Vision-Language Models in ECG Interpretation: An Exploratory Study

Sileshi Nibret Zeleke[1,*], Mario Bochicchio[1,2]

[1]Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
[2]Digital Health National Lab, CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Roma, Italy

Abstract

Electrocardiogram (ECG) interpretation remains a critical yet complex task in cardiovascular diagnostics. With the rise of multimodal learning, vision-language models (VLMs) offer a promising new paradigm for automating and explaining ECG analysis, as well as providing explainable decision support. In this study, we evaluate three state-of-the-art VLMs—GPT-4o, PULSE (zero-shot), and a fine-tuned PULSE via Low-Rank Adaptation (LoRA)—on three benchmark datasets: MIT-BIH, Chapman-Shaoxing, and CPSC-2018. We convert raw ECG signals into high-resolution printout-style images and assess not only abnormality classification but also explanation quality, including factual accuracy, completeness, contextual understanding, and hallucination, using an automated GPT-4–based evaluator. Moreover, our experiments using case studies demonstrate the effectiveness of these models in aligning visual ECG signals with clinical language, generating accurate diagnostic summaries, and providing explanations for uncertain predictions. The findings highlight both the strengths and limitations of selected medical and general-purpose VLMs, offering insights into their readiness in application.

Keywords

Electrocardiogram, Vision-Language Model, Explainable AI, Low-Rank Adaptation

1. Introduction

The electrocardiogram is a fundamental and widely utilized diagnostic tool in healthcare for monitoring the electrical activity of the heart. Its non-invasive and cost-effective nature makes it a crucial component in the initial assessment of various cardiac conditions [1]. Interpreting ECG readings, however, often demands substantial medical expertise to accurately analyze complex signals and integrate them with patient-specific information [2]. This process can be resource-intensive and prone to errors, particularly in settings with limited access to specialized medical personnel.

Recent advancements in artificial intelligence (AI), particularly in the field of VLMs, present a promising avenue for enhancing ECG interpretation. VLMs, which combine computer vision and natural language processing, have demonstrated exceptional performance across various multimodal tasks by integrating visual and linguistic information. There is a growing interest in leveraging VLMs for medical image analysis in the medical domain, aiming to achieve more intelligent and efficient multi-task processing. Despite this potential, the application of VLMs to ECG data remains underexplored. Meanwhile, VLMs have transformed fields from radiology to pathology by interpreting images and generating human-quality reports, raising an obvious question: **can these same foundations tackle ECG interpretation?** ECG signals possess unique characteristics, including temporal dependencies and lead-specific information, which differ significantly from the data types VLMs are typically trained on. Moreover, the effectiveness of fine-tuning strategies, such as Low-Rank Adaptation (LoRA), in adapting VLMs to the nuances of ECG data has not been thoroughly investigated.

Prior work in ECG AI largely focuses on time-series models or bespoke ECG-text systems MEIT [3], and ECG-Chat [4], which either ignore the visual layout of standard 12-lead printouts or require massive domain-specific corpora. General-purpose VLMs such as GPT-4o and ECG-specific PULSE

EXPLIMED 2025 - Second Workshop on Explainable Artificial Intelligence for the Medical Domain - 25–30 October 2025, Bologna, Italy

*Corresponding author.
sileshi.zeleke@uniba.it (S. N. Zeleke); mario.bochiccio@uniba.it (M. Bochicchio)
0009-0006-8172-9646 (S. N. Zeleke); 0000-0002-9122-6317 (M. Bochicchio)

[5], however, are pretrained on image–text pairs and can already "read" chart-style data. Yet rigorous evaluation of how off-the-shelf VLMs cope with waveform artifacts, grid lines, and multi-panel layouts, or what gains lightweight fine-tuning might unlock.

In this work, we ask:

- **Baseline capability:** How do leading VLMs perform on ECGs without any adaptation?
- **Efficient adaptation:** Can low-rank LoRA updates boost accuracy and interpretability with minimal compute and data?
- **Clinical relevance:** Do qualitative case studies reflect genuine understanding of complex arrhythmias, and where do these models still fail?

Moreover, we evaluate the performance of GPT-4o, PULSE, and a fine-tuned PULSE using LoRA on three benchmark ECG datasets: MIT-BIH, Chapman-Shaoxing, and CPSC-2018. By addressing these aspects, we aim to advance the integration of multimodal AI models into clinical ECG interpretation workflows.

2. Related Work

Recent years have witnessed significant progress in AI-driven ECG analysis with three emerging trajectories: clinical reliability, explainable AI, and multimodal integration of ECG data. Although the use of VLMs in medical images [6], such as in radiology, has been thoroughly investigated, their application for ECG interpretation is more recent. Compared to the analysis of static images, the special characteristics of ECG, which is usually represented as a waveform, present both special potential and particular obstacles for the implementation of VLM approaches [7].

Instruction tuning has demonstrated significant efficacy in the multimodal domain, particularly in VLMs such as LLaVA [6], MiniGPT-4 [8], and InstructBLIP [9]. These models exhibit impressive generalizability across various visual understanding and reasoning tasks. While multimodal instruction tuning has been applied to general medical imaging tasks, its application to ECG images remains largely unexplored. A recent study introduced the framework [3], fine-tuning existing open-source large language models for ECG report generation. However, this approach is limited by a single-task instruction dataset focused solely on report generation, potentially constraining its adaptability to other ECG-related tasks. Moreover, their work treats ECG data as temporal signals, whereas PULSE [5] focuses on encoding ECG images with multimodal large language models, which is more applicable to real-world scenarios where only printed or digital ECG images are available.

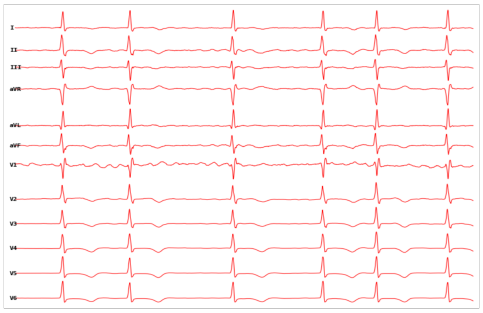
ECG-CoCa [4], an ECG encoder trained on ECG-text pairs, alongside ECG-Chat, a modified LLaVA model capable of processing ECG time series. Moreover, [3] created a framework for instruction tuning that converts ECG-text pairs into chatbot-style instructions and optimizes the linear layers of the LLM for automated ECG report creation. A specifically designed model employing an improved ResNet-18 architecture-based ECG encoder transforms raw ECG signals into a high-dimensional feature space. This feature space is then carefully aligned with the textual feature space derived from a large language model [10].

3. Methodology

We selected three models to represent a spectrum of capabilities: (1) GPT-4o: A state-of-the-art, general-purpose multimodal LLM, chosen to establish a powerful zero-shot baseline and assess the out-of-the-box capability of foundational models on ECG data. (2) PULSE: A recently published VLM specifically pre-trained on ECG-image-text pairs, chosen to represent the current state-of-the-art in domain-specific VLMs for ECG interpretation in a zero-shot setting. (3) PULSE+LoRA: Our fine-tuned version of PULSE. This choice allows us to directly isolate and measure the benefit of efficient parameter fine-tuning on a domain-specific model, answering whether lightweight adaptation can bridge the gap between general pre-training and specialized clinical performance.

The fine-tuning process aims to enhance PULSE's capability in understanding and interpreting ECG images, especially for arrhythmia interpretation. We evaluate the performance of the fine-tuned PULSE model both quantitatively and qualitatively, comparing it against the original PULSE and GPT-4o models. Subsection 3.1 presented a fine-tuning data preparation procedure, and Subsection 3.2 discusses about ECG signal-to-image conversion method. Finally, Subsection 3.3 is about the fine-tuning process.

Instruction
Evaluate the ECG and determine what is the type from the choice given and provide a detailed explanation for your choice:
Options:
A) Sinus Rhythm
B) Sinus brcycardia
C) Atrial fibrillation
D) Supraventricular tachycardia
Response:
C) Atrial fibrillation. The ECG shows an irregularly irregular narrow-complex tachycardia with absence of discrete P waves, consistent with atrial fibrillation. The ventricular rate is approximately 130 bpm. There is no evidence of acute ischemia or conduction blocks.



(a) Fine-tuning data sample from Chapman-Shaoxing dataset; Multi-choice QA

Instruction
You are provided with an ECG signal image. Please classify the ECG image into the following abnormality:
Options:
A) Left bundle branch block
B) Aberrated atrial premature
C) Nodal junction escape
D) Fusion of ventricular and normal beats
Response:
A) Left bundle branch block. Because, it shows wide QRS with M-shaped R wave and deep S waves and absence septal Q waves. Moreover, ST-T waves opposite to QRS direction and delayed left ventricular action indicate the condition is left bundle branch block.

(b) Fine-tuning data sample from MIT-BIH dataset; Multi-choice QA

Figure 1: Fine-tuning data image-text pair sample.

3.1. Fine-tuning Data Preparation

For fine-tuning, we constructed a new dataset composed of three widely used ECG datasets: the single-lead MIT-BIH arrhythmia database [11], the China Physiological Signal Challenge (CPSC-2018) [12], and the Chapman-Shaoxing ECG dataset [13]. We complemented the ECG signal with comprehensive textual descriptions selected from reputable medical sources, including clinical textbooks and ECG interpretation guidelines, to add clinically relevant context to the training data [14]. These textual annotations capture both general cardiac features and particular waveform qualities. For instance, *"60–100 bpm, upright P waves preceding each QRS complex, normal PR intervals, narrow QRS duration (<100 ms, and normal T waves indicating preserved conduction"* are characteristics of a typical sinus rhythm.

The ultimate fine-tuning dataset was formatted in a **multiple-choice question** style as shown in Figure 1, with a single correct description and a set of plausible distractors for every ECG sample. To prevent the model from overfitting to a fixed set of abnormality types, we randomly sampled subsets of

applicable abnormalities for each input. This forces the model to pick up on fine-grained morphological distinctions while maintaining generalization across clinically similar conditions. Moreover, the dataset, while constructed from multiple public sources and enriched with clinical context, remains limited in scale compared to the vast datasets typically used for pre-training general-purpose VLMs.

3.2. ECG Signal to Image Conversion

Given that many vision language models are designed to process image data, converting ECG signals into an image format is a crucial step in leveraging these models for ECG interpretation. Various techniques have been explored for this conversion, allowing the application of established image processing methods and semantic segmentation tools. One straightforward approach involves directly encoding each signal and serializing it into an image of the desired dimensions. We utilize the Python ECG plot library to facilitate the generation of 12-lead ECG images that resemble clinical displays and printouts directly from signal data.

3.3. LoRA Fine-Tuning

Due to the limited size of our ECG-text dataset and to avoid catastrophic forgetting and overfitting, we perform parameter LoRA fine-tuning, as shown in Figure 2. LoRA trains smaller weights while keeping the original model unchanged. Each LoRA adapter introduces a low-rank learnable matrix of the same dimensionality as the target weight matrix to enable efficient adaptation while keeping the pre-trained model parameters frozen. This matrix is integrated into specific model layers to adjust their behavior without modifying the original parameters.

To enable efficient fine-tuning, we modify the original weight matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ by adding a low-rank adaptation in the form of a rank-constrained matrix $\Delta\mathbf{W}$, which is a product of $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times k}$. Given r, the rank of the adaptation acts as a tunable hyperparameter that controls the trade-off between model capacity and efficiency, with $r \ll \min(d, k)$, where d and k are the working dimensions of $\mathbf{W}$. During fine-tuning, the base model parameters $\mathbf{W}$ are kept frozen, while only the parameters of $\mathbf{A}$ and $\mathbf{B}$ are updated. We utilize four weight matrices, denoted as $\mathbf{W_Q}$, $\mathbf{W_K}$, $\mathbf{W_V}$, and $\mathbf{W_O}$, which correspond to the weights of the query, key, value, and output projections within the multi-head self-attention module, respectively.

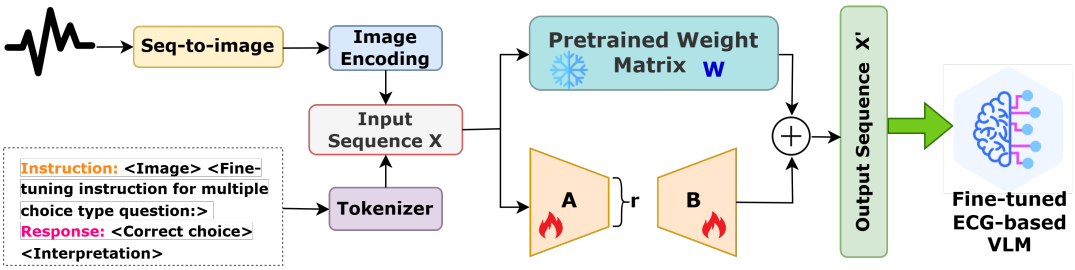


Figure 2: Illustration of the LoRA fine-tuning procedure.

4. Experimental Setting

4.1. Experimental Setting and Implementation

During the fine-tuning phase, the rank hyperparameter r is empirically optimized to $r = 4$ based on common practices in the literature for similar model sizes, achieving an optimal balance between the augmentation of parameters and model performance. The scaling factor α plays a crucial role in mediating the stability and adaptability of low-rank metrics; it is methodically set to a value of 8 before integration with $\mathbf{W}$. These parametric values help to efficiently fine-tune the pre-trained model,

minimizing trainable parameters while maintaining adequate model expressiveness through appropriately scaled updates. Nevertheless, to enhance the modality and tolerance to context shifts between the pretrained representation and ECG patterns, the LoRA bias is adjusted to *lora_only*. All models were then optimized to achieve the best performance across all datasets, including hyperparameter optimization. Training and inference were both conducted on four NVIDIA RTX A6000 GPUs, each with 48 GB of VRAM, enabling parallel experimentation and efficient handling of high-dimensional ECG signals. Experiments were performed in PyTorch and executed on a multi-GPU workstation with CUDA acceleration.

4.1.1. Evaluation Metrics

We employ several evaluation metrics to assess the performance of VLMs, encompassing ECG signal analysis and ECG abnormality detection. Given the class imbalance in the datasets, balanced accuracy, weighted F1, weighted recall, and weighted precision were used to provide a more reliable performance measure.

Interpretation Evaluation Metrics To rigorously assess the quality of the textual explanations generated, we adopt an automated evaluator using GPT-4. The interpretation generated is judged against the ground truth on five criteria. Grounded in clinical requirements and recent research on explainable AI in cardiology [15], the metrics are (i) factual accuracy: to verify that the stated facts correctly reflect the ECG; (ii) completeness: measures whether the interpretation mentions each key feature highlighted in the ground truth; (iii) detail quality: assesses the specificity and precision of descriptions; (iv) context understanding: evaluates the interpretation of the overall cardiac context; and (v) hallucination: penalizes mentions of elements not present in the ECG image. The GPT-based evaluator scores responses on a scale from 1 to 5, as found efficient in related studies [16], with 1 indicating poor quality and 5 indicating high quality for the first four metrics; however, for hallucination, the lower the score, the better.

5. Results and Discussion

5.1. Abnormality Classification Evaluation Result

We test GPT-4o, PULSE, and our fine-tuned PULSE on three benchmark datasets. As shown in Table 1, GPT-4o outperforms ECG-based pre-trained and fine-tuned models on the MIT-BIH dataset in terms of accuracy and precision. However, its performance dropped on the CPSC-2018 dataset. PULSE+LoRA (finetuned) consistently improved over PULSE in most metrics across all datasets, particularly on Chapman-Shaoxing, where it achieved the highest F1 score of 43.28. This result suggests that lightweight fine-tuning using LoRA helps models better adapt to domain-specific challenges, including the integration of spatial and temporal features present in multiple lead signals. The benefit of such targeted adaptation is especially evident in more complex 12-lead datasets. Despite the improvements, all models struggle with low performance, highlighting the complexity of ECG and suggesting that current vision-language models are not yet fully optimized for this task.

Despite these gains, the overall performance of all models remains low across the board. The classification performance across all tasks indicates substantial challenges in achieving robust ECG classification with current VLMs. This underperformance reinforces the notion that ECG signals present unique challenges, such as temporal complexity, noise variability, and inter-patient heterogeneity, that current VLM architectures are not yet equipped to handle effectively.

5.2. Evaluation of Interpretability

The ability of VLMs to provide accurate, complete, and contextually relevant explanations directly impacts trust and usability. We selected 150 random test samples to evaluate the interpretation based

Table 1
Performance Comparison of Vision-Language Models on Three ECG Datasets

Metrics	MIT-BIH			Chapman-Shaoxing			CPSC-2018		
	GPT-4o	PULSE	Finetuned	GPT-4o	PULSE	Finetuned	GPT-4o	PULSE	Finetuned
Accuracy	28.95	19.00	21.81	43.30	38.00	42.44	22.89	22.00	22.00
Precision	51.01	19.38	20.96	47.63	41.79	43.90	30.00	20.24	26.54
Recall	28.95	19.00	21.81	41.24	38.00	42.44	22.89	22.00	22.00
F1-score	27.82	15.71	17.93	42.01	38.49	43.28	19.10	20.28	20.17

on a comprehensive prompt that aligns with established principles for assessing explainability. The prompt used to evaluate the model offers actionable insights into model performance by systematically assessing dimensions such as factual accuracy, quality of detail, completeness, and hallucination.

Table 2 presents a comparative evaluation of explanation quality using GPT-4o as an automatic evaluator across two models—Finetuned and PULSE—on three ECG datasets. The fine-tuned model consistently outperforms PULSE in factual accuracy, completeness, and detail quality, particularly on the CPSC-2018 dataset, indicating a stronger alignment with clinical ground truth and richer explanatory depth. While both models achieve high Context Understanding scores on CPSC-2018, PULSE exhibits elevated hallucination rates, especially on MIT-BIH, suggesting limitations in generating trustworthy outputs. The results highlight the effectiveness of fine-tuning in enhancing explanation fidelity and reducing model hallucination, with implications for improving the reliability of VLMs in medical AI applications.

Furthermore, the explanations were found to be factually accurate, achieving a mean rating of 3.78 on a 5-point scale, which reflects strong alignment with the ground truth. The comprehensive quality evaluation results indicate that the generated explanations are consistent with the corresponding classification outcomes, demonstrating coherent and reliable interpretability, as illustrated in Figure 3. Lower hallucination scores reflect better restraint, and fine-tuning reduces hallucinations. The biggest drop on MIT-BIH again highlights that adaptation helps the model avoid inventing findings on complex tracings; however, non-zero hallucination remains, which is clinically unacceptable.

Table 2
GPT-4 as evaluator performance score mean±standard deviation (↑: the higher the better, ↓: the lower the better)

	PULSE			Finetuned		
Metrics	Chapman	MIT-BIH	CPSC-2018	Chapman	MIT-BIH	CPSC-2018
Factual accuracy (↑)	3.49 ± 1.15	2.94 ± 1.21	4.33 ± 0.73	3.78 ± 1.05	3.02 ± 1.09	4.51 ± 0.30
Completeness (↑)	2.66 ± 1.01	2.18 ± 1.00	3.75 ± 0.83	2.79 ± 0.98	2.72 ± 1.01	3.89 ± 0.69
Detail quality (↑)	3.66 ± 0.88	2.70 ± 0.90	4.03 ± 0.76	3.71 ± 0.99	3.30 ± 0.75	4.42 ± 0.60
Context understanding (↑)	3.29 ± 1.33	2.78 ± 1.29	4.40 ± 0.61	3.23 ± 1.18	3.22 ± 1.16	4.46 ± 0.54
Hallucination (↓)	2.04 ± 1.32	2.71 ± 1.61	1.72 ± 0.76	1.75 ± 1.06	1.94 ± 1.25	1.53 ± 0.51

5.3. Case Studies

Figure 3 provides a qualitative comparison of interpretability across three vision-language models when prompted to analyze a 12-lead ECG and select the correct abnormality type with an explanation. The ECG case corresponds to a Right Bundle Branch Block (RBBB), providing a benchmark for evaluating each model's reasoning fidelity and clinical awareness. Despite the visual richness of ECG data, the interpretability results reveal important differences in model behavior. PULSE Figure 3 (c) misclassifies the ECG as Left Bundle Branch Block (LBBB). Although its explanation highlights prolonged QRS duration and a QS complex in lead V2, which are partially relevant, it overlooks hallmark features of RBBB, such as a terminal R' wave in V1 and wide S waves in leads I and V6. This demonstrates limited

generalization in the pretrained model and suggests that it captures superficial waveform cues without robust clinical grounding.

In contrast, the fine-tuned PULSE+LoRA model shown in Figure 3 (d) correctly identifies RBBB and generates a clinically sound explanation, referencing a **wide S wave in lead I** and *a wide R wave in lead V1*, both diagnostic hallmarks of RBBB. It also correctly cites **prolonged QRS duration**, offering a more complete and accurate clinical interpretation. This highlights the efficacy of lightweight fine-tuning in grounding model outputs in domain-specific knowledge and improving both classification accuracy and explanation quality. Surprisingly, GPT-4o Figure 3 (e), a general-purpose multimodal LLM, also misclassifies the case as LBBB. While its explanation is structurally coherent and references classic signs of LBBB (e.g., notched R waves in V5 and V6), these features do not apply to the given ECG. This suggests that GPT-4o may be relying on memorized text patterns or visual heuristics, rather than integrating the prompt with accurate visual signal interpretation. The model's high performance on simpler, single-lead tasks like MIT-BIH may reflect this shallow visual-text alignment, which becomes a liability in more complex, multi-lead cases requiring fine-grained spatiotemporal reasoning.

Overall, these results emphasize that interpretation quality matters as much as classification accuracy, especially in clinical settings where trust and transparency are critical. Fine-tuning with domain data not only improves predictive performance but also enhances the interpretability of model decisions.

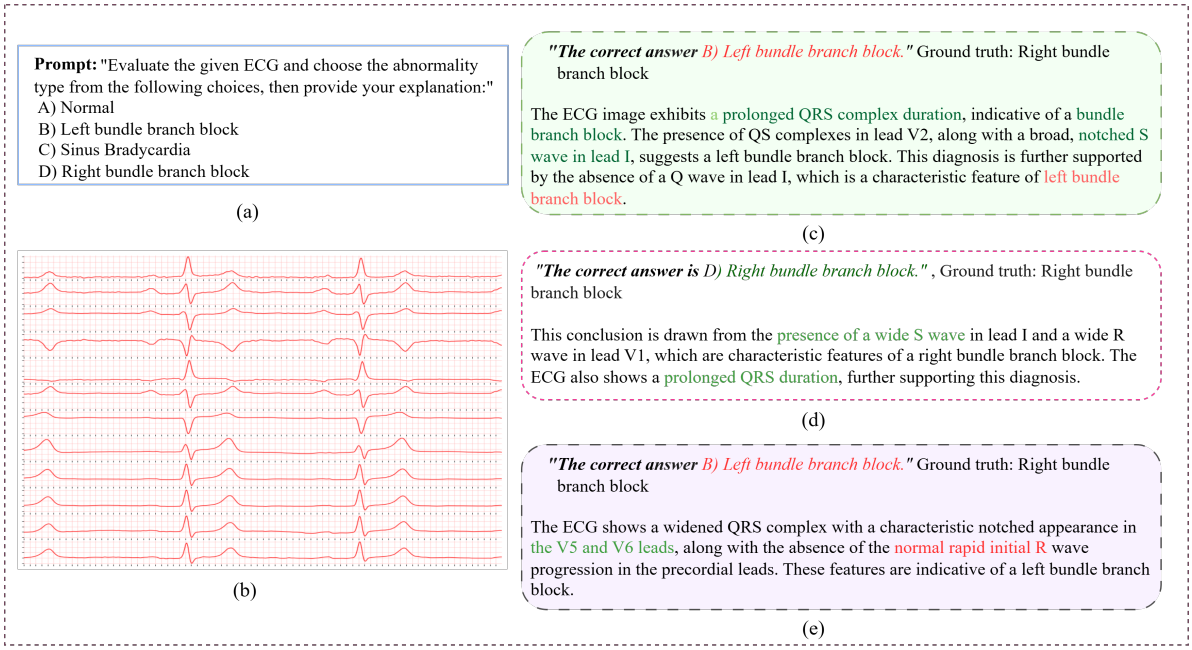


Figure 3: Illustration of the interpreting ability for interpretation: (a) Prompt for inference; (b) input ECG and prompt; (c) PULSE; (d) Finetuned version; (e) GPT-4o.

6. Conclusion

In this study, we investigated the potential of multimodal large language models for ECG interpretation. We test three state-of-the-art VLM models available for ECG interpretation on three baseline public datasets. Our findings underscore the promise of VLMs in enhancing diagnostic and signal interpretation. The primary contribution of this work is not to present a clinically viable tool, but rather to rigorously benchmark the zero-shot and lightly-tuned capabilities of these models on a complex medical task for which they were not specifically designed. The low scores are a significant finding in themselves, highlighting a critical performance gap and the unique challenges of ECG data that VLMs must overcome. Furthermore, the observed model fragility across datasets emphasizes the importance of robust validation across diverse and representative clinical benchmarks.

Future work should focus on designing VLMs that are specifically tailored to the unique characteristics of ECG signals, including the integration of mechanisms capable of effectively capturing temporal dynamics. This involves not only exploring novel LLM architectures but also developing specialized, efficient time-series encoders suited for ECG data. Critically, establishing the clinical utility, safety, and reliability of these models through rigorous, real-world validation is essential before their integration into healthcare practice. While current benchmarks and experimental results are encouraging, the true impact of VLMs on ECG interpretation will depend on comprehensive clinical evaluation and regulatory acceptance. Although the evaluation of explanations using GPT-4 provided a scalable and consistent metric, it represents a form of AI self-assessment that may inherit biases and is not a substitute for clinical expertise

Acknowledgments

This work was supported by the Age-It project, which is part of the National Recovery and Resilience Plan (PNRR) program funded by NextGenerationEU.

Declaration of Generative AI Usage

During the preparation of this work, the authors used ChatGPT and Grammarly to perform grammar and spell checks, as well as to paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Seki, et al., Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ecg image interpretation, medRxiv (2024). URL: https://doi.org/10.1101/2024.03.19.24304442. doi:10.1101/2024.03.19.24304442, preprint.

[2] G. Quer, E. J. Topol, The potential for large language models to transform cardiovascular medicine, The Lancet Digital Health 6 (2024) e767–e771. URL: https://doi.org/10.1016/S2589-7500(24)00151-1. doi:10.1016/S2589-7500(24)00151-1.

[3] Z. Wan, C. Liu, X. Wang, C. Tao, H. Shen, Z. Peng, J. Fu, R. Arcucci, H. Yao, M. Zhang, Meit: Multi-modal electrocardiogram instruction tuning on large language models for report generation, 2024. URL: https://arxiv.org/abs/2403.04945. arXiv:2403.04945.

[4] Y. Zhao, J. Kang, T. Zhang, P. Han, T. Chen, Ecg-chat: A large ecg-language model for cardiac disease diagnosis, 2025. URL: https://arxiv.org/abs/2408.08849. arXiv:2408.08849.

[5] R. Liu, Y. Bai, X. Yue, P. Zhang, Teach multimodal llms to comprehend electrocardiographic images, 2024. URL: https://arxiv.org/abs/2410.19008. arXiv:2410.19008.

[6] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, 2023. URL: https://arxiv.org/abs/2304.08485. arXiv:2304.08485.

[7] H. Liu, H. Kamarthi, Z. Zhao, S. Xu, S. Wang, Q. Wen, T. Hartvigsen, F. Wang, B. A. Prakash, How can time series analysis benefit from multiple modalities? a survey and outlook, 2025. URL: https://arxiv.org/abs/2503.11835. arXiv:2503.11835.

[8] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, Minigpt-4: Enhancing vision-language understanding with advanced large language models, arXiv preprint arXiv:2304.10592 (2023).

[9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, arXiv preprint arXiv:2305.06500 (2023).

[10] K. Yang, M. Hong, J. Zhang, Y. Luo, S. Zhao, O. Zhang, X. Yu, J. Zhou, L. Yang, P. Zhang, M. Qiao, Z. Nie, Ecg-lm: Understanding electrocardiogram with a large language model, Health Data Science 5 (2025) 0221. doi:10.34133/hds.0221.

[11] G. B. Moody, R. G. Mark, The impact of the mit-bih arrhythmia database, IEEE Engineering in Medicine and Biology Magazine 20 (2001) 45–50. doi:10.1109/51.932724.

[12] F. F. Liu, C. Y. Liu, L. N. Zhao, X. Y. Zhang, X. L. Wu, X. Y. Xu, Y. L. Liu, C. Y. Ma, S. S. Wei, Z. Q. He, J. Q. Li, N. Y. Kwee, An open access database for evaluating the algorithms of ecg rhythm and morphology abnormal detection, Journal of Medical Imaging and Health Informatics 8 (2018) 1368–1373. *C. Y. Liu is the corresponding author.

[13] J. Zheng, H. Guo, H. Chu, A large-scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0), *PhysioNet*, 2022. URL: https://doi.org/10.13026/wgex-er52. doi:10.13026/wgex-er52.

[14] A. Luthra, ECG Made Easy, Jaypee Brothers Medical Publishers, 2019.

[15] G. Silva, P. Silva, G. Moreira, V. Freitas, J. Gertrudes, E. Luz, A systematic review of ecg arrhythmia classification: Adherence to standards, fair evaluation, and embedded feasibility, 2025. URL: https://arxiv.org/abs/2503.07276. arXiv:2503.07276.

[16] J. He, P. Li, G. Liu, S. Zhong, Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding, 2024. URL: https://arxiv.org/abs/2410.23822. arXiv:2410.23822.