lasting 10 seconds. We use 23,026 recordings with 38 labels, excluding "unknown" annotations, following MERL. The dataset is split into 70% training, 10% validation, and 20% testing.

Implementation Detail. We down-sample the ECG signals to 100 Hz before fine-tuning CLEAR-HUG, using a pre-trained tokenizer to generate 16 tokens per lead. We set the HUG head, and the linear classifier is trainable. The HUG head and linear classifier are trainable, with the CLEAR Encoder initialized from pre-trained parameters. To compare with previous methods under low-resource conditions, we follow work (Jin et al. 2025) using 1%, 10%, and 100% of the training data for each task to finetune the modules. The learning rate is set as 5×10^{-3}, and the training converged

Method	Source	PTBXL-Sub			PTBXL-Super			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
		1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
SimCLR	ICML2020	60.84	68.27	73.39	63.41	69.77	73.53	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	NeurIPS2020	57.16	67.44	71.64	71.70	73.83	76.45	48.73	61.63	70.82	41.99	74.40	77.17	60.88	74.42	78.75	54.20	71.92	74.69
MoCo-v3	ICCV2021	55.88	69.21	76.69	73.19	76.65	78.26	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	CVPR2021	62.52	69.31	76.38	73.15	72.70	75.63	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	IJCAI2021	53.54	66.98	77.87	70.73	75.88	78.91	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	ICML2021	57.94	72.55	76.24	68.94	73.36	76.31	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	TNNLS2024	61.86	68.77	76.51	72.51	77.31	81.02	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	TNNLS2023	61.98	70.82	78.67	69.68	78.24	77.24	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	ICLR2024	54.12	57.86	63.59	61.12	66.87	71.36	55.71	59.99	66.07	51.12	65.44	74.85	56.69	63.32	70.39	59.77	66.87	71.36
Heartlang	ICLR2025	64.68	79.34	88.91	78.94	85.59	87.52	58.70	63.99	80.23	62.08	76.22	90.34	60.44	66.26	77.87	57.94	68.93	82.49
CLEAR	Ours	73.86	80.11	89.3	78.59	85.43	88.68	61.00	69.96	80.34	79.24	86.72	93.66	64.82	78.10	89.59	62.88	76.33	86.75
Gains(%)	-	**+9.18**	**+0.77**	**+0.39**	-0.35	-0.16	**+1.16**	**+2.30**	**+5.97**	**+0.11**	**+17.16**	**+10.5**	**+3.32**	**+2.69**	**+0.32**	**+7.56**	**+3.11**	**+2.63**	**+4.26**
CLEAR-HUG	Ours	76.66	84.59	91.44	79.61	86.85	90.24	61.01	75.19	82.86	79.48	90.57	94.08	66.97	82.59	91.17	72.09	82.14	89.93
Gains(%)	-	**+11.98**	**+5.25**	**+2.53**	**+0.67**	**+1.26**	**+2.72**	**+2.31**	**+11.2**	**+2.63**	**+17.4**	**+14.35**	**+3.74**	**+4.84**	**+4.81**	**+9.14**	**+12.32**	**+8.44**	**+7.44**

Table 1: Comparison of proposed framework with other eSSL methods on six downstream tasks. We designed two settings of our proposed framework, the first is the pretrained CLEAR, followed by linear probing. The second is the fully CLEAR-HUG framework for linear probing. The SOTA results are underlined and gains for both settings are compared with them.

Method		PTBXL-Sub			PTBXL-Super			PTBXL-Form			PTBXL-Rhythm		
Pre-training	Downstream	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
w/o pretraining	N/A	65.16	77.10	85.89	76.64	83.77	87.16	54.43	63.98	73.68	51.51	77.80	86.65
Baseline	N/A	67.49	77.68	89.00	74.49	84.07	87.33	55.68	64.13	77.06	73.73	83.19	91.52
Baseline	HUG	75.41	83.76	90.74	78.70	85.57	88.76	56.08	70.38	78.39	73.76	86.68	91.60
CLEAR	N/A	73.86	80.11	89.30	78.59	85.43	88.68	61.00	69.96	80.34	79.24	86.72	93.66
CLEAR	HUG	**76.66**	**84.59**	**91.44**	**79.61**	**86.85**	**90.24**	**61.01**	**75.19**	**82.86**	**79.48**	**90.57**	**94.08**

Table 2: Results of the ablation study of modules for CLEAR-HUG. Best scores are in bold. The pre-training baseline is a masked autoencoder. 'w/o pretraining' denotes direct finetuning from scratch with the downstream head; SLG denotes lead grouping without hierarchy. 'N/A' (downstream) denotes linear probing on the frozen pre-trained backbone.

within 100 epochs. Besides, we scaled the ECG recordings to $[-3, 3]$ for CPSC2018 and CSN datasets, obtained all test results using the trained parameters with the best validation results, and used the macro AUC metric for evaluation, to maintain a fair comparison with previous works (Zhang et al. 2023; Na et al. 2024a; Jin et al. 2025).

Results and Analysis

Primary Results

Table 1 shows the comparison of our CLEAR-HUG with the SOTA methods. To present a fair comparison, we designed two settings, where the first directly uses the linear probing layer for downstream tasks using features from the pretrained CLEAR Encoder. While the second introduced our proposed HUG head between the pretrained backbone and the linear probing layer to integrate features of 12 leads further. Compared with the SOTA methods, both CLEAR and CLEAR-HUG demonstrated significant advantages in almost all downstream tasks, with 3.94% and 6.84% improvements in the average separately. Notably, in the PTBXL-Rhythm task, both CLEAR and CLEAR-HUG outperform the SOTA results by over 17% and 10%, respectively, when fine-tuned on 1% and 10% of the training data. This result further confirms the advantages of our design for heartbeat conduction in representing and analyzing heartbeat rhythms.

Further comparing the CLEAR and CLEAR-HUG, we found that the designed lead fusion strategy can better overcome the challenges of more diverse category classification in ECG understanding tasks. For example, in the CSN dataset, when introducing our designed HUG, our CLEAR-HUG gains an additional performance improvement of over 9.21%, 5.81%, and 3.18% when finetuning on 1%, 10%, and 100% of the training data, compared to CLEAR. Furthermore, the performance of CLEAR-HUG is improved in all downstream task settings by a clear margin. We attribute this result to the fact that the hierarchical grouping strategy is more in line with the clinical diagnosis process and can better aggregate lead features to achieve adaptation to more complex scenarios and more diverse categories.

Ablation Study

Designed Modules To explore the potential necessity of the CLEAR pre-training and the HUG finetuning, we designed several ablation studies as shown in Table 2. In the experiment, we considered four combinations of with/without CLEAR pre-training and with/without HUG head. Notably, our baseline is the native masked autoencoder model, having the same learnable parameters as CLEAR, but using full attention instead of the designed sparse attention.

From the perspective of CLEAR pretraining, experiments show our designed conduction-view guidance can effectively improve model performance, independent of the impact of the HUG head. We attribute this improvement to CLEAR pretraining's ability to better integrate both common and lead-specific information within each lead. This is crucial for tasks involving the identification of heartbeat patterns and rhythm regularity, such as heart rhythm analysis, where we observed a 17.4% improvement with 1% of the training data. And for the Hierarchical Unified Group

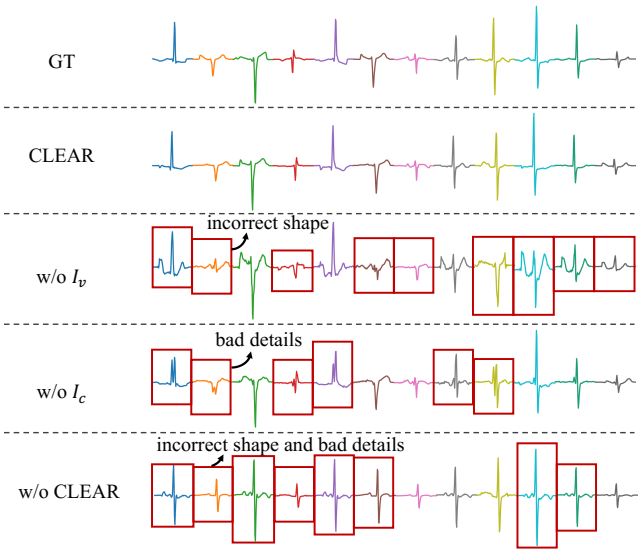


Figure 4: Visualization of reconstructed tokens for CLEAR variants w/o components: 12 leads in different colors, with poor reconstructions marked by red boxes.

(HUG) head, its introduction improves model performance overall, with more pronounced benefits in fine-grained tasks such as sub-class categorization. We attribute this improvement to the grouping strategy, whose hierarchical structure better aligns with clinical diagnosis processes while enabling more comprehensive utilization of 12-lead information. We will further analyze the internal design of our model and provide a series of in-depth analyses in the following.

Method	CSN 1%	CSN 10%	CSN 100%
N/A (Averaged)	62.88	76.33	86.75
Weighted Averaged	66.73	80.02	88.11
Single-level Grouping	68.70	80.21	89.35
HUG Head	**72.09**	**82.14**	**89.93**

Table 3: Comparison of different designs to integrate 12 leads for downstream tasks.

Insights of CLEAR-HUG

I_v and I_c visualization of pretrained model. We visualize heartbeat reconstruction under four settings: CLEAR, w/o I_v, w/o I_c, and w/o CLEAR. As shown in Fig. 4, CLEAR yields the best reconstruction. We further analyze the reconstruction effect as follows: (1) When removing I_v (only conduction-guided information), the reconstructed signal shows an incorrect wave shape, consistent with I_v encoding the global lead context and overall waveform contour. (2) When removing I_c (only view-guided information), the global shape is preserved but fine details are lost, supporting that I_c captures heartbeat-specific details. (3) When both I_c and I_v are removed, neither shape nor details are correct, further validating our design of the two guidance terms.

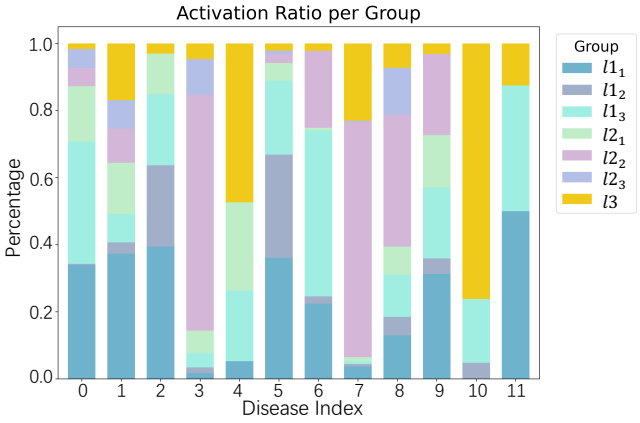


Figure 5: The visualization of activation ratios on 7 group combinations from HUG on different diseases.

Grouping in finetuning. To further investigate the impact of combining the three groups in HUG, we visualize the activation patterns of HUG on PTBXL-form in Fig. 5. The visualization demonstrates that HUG exhibits diagnosis-specific activation patterns across its different group combinations, varying with distinct types of cardiac abnormalities. This differential activation directly corresponds to the feature of different cardiac conditions, confirming the ability of ClEAR-HUG to autonomously extract diagnosis-critical features. Further analysis can be found in the appendix.

Hierarchical lead integration. We also evaluate different integration strategies for grouping 12 leads. We select the CSN dataset as our benchmark and compare our HUG design with the average, weighted average, and single-level grouping methods. We note that single-level grouping integrates 3 groups through full combination and generates 7 combinations as HUG. Further ablation studies (Table 3) demonstrate that the hierarchical lead-grouping (HUG) design outperforms alternative approaches by a clear margin. We attribute this to the fact that the hierarchical design is closer to clinical diagnosis and can more effectively utilize the information of the 12 leads. Results of other datasets are available in the appendix.

Conclusion and Limitation

In this paper, we propose CLEAR-HUG, which comprises: (1) CLEAR (Conduction-LEAd Reconstructor), which uses conduction-inspired sparse attention to guide masked 12-lead ECG token reconstruction; (2) HUG (Hierarchical lead-Unified Group head), which employs a hierarchical grouping strategy to mimic clinical diagnosis. Experiments on 6 datasets show that CLEAR-HUG surpasses existing methods across diverse ECG analysis tasks, and ablations verify the contribution of each component. **Limitations.** CLEAR-HUG could be extended to more downstream tasks. Moreover, HUG is tightly coupled with our pretraining design, making it incompatible with prior ECG foundation models that use only a single *cls* token.

Acknowledgments

This work was supported by Fudan University AI4S Project (FudanX24AI056), Pujiang Talent Program (24PJD014), and Shanghai Municipal Science and Technology Major Project (2023SHZDZX02 and 2017SHZDZX01 to L.J.). The computations in this research were performed using the CFFF platform of Fudan University.

References

Alimbayeva, Z.; Alimbayev, C.; Ozhikenov, K.; Bayanbay, N.; and Ozhikenova, A. 2024. Wearable ECG device and machine learning for heart monitoring. *Sensors*, 24(13): 4201.

Auer, R.; Bauer, D. C.; Marques-Vidal, P.; Butler, J.; Min, L. J.; Cornuz, J.; Satterfield, S.; Newman, A. B.; Vittinghoff, E.; Rodondi, N.; et al. 2012. Association of major and minor ECG abnormalities with coronary heart disease events. *Jama*, 307(14): 1497–1505.

Berkaya, S. K.; Uysal, A. K.; Gunal, E. S.; Ergin, S.; Gunal, S.; and Gulmezoglu, M. B. 2018. A survey on ECG analysis. *Biomedical Signal Processing and Control*, 43: 216–235.

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

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

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

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Gow, B.; Pollard, T.; Nathanson, L. A.; Johnson, A.; Moody, B.; Fernandes, C.; Greenbaum, N.; Waks, J. W.; Eslami, P.; Carbonati, T.; et al. 2023. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *Type: dataset*, 6: 13–14.

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

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. R.; and Xu, D. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, 272–284. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Jin, J.; Wang, H.; Li, H.; Li, J.; Pan, J.; and Hong, S. 2025. Reading Your Heart: Learning ECG Words and Sentences via Pre-training ECG Language Model. In *The Thirteenth International Conference on Learning Representations*.

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

Kligfield, P.; Gettes, L. S.; Bailey, J. J.; Childers, R.; Deal, B. J.; Hancock, E. W.; Van Herpen, G.; Kors, J. A.; Macfarlane, P.; Mirvis, D. M.; et al. 2007. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. *Journal of the American College of Cardiology*, 49(10): 1109–1127.

Liu, C.; Wan, Z.; Ouyang, C.; Shah, A.; Bai, W.; and Arcucci, R. 2024. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, 31949–31963.

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

Mirvis, D. M.; and Goldberger, A. L. 2001. Electrocardiography. *Heart disease*, 1: 82–128.

Na, Y.; Park, M.; Tae, Y.; and Joo, S. 2024a. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. In *The Twelfth International Conference on Learning Representations*.

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

Pan, T.; Tan, Z.; Guo, K.; Xu, D.; Xu, W.; Jiang, C.; Guo, X.; Qi, Y.; and Cheng, Y. 2025. Structure-aware Semantic Discrepancy and Consistency for 3D Medical Image Self-supervised Learning. *arXiv preprint arXiv:2507.02581*.

Poterucha, T. J.; Jing, L.; Ricart, R. P.; Adjei-Mosi, M.; Finer, J.; Hartzel, D.; Kelsey, C.; Long, A.; Rocha, D.; Ruhl, J. A.; et al. 2025. Detecting structural heart disease from electrocardiograms using AI. *Nature*, 1–10.

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2020. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Kreiseler, D.; Lunze, F. I.; Samek, W.; and Schaeffter, T. 2020. PTB-XL,

a large publicly available electrocardiography dataset. *Scientific data*, 7(1): 1–15.

Wang, H.; Guo, X.; Deng, Z.-H.; and Lu, Y. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16041–16050.

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

Wang, Y.; Li, Z.; Mei, J.; Wei, Z.; Liu, L.; Wang, C.; Sang, S.; Yuille, A. L.; Xie, C.; and Zhou, Y. 2023b. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 486–496. Springer.

Wu, L.; Zhuang, J.; and Chen, H. 2024. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22873–22882.

Xia, H.; Asif, I.; and Zhao, X. 2013. Cloud-ECG for real time ECG monitoring and analysis. *Computer methods and programs in biomedicine*, 110(3): 253–259.

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

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

Zheng, J.; Zhang, J.; Danioko, S.; Yao, H.; Guo, H.; and Rakovski, C. 2020. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1): 48.

Zhou, S.; Huang, X.; Liu, N.; Zhang, W.; Zhang, Y.-T.; and Chung, F.-L. 2024. Open-world electrocardiogram classification via domain knowledge-driven contrastive learning. *Neural Networks*, 179: 106551.