Robust Transfer Learning for Paper ECG Recognition

Yinghao Xie* **Zhenbang Dai[1,*]** **Haojun Wang[2,3]** **Jinyu Cai** **Fabio Bonassi[1]**
Hongwu Chen[2,3] **Johan Sundström[1]** **Jiawei Li[1,†]** **Antônio H. Ribeiro[1]**

[1]Uppsala University, Sweden [2]Nanjing Medical University, China
[3]The First Affiliated Hospital with Nanjing Medical University, China

Abstract

Paper ECG recognition is challenging because real-world ECG images vary in layout, physical artifacts, and label availability. We introduce RobECG-CL, a rank-aware contrastive learning framework for robust paper ECG representation learning. Starting from standard 12-lead ECG recordings, we construct progressively degraded paper ECG views with heterogeneous layouts and train the model to balance same-recording invariance with degradation-aware ordering. Across synthetic stress tests on CODE-II and EchoNext, RobECG-CL improves robustness under severe degradation and few-shot transfer, outperforming contrastive learning baselines and surpassing the waveform-based foundation model, ECG-FM, in the 1% labeled setting. On 312 samples of hospital data with 37 labels, RobECG-CL achieves the best macro AUROC.

1 Introduction and related work

Paper electrocardiograms (paper ECGs) are ECG recordings that are printed, scanned, or photographed after acquisition. Because ECG machines and hospital archiving systems often retain only reports or images, the original digital waveform may be discarded or overwritten, leaving the paper ECG as the only available record.

Recognizing paper ECGs is difficult because the model should be robust to both clinical and visual variation. In real-world archives, ECG images vary across lead layouts, rhythm strips, synchronous or asynchronous displays, paper handling artifacts, and annotation availability. These factors make paper ECG recognition a transfer problem: a good representation learned by deep learning models should preserve ECG morphology while remaining stable under layout shift, image degradation, and few-shot downstream supervision.

Large ECG foundation models have shown strong transfer ability from digital waveforms [10, 12, 7], and image-based ECG models can predict clinical labels directly from rendered or scanned ECG images [15, 13, 14]. The PhysioNet Challenge 2024 focused on ECG image digitization and benchmarked classification as a related task [13]. However, waveform foundation models cannot be applied when only paper ECG images are available, and the robustness of image-based models under degradation, heterogeneous layouts, and limited labels remains unclear. Our contributions are:

(a) We construct a progressive degradation pipeline for the training and evaluation. We render clean paper ECGs with 3×4 and 6×2 layouts, synchronous or asynchronous displays, and lead-II rhythm strips, then generate matched views with increasing degradation severity.

*Equal contribution.
†Correspondence: jiawei.li@it.uu.se

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences Workshop.

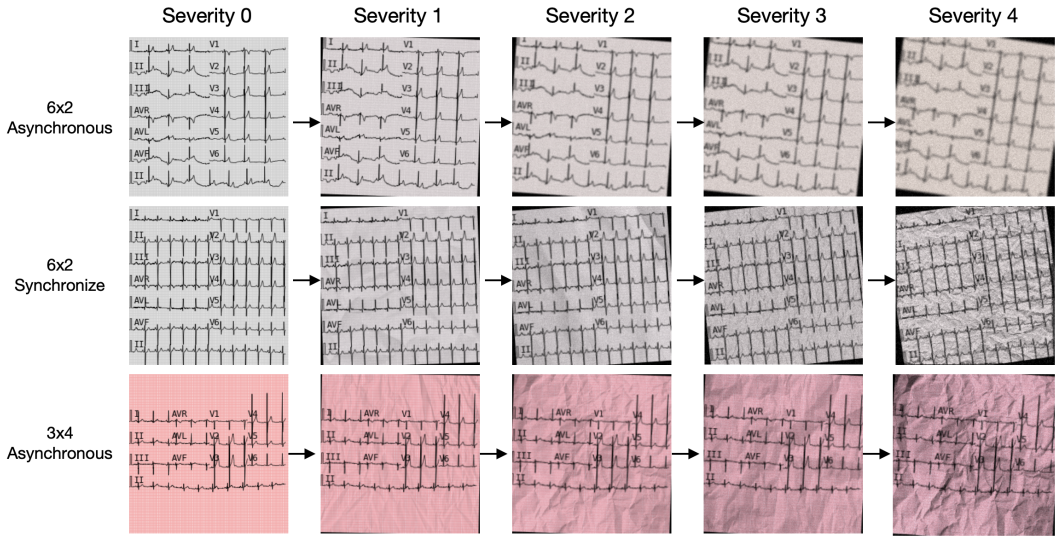

Figure 1: Progressive degradation. For each ECG recording, we first render a clean paper ECG using a randomly sampled layout, and then generate progressively degraded views by applying cumulative visual corruptions, including rotation, geometric perturbation, blur, color shift, noise, and artifacts.

(b) We propose RobECG-CL for robust representation learning. RobECG-CL uses rank-aware contrastive learning to keep degraded views from the same ECG close in representation space, while encouraging mildly degraded views to stay closer to the clean anchor than severely degraded views.

(c) We evaluate transfer under synthetic stress tests. Across sex prediction and structural heart disease (SHD) classification, RobECG-CL achieves the strongest overall robustness across four degradation severity levels and outperforms waveform foundation models in the 1% few-shot setting.

(d) We evaluate transfer under real-world stress tests. We evaluated RobECG-CL on a small paper ECG dataset from a hospital with 37 labels. RobECG-CL achieves the best macro AUROC on all labels and the second best on the top 12 most frequent labels.

2 Study design and methods

Study design. We study whether self-supervised pretraining on progressively degraded paper ECGs improves transfer under layout shift, image degradation, and limited labels. For each standard 12-lead ECG recording, we first render a clean paper ECG image $\mathbf{x}_i^{(0)} \in \mathbb{R}^{3\times224\times224}$ using the pmecg Python library[1], with rendering parameters specified by ℓ_i. The ℓ_i is sampled once and shared by all views from the same recording: we sample a 3×4 or 6×2 layout with equal probability, include a lead-II rhythm strip at the bottom, and use synchronous or asynchronous lead display with equal probability. For upstream pretraining, we use PTB-XL [16]. For controlled transfer, we apply the same rendering and degradation definitions to CODE-II for sex classification and age regression [1], and to EchoNext for SHD classification [5].

Progressive degradation. As illustrated in Figure 1, we generate an ordered set of degradation severity levels from each clean-rendered view:

$$\begin{aligned}\mathbf{u}_i^{(k)} &= (1-\mathbf{M}_i^{(k)}) \odot \mathcal{C}_i^{(k)}\left(\mathcal{W}_{\mathbf{A}_i^{(k)}}(\mathbf{x}_i^{(0)})\right) + \mathbf{M}_i^{(k)} \odot \mathbf{P}_i^{(k)},\\ \mathbf{x}_i^{(k)} &= \mathrm{clip}\left[\mathbf{K}_i^{(k)} * \mathbf{u}_i^{(k)} + \boldsymbol{\epsilon}_i^{(k)}\right], \qquad r(\mathbf{x}_i^{(k)}) = k.\end{aligned} \tag{1}$$

Here, $\mathcal{W}_{\mathbf{A}_i^{(k)}}$ is a geometric warp parameterized by an affine matrix $\mathbf{A}_i^{(k)}$, covering rotation and cropping. $\mathcal{C}_i^{(k)}$ is a sampled color transform, $\mathbf{M}_i^{(k)}$ is an artifact mask, $\mathbf{P}_i^{(k)}$ is a sampled paper artifact

[1] https://pmecg.readthedocs.io/en/stable/intro.html

Table 1: AUROC across severity levels on CODE-II sex classification. Higher severity levels denote progressively stronger degradation. Values are mean ± standard deviation over five fine-tuning seeds.

Method	Clean Paper ECG		Severity 1		Severity 2		Severity 3		Severity 4		Avg.	
	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
PTB-XL → downstream												
RobECG-CL	**0.766 ± 0.006**	0.798 ± 0.003	**0.766 ± 0.005**	0.797 ± 0.003	**0.763 ± 0.005**	**0.793 ± 0.003**	**0.758 ± 0.006**	**0.786 ± 0.003**	**0.742 ± 0.005**	**0.768 ± 0.003**	**0.759 ± 0.005**	**0.788 ± 0.003**
MoCo	0.756 ± 0.007	**0.806 ± 0.002**	0.740 ± 0.006	**0.798 ± 0.002**	0.717 ± 0.009	0.776 ± 0.004	0.682 ± 0.013	0.745 ± 0.004	0.639 ± 0.016	0.705 ± 0.006	0.710 ± 0.007	0.768 ± 0.003
DINO	0.707 ± 0.013	0.782 ± 0.003	0.706 ± 0.014	0.780 ± 0.003	0.684 ± 0.019	0.766 ± 0.003	0.651 ± 0.019	0.741 ± 0.004	0.614 ± 0.017	0.704 ± 0.005	0.675 ± 0.015	0.756 ± 0.003
BYOL	0.739 ± 0.006	0.803 ± 0.001	0.724 ± 0.007	0.793 ± 0.001	0.698 ± 0.009	0.772 ± 0.003	0.663 ± 0.011	0.743 ± 0.003	0.623 ± 0.013	0.708 ± 0.003	0.693 ± 0.008	0.766 ± 0.002
SimCLR	0.723 ± 0.010	0.754 ± 0.003	0.716 ± 0.008	0.750 ± 0.003	0.698 ± 0.007	0.741 ± 0.004	0.669 ± 0.006	0.723 ± 0.004	0.633 ± 0.007	0.689 ± 0.005	0.690 ± 0.006	0.732 ± 0.003
Pretrained (image-based) → downstream												
PULSE	0.770 ± 0.011	0.822 ± 0.003	0.764 ± 0.013	0.816 ± 0.002	**0.750 ± 0.016**	0.802 ± 0.004	0.727 ± 0.017	0.781 ± 0.005	0.695 ± 0.017	0.752 ± 0.004	0.742 ± 0.014	0.795 ± 0.003
GEM	**0.772 ± 0.009**	**0.821 ± 0.009**	**0.766 ± 0.010**	**0.818 ± 0.007**	0.750 ± 0.013	**0.805 ± 0.006**	**0.729 ± 0.013**	**0.786 ± 0.006**	**0.697 ± 0.013**	**0.757 ± 0.007**	**0.743 ± 0.010**	**0.797 ± 0.007**
Pretrained (waveform-based) → downstream												
CSFM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**0.743 ± 0.026**	0.864 ± 0.004
ECGFounder	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.644 ± 0.040	0.860 ± 0.007
ECG_FM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.681 ± 0.034	**0.866 ± 0.001**

texture, $\mathbf{K}_i^{(k)}$ is a blur kernel, $*$ denotes convolution, $\odot$ denotes element-wise multiplication, and $\boldsymbol{\epsilon}_i^{(k)}$ is additive acquisition noise. The degradation state is updated cumulatively. Specifically,

$$\begin{aligned} \mathbf{A}_i^{(k)} &= \Delta\mathbf{A}_i^{(k)}\mathbf{A}_i^{(k-1)}, & \mathcal{C}_i^{(k)} &= \Delta\mathcal{C}_i^{(k)} \circ \mathcal{C}_i^{(k-1)}, \\ \mathbf{M}_i^{(k)} &= \mathbf{M}_i^{(k-1)} + \Delta\mathbf{M}_i^{(k)}, & \beta_i^{(k)} &= \beta_i^{(k-1)} + \Delta\beta_i^{(k)}, \\ \sigma_i^{(k)} &= \sigma_i^{(k-1)} + \Delta\sigma_i^{(k)}, & \mathbf{K}_i^{(k)} &= \mathrm{BlurKernel}(\beta_i^{(k)}). \end{aligned} \tag{2}$$

Here, $\Delta\mathbf{A}_i^{(k)}$ and $\Delta\mathcal{C}_i^{(k)}$ are newly sampled geometric and color increments, while $\Delta\mathbf{M}_i^{(k)}$, $\Delta\beta_i^{(k)}$, and $\Delta\sigma_i^{(k)}$ increase the artifact, blur, and noise states. Thus, larger k denotes stronger degradation.

Rank-aware contrastive pretraining. RobECG-CL uses a CvT-13 [17] image encoder f_θ and a two-layer MLP g_ϕ to map each view to a normalized embedding $\mathbf{z}_i^{(k)} = g_\phi(f_\theta(\mathbf{x}_i^{(k)}))$. For each recording, the clean-rendered view is used as the anchor, all degraded views from the same recording are treated as positives, and views from other recordings are negatives. Let $s(\cdot,\cdot)$ denote cosine similarity and let $\mathcal{V}_i$ be the degraded views of recording i. The set-level multi-positive InfoNCE loss is

$$\mathcal{L}_{\mathrm{set}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\sum_{\mathbf{v}\in\mathcal{V}_i}\exp(s(\mathbf{z}_i^{(0)},\mathbf{v})/\tau)}{\sum_{\mathbf{q}\in\mathcal{Q}_i}\exp(s(\mathbf{z}_i^{(0)},\mathbf{q})/\tau)}, \tag{3}$$

where $\mathcal{Q}_i$ contains all candidate views in the mini-batch B except the anchor. The InfoNCE loss treats all degraded views from the same ECG as equally positive, although severe degradation can obscure fine ECG morphology which might be crucial for diagnosis. We therefore encourage a degradation-aware ordinal representation, where mildly degraded views stay closer to the clean-rendered anchor than severely degraded views. Inspired by rank-aware learning [18], we construct ordinal pairs using two distances: a Canny-based geometric distance d_g and a CIELAB color-based distance d_c. A pair (p,q) is used as an ordinal edge if view p is consistently closer to the anchor than view q:

$$\mathcal{E}_i = \left\{(p,q) : d_g(\mathbf{x}_i^{(0)},\mathbf{x}_i^{(p)}) \le (1-\eta)d_g(\mathbf{x}_i^{(0)},\mathbf{x}_i^{(q)}),\ d_c(\mathbf{x}_i^{(0)},\mathbf{x}_i^{(p)}) \le (1-\eta)d_c(\mathbf{x}_i^{(0)},\mathbf{x}_i^{(q)})\right\}. \tag{4}$$

Here, $\eta \in [0,1)$ controls the relative separation required to construct an ordinal edge: the distances from view p to the clean anchor must be no greater than $(1-\eta)$ times the corresponding distances from view q under both image criteria. Larger η imposes a stricter separation requirement. The less degraded view should be more similar to the anchor than the more degraded view:

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{|\mathcal{E}_i|}\sum_{(p,q)\in\mathcal{E}_i}\left[m - s(\mathbf{z}_i^{(0)},\mathbf{z}_i^{(p)}) + s(\mathbf{z}_i^{(0)},\mathbf{z}_i^{(q)})\right]_+^2, \tag{5}$$

where $\mathcal{I} = \{i : |\mathcal{E}_i| > 0\}$, $[\cdot]_+ = \max(\cdot, 0)$, and m is the margin specifying the desired gap between the cosine similarities of the clean anchor to the nearer and farther views. The final objective is $\mathcal{L} = \mathcal{L}_{\mathrm{set}} + \lambda\mathcal{L}_{\mathrm{rank}}$, where λ controls the contribution of the ranking regularizer relative to the set-level contrastive loss. The model learns both invariance across degradation severity levels and a weak ordering of image degradation.

3 Results

We evaluate against three baseline categories: (1) contrastive learning methods pretrained on PTB-XL (MoCo [8], DINO [2], BYOL [6], SimCLR [3]); (2) pretrained paper ECG foundation mod-

Table 2: AUROC across severity levels on EchoNext SHD classification. Higher severity levels denote progressively stronger degradation. Values are mean ± standard deviation over five runs.

Method	Clean Paper ECG		Severity 1		Severity 2		Severity 3		Severity 4		Avg.	
	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
PTB-XL → downstream												
RobECG-CL	0.692 ± 0.006	0.738 ± 0.002	0.691 ± 0.006	0.735 ± 0.002	**0.689 ± 0.006**	0.734 ± 0.002	**0.684 ± 0.006**	**0.729 ± 0.002**	**0.673 ± 0.006**	**0.715 ± 0.002**	**0.686 ± 0.006**	**0.730 ± 0.002**
MoCo	**0.722 ± 0.002**	**0.768 ± 0.004**	**0.709 ± 0.004**	0.756 ± 0.005	0.681 ± 0.007	0.726 ± 0.007	0.642 ± 0.010	0.690 ± 0.006	0.600 ± 0.010	0.644 ± 0.007	0.671 ± 0.005	0.717 ± 0.005
DINO	0.709 ± 0.005	0.766 ± 0.004	0.695 ± 0.006	**0.758 ± 0.004**	0.670 ± 0.011	**0.743 ± 0.005**	0.631 ± 0.012	0.708 ± 0.005	0.594 ± 0.014	0.667 ± 0.004	0.660 ± 0.009	0.728 ± 0.004
BYOL	0.711 ± 0.006	0.764 ± 0.004	0.689 ± 0.007	0.749 ± 0.004	0.663 ± 0.008	0.723 ± 0.004	0.626 ± 0.010	0.694 ± 0.005	0.586 ± 0.012	0.657 ± 0.006	0.655 ± 0.006	0.717 ± 0.004
SimCLR	0.664 ± 0.010	0.698 ± 0.003	0.657 ± 0.010	0.691 ± 0.003	0.643 ± 0.012	0.676 ± 0.003	0.620 ± 0.010	0.651 ± 0.005	0.582 ± 0.008	0.615 ± 0.003	0.633 ± 0.010	0.666 ± 0.003
Pretrained (image-based) → downstream												
PULSE	**0.766 ± 0.010**	**0.793 ± 0.008**	**0.751 ± 0.010**	0.785 ± 0.008	**0.734 ± 0.009**	0.774 ± 0.008	**0.712 ± 0.012**	0.758 ± 0.007	**0.677 ± 0.010**	**0.726 ± 0.007**	**0.728 ± 0.009**	**0.767 ± 0.006**
GEM	0.759 ± 0.014	0.792 ± 0.009	0.745 ± 0.013	**0.786 ± 0.009**	0.727 ± 0.014	**0.775 ± 0.010**	0.705 ± 0.016	**0.759 ± 0.009**	0.667 ± 0.018	0.721 ± 0.011	0.721 ± 0.013	0.766 ± 0.009
Pretrained (waveform-based) → downstream												
CSFM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.651 ± 0.078	0.768 ± 0.003
ECGFounder	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.615 ± 0.056	0.764 ± 0.004
ECG_FM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	**0.657 ± 0.095**	**0.795 ± 0.004**

Table 3: Per-label AUROC and F1 on the 12 diagnostic labels with more than 20 samples in a real-world paper ECG dataset. Results are pooled out-of-fold performance under 5-fold evaluation. Full results on 37 labels are provided in Supplementary Table 7.

Method	1AVB		AF		CRBBB		IRBBB		ISC		LAE		LVH		NORM		OMI		PAC		PVC		SB		Avg.	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
RobECG-CL	**0.581**	**0.212**	0.753	0.278	**0.928**	**0.691**	0.626	0.140	0.653	0.271	**0.830**	**0.429**	0.952	0.615	**0.841**	**0.299**	0.557	0.177	0.699	**0.219**	0.842	0.590	**0.912**	**0.635**	0.765	**0.380**
MoCo	0.541	0.171	0.752	**0.393**	0.877	0.597	**0.645**	0.192	0.736	0.346	0.777	0.226	0.944	0.516	0.755	0.246	0.628	0.186	0.764	0.156	0.899	0.548	0.892	0.558	**0.767**	0.345
DINO	0.499	0.171	0.778	0.295	0.880	0.571	0.604	**0.213**	0.750	0.293	0.702	0.222	0.877	0.324	0.839	0.293	**0.628**	**0.212**	**0.799**	0.203	0.903	0.635	0.848	0.467	0.759	0.325
BYOL	0.500	0.134	0.617	0.273	0.848	0.522	0.574	0.139	0.657	0.259	0.745	0.203	**0.960**	**0.636**	0.792	0.250	0.575	0.115	0.729	0.216	0.884	0.577	0.889	0.454	0.731	0.315
SimCLR	0.510	0.148	0.618	0.256	0.722	0.278	0.540	0.136	**0.788**	**0.351**	0.790	0.189	0.928	0.415	0.778	0.233	0.598	0.182	0.691	0.216	0.813	0.463	0.840	0.500	0.718	0.281
PULSE	0.531	0.131	0.750	0.275	0.895	0.597	0.569	0.162	0.688	0.205	0.737	0.138	0.859	0.579	0.774	0.222	0.493	0.057	0.635	0.000	**0.915**	**0.685**	0.816	0.314	0.722	0.280
GEM	0.532	0.140	**0.818**	0.244	0.908	0.677	0.607	0.194	0.726	0.270	0.721	0.231	0.816	0.474	0.791	0.118	0.514	0.179	0.635	0.138	0.908	0.683	0.811	0.406	0.732	0.313

els (PULSE [11], GEM [9]); and (3) pretrained waveform foundation models (CSFM [7], ECG-founder [10], ECG-FM [12]). Transfer learning is assessed on sex and SHD classification using 1% and 10% of the training data (Tables 1 and 2), with CODE-II age prediction results deferred to Appendix A.1. Finally, we validate on 312 real-world paper ECGs from The First Affiliated Hospital with Nanjing Medical University.

Robustness under limited labeled data and progressive degradation. RobECG-CL achieves the best macro AUROC on both sex classification and SHD classification among models pretrained on PTB-XL paper ECGs. The advantage is more evident at severe degradation levels (i.e., severity 4), indicating improved robustness of the complete training framework under the evaluated progressive degradations.

Comparison with waveform foundation models. Although paper ECG images provide only an indirect visual observation of the underlying waveform, RobECG-CL outperforms ECG-FM under 1% training data on both sex and SHD classifications.

Comparison with image-based foundation models. PULSE and GEM show strong transferability to degraded paper ECG images, even at severe degradation levels. However, they underperform RobECG-CL on sex classification and age prediction when using 1% labeled training data (Table 4).

Analysis on Real-World Paper ECGs from a Hospital. The images were annotated by an experienced cardiologist and captured from printed ECG records using a RICOH 8001 scanner. Additional details are provided in the Appendix C. As shown in Table 3, among the 12 labels with more than 20 positive samples, RobECG-CL achieves the second-best macro AUROC and the best macro F1 compared with other contrastive learning baselines and image-based foundation models.

Visualization of Robust Representations. As illustrated in Figure 2, our method generally aligns the latent representations closely with the clean anchor across varying degradation severities. It also flexibly permits a small subset of severely degraded samples to deviate to avoid unnatural alignments.

4 Conclusion and limitation

In this study, we propose RobECG-CL, a robust self-supervised learning framework designed to extract degradation-aware representations from paper ECG images. By leveraging a synthetic dataset with progressive degradations, our method effectively aligns features across varying severity levels without forcing unnatural alignments for severely degraded samples. Consequently, RobECG-CL demonstrates superior classification performance and strong transferability to real-world clinical records. A limitation of our study is that the baselines pretrained on PTB-XL rely on augmentation

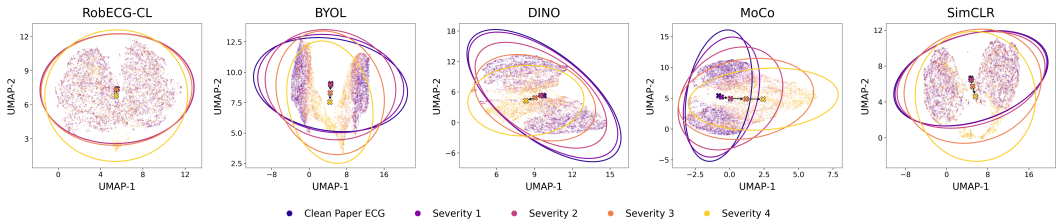

Figure 2: UMAP visualization of representations for sex prediction with 10% labeled training data. RobECG-CL learns degradation-robust representations while preserving the relative structure among degraded views. The severely corrupted views are not forced to be close to the clean view.

pipelines for clean ECGs, rather than treating progressively degraded views as positive pairs. Thus, the baseline comparisons do not disentangle the effects of degradation exposure from those of the learning objective. In the frozen-encoder ablation (Appendix F), adding the ranking term increases average AUROC from 0.825 to 0.829. This modest improvement suggests a potential incremental benefit, but does not establish its consistency across pretraining seeds or explain the full gains in downstream fine-tuning.Moreover, some degraded views might lose the clinical features required for accurate diagnosis, and we did not manually filter these cases. Finally, our real-world validation set is limited to a single center.

5 Acknowledgments and Disclosure of Funding

Jiawei Li and Antônio H. Ribeiro are financially supported by the eSSENCE and SciLifeLab, with the project "Digital Biomarkers from the Electrocardiogram using Artificial Intelligence"; and, by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation. This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme through grant agreement no. 101054643. Computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

[1] Petrus E. O. G. B. Abreu, Gabriela M. M. Paixão, Jiawei Li, Paulo Rodrigues Gomes, Peter W. Macfarlane, Ana C. S. Oliveira, Vinicius T. Carvalho, Thomas B. Schön, Antônio Luiz P. Ribeiro, and Antônio H. Ribeiro. Code-II: A large-scale dataset for artificial intelligence in ecg analysis. *npj Digital Medicine*, 2025.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

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

[4] Lovedeep Singh Dhingra, Arya Aminorroaya, Veer Sangha, Aline F. Pedroso, Sumukh Vasisht Shankar, Andreas Coppi, Murilo Foppa, Luisa C. C. Brant, Sandhi M Barreto, Antônio Luiz Pinho Ribeiro, Harlan M Krumholz, Evangelos K. Oikonomou, and Rohan Khera. Ensemble deep learning algorithm for structural heart disease screening using electrocardiographic images: Present shd. *Journal of the American College of Cardiology*, 85:1302 – 1313, 2025.

[5] Pierre Elias and Joshua Finer. EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs. *PhysioNet*, 2026. Version 1.1.1.

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own

latent a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank Narayana Gowda, Antônio H. Ribeiro, Patrick Schwab, Kim Branson, Lei A. Clifton, Antônio Luiz Pinho Ribeiro, Zhangdaihong Liu, and David A. Clifton. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8:220 – 233, 2026.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019.

[9] Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. Gem: Empowering mllm for grounded ecg understanding with time series and images. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:94421–94455, 2026.

[10] Jun Li, Aaron D Aguirre, Valdery F. Moura, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari D. Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2, 2025.

[11] Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teaching multimodal llms to comprehend 12-lead electrocardiographic images. *npj Digital Medicine*, 9(1):349, 2026.

[12] Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: an open electrocardiogram foundation model. *JAMIA Open*, 8, 2025.

[13] Matthew A. Reyna, Deepanshi Deepanshi, James Tyler Weigle, Zuzana Koscova, Kiersten S. Campbell, Salman Seyedi, Andoni Elola, Ali Bahrami Rad, Amit Jasvant Shah, Neal Kumar Bhatia, Gari D. Clifford, and Reza Sameni. Digitization and classification of ecg images: The george b. moody physionet challenge 2024. *Computing in Cardiology Conference (CinC)*, 2024.

[14] Matthew A. Reyna, Deepanshi, Audrey Tyler Weigle, Zuzana Koscova, Kiersten Campbell, Kshama Kodthalu Shivashankara, Soheil Saghafi, Sepideh Nikookar, Mohsen Motie-Shirazi, Yashar Kiarashi, Salman Seyedi, Mohammadsina Hassannia, Agnar Martin Bjørnstad, Elias Stenhede, Arian Ranjbar, Gari D. Clifford, and Reza Sameni. Ecg-image-database: large-scale paired ecg images and time-series with real-world artifacts; a foundation for computerized ecg digitization and analysis. *Physiological Measurement*, 47, 2026.

[15] Veer Sangha, B. Mortazavi, Adrian D. Haimovich, Antônio H. Ribeiro, Cynthia A. Brandt, Denise L. Jacoby, Wade L. Schulz, Harlan M. Krumholz, Antônio Luiz P. Ribeiro, and Rohan Khera. Automated multilabel diagnosis on electrocardiographic images and signals. *Nature Communications*, 13, 2021.

[16] Patrick Wagner, Nils Strodthoff, Ralf Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7, 2020.

[17] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, October 2021.

[18] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Table of Contents

A Supplementary results **8**
- A.1 Supplementary age prediction results 8
- A.2 Supplementary AUPRC results 8
- A.3 Supplementary real-world paper ECG results 8

B Synthetic dataset curation **10**

C Real-world dataset curation **11**

D Baselines and implementations **11**

E Distances for ordinal pairs **12**

F Ablation study **13**

G Supplementary related work **13**

A Supplementary results

A.1 Supplementary age prediction results

RobECG-CL achieves the best age prediction performance among the PTB-XL-pretrained methods under both the 1% and 10% training settings (Table 4). Increasing the labeled training data from 1% to 10% reduces its average MAE and increases the average R^2. Additionally, RobECG-CL exhibits a smaller increase in MAE as degradation severity increases. Compared with PULSE and GEM, RobECG-CL achieves better performance in the more challenging 1% labeled training data setting and remains competitive when 10% labeled training data are available, although PULSE and GEM obtain lower absolute errors in the latter setting.

Table 4: Transfer performance across degradation severity levels on CODE-II age prediction. Higher severity levels denote progressively stronger degradation. Values are mean ± standard deviation over five runs.

Method	Clean Paper ECG MAE		Severity 1 MAE		Severity 2 MAE		Severity 3 MAE		Severity 4 MAE		Avg. MAE		Avg. RMSE		Avg. R^2	
	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
RobECG-CL	12.66 ± 0.22	11.75 ± 0.15	**12.67 ± 0.19**	11.78 ± 0.17	**12.76 ± 0.21**	11.87 ± 0.18	**12.90 ± 0.23**	12.07 ± 0.18	**13.16 ± 0.25**	12.45 ± 0.19	**12.83 ± 0.21**	11.98 ± 0.17	**16.00 ± 0.31**	15.07 ± 0.23	**0.102 ± 0.035**	0.204 ± 0.024
MoCo	13.10 ± 0.10	12.18 ± 0.08	13.00 ± 0.09	12.26 ± 0.07	13.25 ± 0.13	12.50 ± 0.05	13.61 ± 0.22	12.87 ± 0.06	13.88 ± 0.22	13.22 ± 0.06	13.37 ± 0.14	12.60 ± 0.05	16.48 ± 0.16	15.66 ± 0.07	0.047 ± 0.019	0.140 ± 0.007
DINO	13.28 ± 0.20	12.10 ± 0.05	13.27 ± 0.22	12.14 ± 0.05	13.42 ± 0.23	12.38 ± 0.05	13.66 ± 0.15	12.76 ± 0.06	13.90 ± 0.10	13.17 ± 0.07	13.51 ± 0.17	12.51 ± 0.05	16.64 ± 0.18	15.47 ± 0.06	0.028 ± 0.021	0.160 ± 0.006
BYOL	13.14 ± 0.12	12.12 ± 0.05	13.25 ± 0.15	12.26 ± 0.04	13.46 ± 0.20	12.50 ± 0.03	13.69 ± 0.19	12.87 ± 0.02	13.88 ± 0.13	13.22 ± 0.02	13.48 ± 0.15	12.59 ± 0.01	16.66 ± 0.16	15.65 ± 0.03	0.027 ± 0.019	0.141 ± 0.003
SimCLR	13.24 ± 0.09	12.60 ± 0.08	13.28 ± 0.06	12.65 ± 0.09	13.52 ± 0.07	12.75 ± 0.07	13.72 ± 0.09	13.02 ± 0.03	13.86 ± 0.13	13.31 ± 0.06	13.52 ± 0.08	12.87 ± 0.05	16.68 ± 0.10	15.94 ± 0.06	0.025 ± 0.011	0.108 ± 0.007
PULSE	**12.65 ± 0.39**	10.98 ± 0.19	12.74 ± 0.32	11.13 ± 0.22	12.89 ± 0.27	**11.38 ± 0.19**	13.18 ± 0.22	**11.72 ± 0.17**	13.65 ± 0.30	**12.19 ± 0.17**	13.02 ± 0.26	**11.48 ± 0.18**	16.14 ± 0.31	**14.40 ± 0.20**	0.086 ± 0.035	**0.271 ± 0.020**
GEM	12.86 ± 0.29	**10.86 ± 0.26**	12.96 ± 0.15	**11.06 ± 0.27**	13.05 ± 0.11	11.42 ± 0.27	13.25 ± 0.10	11.84 ± 0.34	13.53 ± 0.09	12.38 ± 0.38	13.13 ± 0.12	11.52 ± 0.30	16.26 ± 0.09	14.45 ± 0.37	0.073 ± 0.011	0.266 ± 0.038

A.2 Supplementary AUPRC results

CODE-II sex classification. Among models pretrained on PTB-XL paper ECGs, RobECG-CL achieves the highest average AUPRC under both training fractions, as shown in Table 5. Although MoCo performs better on clean paper ECGs, its performance decreases under progressive degradation. From clean paper ECGs to Severity 4, the AUPRC of RobECG-CL decreases by only 0.024 and 0.034 under the 1% and 10% settings; by contrast, it is 0.128 and 0.118 for MoCo. Additionally, RobECG-CL also outperforms PULSE, GEM, and the waveform foundation models in average AUPRC under the 1% setting.

EchoNext SHD classification. RobECG-CL also achieves the highest average AUPRC among the PTB-XL-pretrained methods on EchoNext (Table 6). From clean paper ECGs to Severity 4, RobECG-CL shows AUPRC decreases of 0.023 and 0.029, whereas the corresponding decreases for MoCo are 0.135 and 0.143. PULSE and GEM exhibit higher average AUPRC but larger degradation-related decreases. These results suggest that RobECG-CL preserves strong robustness under progressive paper ECG degradation.

Table 5: AUPRC across degradation severity levels on CODE-II sex classification. Higher severity levels denote progressively stronger degradation. Values are mean ± standard deviation over five runs.

Method	Clean Paper ECG		Severity 1		Severity 2		Severity 3		Severity 4		Avg.	
	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
PTB-XL → downstream												
RobECG-CL	0.695 ± 0.006	0.739 ± 0.007	**0.695 ± 0.005**	0.736 ± 0.006	**0.692 ± 0.006**	**0.731 ± 0.006**	**0.688 ± 0.006**	**0.725 ± 0.005**	**0.671 ± 0.005**	**0.705 ± 0.005**	**0.688 ± 0.005**	**0.727 ± 0.006**
MoCo	**0.696 ± 0.014**	**0.755 ± 0.005**	0.678 ± 0.012	**0.744 ± 0.005**	0.650 ± 0.012	0.717 ± 0.006	0.611 ± 0.014	0.680 ± 0.006	0.568 ± 0.015	0.637 ± 0.005	0.645 ± 0.011	0.710 ± 0.004
DINO	0.648 ± 0.014	0.722 ± 0.007	0.647 ± 0.013	0.720 ± 0.005	0.622 ± 0.016	0.707 ± 0.003	0.583 ± 0.018	0.680 ± 0.003	0.541 ± 0.016	0.640 ± 0.003	0.612 ± 0.013	0.696 ± 0.004
BYOL	0.665 ± 0.008	0.745 ± 0.004	0.650 ± 0.006	0.736 ± 0.004	0.624 ± 0.006	0.713 ± 0.004	0.590 ± 0.009	0.680 ± 0.003	0.549 ± 0.011	0.640 ± 0.003	0.619 ± 0.005	0.706 ± 0.003
SimCLR	0.652 ± 0.009	0.695 ± 0.003	0.644 ± 0.008	0.691 ± 0.004	0.623 ± 0.008	0.680 ± 0.005	0.592 ± 0.008	0.659 ± 0.005	0.554 ± 0.009	0.623 ± 0.005	0.615 ± 0.007	0.671 ± 0.003
Pretrained (image-based) → downstream												
PULSE	**0.709 ± 0.023**	0.774 ± 0.005	**0.705 ± 0.021**	0.766 ± 0.008	**0.692 ± 0.022**	0.749 ± 0.011	**0.666 ± 0.021**	0.725 ± 0.010	0.627 ± 0.021	0.692 ± 0.010	**0.681 ± 0.019**	0.742 ± 0.009
GEM	0.708 ± 0.021	**0.776 ± 0.010**	0.700 ± 0.024	**0.770 ± 0.009**	0.686 ± 0.023	**0.756 ± 0.008**	0.663 ± 0.023	**0.734 ± 0.007**	**0.630 ± 0.021**	**0.701 ± 0.008**	0.678 ± 0.021	**0.748 ± 0.008**
Pretrained (waveform-based) → downstream												
CSFM	–	–	–	–	–	–	–	–	–	–	**0.676 ± 0.044**	**0.827 ± 0.004**
ECGFounder	–	–	–	–	–	–	–	–	–	–	0.543 ± 0.036	0.813 ± 0.009
ECG_FM	–	–	–	–	–	–	–	–	–	–	0.602 ± 0.045	0.820 ± 0.006

A.3 Supplementary real-world paper ECG results

We further evaluate all methods on 312 real-world paper ECGs including 37 diagnostic labels. All results are pooled out-of-fold predictions obtained using five-fold evaluation.

Performance across all diagnostic labels. RobECG-CL achieves the best average macro AUROC across all 37 diagnostic labels. RobECG-CL also remains comparable to the strongest baselines in macro AUPRC and F1, compared with MoCo and GEM, and shows strong performance for

Table 6: AUPRC across degradation severity levels on EchoNext structural heart disease classification. Higher severity levels denote progressively stronger degradation. Values are mean ± standard deviation over five runs. For baseline methods, severity-specific AUPRC is not available in the source table and is marked as missing.

Method	Clean Paper ECG		Severity 1		Severity 2		Severity 3		Severity 4		Avg.	
	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
PTB-XL → downstream												
RobECG-CL	0.624 ± 0.008	0.683 ± 0.003	0.622 ± 0.009	0.679 ± 0.003	**0.621 ± 0.009**	**0.679 ± 0.004**	**0.617 ± 0.010**	**0.673 ± 0.003**	**0.601 ± 0.009**	**0.654 ± 0.004**	**0.617 ± 0.009**	**0.674 ± 0.003**
MoCo	**0.663 ± 0.004**	**0.715 ± 0.005**	**0.644 ± 0.005**	0.698 ± 0.005	0.611 ± 0.007	0.665 ± 0.007	0.571 ± 0.014	0.626 ± 0.006	0.528 ± 0.011	0.572 ± 0.006	0.609 ± 0.005	0.661 ± 0.005
DINO	0.650 ± 0.006	0.712 ± 0.003	0.628 ± 0.007	**0.701 ± 0.003**	0.600 ± 0.013	0.679 ± 0.003	0.555 ± 0.017	0.636 ± 0.005	0.520 ± 0.016	0.594 ± 0.005	0.595 ± 0.010	0.669 ± 0.003
BYOL	0.654 ± 0.005	0.709 ± 0.003	0.624 ± 0.008	0.690 ± 0.003	0.597 ± 0.009	0.665 ± 0.004	0.558 ± 0.014	0.631 ± 0.005	0.516 ± 0.013	0.588 ± 0.005	0.593 ± 0.007	0.660 ± 0.003
SimCLR	0.577 ± 0.015	0.613 ± 0.004	0.568 ± 0.013	0.604 ± 0.005	0.553 ± 0.014	0.590 ± 0.004	0.533 ± 0.012	0.564 ± 0.006	0.498 ± 0.005	0.528 ± 0.005	0.546 ± 0.013	0.581 ± 0.004
Pretrained (image-based) → downstream												
PULSE	**0.708 ± 0.011**	0.747 ± 0.008	**0.689 ± 0.017**	0.740 ± 0.007	**0.668 ± 0.016**	0.726 ± 0.005	**0.642 ± 0.021**	0.705 ± 0.005	**0.607 ± 0.019**	**0.668 ± 0.010**	**0.664 ± 0.016**	0.717 ± 0.005
GEM	0.692 ± 0.019	**0.748 ± 0.010**	0.673 ± 0.021	**0.741 ± 0.010**	0.656 ± 0.017	**0.727 ± 0.013**	0.635 ± 0.017	**0.705 ± 0.015**	0.594 ± 0.024	0.663 ± 0.018	0.651 ± 0.018	**0.717 ± 0.014**
Pretrained (waveform-based) → downstream												
CSFM	–	–	–	–	–	–	–	–	–	–	0.150 ± 0.037	0.238 ± 0.003
ECGFounder	–	–	–	–	–	–	–	–	–	–	0.125 ± 0.018	0.215 ± 0.007
ECG_FM	–	–	–	–	–	–	–	–	–	–	**0.173 ± 0.052**	**0.272 ± 0.007**

several individual labels (Table 7). Specifically, RobECG-CL performs an AUROC/AUPRC/F1 of 0.928/0.724/0.691 for complete right bundle branch block (CRBBB), 0.912/0.552/0.635 for sinus bradycardia (SB), and 0.841/0.404/0.299 for normal ECGs (NORM). Although the best-performing model varies across different labels, RobECG-CL demonstrates the most consistent overall discrimination across the entire diagnostic label space.

Table 7: Per-label AUROC, AUPRC, and F1 on all 37 diagnostic labels in the real-world paper ECG set collected from The First Affiliated Hospital with Nanjing Medical University. Results are pooled out-of-fold performance under strict patient-aware 5-fold evaluation.

Label	RobECG-CL			MoCo			DINO			BYOL			SimCLR			PULSE			GEM		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1
1AVB (36)	**0.581**	0.149	**0.212**	0.541	0.158	0.171	0.499	0.124	0.171	0.500	0.167	0.134	0.510	0.129	0.148	0.531	0.136	0.131	0.532	**0.176**	0.140
AF (28)	0.753	0.278	0.278	0.752	0.325	**0.393**	0.778	0.309	0.295	0.617	0.205	0.273	0.618	0.220	0.256	0.750	0.264	0.275	**0.818**	**0.385**	0.244
CRBBB (36)	**0.928**	**0.724**	**0.691**	0.877	0.656	0.597	0.880	0.581	0.571	0.848	0.563	0.522	0.722	0.248	0.278	0.895	0.676	0.597	0.908	0.693	0.677
IRBBB (24)	0.626	0.157	0.140	**0.645**	**0.259**	0.192	0.604	0.209	**0.213**	0.574	0.139	0.139	0.540	0.128	0.136	0.569	0.159	0.162	0.607	0.198	0.194
ISC (41)	0.653	0.240	0.271	0.736	0.249	0.346	0.750	0.317	0.293	0.657	0.211	0.259	**0.788**	**0.428**	**0.351**	0.688	0.246	0.205	0.726	0.264	0.270
LAE (20)	**0.830**	0.246	**0.429**	0.777	0.259	0.226	0.702	0.146	0.222	0.745	0.268	0.203	0.790	0.200	0.189	0.737	0.268	0.138	0.721	**0.309**	0.231
LVH (23)	0.952	**0.699**	0.615	0.944	0.572	0.516	0.877	0.447	0.324	**0.960**	0.655	**0.636**	0.928	0.604	0.415	0.859	0.556	0.579	0.816	0.512	0.474
NORM (23)	**0.841**	**0.404**	**0.299**	0.755	0.322	0.246	0.839	0.311	0.293	0.792	0.258	0.250	0.778	0.224	0.233	0.774	0.230	0.222	0.791	0.235	0.118
OMI (26)	0.557	0.104	0.177	0.628	0.189	0.186	**0.628**	**0.225**	**0.212**	0.575	0.112	0.115	0.598	0.161	0.182	0.493	0.127	0.057	0.514	0.165	0.179
PAC (20)	0.699	0.185	**0.219**	0.764	0.160	0.156	**0.799**	0.194	0.203	0.729	0.144	0.216	0.691	**0.241**	0.216	0.635	0.102	0.000	0.635	0.115	0.138
PVC (41)	0.842	0.671	0.590	0.899	0.661	0.548	0.903	0.621	0.635	0.884	0.612	0.577	0.813	0.436	0.463	**0.915**	**0.754**	**0.685**	0.908	0.727	0.683
SB (42)	**0.912**	0.552	**0.635**	0.892	0.545	0.558	0.848	0.423	0.467	0.889	0.467	0.454	0.840	**0.553**	0.500	0.816	0.392	0.314	0.811	0.414	0.406
2AVB (1)	0.707	0.011	0.000	0.714	0.011	0.000	**0.891**	**0.029**	0.000	0.637	0.009	0.000	0.492	0.006	0.000	0.248	0.004	0.000	0.595	0.008	0.000
3AVB (3)	0.963	0.177	0.000	0.810	0.035	0.000	0.926	0.078	0.000	**0.965**	**0.278**	**0.444**	0.538	0.022	0.000	0.817	0.040	0.000	0.946	0.131	0.000
AFL (6)	0.551	0.045	0.065	0.760	0.225	0.085	0.836	0.116	**0.105**	**0.849**	**0.265**	0.088	0.789	0.068	0.041	0.813	0.147	0.000	0.817	0.177	0.000
AMI (8)	0.781	0.206	0.150	0.726	0.206	0.113	0.710	0.161	0.109	0.690	0.125	0.092	0.640	0.049	0.066	0.771	0.231	0.182	**0.785**	**0.393**	**0.400**
ASMI (10)	0.724	0.189	0.070	0.846	**0.394**	0.244	**0.884**	0.330	**0.320**	0.840	0.348	0.207	0.710	0.219	0.126	0.820	0.324	0.167	0.800	0.283	0.182
AT (2)	**0.731**	0.029	0.000	0.656	0.013	0.000	0.255	0.006	0.000	0.592	0.011	0.000	0.642	0.013	0.000	0.703	0.017	0.000	0.581	**0.059**	0.000
CLBBB (5)	**0.857**	0.148	**0.081**	0.851	0.086	0.064	0.776	0.043	0.042	0.850	**0.171**	0.073	0.484	0.019	0.038	0.711	0.058	0.000	0.771	0.081	0.000
EAMI (10)	0.803	0.219	0.167	0.845	0.367	0.316	0.708	0.213	0.207	0.675	0.312	0.081	0.655	0.171	0.091	**0.858**	0.429	**0.462**	0.856	**0.477**	**0.462**
ILBBB (1)	0.563	0.007	0.000	0.685	0.010	0.000	0.064	0.003	0.000	0.289	0.005	0.000	0.148	0.004	0.000	**0.846**	**0.020**	0.000	0.129	0.004	0.000
IMI (5)	0.560	0.034	0.031	**0.791**	0.058	0.054	0.745	**0.074**	**0.066**	0.643	0.028	0.038	0.582	0.035	0.019	0.619	0.069	0.000	0.615	0.027	0.000
IVCD (13)	0.522	0.057	0.078	0.556	0.051	0.079	0.552	0.047	**0.083**	0.347	0.038	0.054	0.426	0.036	0.063	**0.631**	**0.106**	0.000	0.606	0.095	0.000
JE (2)	0.981	0.196	0.000	**0.992**	0.417	**0.286**	0.898	0.044	0.000	0.990	**0.625**	0.222	0.903	0.074	0.000	0.947	0.079	0.000	0.897	0.068	0.000
LAFB (13)	**0.932**	0.433	0.364	0.799	0.241	0.222	0.771	0.194	0.167	0.759	0.159	0.270	0.849	0.255	0.125	0.878	0.405	0.381	0.865	**0.470**	**0.455**
LNGQT (6)	**0.665**	0.041	**0.039**	0.499	0.021	0.022	0.447	0.022	0.029	0.533	0.023	0.025	0.572	0.041	0.015	0.629	0.048	0.000	0.570	**0.070**	0.000
PACE-A (1)	0.370	0.005	0.000	0.113	0.004	0.000	**0.826**	**0.018**	0.000	0.318	0.005	0.000	0.158	0.004	0.000	0.077	0.003	0.000	0.087	0.004	0.000
PACE-D (5)	**0.747**	**0.069**	0.050	0.639	0.028	0.023	0.352	0.017	0.015	0.494	0.020	0.038	0.609	0.049	**0.087**	0.730	0.041	0.000	0.694	0.033	0.000
PACE-V (12)	0.877	0.342	0.296	**0.923**	**0.444**	**0.429**	0.747	0.163	0.093	0.897	0.364	0.324	0.878	0.319	0.316	0.778	0.285	0.381	0.779	0.258	0.267
RVH (4)	0.732	0.078	0.031	0.839	0.061	**0.070**	0.635	0.037	0.028	0.592	0.039	0.042	0.209	0.010	0.016	0.727	**0.105**	0.000	**0.867**	0.080	0.000
SA (1)	0.354	0.005	0.000	0.486	0.006	0.000	0.367	0.005	0.000	**0.958**	**0.071**	0.000	0.704	0.011	0.000	0.391	0.005	0.000	0.272	0.004	0.000
SArrh (4)	0.304	0.011	0.013	0.211	0.010	0.011	0.250	0.010	**0.019**	0.189	0.009	0.000	**0.506**	**0.016**	0.016	0.421	0.014	0.000	0.312	0.011	0.000
STach (10)	0.919	0.369	0.364	0.945	0.411	0.357	**0.976**	0.408	**0.600**	0.949	**0.458**	0.429	0.919	0.342	0.421	0.959	0.370	0.476	0.951	0.363	0.273
STTC (13)	0.584	0.061	0.089	0.645	0.082	0.095	0.698	0.077	0.094	0.545	0.052	0.066	**0.757**	**0.114**	**0.128**	0.629	0.059	0.000	0.587	0.062	0.000
VE (1)	**0.550**	**0.007**	0.000	0.222	0.004	0.000	0.424	0.006	0.000	0.151	0.004	0.000	0.013	0.003	0.000	0.106	0.004	0.000	0.167	0.004	0.000
VT (3)	**0.865**	0.359	0.286	0.725	0.347	0.222	0.824	0.138	0.000	0.783	0.363	0.400	0.574	0.342	0.222	0.775	0.354	0.000	0.761	**0.389**	**0.500**
WPW (5)	0.762	0.108	0.154	**0.767**	0.146	0.087	0.759	0.079	0.211	0.622	0.146	0.025	0.648	0.095	0.121	0.696	**0.251**	**0.333**	0.731	0.086	0.000
Avg.	**0.718**	0.206	0.186	0.710	**0.217**	**0.187**	0.687	0.168	0.165	0.674	0.209	0.181	0.622	0.159	0.143	0.682	0.199	0.155	0.671	0.212	0.170

The real-world semantic prediction space. Figure 3 visualizes the diagnostic-label centroids calculated from the pooled out-of-fold predictions. For RobECG-CL, CRBBB, SB, and LVH are clearly separable, resulting in strong classification performance. The left ventricular hypertrophy (LVH) and sinus bradycardia (SB) separated from other diagnostic labels. Atrial fibrillation (AF) and premature atrial contractions (PAC) located close to each other. From a clinical perspective, atrial arrhythmias occur a continuum from premature atrial contractions to flutter and fibrillation, and these rhythm abnormalities often coexist. The centroid distance of PVC and left atrial enlargement (LAE) were close either. Clinically, patients with left ventricular dysfunction may present with higher PVC burden and left atrial enlargement. Such coexistence of multiple cardiac abnormalities may reduce the model's ability to distinguish ECG categories as clearly as in cases with a single diagnostic label.

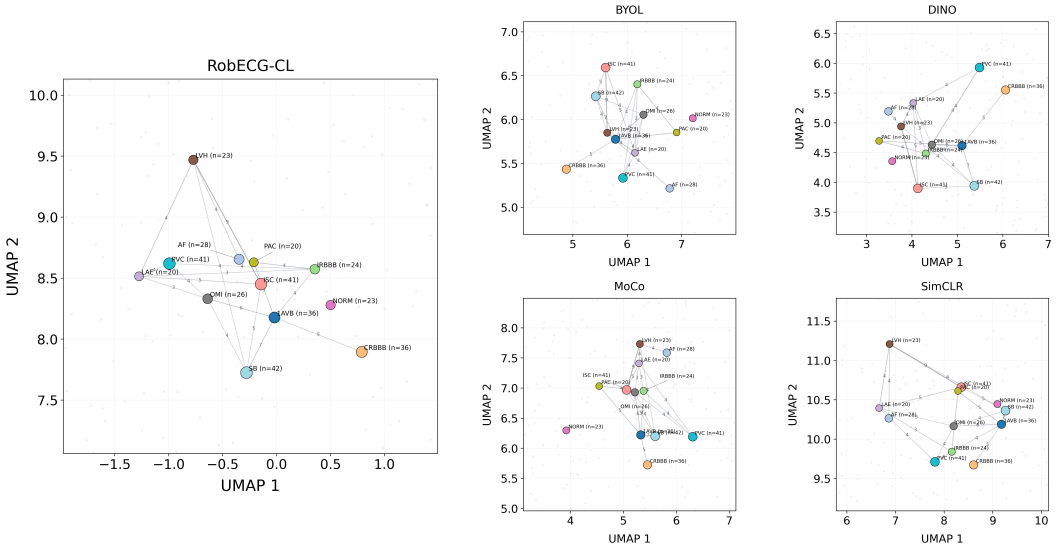

Figure 3: Diagnostic-label centroids in the semantic UMAP spaces of RobECG-CL, BYOL, DINO, MoCo, and SimCLR. Here, n denotes the number of ECGs positive for each label. For each method, the standardized per-label logits are projected into a two-dimensional UMAP space. Each colored point is then calculated as the arithmetic mean of the UMAP coordinates of all ECGs positive for the corresponding label. Edges connect label pairs with the strongest true-label co-occurrence, while the number on each edge indicates the number of ECGs positive for both labels.

Impact of label imbalance. The complete 37-label evaluation includes several extremely rare samples. In particular, 2AVB, ILBBB, PACE-A, SA, and VE contain only one positive sample, while 3AVB, AT, JE, CLBBB, IMI, PACE-D, RVH, SArrh, VT, and WPW contain only two to five positive samples. Therefore, results for these labels are sensitive to indivisual samples, and high AUROC values may co-exist with low AUPRC or zero F1.

Grad-CAM. Figure 4 shows representative Grad-CAM visualizations on real-world paper ECGs. Across different diagnostic labels, RobECG-CL attends to localized ECG regions rather than background artifacts. The highest interpretability was observed for paced rhythm, premature ventricular contraction (PVC), and left ventricular hypertrophy (LVH). The first row of Figure 4 derived from a single patient with multiple diagnostic labels. Although PVC, ASMI, and OMI were all identified accurately, the presence of multiple labels resulted in interactions among different diagnostic features, leading to overlapping and mixed highlighted regions. The second row of Figure 4 was derived from three other patients. The visualization for CLBBB was clearly more accurate than that for CRBBB. For CRBBB, the V1 and V2 leads were not highlighted, although it's relatively easy to recognize in routine ECG interpretation. The highlighted regions for STTC showed no clear specificity. Overall, the Grad-CAM maps indicate that the model mainly focused on the precordial leads rather than the limb leads, with the highlighted regions mainly focused on QRS complexes.

B Synthetic dataset curation

Our synthetic ECG dataset comprises four progressive degradation stages. The dataset maintains a balanced distribution across both lead layouts (50% 2×6 with red grids, 50% 4×3 with black-and-white grids) and segmentation strategies (50% synchronous, 50% asynchronous).

Clean paper ECGs (Severity 0) consist of 224×224 images rendered directly from raw WFDB signals. For Severities 1 to 4, we employ a cumulative degradation strategy that progressively injects realistic artifacts encompassing rotation, color temperature shifts, cropping, noise, Gaussian blur, and wrinkling. The degradation severity monotonically increases at each stage. Specifically, the continuous parameters inherited from the previous stage are incremented by sampling from Gaussian distributions. A comprehensive summary of the parameter configurations, their physical interpretations, and corresponding implementations is detailed in Table 8.

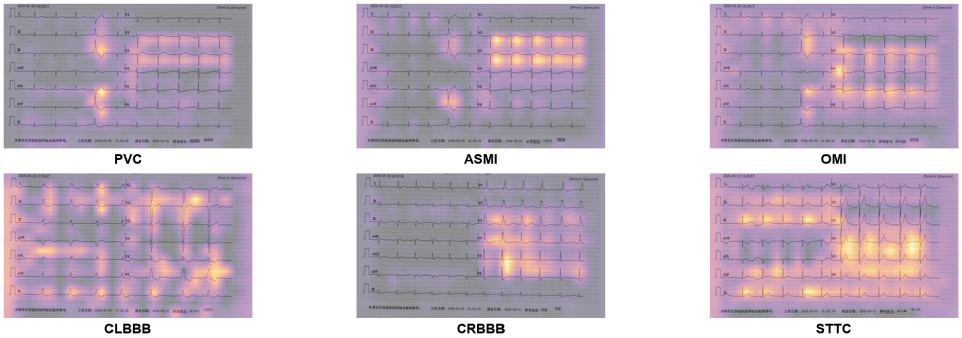

Figure 4: Grad-CAM heatmap visualizations on real-world paper ECGs. Each panel shows the attribution map for one diagnostic label from the external hospital dataset. RobECG-CL highlights localized ECG regions across diverse diagnoses, suggesting that the learned representation remains interpretable under real-world paper ECG degradation.

Table 8: Summary of progressive degradation methods, parameter step distributions, physical interpretations, and corresponding implementations.

Method	**Step Distribution (ϵ)**	**Physical Interpretation**	**Implementation**
Rotation	$\mathcal{N}(0, 5)$	Paper misalignment during scanning/photography.	`iaa.Affine`
Color Temperature	$\mathcal{N}(0, 300)$	Variations in white-balance and ambient illumination.	`iaa.ChangeColorTemperature`
Cropping	$\mathcal{N}(0.002, 0.01)$	Partial border loss or automatic edge trimming.	`iaa.Crop`
Noise	$\mathcal{N}(2, 6)$	Pixel-level noise introduced by camera/scanner sensors.	`iaa.AdditiveGaussianNoise`
Gaussian Blur	$\mathcal{N}(1, 3)$	Defocus, motion blur, and image-softening effects.	`cv2.GaussianBlur`

C Real-world dataset curation

We collected a total of 312 paper ECGs from The First Affiliated Hospital with Nanjing Medical University between March 2026 and June 2026. The images were obtained using a Ricoh 8001 scanner. To ensure data de-identification and protect patient privacy, the top section of each scanned image was cropped to remove machine-generated ECG parameters and personally identifiable information (PII), including patient names, inpatient numbers, and ages. The processed images feature an asynchronous 6×2 layout, retaining the complete 12-lead ECG waveforms, and the acquisition timestamps. Two examples of these images are illustrated in Figure 5.

The preliminary diagnostic opinions were provided by senior cardiologists. The image acquisition procedure was executed by a second-year graduate student, while the final data annotation and labelling were conducted by a Chief Physician. Each original image was transformed to RGB and resized to fit within a 384×224 pixel canvas while preserving its original aspect ratio. We evaluated our model using a nested five-fold cross-validation strategy. For each outer iteration, one fold was held out for testing, while the remaining data underwent an another five-fold split to form the training and validation sets.

This retrospective study was conducted in accordance with the Declaration of Helsinki and received approval from the Jiangsu Province Hospital Institutional Review Board (Approval number: 2025-SR-738), which waived the requirement for written informed consent.

D Baselines and implementations

The pretraining experiments comprised five baseline methods. SimCLR, MoCo, DINO, and BYOL were implemented as self-supervised learning baselines. All methods were pretrained using the generated PTB-XL ECG image dataset with an input resolution of 224×224. CvT-13 was adopted as the image encoder and initialized with ImageNet-pretrained weights. The random seed was fixed at 3407, and the training batch size was set to 64. All pretraining experiments were conducted on a single NVIDIA RTX PRO 6000 GPU with 96 GB of memory.

For downstream fine-tuning, in addition to the five pretrained baseline methods, we further included two image models (PULSE and GEM), each containing 303.5M parameters. All methods were

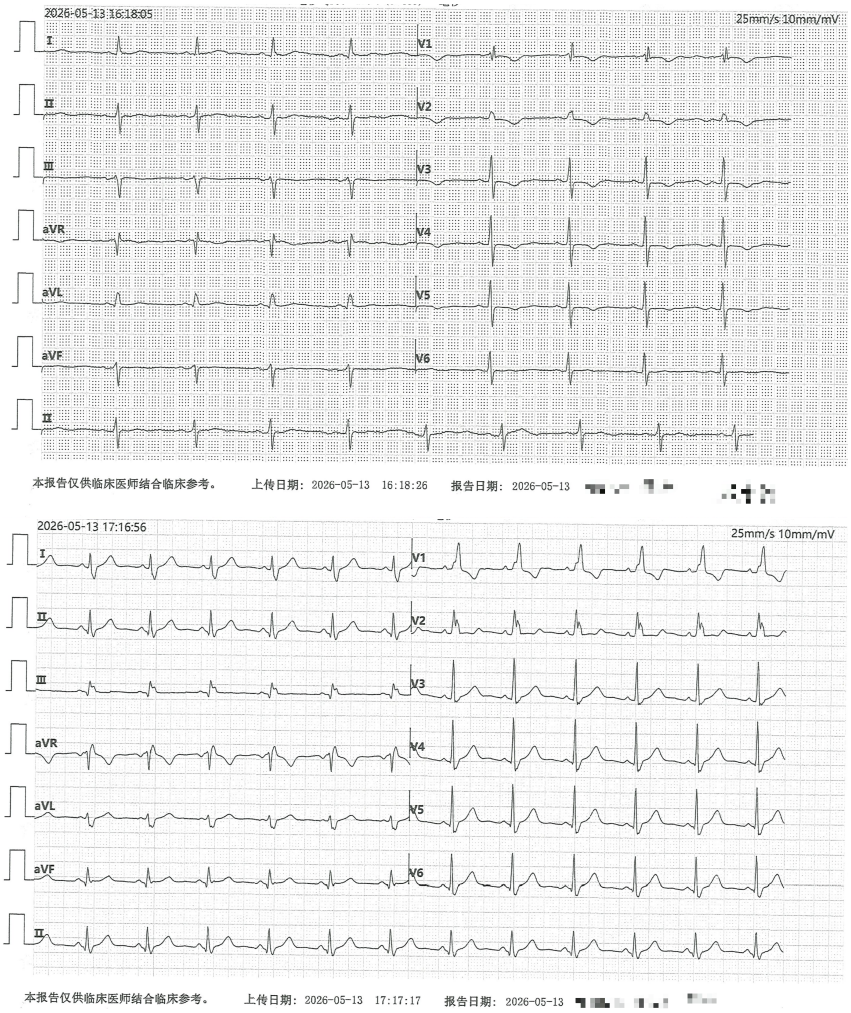

Figure 5: Examples of anonymized paper ECG images from the real-world cohort.

Table 9: Pretraining settings of baseline methods.

Method	Max epoch	Actual epoch	Pretraining LR	Optimizer	Early stopping strategy
SimCLR	50	38	3×10^{-2}	SGD	Patience = 5; validation contrastive loss
MoCo	80	26	3×10^{-2}	SGD	Patience = 5; validation contrastive loss
DINO	50	50	5×10^{-4}	AdamW	Patience = 5; validation self-distillation loss
BYOL	50	16	3×10^{-2}	SGD	Patience = 8; validation BYOL loss

fine-tuned on the same downstream datasets with 1% and 10% training data, using five random seeds from 3407 to 3411. For each method pretrained on PTB-XL, all five downstream runs use the same checkpoint obtained with pretraining seed 3407. The reported standard deviations therefore capture downstream fine-tuning variability, but not variability across independent pretraining runs.

E Distances for ordinal pairs

Geometric distance. Given an image $\mathbf{x}$, we extract its binary Canny edge map, where $\mathcal{P}$ denotes the set of edge-pixel coordinates. The geometric discrepancy between images $\mathbf{x}_a$ and $\mathbf{x}_b$ is computed via

the symmetric Chamfer distance:

$$d_g(\mathbf{x}_a, \mathbf{x}_b) = \frac{1}{2D}\left(\frac{1}{|\mathcal{P}_a|}\sum_{\mathbf{u}\in\mathcal{P}_a}\min_{\mathbf{v}\in\mathcal{P}_b}\|\mathbf{u}-\mathbf{v}\|_2 + \frac{1}{|\mathcal{P}_b|}\sum_{\mathbf{v}\in\mathcal{P}_b}\min_{\mathbf{u}\in\mathcal{P}_a}\|\mathbf{v}-\mathbf{u}\|_2\right), \tag{6}$$

where D is a distance normalization factor.

CIELAB distance. To evaluate variations in paper color, illumination, and contrast, we map images into the CIELAB color space. For each channel $c \in \{L, a, b\}$, we compute the one-dimensional Wasserstein-1 distance, denoted as $W_{1,c}$, between their marginal cumulative histograms. The final appearance distance is the average across the three channels:

$$d_c(\mathbf{x}_a, \mathbf{x}_b) = \frac{1}{3}\sum_{c\in\{L,a,b\}} W_{1,c}(\mathbf{x}_a, \mathbf{x}_b). \tag{7}$$

This serves as a computationally efficient metric rather than an exact joint three-dimensional optimal-transport distance.

F Ablation study

We ablate the rank-loss weight λ while keeping all other pretraining and evaluation settings fixed. Specifically, all experiments use `set` positive aggregation, a set-level multi-positive InfoNCE temperature of 0.10, a ranking margin of $m = 0.10$, an ordinal-edge separation parameter of $\eta = 0.05$, an anchor batch size of 64, and seed 3407. The encoder is optimized using an initial learning rate of 0.03, a minimum learning rate of 10^{-5}, one warm-up epoch, momentum of 0.9, and weight decay of 10^{-4}. Online augmentation is disabled. We assess the representation quality by linear probing on the five PTB-XL diagnostic superclasses (CD, HYP, MI, NORM, and STTC). We keep the encoder frozen in evaluation mode, and then we train a single linear head on clean and Severity 1–4 images from folds 1–8, using the binary cross-entropy with logits and SGD at a learning rate of 0.1. Fold 9 is used for checkpoint selection and per-class F1 threshold calibration, pooling all severity levels. We report performance separately at each severity level on fold 10. Table 10 shows that a moderate rank-loss weight of $\lambda = 0.5$ provides the best overall balance between discriminative performance and robustness to progressive degradation.

Table 10: Ablation study.

Pareto weight	Clean Paper ECG			Severity 1			Severity 2			Severity 3			Severity 4			Avg.		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1
0	0.835	0.637	0.611	0.836	0.638	0.611	0.831	0.631	0.607	0.820	0.618	0.597	0.803	0.592	0.576	0.825	0.623	0.600
0.5 (RobECG-CL)	**0.841**	**0.649**	0.619	0.839	**0.649**	**0.620**	**0.837**	**0.644**	**0.617**	**0.823**	**0.624**	**0.607**	**0.805**	**0.599**	0.575	**0.829**	**0.633**	**0.607**
1.0	0.838	0.644	0.617	0.836	0.641	0.614	0.834	0.638	0.614	**0.823**	0.621	0.599	0.802	0.590	**0.581**	0.827	0.627	0.605
2.0	0.840	0.648	**0.620**	**0.840**	0.646	0.617	0.836	0.640	0.615	0.820	0.617	0.594	0.787	0.570	0.551	0.825	0.624	0.599

G Supplementary related work

Large pretrained models have improved transfer learning for waveform-based ECG analysis. ECG-FM combines masked reconstruction and contrastive learning to learn label-efficient representations [12]. ECGFounder uses more than ten million cardiologist-annotated ECG recordings to support transfer across clinical tasks and data sources [10]. CSFM further uses masked generative pretraining on heterogeneous cardiac signals and associated text, enabling transfer across tasks, lead configurations, and sensing devices [7]. These models provide strong references for low-label transfer, operating on digital waveforms.

Image-based ECG models instead learn directly from ECG images. ECG-DualNet performs multilabel prediction from digitally rendered ECG images, while PRESENT-SHD uses ECG images generated with many predefined formats for structural heart disease screening [15, 4]. The PhysioNet Challenge 2024 extended ECG image analysis to synthetic and real images with scanning and imaging artifacts [13, 14]. In parallel, visual self-supervised learning provides a general approach to transferable representation learning: SimCLR and MoCo learn augmentation-invariant features using contrastive objectives, and BYOL and DINO learn from paired views without explicit negative samples [3, 8, 6, 2]. RobECG-CL learns consistent representations for progressively degraded views of the same ECG recording while preserving their relative degradation relationships.