ECG-CoCa: Physiology-Informed Contrastive and Generative Framework for ECG Interpretation

Xinyue Liu*[1,2][0009−0002−7311−938X], Yiming Lei*[1][0009−0000−8049−5290], Shiwei Li†[2][0000−0002−8765−9283], Zhiyi Zhu[1][0009−0005−4219−0046], Zhenghui Hu[2][0000−0002−6106−0416], and Qingjie Liu[1,2][0000−0002−5181−6451]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
{xinyueliu,ymlei,22230620,qingjie.liu}@buaa.edu.cn
[2] Hangzhou Innovation Institute, Beihang University, Hangzhou 310051, China
{shiweili93,zhenghuihu2021}@buaa.edu.cn

Abstract. Electrocardiogram (ECG) is a quintessential non-invasive modality for cardiovascular diagnosis. However, existing automated interpretation systems often rely on reconstructed ECG images, leading to a significant loss of fine-grained morphological details, or suffer from long-tail distribution and diagnostic drift in complex clinical scenarios. To address these issues, we propose **ECG-CoCa**, a unified framework that synergistically integrates contrastive and generative learning for direct interpretation of raw ECG signals. To capture essential ecg features, we design the Physiology-Informed Dual-Granularity Encoder (**PI-DGE**), which extracts multi-scale morphological details and global rhythmic patterns, providing both global contrastive alignment and token-level generative modeling. To mitigate long-tail bias, we develop Clinical-Concept Aware Reweighting (**CCAR**), which dynamically optimizes the generative objective by shifting from simple empirical risk to concept-rarity-informed reweighted risk. Furthermore, the Diagnostic-Context Memory (**DCM**) module bridges the semantic gap by injecting global diagnostic anchors into the multimodal decoder memory, ensuring clinical consistency in report generation. Results across diverse tasks demonstrate that ECG-CoCa effectively captures intricate ECG features, showcasing its versatility for automated diagnostic systems and its potential to enhance the efficiency of large-scale cardiovascular screening.

Keywords: Contrastive Learning · Report Generation · ECG Analysis.

1 Introduction

Cardiovascular diseases remain the preeminent cause of global mortality [2]. The 12-lead ECG is the quintessential non-invasive modality for cardiac monitoring, requiring synergistic analysis of rhythmic patterns and subtle P-QRS-ST morphological nuances [1]. However, expert interpretation is constrained by intensive training and subjectivity, with error rates reaching 25%–31% due to

* These authors contributed equally to this work. † Corresponding author.

complex pathological variability [6]. This necessitates automated systems that deliver consistent, diagnostically faithful, and clinically interpretable analysis.

Early efforts in automated ECG interpretation primarily relied on expert-defined rules and hand-crafted features, which often struggled with the non-stationary nature of physiological signals and lacked robustness against noise. The advent of deep learning has fundamentally reshaped this landscape. Supervised architectures, including 1D Convolutional Neural Networks [31], Recurrent Neural Networks [28], and Transformers [26], have demonstrated expert-level accuracy in specific arrhythmia classification tasks. However, these discriminative paradigms lack physiological interpretability as they provide only isolated categorical labels. Furthermore, their performance is constrained by an unsustainable dependency on large-scale, labor-intensive expert annotations [5].

To overcome the limitations of isolated categorical predictions, the frontier of ECG analysis has shifted toward multimodal learning. Inspired by the remarkable Success of CLIP [20], contrastive learning has emerged as a dominant framework for aligning ECG signals with unstructured clinical narratives in a shared latent space. Models such as ECG-CLIP [34] and METS [11] leverage these objectives to facilitate zero-shot diagnostic reasoning and robust feature representation. However, these contrastive-only paradigms primarily focus on global alignment, often struggling to capture the fine-grained morphological details required for precise localized diagnosis. Concurrently, generative modeling has transformed interpretation by moving toward automated report generation. Recent advances, including ECG-Chat [32] and HeartLLM [27], explore the synergy between complex signal dynamics and language decoders to generate clinically interpretable narratives. Nevertheless, these generative models remain susceptible to hallucinations due to the significant semantic gap between high-dimensional waveforms and discrete clinical text in practical scenarios.

However, advancing these multimodal frameworks into reliable clinical tools is hindered by three persistent bottlenecks. First, **Inadequate Granularity Coupling** occurs as models often rely on image-based proxies or global alignment, failing to disentangle fine-grained morphological nuances from holistic rhythmic patterns. Second, **Long-Tail Concept Bias** remains prevalent; standard empirical risk minimization forces models to converge toward majority classes like sinus rhythm, thereby neglecting rare but critical pathological concepts. Finally, the absence of diagnostic constraints leads to **Diagnostic Semantic Drift**, where generative decoders produce textual descriptions deviating from physiological evidence, compromising clinical consistency.

To address these issues, we propose ECG-CoCa, a unified foundation framework that synergistically integrates contrastive and generative learning to achieve a comprehensive and robust understanding of complex, multi-channel raw ECG signals. We design the **Physiology-Informed Dual-Granularity Encoder (PI-DGE)**, which disentangles morphology-level and rhythm-level representations via multi-scale convolutions and Transformer layers, yielding synergistic embeddings for both contrastive and generative tasks. To mitigate the inherent clinical conservatism, we develop **Clinical-Concept Aware Reweighting**

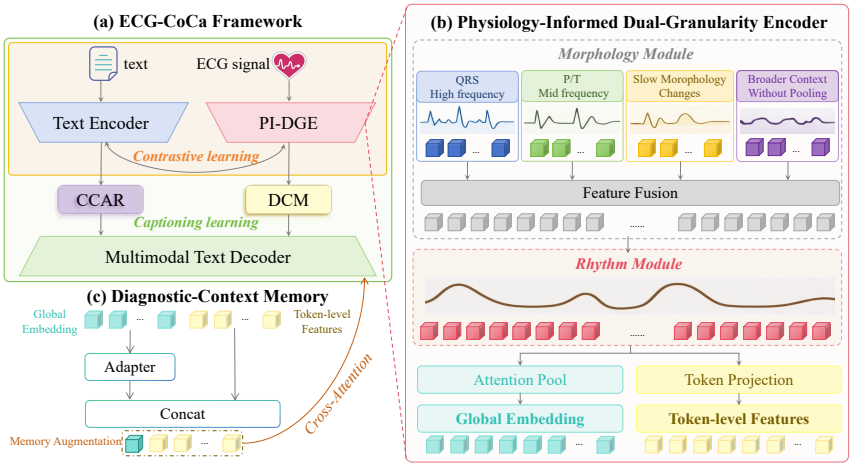

Fig. 1. (a) Overview of the ECG-CoCa framework. (b) The PI-DGE for multi-scale feature extraction. (c) The DCM module for enhancing report generation.

(CCAR), which reformulates the training objective by shifting from naive empirical risk minimization to a concept-rarity-informed reweighted risk, effectively prioritizing rare yet critical pathological conditions. Furthermore, we propose a **Diagnostic-Context Memory (DCM)** module that injects global diagnostic anchors into the decoder's cross-attention space. By providing persistent semantic guidance without compromising the decoder's language priors, DCM ensures diagnostic fidelity and mitigates hallucinations in report generation.

In summary, our work makes three primary contributions to ECG interpretation. 1) We propose **ECG-CoCa**, which integrates contrastive and generative learning through a dual-granularity encoder (**PI-DGE**), bridging the gap between local morphology and global rhythm. 2) We introduce **CCAR**, a concept-rarity-informed reweighting strategy to mitigate long-tail bias, and **DCM**, a memory-augmentation module that ensures diagnostic consistency through global semantic anchoring. 3) Extensive experiments across cross-modal retrieval, zero-shot classification, and report generation show that ECG-CoCa consistently achieves state-of-the-art performance with superior clinical fidelity.

2 Method

2.1 Problem Formulation

We formulate ECG-CoCa as a unified multi-modal framework designed to reconcile global diagnostic alignment with granular clinical narrative generation. Let $\mathcal{D} = \{(x_i, t_i, c_i)\}_{i=1}^{N}$ denote a dataset of N clinical records, where each $x \in \mathbf{R}^{12\times5000}$ represents a standard 10-second, 12-lead ECG recording sampled at 500 Hz. For each record, we derive two distinct textual modalities to facilitate

multi-objective learning. First, a contrastive text (c_i) is constructed by augmenting diagnostic labels with objective physiological biomarkers (e.g., RR and QTc intervals) extracted via NeuroKit2, providing a low-entropy, template-driven anchor for contrastive alignment. Second, a generative caption (t_i) is synthesized as a high-entropy clinical narrative through expert-curated templates, mirroring the semantic complexity of medical discourse to support report generation.

2.2 ECG-CoCa

The ECG-CoCa (Fig. 1(a)) comprises three modules: a PI-DGE for physiology-informed dual-granularity signal encoding, a pre-trained BiomedBERT [3] for text encoding, and a multimodal Transformer decoder for report generation.

Physiology-Informed Dual-Granularity Encoder. To resolve the representational tension between fine-grained waveform morphology and long-range rhythmic patterns, we propose PI-DGE (Fig. 1(b)). The morphology module employs Multi-Scale Convolutional Blocks to capture frequency geometries:

$$H_{\mathrm{m}} = \mathrm{GELU}\Big(\mathrm{Merge}\big(\mathrm{Conv}_k(x)\big) + \mathrm{Res}(x)\Big), \quad k \in \{7, 15, 31, 15_{\mathrm{dilated}}\} \tag{1}$$

where $k = 7$ resolves high-frequency QRS complexes, while larger and dilated kernels distill slow morphological variations in P and T waves. The resulting dense feature map H_{m} is then projected via a strided tokenizer $\mathcal{T}(\cdot)$ and augmented with sinusoidal positional encodings P to form rhythm tokens. These tokens are processed by a Rhythm Transformer to model macroscopic temporal dynamics, such as RR-interval variability and complex conduction delays.

To support synergistic multimodal objectives, PI-DGE employs a dual-pathway projection. For contrastive alignment, a Global Diagnostic Embedding $E_g = \mathrm{LN}(\mathbf{W}_g \cdot \sum_i \alpha_i h_i)$ is aggregated via attention pooling over hidden states h_i. Concurrently, the unpooled sequence is projected into Granular ECG Tokens $E_l = \mathbf{W}_t \cdot H_{\mathrm{rhy}}$ to serve as high-fidelity anchors for generative cross-attention.

Clinical-Concept Aware Reweighting. To mitigate the diagnostic bias induced by the extreme long-tail distribution of cardiac pathologies, we introduce CCAR to shift the generative training focus from statistical prevalence to clinical significance. In large-scale ECG datasets, majority classes often overshadow rare but high-stakes conditions, causing the decoder to exhibit classification laziness. CCAR addresses this by quantifying the rarity of clinical concepts and dynamically elevating the optimization priority of samples containing critical, low-frequency findings. This strategy forces the model to transcend empirical statistical biases and capture the fine-grained features essential for rare diagnostic intents. Specifically, we estimate the global occurrence probability p_k for each concept k in a clinical vocabulary $\mathcal{V}_c$. For a report t_i containing concepts $\mathcal{K}_i \subseteq \mathcal{V}_c$, we define a sample-specific weight ω_i using a max-pooling strategy:

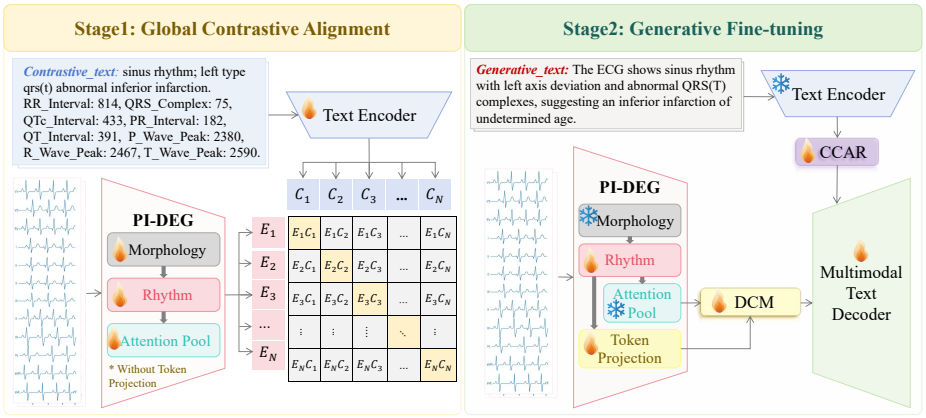

Fig. 2. The two-stage training strategy of ECG-CoCa.

$$\omega_i = \min\left(\max\left(\max_{k\in\mathcal{K}_i}\left(\frac{1}{p_k+\epsilon}\right)^{\alpha}, W_{\min}\right), W_{\max}\right) \tag{2}$$

where α is a smoothing factor, ϵ is a damping term, and $W_{\min}, W_{\max}$ are boundary constraints. This formulation ensures that even a single rare pathology can significantly impact the generative loss. By transforming the generative objective into a concept-aware risk, CCAR effectively prioritizes clinically salient features over simple frequency-based patterns during report generation.

Diagnostic-Context Memory. To address the challenge of diagnostic drift in automated report generation, we propose the DCM module (Fig. 1(c)). Unlike vanilla decoders that rely solely on local waveform tokens, DCM explicitly injects the high-level diagnostic priors E_g into the generative memory space. Specifically, a lightweight adapter projects the global diagnostic embedding E_g into decoder latent space to form a specialized diagnostic anchor token p_{diag}:

$$p_{\text{diag}} = \text{Dropout}(\text{GELU}(\mathbf{W}_p E_g + b_p)) \tag{3}$$

where p_{diag} serves as a stable semantic reference. To provide the decoder with global context and fine-grained details, we redefine the memory bank by prepending p_{diag} to the granular ECG token sequence $E_l = \{e_1, \ldots, e_L\}$:

$$\text{Memory}_{\text{aug}} = [\, p_{\text{diag}},\, e_1,\, e_2,\, \ldots,\, e_L \,] \tag{4}$$

During each decoding step, the cross-attention mechanism is forced to attend to this augmented memory. By placing p_{diag} at the primary position of the sequence, DCM ensures that every generated token is grounded by the initial diagnostic intent. This persistent anchoring effectively prevents the forgetting of core clinical findings as the sequence length increases, ensuring that the final narrative remains strictly consistent with the ECG's physiological evidence.

Table 1. Cross-modal Retrieval Performance on MIMIC-IV-ECG and PTB-XL. Best results are shown in bold and second-best results are underlined.

Model	MIMIC-IV-ECG						PTB-XL					
	ECG → Text			Text → ECG			ECG → Text			Text → ECG		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Image-Text Contrastive Method												
CLIP [20]	0.84	3.04	5.14	1.06	4.02	6.44	0.82	3.28	5.74	0.91	3.00	5.05
CoCa [29]	0.02	0.14	0.22	0.02	0.14	0.22	0.05	0.32	0.41	0.09	0.18	0.46
UniMed-CLIP [8]	0.02	0.08	0.24	0.00	0.08	0.10	0.05	0.27	0.50	0.09	0.23	0.36
BiomedCLIP [30]	0.04	0.16	0.32	0.04	0.12	0.20	0.05	0.18	0.32	0.05	0.27	0.55
GEM [9]	0.02	0.12	0.20	0.00	0.08	0.14	0.05	0.23	0.36	0.00	0.27	0.73
Signal-Text Contrastive Method												
Human-CLAP [22]	0.02	0.06	0.14	0.00	0.08	0.22	0.05	0.18	0.50	0.14	0.27	0.50
Clip-TTS [17]	22.67	43.74	53.11	23.50	44.85	54.79	1.41	4.73	7.83	2.18	5.60	9.10
MERL [13]	0.52	1.74	2.74	0.04	0.14	0.20	0.41	1.04	1.63	0.46	1.37	2.18
MELP [25]	0.04	0.08	0.14	0.96	2.42	3.18	0.04	0.13	0.31	0.09	0.27	0.31
ECG-Chat [32]	86.57	96.09	97.53	86.53	96.07	97.75	50.25	73.65	80.52	57.35	79.79	85.98
MAUnet [4]	81.64	95.81	97.79	82.70	96.33	97.85	35.55	65.09	74.74	43.33	72.74	81.75
Ecgmamba [19]	82.62	96.27	98.16	82.80	96.37	98.10	36.46	63.91	73.69	42.79	72.74	82.61
ECG-CoCa & Backbone Ablations												
1D ViT	75.19	92.82	95.92	76.03	92.54	95.21	8.92	22.17	28.86	9.51	25.94	35.05
CNN-Global	51.54	80.99	89.01	48.46	77.29	85.97	8.78	22.85	32.27	10.33	26.58	39.33
Ours (Hybird)	**94.93**	**98.79**	**99.31**	**94.91**	**98.88**	**99.42**	**89.12**	**97.82**	**98.73**	**89.30**	**97.82**	**98.73**

2.3 Training Strategy

Stage 1: Global Contrastive Alignment. Following Fig. 2, a contrastive loss optimizes the PI-DGE and text encoder by maximizing the cosine similarity between ECG embeddings E_g and text embeddings C_g for each batch B:

$$\mathcal{L}_{\text{Con}} = \frac{1}{2B}\sum_{i=1}^{B}\left(\log\frac{e^{\tau E_{g,i}^{\top}C_{g,i}}}{\sum_{j=1}^{B}e^{\tau E_{g,i}^{\top}C_{g,j}}} + \log\frac{e^{\tau C_{g,i}^{\top}E_{g,i}}}{\sum_{j=1}^{B}e^{\tau C_{g,j}^{\top}E_{g,i}}}\right) \tag{5}$$

Stage 2: Generative Fine-tuning. We activate the multimodal text decoder and DCM, unfreezing only the rhythm module and projection layers to maintain stability. The generative objective is optimized via the CCAR loss:

$$\mathcal{L}_{\text{Gen}} = -\frac{1}{B}\sum_{i=1}^{B}\omega_i\sum_{m=1}^{M}\log P\big(y_{i,m} \mid y_{i,<m}, \text{Memory}_{\text{aug},i}\big) \tag{6}$$

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate ECG-CoCa across a large-scale, multi-source corpus to ensure robust generalization. Training utilizes MIMIC-IV-ECG [7], PTB-XL [23],

Table 2. Zero-Shot Performance Comparison on CPSC2018 and CODE-15% Datasets.

Model	CPSC2018			CODE-15%		
	ACC↑	F1↑	AUC↑	ACC↑	F1↑	AUC↑
Image-Text Contrastive Method						
CLIP [20]	11.67	20.42	55.32	14.56	3.93	54.19
CoCa [29]	19.59	20.68	48.85	71.32	4.84	49.36
UniMed-CLIP [8]	11.67	20.42	50.00	43.66	5.81	53.01
BiomedCLIP [30]	11.67	20.42	50.00	30.69	4.62	43.87
GEM [9]	12.93	20.47	50.23	31.68	4.60	46.55
Signal-Text Contrastive Method						
Human-CLAP [22]	11.67	20.42	51.32	14.45	16.83	49.08
Clip-TTS [17]	75.15	36.08	73.81	76.14	<u>56.32</u>	<u>76.65</u>
MERL [13]	34.63	20.94	60.89	60.80	16.28	52.05
MELP [25]	46.41	27.45	50.93	20.41	16.86	34.62
ECG-Chat [32]	<u>81.26</u>	<u>47.61</u>	<u>81.52</u>	<u>87.36</u>	20.14	72.69
MAUnet [4]	80.30	40.18	76.77	70.78	14.48	73.15
Ecgmamba [19]	75.19	29.76	67.88	61.16	16.88	60.11
ECG-CoCa & Backbone Ablations						
1D ViT	78.30	41.50	78.03	61.03	20.43	63.87
CNN-Global	77.96	35.48	74.01	57.85	16.54	58.94
Ours (Hybrid)	**86.63**	**52.82**	**85.30**	**93.66**	**56.83**	**80.93**

SPH [16], and CSN [33]. We perform in-domain evaluation on MIMIC-IV-ECG [7] and PTB-XL [23], while CODE-15% [21] and CPSC2018 [14] as cross-dataset benchmarks. All splits follow patient-level or official protocols to prevent leakage.

Implementation Details. ECG-CoCa is implemented in PyTorch and trained on 8 NVIDIA RTX A6000 GPUs. Both stages use the AdamW optimizer ($lr = 1 \times 10^{-4}$, weight decay=0.05, clipping=0.5). Stage 1 optimizes the encoders for 30 epochs with a global batch size of 256. Stage 2 initializes from the optimal Stage 1 checkpoint and fine-tunes for another 30 epochs to ensure convergence.

3.2 Experimental Results

Cross-modal Retrieval. Table 1 demonstrates that ECG-CoCa establishes a new state-of-the-art across all retrieval metrics, significantly outperforming existing image-text and signal-text paradigms. On the MIMIC-IV-ECG benchmark, our framework achieves an exceptional R@1 of 94.93%, surpassing the strongest baseline, ECG-Chat, by approximately 8.3%. More importantly, on the clinically complex PTB-XL dataset, ECG-CoCa maintains high diagnostic fidelity with an R@1 of 89.12%. These gains underscore the structural superiority of the PI-DGE module; by decoupling micro-scale morphology from macro-scale rhythms, it cultivates a discriminative latent space that bridges raw signals and professional descriptors. Clinically, this robust alignment facilitates precise report indexing and rapid longitudinal search for evidence-based analysis.

Table 3. Report Generation Performance across NLP (B-4: BLEU-4, M: METEOR, R-L: ROUGE-L) and Clinical Efficacy (CP: Clinical_P, NACP: NAClinical_P) Metrics.

Model	MIMIC-IV-ECG					PTB-XL				
	B-4↑	M↑	R-L↑	CP↑	NACP↑	B-4↑	M↑	R-L↑	CP↑	NACP↑
General and Medical Multimodal Models										
BLIP2 [12]	0.007	0.048	0.082	0.096	0.096	0.008	0.051	0.084	0.076	0.076
CoCa [29]	0.012	0.100	0.174	0.534	0.532	0.015	0.114	0.180	<u>0.751</u>	<u>0.747</u>
LLaVA-v1.6 [15]	0.002	0.077	0.084	0.553	0.553	0.002	0.069	0.069	0.664	0.664
LLaVA-Med [10]	0.005	0.123	0.129	0.520	0.520	0.004	0.119	0.120	0.721	0.721
MedVLM-R1 [18]	0.005	0.079	0.086	0.211	0.000	0.002	0.034	0.075	0.000	0.000
Signal-Text Generative Models										
HeartLLM [27]	0.046	0.127	0.265	0.628	0.628	0.012	0.063	0.179	0.596	0.596
ECG-Chat [32]	0.005	0.074	0.084	0.189	0.189	0.004	0.069	0.080	0.220	0.220
GEM [9]	0.088	0.162	0.166	0.533	0.533	0.007	0.137	0.148	0.746	0.742
MEIT [24]	<u>0.597</u>	<u>0.764</u>	<u>0.779</u>	0.713	0.749	<u>0.448</u>	<u>0.656</u>	<u>0.587</u>	0.693	0.686
ECG-CoCa & Module Ablations										
Ours	**0.664**	**0.832**	**0.817**	**0.897**	**0.864**	**0.577**	**0.701**	**0.689**	**0.807**	**0.796**
-CARR	0.423	0.681	0.635	<u>0.833</u>	<u>0.798</u>	0.319	0.516	0.411	0.666	0.702
-DCM	0.361	0.599	0.600	0.701	0.644	0.170	0.402	0.379	0.548	0.683
-CARR -DCM	0.203	0.436	0.514	0.667	0.535	0.096	0.351	0.283	0.505	0.628

Zero-shot Classification. Table 2 presents the zero-shot performance of ECG-CoCa on external datasets, demonstrating its exceptional generalization capability. The significant F1-score improvement on CODE-15% (56.83%) particularly highlights the model's robustness against the domain shifts inherent in diverse sampling rates and regional clinical populations. This superior zero-shot transferability stems from the PI-DGE's ability to extract invariant electrophysiological features rather than overfitting to dataset-specific noise. Such robust cross-domain transferability is critical for deploying automated screening tools across heterogeneous healthcare systems, ensuring consistent diagnostic fidelity without the requisite for local fine-tuning or extensive site-specific data re-annotation.

Report Generation. Table 3 evaluates report generation using linguistic metrics and two specialized indicators: Clinical Precision (CP) for diagnostic accuracy and Negation-Aware CP (NACP) for distinguishing positive from negative findings. The performance margin over competitors like MEIT is more pronounced in clinical metrics, indicating that improvements extend beyond textual similarity to accurate pathological grounding. Ablations confirm the complementary contributions of CCAR and DCM: removing CCAR degrades performance under long-tail distributions, while excluding DCM causes larger drops in CP and NACP, signaling increased diagnostic drift. ECG-CoCa generates physiologically grounded, clinically consistent reports with reduced hallucinations.

4 Conclusion

In this paper, we presented ECG-CoCa, a unified foundation framework that synergistically couples contrastive and generative learning for automated ECG interpretation. By leveraging the PI-DGE module to disentangle multi-scale morphology from rhythm, the CCAR strategy to mitigate long-tail pathology bias, and the DCM module to eliminate diagnostic drift, our approach ensures clinical fidelity in report generation. Evaluations across three benchmarks demonstrate state-of-the-art performance in cross-modal retrieval, zero-shot classification, and report generation. ECG-CoCa effectively bridges the gap between raw physiological signals and clinical narratives, offering a robust, clinician-aligned solution for automated cardiovascular diagnosis.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 62506025).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Camm, A.J., et al.: Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. Circulation **93**(5), 1043–1065 (1996)
2. of Cardiovascular Diseases, G.B., Collaborators, R..: Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990-2023. Journal of the American College of Cardiology **86**(22), 2167–2243 (2025)
3. Chakraborty, S., et al.: Biomedbert: A pre-trained biomedical language model for qa and ir. In: Proceedings of the 28th international conference on computational linguistics. pp. 669–679 (2020)
4. Chen, M., et al.: Diagnostic-wise hybrid loss for ecg signals reconstruction based on wavelet decomposition. IEEE Transactions on Instrumentation and Measurement (2025)
5. Ebrahimi, Z., et al.: A review on deep learning methods for ecg arrhythmia classification. Expert Systems with Applications: X **7**, 100033 (2020)
6. Fu, G., et al.: Cardiogpt: An ecg interpretation generation model. IEEE Access **12**, 50254–50264 (2024)
7. Gow, B., et al.: MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet (Sep 2023). https://doi.org/10.13026/4nqg-sb35, https://doi.org/10.13026/4nqg-sb35, version 1.0
8. Khattak, M.U., et al.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv preprint arXiv:2412.10372 (2024)
9. Lan, X., et al.: Gem: Empowering mllm for grounded ecg understanding with time series and images. In: Advances in Neural Information Processing Systems (NeurIPS) 2025 (2025)

10. Li, C., et al.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
11. Li, J., et al.: Frozen language model helps ecg zero-shot learning. In: Medical Imaging with Deep Learning. pp. 402–415. PMLR (2024)
12. Li, J., et al.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
13. Liu, C., et al.: Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In: Forty-first International Conference on Machine Learning (2024)
14. Liu, F., et al.: The china physiological signal challenge 2018: automatic identification of the rhythm/morphology abnormalities in 12-lead ecgs. In: The 7th International Conference on Biomedical Engineering and Biotechnology (ICBEB 2018) (2018)
15. Liu, H., et al.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
16. Liu, H., et al.: A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. Scientific data **9**(1), 272 (2022)
17. Liu, T.: Clip-tts: Contrastive text-content and mel-spectrogram, a high-quality text-to-speech method based on contextual semantic understanding. arXiv preprint arXiv:2502.18889 (2025)
18. Pan, J., et al.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
19. Qiang, Y., et al.: Ecgmamba: Towards ecg classification with state space models. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 6498–6505. IEEE (2024)
20. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Ribeiro, A.H., et al.: Code-15%: A large scale annotated dataset of 12-lead ecgs. Zenodo, Jun **9**, 10–5281 (2021)
22. Takano, T., et al.: Human-clap: Human-perception-based contrastive language-audio pretraining. Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) pp. 131–136 (2025)
23. Wagner, P., et al.: PTB-XL, a large publicly available electrocardiography dataset. PhysioNet (Apr 2020). https://doi.org/10.13026/x4td-x982, https://doi.org/10.13026/x4td-x982, version 1.0.1
24. Wan, Z., et al.: Meit: Multimodal electrocardiogram instruction tuning on large language models for report generation. In: Findings of the association for computational linguistics: ACL 2025. pp. 14510–14527 (2025)
25. Wang, F., et al.: From token to rhythm: A multi-scale approach for ecg-language pretraining. In: Proceedings of the 42nd International Conference on Machine Learning (ICML 2025) (2025)
26. Xia, Y., et al.: Generative adversarial network with transformer generator for boosting ecg classification. Biomedical Signal Processing and Control **80**, 104276 (2023)
27. Yang, J., et al.: Diagecg: An llm-driven framework for diagnostic reasoning via discretized ecg tokenization. arXiv preprint arXiv:2508.15338 (2025)

28. Yildirim, O., et al.: A new approach for arrhythmia classification using deep coded features and lstm networks. Computer methods and programs in biomedicine **176**, 121–133 (2019)
29. Yu, J., et al.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
30. Zhang, S., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024). https://doi.org/10.1056/AIoa2400640, https://ai.nejm.org/doi/full/10.1056/AIoa2400640
31. Zhang, Y., et al.: A cnn model for cardiac arrhythmias classification based on individual ecg signals. Cardiovascular Engineering and Technology **13**(4), 548–557 (2022)
32. Zhao, Y., et al.: Ecg-chat: A large ecg-language model for cardiac disease diagnosis. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
33. Zheng, J., et al.: A large scale 12-lead electrocardiogram database for arrhythmia study. PhysioNet (Aug 2022). https://doi.org/10.13026/wgex-er52, https://doi.org/10.13026/wgex-er52, version 1.0.0
34. Zhou, X., et al.: Diagnosis of cardiac conditions from 12-lead electrocardiogram through natural language supervision. npj Digital Medicine **8**(1), 697 (2025)