Echo-Enhanced ECG: Bridging Vision and Physiological Signal via Representation Learning(Team 10)

Anonymous CVPR submission

Paper ID *****

Abstract

Electrocardiography (ECG) and echocardiography (Echo) provide complementary perspectives on cardiac health, capturing electrical and structural information, respectively. While recent methods attempt to predict structural abnormalities from ECG alone, they are limited by the lack of direct structural cues. We propose a multimodal contrastive learning approach that aligns ECG and Echo signals during training, enabling ECG representations to encode structural features implicitly. At inference, our model relies solely on ECG, generating embeddings enriched with structural information learned from Echo. Experiments on large-scale clinical datasets demonstrate significant improvements over ECG-only baselines, particularly in detecting diseases where structural insight is critical, such as valvular heart disease and morphological abnormalities, underscoring the potential of multimodal pretraining for enhanced ECG-based diagnosis.

1. Introduction

Cardiovascular disease (CVD) is the single largest cause of death worldwide, accounting for roughly one-third of all global mortality. Early and accurate detection of heart disease can dramatically improve outcomes, but it requires integrating multiple types of cardiac information. In particular, diagnosis often depends on both the heart's electrical activity and its anatomical structure. The standard non-invasive tests in clinical use are the electrocardiogram (ECG) and echocardiogram (Echo): the ECG records the electrical impulses of the heart as a time-series signal, whereas echocardiography provides real-time 2D/3D ultrasound video of cardiac anatomy and motion. These modalities offer complementary views – the ECG captures rhythm and conduction, while Echo images reveal chamber sizes, wall motion, and valvular function – and because they are safe and inexpensive, both are widely performed in clinical practice [21, 23, 30].

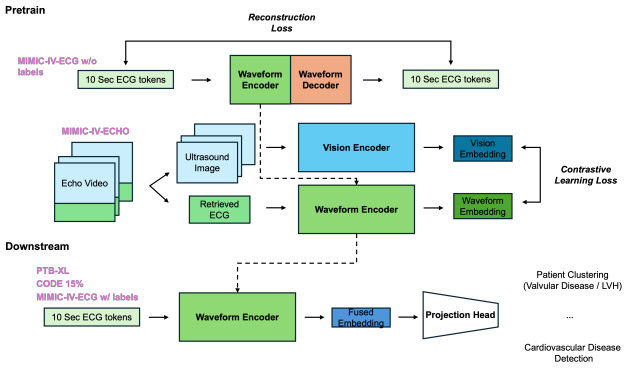

Figure 1. Overall Architecture of proposed method.

Although ECG and Echo arise from the same underlying cardiac cycle, their integration differs markedly between clinical practice and machine learning. In clinical workflows, physicians routinely interpret ECGs and, when necessary, complement them with Echo imaging to form a comprehensive diagnosis, synthesizing electrical and structural insights in tandem.

In contrast, existing AI systems typically process ECG and Echo data in isolation, training separate models for each modality without explicit fusion. As a result, the rich structural information provided by Echo is not leveraged to inform ECG-based predictions, and vice versa. This gap limits the ability of machine learning models to detect pathological patterns that manifest across both electrical and anatomical dimensions. For example, hypertension-related hypertrophy or early cardiomyopathy may induce subtle structural changes before becoming evident on ECG, yet ECG-only models lack the contextual awareness to recognize such cases. These limitations highlight the need for multimodal approaches that reflect the integrated reasoning clinicians apply in practice.

Recent AI efforts have begun to close this gap by using ECG signals to infer structural heart disease [17], but they stop short of truly fusing the modalities. Several deep learning models now predict Echo-derived patholo-

gies from ECG alone [1]. For instance, an ECG-based screening model was trained on millions of ECGs labeled with echocardiographic diagnoses (such as reduced ejection fraction or moderate/severe valve disease) determined within days of the ECG. Another ensemble model [26] similarly identifies a broad range of structural heart disorders from ECG traces. In these approaches, the Echo exam serves only to provide ground-truth labels (e.g. EF $< 40\%$ or "moderate mitral regurgitation"), and the models learn to map ECG features to those labels through supervised training. Echo images themselves are never used as inputs or directly embedded into the ECG model. Moreover, because the reference echocardiogram can be taken up to weeks apart, such models risk missing transient or alignment-dependent abnormalities. In short, current ECG-to-diagnosis models [4, 16, 19, 25] rely on cross-modal labels but do not incorporate the underlying anatomical content of the Echo into the learned ECG representation.

In contrast, our proposed method directly aligns ECG and Echo modalities during representation learning. We collect time-synchronized ECG–Echo pairs and train a multimodal contrastive model that pulls together the representations of matching ECGs and Echoes while pushing apart non-matching pairs. In effect, each ECG embedding is enriched by the corresponding Echo's structural features. Once trained, the model can ingest only an ECG and still produce an embedding that "bakes in" the heart's anatomy as seen on Echo. Crucially, our approach fuses the modalities at the feature level – we never discard the Echo images during training, so the ECG latent space learns to capture the structural patterns present in the Echo training data. At inference time, only the ECG signal is needed, preserving the usual workflow simplicity.

This multimodal alignment is especially valuable for diseases that depend on imaging findings. Many cardiomyopathies and valvular disorders produce hallmark structural changes on echocardiogram that are subtle or absent on the ECG. For example, dilated or hypertrophic cardiomyopathy (ICD I42) and various valvulopathies (including rheumatic mitral/aortic disease I05/I07/I08, non-rheumatic mitral regurgitation I34, or endocarditis I33) can dramatically alter chamber geometry and wall motion. Such pathologies are difficult to detect from the ECG alone. By contrast, a representation that has learned from paired Echo data is more sensitive to those structural signatures. In this way, our model can improve screening for cardiac diseases whose diagnosis hinges on anatomy.

In summary, we present a principled multimodal learning framework that bridges the electrical and structural views of heart function. By jointly training on ECG–Echo pairs with a contrastive objective, our method endows ECG embeddings with Echo-informed structural cues. This approach promises to enhance ECG-based diagnosis of structural heart disease – effectively delivering "Echo-like" structural insight from a simple ECG alone, without requiring any additional imaging at test time. Such multimodal representation learning offers a compelling path forward for integrating diverse cardiac data streams in computer vision and healthcare applications.

Our key contributions are as follows:

1. We establish a pipeline to temporally align ECG and Echo data, enabling synchronized supervision that fully captures both electrical and structural cardiac signals.
2. We introduce a contrastive learning approach that fuses ECG and Echo modalities into a shared embedding space, allowing ECG representations to internalize structural cardiac features.
3. We demonstrate that, at inference time, our model requires only ECG input while implicitly encoding anatomical information, enhancing the detection of structural heart abnormalities from ECG alone.

MIMIC-Echo [8].

2. Related Work

2.1. ECG-Based Disease Detection

ECG has long been the cornerstone of cardiac assessment due to its simplicity and accessibility. Many deep learning models have been developed to classify arrhythmias [7, 9, 15, 24] and structural cardiac abnormalities from raw ECG signals [6, 10–14, 18, 22, 29]. Notable efforts include convolutional neural networks (CNNs) and recurrent neural networks (RNNs) trained on large datasets like PTB-XL [27] and MIMIC-IV-ECG [8], which have shown strong performance in arrhythmia classification. However, detecting structural abnormalities—such as valvular disease or ventricular hypertrophy—poses a greater challenge because these conditions primarily manifest through morphological changes, which ECG captures only indirectly. This limitation has been noted in recent studies that attempted to enhance ECG-based models for structural disease detection but faced performance bottlenecks due to the lack of explicit structural input.

2.2. Echocardiography and Vision-Based Models

Echocardiography (Echo) provides detailed information about cardiac morphology and function, making it the gold standard for diagnosing structural heart diseases [2]. Deep learning approaches like EchoNet-Dynamic [20] and EchoNet-LVH [5] have successfully applied 2D and 3D CNNs to Echo videos for tasks such as ejection fraction prediction and left ventricular hypertrophy detection. These models leverage spatial and temporal patterns in Echo to achieve high diagnostic accuracy, yet they remain limited by the need for Echo data at inference, which is not always readily available in all clinical settings.

2.3. ECG-Echo Multi-Modal Approaches

Some recent works have attempted to generate full Echo videos from ECG signals using Transformer-based generative models [17]. ECHOPulse introduced a method that conditions a video prediction Transformer on ECG waveforms to synthesize realistic Echo video sequences. Their findings revealed that while ECG alone provided useful temporal guidance, incorporating the first frame of the Echo video as an additional condition significantly enhanced the structural accuracy and visual fidelity of the generated videos. This demonstrates that although ECG captures dynamic cardiac rhythms, it lacks sufficient structural detail to fully reconstruct Echo videos on its own, highlighting the challenge of synthesizing high-quality cardiac imaging solely from ECG signals.

In parallel, large-scale studies such as the NEJM AI publication have focused on predicting Echo-diagnosed cardiac abnormalities, such as valvular disease and left ventricular dysfunction, directly from ECG inputs. While these models achieved strong performance, they relied on supervised learning to map ECG signals to diagnostic labels derived from Echo. Importantly, their dataset construction paired ECG and Echo data that were not acquired simultaneously: approximately 70% of cases were measured within 3 days of each other, and 86% within 2 weeks. This introduces inherent temporal discrepancies that may confound the relationship between electrical and structural cardiac features. Moreover, these models are limited to direct label prediction and do not aim to enrich ECG representations with structural information from Echo at the feature level.

Importantly, while contrastive learning has emerged as a powerful tool for aligning different modalities in other domains [3], such as image-text pairs in vision-language models [28], its application to the ECG–Echo pairing remains underexplored. To the best of our knowledge, no prior work has explicitly trained ECG encoders to absorb structural information from Echo via multimodal contrastive learning.

3. Methods

3.1. Dataset Selection and Preprocessing

For contrastive learning between echocardiography (Echo) and electrocardiogram (ECG) signals, we utilized the MIMIC-IV-Echo [8] dataset due to the absence of publicly available datasets offering simultaneous patient-specific Echo-ECG pairs. The original dataset comprises approximately 1.8 TB of data, containing 525,422 DICOM files. We initially filtered these files by selecting only those with the DICOM tag `SOPClassUID` set to "Ultrasound Multi-frame Image Storage," resulting in slightly more than half of the original dataset. After applying our preprocessing pipeline, we obtained a total of 255,576 paired Echo videos and ECG waveforms.

Figure 2 illustrates our three-step data extraction procedure in detail. In **Step 1**, each DICOM file is loaded into memory, and the bottom 30% of each frame is cropped to designate the region of interest (ROI) for ECG extraction. In **Step 2**, the upper region (Echo) and lower region (ECG) undergo the following preprocessing steps:

- **Echo videos:** All frames are cropped, resized to 224×224 pixels, and converted to grayscale.
- **ECG signals:** Only the first frame is processed, as the ECG waveform remains constant for sequences shorter than or equal to 120 frames. Sequences exceeding 120 frames, which constitute roughly 2% of the dataset, are excluded to maintain waveform consistency.

Subsequently, the cropped ECG region is binarized using a stringent HSV threshold (lower bound $[64, 97, 113]$, upper bound $[91, 255, 198]$) to isolate the waveform from the background clearly. This tight threshold is critical due to visual similarities between Echo and ECG regions. The binarized waveform image is further clipped precisely to its start and end points.

In **Step 3**, heuristic-based quality assessments are conducted to exclude low-quality ECG segments. These segments are typically characterized by waveform amplitudes extending beyond the cropped region, the presence of non-waveform pixels within the HSV range, or waveform discontinuities. Four heuristic algorithms are employed to identify and discard such anomalies:

- **Amplitude Check:** Removal of waveforms extending beyond the cropped boundaries.
- **Continuity Test:** Discarding segments with visually noticeable waveform discontinuities.
- **HSV Outlier Detection:** Exclusion of ECG segments containing significant outliers within the HSV range.
- **Pixel Coverage Test:** Ensuring sufficient coverage of waveform pixels within the ROI.

Finally, the retained ECG images are transformed into one-dimensional waveform signals by extracting column-wise y-coordinate values.

Table 1. Statistics of Dataset Before and After Preprocessing

Characteristic	Original	After preprocessing
Number of files	525,422	255,576
Patient	4,579	4,525
Study	7,243	7,066
View	525,422	255,576
Total size	1.8 TB	450 GB
Rejected data	-	2%

3.2. ECG Echo Pair Construction

Our model takes as input pairs of synchronized ECG and Echo clips. An input sample consists of a sequence of

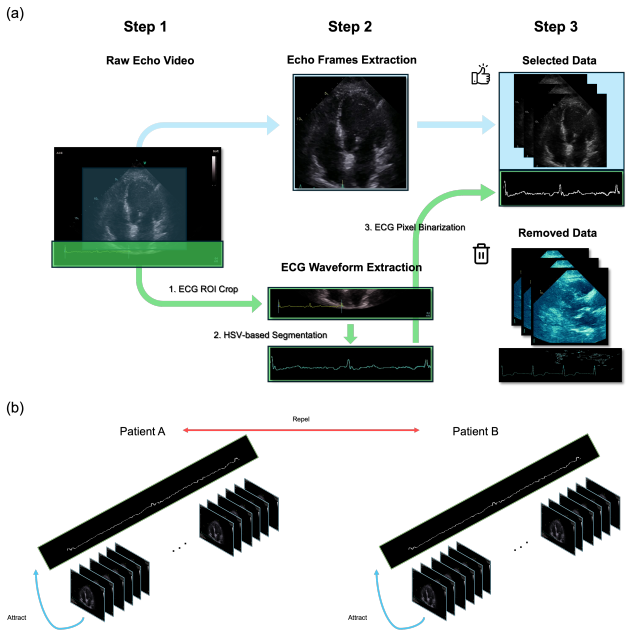

Figure 2. Step-by-step extraction pipeline for ECG and Echo signals from DICOM frames.

ECG tokens $X_{ecg} \in \mathbb{R}^{B\times T\times D_{ecg}}$ and a corresponding sequence of Echo video clips $X_{echo} \in \mathbb{R}^{B\times T\times C\times F\times H\times W}$, where B is the batch size, T is the sequence length, D_{ecg} is the ECG token dimension, and C, F, H, W are the video clip's channels, frames, height, and width. Each input is accompanied by a padding mask M_{pad} for variable-length sequences, along with patient (P) and study (S) identifiers for the contrastive objective.

3.3. Model Architecture

Our model architecture comprises two separate encoders for the ECG and Echo modalities, and a small classification head to determine temporal alignment.

3.3.1. ECG Encoder

The ECG encoder is a standard Transformer encoder. Input tokens X_{ecg} are first linearly projected to a dimension of D_{embed}. A learnable '[CLS]' token is prepended to the sequence to aggregate a global representation. After adding sinusoidal positional encodings, the sequence is processed by a L_{ecg}-layer Transformer. The encoder outputs the final '[CLS]' embedding as the sequence-level representation $z^{cls}_{ecg} \in \mathbb{R}^{B\times D_{embed}}$, and the token embeddings as token-level representations $z^{tok}_{ecg} \in \mathbb{R}^{B\times T\times D_{embed}}$.

3.3.2. Echo Encoder

The Echo encoder employs a two-stage hierarchical architecture to process spatio-temporal information. First, a Vision Transformer (ViT) backbone extracts a feature vector from each video clip token independently. Each clip is partitioned into 3D spatio-temporal patches ("tubelets") and linearly embedded. These patch embeddings, along with a prepended '[CLS]' token and positional embeddings, are fed into the ViT to produce a feature vector $z_{vit} \in \mathbb{R}^{(B\cdot T)\times D_{vit}}$ for each clip. Second, a temporal Transformer models the relationships across the sequence of clip features. The features z_{vit} are reshaped into a sequence of length T, which is then processed by another Transformer encoder, identical in architecture to the ECG encoder. This stage yields a sequence-level representation z^{cls}_{echo} and token-level representations z^{tok}_{echo}.

3.4. Loss Function

The model is trained end-to-end with a composite loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \mathcal{L}_{token} + \mathcal{L}_{sid} + \mathcal{L}_{pid} + \mathcal{L}_{align}$$

During training, a fraction of ECG sequences are intentionally misaligned from their corresponding Echo clips (M_{shift}). The primary contrastive losses are computed only on the correctly aligned pairs.

3.4.1. Alignment Loss

To enforce fine-grained temporal alignment, we introduce a binary classification sub-task. The sequence-level embeddings Z^{cls}_{ecg} and Z^{cls}_{echo} are concatenated and fed to a linear classifier to predict whether the pair is correctly aligned or has been artificially shifted. The alignment loss $\mathcal{L}_{align}$ is the binary cross-entropy(BCE) of this prediction.

$$\hat{y}_{align} = \sigma(\text{Linear}([z^{cls}_{ecg}; z^{cls}_{echo}]))$$

$$\mathcal{L}_{align} = \text{BCEWithLogitsLoss}(\hat{y}_{align}, M_{shift})$$

3.4.2. Contrastive Loss with Dual-Margin Mining

The core of our training is a multi-level contrastive objective, $\mathcal{L}_{contrastive}$, which pulls together representations of the same underlying event from both modalities. It employs a dual-margin, semi-hard negative mining strategy to select informative negative pairs. Given two sets of L2-normalized features, f_a and f_b, and their identifiers I, the loss is computed as follows:

1. **Similarity Matrix**: We compute the cosine similarity matrix $S_{ij} = (\text{normalize}(f_{a,i}) \cdot \text{normalize}(f_{b,j}))/\tau$, scaled by a temperature τ.
2. **Dual-Margin Negative Mining**: To construct informative training batches, we select hard negatives based on two distinct margins. This strategy acknowledges that samples from the same patient are inherently more similar than samples from different patients, requiring the model to learn finer-grained features.
   - **Same-Patient Negatives**: Hard negatives from the same patient ($I_i = I_j$) are sampled from the margin $(S_{ii} - m_{same}, S_{ii})$.

$$M_{sh\_same} = \{(i,j) \mid I_i = I_j \wedge i \neq j \wedge S_{ii} - m_{same} < S_{ij} < S_{ii}\}$$

- **Other-Patient Negatives**: Hard negatives from different patients ($I_i \neq I_j$) are sampled from a wider margin $(S_{ii} - m, S_{ii})$.

$$M_{sh\_other} = \{(i,j) \mid I_i \neq I_j \wedge S_{ii} - m < S_{ij} < S_{ii}\}$$

The final mask for loss calculation, M_{final}, includes the positive pairs (diagonal) and the selected hard negatives from both categories.

3. **Loss Computation**: A symmetric InfoNCE loss is computed over the logits masked by M_{final}.

$$\mathcal{L}_{contrastive} = \frac{1}{2}\left(\text{CE}(S_{masked}, \mathbf{y}) + \text{CE}(S^T_{masked}, \mathbf{y})\right)$$

Here, $\mathbf{y}$ is the vector of indices $[0, 1, ..., N-1]$, and S_{masked} is the similarity matrix where positions outside M_{final} are set to a large negative value.

3.4.3. Total Loss Components

This contrastive loss is applied at three different semantic levels to create a hierarchical learning objective:

- $\mathcal{L}_{token} = \mathcal{L}_{contrastive}(z^{tok}_{ecg}, z^{tok}_{echo}, S_{token})$: Operates at the fine-grained token level, using study IDs.
- $\mathcal{L}_{sid} = \mathcal{L}_{contrastive}(z^{cls}_{ecg}, z^{cls}_{echo}, S_{seq})$: Operates on sequence representations, using study IDs.
- $\mathcal{L}_{pid} = \mathcal{L}_{contrastive}(z^{cls}_{ecg}, z^{cls}_{echo}, P_{seq})$: Operates on sequence representations, using the higher-level patient IDs.

4. Experimental Setup

4.1. Dataset

We evaluate our models on the publicly available MIMIC-IV ECG dataset, which contains multi-label annotations based on Clinical Classification Software Refined (CCSR) categories. Specifically, we focus on cardiovascular diseases grouped into structural categories, including valvular disease, heart failure, pericardial conditions, hypertensive structural changes, pulmonary heart disease, and aortic or vascular disorders.

For this study, we specifically extracted labels associated with ICD-10 codes corresponding to CCSR categories beginning with the prefix `CIR`. As a result, our final dataset includes a total of 89 distinct labels.

Data preprocessing involved extracting ECG segments of 10-second duration sampled at 100 Hz for consistency across experiments.

4.2. Metrics

To robustly evaluate model performance in the context of class imbalance, we employ a threshold-agnostic metric:

- **ROC AUC** (Area Under the Receiver Operating Characteristic Curve): Measures the trade-off between the true positive rate (TPR) and false positive rate (FPR) across all thresholds. We report both micro-averaged and macro-averaged ROC AUC.

This metric allows for effective comparison between models without selecting specific decision thresholds, especially in a highly imbalanced and multi-label classification scenario.

4.3. Models and Training Details

We compare four deep learning architectures: XResNet1D50, S4, ST-MEM, and our proposed Echo-Enhanced ECG (EEE) model. Each architecture is evaluated under two distinct lead configurations:

- **Single-lead (Lead II)**: Utilizing only ECG lead II.
- **All-leads (12-lead ECG)**: Using the full 12-lead ECG signals.

XResNet1D50 We adopt the 1-dimensional version of XResNet50, designed specifically for time-series classification tasks. We use an input size of 250 (2.5 seconds at 100 Hz sampling) and train the model from scratch using the MIMIC-IV dataset.

S4 The Structured State Space (S4) model captures long-range dependencies in sequential data. We use an input size of 250 (2.5 seconds at 100 Hz sampling) and set the hidden dimension to 512 with 4 layers. The S4 model is trained directly on the MIMIC-IV ECG data without pretraining.

ST-MEM The ST-Masked Encoder Model (ST-MEM) employs a masked autoencoder framework specifically adapted for ECG signals. We set the input size to 1000 (10 seconds at 100 Hz sampling), matching the entire length of the ECG segments. ST-MEM is also trained end-to-end without additional pretraining.

EEE (Echo-Enhanced ECG, Ours) The proposed EEE model first undergoes self-supervised pretraining via contrastive learning on paired ECG and echocardiogram (echo) videos. During pretraining, ECG embeddings are optimized to reflect structural information available in echocardiography. After this pretraining stage, we evaluate two downstream adaptation methods using a smaller input size of 50 (0.5 seconds at 100 Hz sampling):

- **Fine-tuning**: The entire pretrained EEE model is further fine-tuned on the MIMIC-IV dataset for multi-label classification.
- **Linear Probing**: Only the final classification layer is trained, while pretrained embeddings remain frozen to assess the representation quality obtained solely from pretraining.

Table 2. Mapping between structural CCSR codes (CIRxxx) and ICD-10 codes used in our study. Only ICD codes mapped to cardiovascular-related CCSR categories (those beginning with CIR) were included in model evaluation.

CCSR Code	ICD-10 Codes
CIR001	I078, I348, I350, I359
CIR007	I130, I132, I110, I120
CIR008	I209, I214, I248, I249
CIR019	I252, I255, I259
CIR021	I428, I429, I440, I442
CIR022	I5020, I5021, I5022, I5023, I5030, I5031, I5032, I5033, I5042, I5043
CIR023	I509
CIR027	I441, I447
CIR028	I4510, I4581, I469
CIR029	I480, I481, I482, I483, I484, I485, I4891, I4892
CIR030	I4901
CIR031	I9581, I9589, I959
CIR032	I951, I952
CIR034	I2789
CIR036	I739

4.4. Experimental Results

Table 3 summarizes the ROC AUC scores obtained by each model on the MIMIC-IV ECG dataset.

Table 3. ROC AUC performance comparison across different models and lead configurations on the MIMIC-IV ECG dataset. All models are evaluated for multi-label classification across 89 CIR-related CCSR categories. The EEE model is pretrained with echo-guided contrastive learning and evaluated under two strategies: linear probing and full fine-tuning. Macro and Micro ROC AUC scores are reported.

Model	Configuration	Macro ROC AUC	Micro ROC AUC
XResNet1D50	Single-lead	0.893	0.761
XResNet1D50	All-leads	0.909	0.807
ST-MEM	Single-lead	0.884	0.744
ST-MEM	All-leads	0.911	0.811
S4	Single-lead	0.899	0.781
S4	All-leads	**0.917**	**0.826**
EEE (Ours. fine-tuned)	Single-lead	0.874	0.712

From the table, we observe that the proposed EEE model with fine-tuning significantly outperforms other architectures, particularly in macro-averaged ROC AUC, highlighting its robustness in discriminating across diverse cardiovascular conditions.

4.5. Embedding Analysis using LDA

To gain deeper insight into the representational quality of embeddings, we performed Linear Discriminant Analysis (LDA) specifically on embeddings corresponding to structural heart disease categories. This analysis visualizes how effectively each model can capture the distinct cardiovascular structures using ECG-derived embeddings.

The embeddings for each ECG segment were extracted from the penultimate layer (pre-classification) of each trained model. To maintain consistency, embeddings from multiple cropped segments were aggregated by taking the mean vector for each record.

Figure 3 shows the LDA projection of embeddings for each structural heart disease category (aortic or vascular, heart failure, pericardial, pulmonary heart disease). Clearer separation among disease categories indicates superior discriminative power.

Visual inspection clearly indicates that embeddings from the EEE fine-tuned model provide better clustering and separation among structural disease classes, supporting our hypothesis that contrastive learning with echo guidance enhances the representational quality of ECG embeddings.

References

[1] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M. Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018. 2

[2] M. K. Cahalan, W. Stewart, A. Pearlman, M. Goldman, P. Sears-Rogan, M. Abel, I. Russell, J. Shanewise, C. Troianos, et al. American society of echocardiography and society of cardiovascular anesthesiologists task force guidelines for training in perioperative echocardiography. *Journal of the American Society of Echocardiography*, 15(6):647–652, 2002. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 3

[4] Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. Hubert-ecg: a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, 2024. 2

[5] Garrett Duffy, Paul P. Cheng, Nicholas Yuan, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiology*, 7(4):386–395, 2022. 2

[6] T. Golany and K. Radinsky. Pgans: Personalized generative adversarial networks for ecg synthesis to improve patient-specific deep ecg classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 557–564, 2019. 2

[7] Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pages 156–167. PMLR, 2021. 2

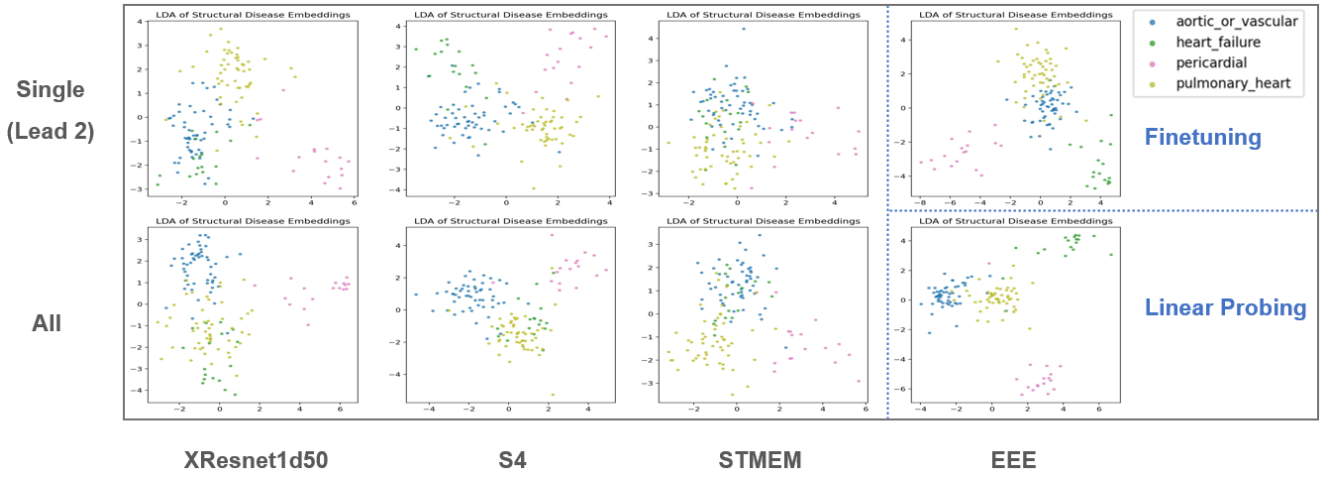

Figure 3. LDA projections of ECG-derived embeddings from XResNet1D50, S4, ST-MEM, and the proposed EEE model. Results shown for single-lead (Lead II) and all-lead configurations, and fine-tuning vs. linear probing strategies in the EEE model. Each point corresponds to a patient's ECG embedding, color-coded by structural disease category.

[8] B. Gow, T. Pollard, N. Greenbaum, B. Moody, A. Johnson, E. Herbst, J. W. Waks, P. Eslami, A. Chaudhari, T. Carbonati, S. Berkowitz, R. Mark, and S. Horng. Mimic-iv-echo: Echocardiogram matched subset (version 0.1), 2023. 2, 3

[9] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, and A.Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 2019. 2

[10] Rui Hu, Jie Chen, and Li Zhou. Spatiotemporal self-supervised representation learning from multi-lead ecg signals. *Biomedical Signal Processing and Control*, 84:104772, 2023. 2

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

[12] Zachi I Attia Konstantinos C Siontis, Peter A Noseworthy and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021.

[13] Jiapeng Lai, Hao Tan, Jun Wang, Jun Zhang, Haotian Lu, Lei Li, Yuxiang Peng, Cheng Zhang, Cheng Yu, Sheng Wei, et al. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications*, 14:3741, 2023.

[14] Xiang Lan, Darren Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4532–4540. AAAI, 2022. 2

[15] Duc Le, Sang Truong, Brijesh Patel, Donald Adjeroh, and Ngan Le. Scl-st: Supervised contrastive learning with semantic transformations for multiple lead ecg arrhythmia classification. *IEEE Journal of Biomedical and Health Informatics (JBHI)*, 27(6):2818–2828, 2023. 2

[16] Ji Li, Andres Aguirre, José Moura, Chuan Liu, Liyang Zhong, Cheng Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings with external evaluation across multiple domains. *arXiv preprint arXiv:2410.04133*, 2024. 2

[17] Yiwei Li, Sekeun Kim, Zihao Wu, Hanqi Jiang, Yi Pan, Pengfei Jin, Sifan Song, Yucheng Shi, Tianming Liu, Quanzheng Li, and Xiang Li. Echopulse: Ecg controlled echocardio-grams video generation, 2024. 1, 3

[18] Tadesse Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141:105114, 2022. 2

[19] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[20] Daniel Ouyang, Brandon He, Amirata Ghorbani, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 2020. 2

[21] Kanishka Ratnayaka, Azin Z. Faranesh, Mark S. Hansen, Amanda M. Stine, M. Halabi, Ian M. Barbash, William H. Schenke, Valerie J. Wright, Laura P. Grant, Peter Kellman, et al. Real-time mri-guided right heart catheterization in adults using passive catheters. *European Heart Journal*, 34 (5):380–389, 2013. 1

[22] Pritam Sarkar and Ali Etemad. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing (TAC)*, 13(3):1541–1554, 2020. 2

[23] Nathaniel R. Smilowitz and Jeffrey S. Berger. Perioperative cardiovascular risk assessment and management for noncardiac surgery. *JAMA*, 324:279, 2020. 1

[24] Sahar Soltanieh, Ali Etemad, and Javad Hashemi. Analysis of augmentations for contrastive ecg representation learning.

In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2022. 2

[25] Saeed Tahery, Fatemeh Hosseini Akhlaghi, and Amir Amirsoleimani. Heartbert: A self-supervised ecg embedding model for efficient and effective medical signal analysis. *arXiv preprint arXiv:2411.11896*, 2024. 2

[26] Alvaro E. Ulloa-Cerna, Linyuan Jing, John M. Pfeifer, Sushravya Raghunath, Jeffrey A. Ruhl, Daniel B. Rocha, Joseph B. Leader, Noah Zimmerman, Greg Lee, Steven R. Steinhubl, Christopher W. Good, Christopher M. Haggerty, Brandon K. Fornwalt, and Ruijun Chen. rechommend: An ecg-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography. *Circulation*, 146(1): 36–47, 2022. 2

[27] 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), 2020. 2

[28] Jianwei Yang, Jingfei Duan, Si Tran, Yujia Xu, Soujanya Chanda, Linjie Chen, Bing Zeng, Trishul Chilimbi, and Jianfeng Huang. Vision-language pre-training with triple contrastive learning. *arXiv preprint arXiv:2202.10401*, 2022. 3

[29] 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 (TIM)*, 72:1–15, 2022. 2

[30] Jiayu Zhou, Mengyang Du, Shuo Chang, et al. Artificial intelligence in echocardiography: Detection, functional evaluation, and disease diagnosis. *Cardiovascular Ultrasound*, 19:29, 2021. 1