NeuroECG-RAG: A Biomimetic Decoupled Retrieval-Augmented Framework for ECG Interpretation

Wen Li, Ziyu Liu, Hangzhou Wang, Ziyang He, and Yinling Wang(✉)

Zhengzhou University, Zhengzhou 450001, China
brut819@qq.com

Abstract. Electrocardiography (ECG) is essential in clinical diagnosis, but existing ECG interpretation methods still face challenges such as medical hallucinations, limited knowledge updating, and weak evidence-grounded reasoning. To address these issues, we propose NeuroECG-RAG, a biomimetic decoupled retrieval-augmented framework for intelligent ECG interpretation. It follows a perception–retrieval–reasoning pipeline: an ECG visual expert first converts raw ECG images into structured waveform descriptions; then, a biomimetic retrieval module performs multi-hop evidence retrieval over an ECG-oriented knowledge graph built from Wikipedia and The ECG Made Easy; finally, Qwen3.5 generates diagnoses and clinical interpretations under explicit evidence constraints. Experiments on the ECG-Bench core subset show that NeuroECG-RAG achieves the best overall performance, with clear advantages in clinical report generation and multi-turn reasoning while remaining competitive in classification tasks. Overall, this work provides a more reliable and interpretable solution for intelligent ECG analysis.

Keywords: Electrocardiography, Multimodal Large Language Models, Retrieval-Augmented Generation, Biomimetic Memory Retrieval, Clinical Reasoning.

1 Introduction

Electrocardiography (ECG) is a non-invasive and low-cost diagnostic tool widely used in clinical practice, especially for detecting cardiac rhythm abnormalities. Unlike natural images, ECG interpretation depends on fine-grained waveform features, such as P waves, QRS complexes, ST-T changes, rhythm regularity, and cross-lead relationships. Therefore, intelligent ECG interpretation is not only a visual recognition task, but also a knowledge-intensive decision-making process that requires medical criteria, diagnostic thresholds, and clinical reasoning.

Existing ECG interpretation has been extensively studied with deep learning methods, including signal-based and image-based diagnostic models [1,2]. These methods can be broadly divided into two categories. The first category is end-to-end methods, which directly learn the mapping from ECG input to diagnostic output within a single parametric model [3,4]. Recent multimodal large language models (MLLMs) have advanced cross-modal alignment, instruction following, and long-range reasoning, enabling ECG interpretation to move beyond fixed-label prediction toward open-ended tasks such as abnormality question answering, evidence explanation, and report generation. ECG-specialized models, including PULSE [5], MEIT [6], ECG-R1 [7], and GEM [8], further improve waveform understanding by adapting MLLMs to ECG-specific signal and image representations.

However, these methods still rely mainly on parametric knowledge and may suffer from insufficient knowledge integration and medical hallucinations in complex reasoning or knowledge-updating scenarios.

The second category is decoupled methods, which separate waveform perception, evidence retrieval, and language generation into different modules. This design reduces error propagation and improves interpretability, knowledge updating, and hallucination suppression. In multimodal research, BLIP-2 [9] connects a frozen vision encoder with a language model through a lightweight bridge, while ViperGPT [10] decouples visual processing from reasoning execution. In the ECG domain, Q-Heart [3] uses historical clinical report retrieval and dynamic prompting to improve ECG question answering.

However, existing decoupled ECG methods still lack a systematic mechanism for converting visual perception outputs into structured queries and connecting them with external knowledge. Retrieval-augmented generation (RAG) [11,12] provides a useful starting point by introducing external evidence into the generation process to improve factual consistency and reasoning. Several RAG extensions, such as IRCoT [13], Self-RAG [14], RAPTOR [15], Adaptive-RAG [16], and HippoRAG [17], have improved evidence utilization and multi-hop reasoning. Nevertheless, most RAG methods are designed for text-query–text-knowledge settings and rely on flat vector retrieval, which is not directly suitable for ECG interpretation, where evidence comes from waveform images and diagnosis often requires multi-hop clinical knowledge integration.

To address this challenge, we propose NeuroECG-RAG, a biomimetic decoupled retrieval-augmented reasoning framework for intelligent ECG interpretation. It follows a clinically inspired workflow: first inspect the waveform, then retrieve diagnostic criteria, and finally generate the diagnosis. Specifically, a visual expert converts ECG images into structured medical descriptions; an open knowledge graph and biomimetic memory retrieval module perform multi-hop evidence retrieval; and Qwen3.5 generates the final diagnosis under the constraints of visual descriptions, retrieved evidence, and task instructions. Compared with existing ECG-RAG methods that mainly rely on handcrafted prompts and flat text retrieval, NeuroECG-RAG integrates ECG-specific visual understanding with structured retrieval-augmented reasoning, supporting more reliable ECG interpretation, report generation, and multi-turn clinical reasoning.

The main contributions of this work are as follows:

1. **A biomimetic decoupled framework for ECG interpretation:** We propose NeuroECG-RAG, which coordinates an ECG visual expert, the Bio-RAG (Artificial Hippocampus), and a reasoning core (Artificial Neocortex) to form a closed-loop pipeline of waveform perception, criterion retrieval, and diagnosis generation.

2. **A large-scale structured ECG knowledge graph for multi-hop evidence retrieval:** We construct a large-scale ECG-oriented open knowledge graph based on Wikipedia and The ECG Made Easy [18], containing phrase nodes, passage nodes, relation

edges, synonym edges, and contextual edges, thereby supporting multi-hop associative retrieval and context-preserving evidence grounding.

3. **A retrieval-augmented reasoning pipeline for grounded diagnosis generation:** We integrate structured visual descriptions with graph-based retrieval and evidence-constrained generation, allowing the model to produce diagnoses and explanations that are more explicitly supported by external clinical knowledge.

4. **Comprehensive evaluation on ECG-Bench core tasks:** Experimental results on the core subset of ECG-Bench show that NeuroECG-RAG achieves strong overall performance, with particular advantages in report generation and multi-turn clinical reasoning, while maintaining competitive classification capability.

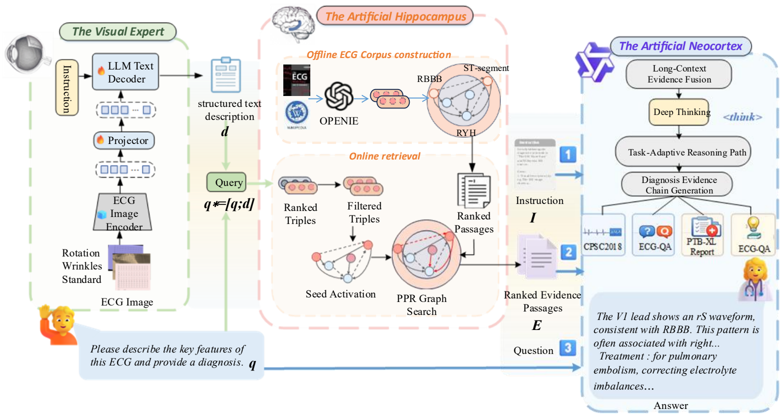

Fig. 1. Framework of NeuroECG-RAG.

2 Method

2.1 Structured Textual Description of ECG Images via the Visual Expert.

As shown in **Fig. 1**, the Visual Expert is built upon ECGInstruct-driven multimodal instruction tuning [5], covering waveform recognition, rhythm analysis, morphological assessment, and report generation. Since this paradigm has shown stronger fine-grained ECG perception and cross-domain generalization than general-purpose multimodal models on ECG-Bench, we adopt it as the front-end perception module.

Instead of reconstructing ECG images into one-dimensional signals before text generation [19], we directly use an ECG-pretrained multimodal model to generate structured natural language descriptions from raw ECG images. This avoids the amplification of grid artifacts, sampling distortion, and blurring noise often introduced during signal reconstruction. It also makes the generated descriptions more semantically aligned with triples and entity relations in the knowledge graph.

For each ECG image, the Visual Expert produces a structured description covering rhythm regularity, P-wave visibility and morphology, PR interval, QRS abnormalities, ST–T changes across leads, and potential pacing or conduction-related findings. These observations serve as objective waveform evidence and are used to form the joint query for subsequent retrieval and reasoning.

2.2 ECG Knowledge Corpus and Knowledge Graph Construction

We construct an offline indexing mechanism that integrates an open knowledge graph, passage-level representations, and online LLM-based filtering. This structure organizes ECG knowledge while preserving both conceptual relations and contextual information.

The knowledge sources include Wikipedia entries related to ECG diagnostic labels, arrhythmias, conduction blocks, and ST–T abnormalities, as well as diagnostic descriptions from The ECG Made Easy [18]. They cover rhythm identification, waveform morphology, lead distribution, and diagnostic criteria. In total, the corpus contains 4,043 passages.

During offline indexing, the corpus is first segmented into semantically complete passages. GPT-4o-mini is then used for Open Information Extraction (OpenIE), producing schema-free triples. The subjects and objects of triples are treated as phrase nodes, predicates as relation edges, and the original passages as passage nodes. This enables joint modeling of conceptual knowledge and source context.

To improve connectivity among different ECG expressions, we compute embedding similarities between phrase nodes using NV-Embed-v2 and add synonym edges. For example, “absence of P wave,” “no visible P wave,” and “P wave not discernible” refer to the same diagnostic cue, while “QRS widening,” “prolonged QRS duration,” and “QRS ≥ 120 ms” may indicate the same abnormal pattern.

The final knowledge graph Gcontains phrase nodes, passage nodes, relation edges, synonym edges, and contextual edges. It supports multi-hop associative reasoning along the chain of concept → feature → diagnostic criterion → source passage.

2.3 Biomimetic Memory-Driven Online Associative Retrieval

In the online retrieval stage, the user query is concatenated with the structured description generated by the Visual Expert to form a joint query q^*. This integrates visual observations with task intent. The retrieval process includes five steps: query-to-triple retrieval, recognition memory, seed assignment, Personalized PageRank propagation, and evidence ranking.

First, NV-Embed-v2 computes the semantic similarity between the joint query and triples in the graph, returning candidate triples as the starting points for graph propagation:

$$s_t = \cos(M(q^*), M(t)), \quad t \in T \tag{1}$$

$$T_{top} = \mathrm{TopM}_{t \in T}\, s_t \tag{2}$$

where $M(\cdot)$ denotes the embedding encoder, $t \in T$ represents candidate triples in the knowledge graph, s_t is the semantic similarity between the joint query and triple t, and $\mathrm{TopM}\ (\cdot)$ returns the top M triples.

Next, GPT-4o-mini filters the candidate triples and retains only those truly relevant to the current query. This step reduces semantic noise from dense retrieval, which is important for ECG interpretation because one feature, such as QRS widening, may correspond to multiple diagnostic directions, including bundle branch block, ventricular rhythm, or pacing.

$$\rho(t|q^*)=p_\psi(\text{relevant}|q^*,t) \tag{3}$$

$$T'=\{t\in T_{top} \mid \rho(t|q^*)>\tau_r\} \tag{4}$$

where $\rho(t \mid q^*)$ denotes the estimated relevance probability of triple t, $p_\psi(\cdot)$ is the filtering model, τ_r is the relevance threshold, and T' is the filtered triple set.

After filtering, restart probabilities are assigned to phrase nodes and passage nodes. The passage node weight scaling factor is set to 0.05, balancing concept-level propagation and contextual preservation. Phrase node weights are computed as follows:

$$T'_v=\{\ t\in T' | v\in t\ \} \tag{5}$$

$$w_v=\frac{1}{|T'_v|}\sum\nolimits_{t\in T'_v} s_t\ s_v \tag{6}$$

Here, T'_v denotes the filtered triples containing phrase node v, s_v is the node-specific weight, and w_v is the initial seed score of the phrase node.

To preserve contextual information from the original passages, we further assign weighted seed scores to passage nodes and construct a personalized restart distribution accordingly:

$$w_p=\beta\cdot \cos(M(q^*),M(p)), \quad p\in V_{pas} \tag{7}$$

$$r_u=\frac{w_u}{\sum_{v\in V_{seed}} w_v} \tag{8}$$

where $p \in V_{\text{pas}}$ denotes a passage node, β is the scaling factor for passage node weights, V_{seed} is the set of all seed nodes, and r_u is the normalized restart probability of node u.

The system then performs Personalized PageRank propagation over the graph:

$$\pi=\alpha W^{T}\pi+(1-\alpha)r \tag{9}$$

where W is the transition matrix, r is the personalized restart distribution, $\alpha = 0.5$ is the damping factor, and π is the final stationary node distribution. Compared with flat vector retrieval, PPR enables multi-hop propagation along graph edges and activates evidence that may not be directly matched but is clinically relevant. Finally, passages are ranked by propagated scores, and the top-k passages are selected as the final evidence set E.

2.4 Evidence-Constrained Diagnosis Generation with Qwen3.5

In the final stage, Qwen3.5-27B is used as the text generation backbone for diagnosis generation and clinical interpretation under explicit evidence constraints. We select this model for its strong long-context modeling capability [20], which allows it to integrate structured waveform descriptions, retrieved evidence passages, and task instructions in retrieval-augmented settings.

Since this stage no longer requires direct image processing, Qwen3.5-27B is deployed in a text-only manner. This design avoids reintroducing a generic visual pathway and focuses computation on reasoning over retrieval-augmented textual inputs.

The input to Qwen3.5-27B consists of system instructions, structured waveform descriptions, retrieved knowledge passages, and the user query. The system prompt requires the model to prioritize the visual expert's descriptions and the retrieved evidence. If the evidence is insufficient, conflicting, or unable to support a confident conclusion, the model is instructed to provide a conservative response rather than unsupported speculation.

Through this design, Qwen3.5-27B generates diagnoses within the semantic boundary defined by waveform descriptions and retrieved evidence. As shown in **Fig. 2**, the output is no longer unconstrained open-ended generation, but evidence-grounded diagnosis generation, reducing the risk of hallucination.

```
Qwen-3.5 Evidence-Constrained Diagnosis Guider
You are a board-certified cardiologist , strictly following the diagnostic protocols in "The ECG Made Easy" and Wikipedia ECG entries
Given:
1. Visual Description (d).
2. Retrieved Evidence (E): e.g.,
   E1: Left bundle branch block is defined as QRS duration ≥ 120 ms... (The ECG Made Easy, Chapter on Bundle Branch Blocks).
   E2: ST-segment depression that is concordant with the QRS direction in LBBB may indicate underlying ischemia or strain …
3. User Query / Task Instruction (q).
Requirements:
- Answer EXCLUSIVELY based on (d) and (E). NEVER add, infer, or hallucinate any external knowledge.
- Explicitly cite evidence by paragraph number .
- Link every diagnostic conclusion to specific waveform features in (d) and corresponding medical criteria in (E).
- If evidence is insufficient for a definitive diagnosis , output "Insufficient evidence for definitive diagnosis"'.
- Use precise, professional medical terminology . Maintain clinician-style structure and transparency.
- Provide a clear reasoning chain: waveform observation → evidence citation → diagnostic conclusion.

#CPSC2018*Perform multi-label abnormality detection . Output format: Abnormalities: [comma-separated standard labels from ECGBench]
;Confidence: [macro-AUC/F1 aligned] ;Evidence-based Reasoning: [1-2 sentence chain]

# ECG-QA*Answer the question with explicit waveform evidence Output format: Answer: [direct answer] ;Key Evidence: [cited
paragraphs] ;Reasoning Chain: [step-by-step link from d to E]

#PTB-XL Report *Generate a complete clinical ECG report in the style of PTB-XL physician-annotated reports . Structure: 1.Rhythm
Analysis;2.Waveform Morphology Description (P/QRS/ST-T per lead);3.Pathological Diagnosis;4.Clinical Impression &Recommendation
Cite evidence paragraphs for each section

#ECG ARENA *Engage in multi-turn clinical dialogue. Maintain full context consistency across turns . In each response: Restate key evidence
from previous turns when referenced; Provide step-by-step clinical reasoning ; End with actionable clinical recommendation if appropriate
```

Fig. 2. ECG Evidence-Guided Prompt Design of Qwen3.5 for Different Tasks.

Overall, Qwen3.5 serves as the reasoning core rather than a general multimodal model. The Visual Expert converts ECG images into structured semantic representations, BioRAG retrieves relevant medical evidence, and Qwen3.5 performs diagnosis determination, conflict resolution, and natural language explanation under evidence constraints. The final generation objective is formulated as:

$$y^* = \arg\max_{y} \log P_\theta (y|q,d,E) \tag{10}$$

where $P_\theta(\cdot)$ denotes the conditional generation probability of Qwen3.5, y denotes a candidate answer, and y^* represents the optimal output conditioned on the visual description d, the retrieved evidence E, and the user query q.

3 Experiments

3.1 Benchmark Datasets

To reduce computational cost and avoid redundant evaluation bias from using the full ECGBench, we evaluate NeuroECG-RAG on a compact core subset covering four representative ECG tasks. The subset includes CPSC2018, ECG-QA, PTB-XL Report, and ECG Arena, with 2,061, 1,317, 500, and 50 samples, respectively.

CPSC2018 is used for multi-label abnormality detection across nine diagnostic categories and evaluates cross-domain diagnostic generalization. ECG-QA contains ECG image-based question–answer pairs with 70 question templates, assessing basic ECG understanding and abnormality recognition. PTB-XL Report includes 500 physician-annotated reports and evaluates long-form clinical report generation, focusing on diagnostic accuracy, waveform description, and structural completeness. ECG Arena consists of 50 manually curated multi-turn clinical dialogues, designed to assess contextual consistency, clinical reasoning, and multi-step decision-making in open-ended ECG interpretation.

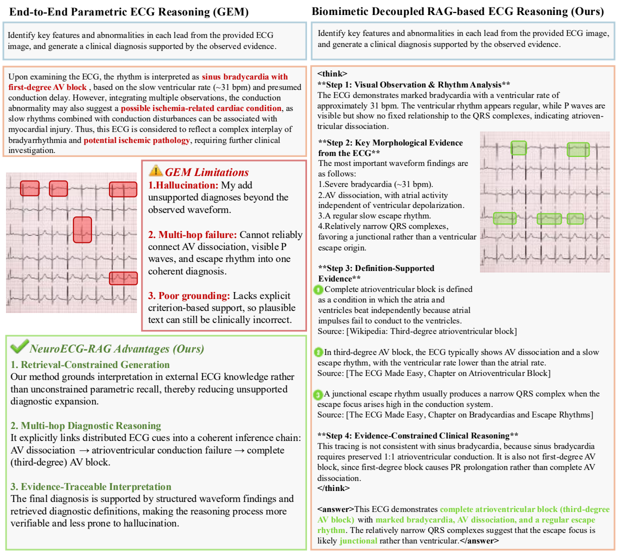

Fig. 3. Example comparison between NeuroECG-RAG and GEM in ECG report generation.

3.2 Evaluation Metrics

For abnormality detection and understanding tasks, we use macro-AUC, macro-F1, and Hamming Loss for CPSC2018, and accuracy for ECG-QA. For generation tasks, GPT-

4o is used to evaluate PTB-XL Report outputs against physician-annotated references from three aspects: diagnostic accuracy, waveform description, and rhythm analysis. Each aspect is scored out of 10, and the averaged score is scaled to 100. For ECG Arena, GPT-4o evaluates multi-turn responses in terms of accuracy, completeness, and instruction adherence, with the final score also scaled to 100.

3.3 Experimental Settings

Experiments were conducted on a server equipped with a server equipped with 8 NVIDIA GeForce RTX 4090 GPUs. The visual perception module employs PULSE [5] to generate structured ECG descriptions; during the graph retrieval stage, the GPT-4o-mini API is used for online triple filtering, while NV-Embed-v2 is used for dense retrieval and similarity computation; the reasoning stage adopts Qwen3.5-27B as the text generation backbone.

In the offline graph construction stage, the synonym edge threshold is set to 0.8; the damping factor of personalized PageRank is set to 0.5; the passage node weight scaling factor is set to 0.05; and 5 phrase nodes are retained as seed nodes during graph propagation initialization.

For decoding strategies, CPSC2018 and ECG-QA adopt the non-thinking/instruct mode with parameters set to temperature = 0.7, top_p = 0.8, and top_k = 20; PTB-XL Report and ECG Arena adopt the thinking mode with parameters set to temperature = 1.0, top_p = 0.95, and top_k = 20. All other parameters remain at their default settings.

3.4 Experimental Results

Table 1 reports the results on the four core tasks. NeuroECG-RAG achieves the best overall performance in knowledge-intensive settings, obtaining the highest scores on PTB-XL Report and ECG Arena, with 67.5 and 42.8, respectively. It also achieves the highest AUC on CPSC2018, reaching 79.2.

On ECG-QA, NeuroECG-RAG obtains an accuracy of 73.4, which is close to PULSE and GEM, with 73.8 and 73.6, respectively.

Compared with general-purpose MLLMs, NeuroECG-RAG shows clear advantages on ECG tasks, indicating that structured ECG perception and evidence-based retrieval are more suitable for fine-grained waveform interpretation.

As shown in **Fig. 3**, NeuroECG-RAG provides a more evidence-grounded and clinically coherent ECG interpretation than GEM, especially in linking waveform observations to diagnostic conclusions.

As shown in **Fig. 4**, compared with ECG-specialized models such as PULSE and GEM, our method performs slightly lower on ECG-QA accuracy and CPSC2018 F1/HL, but achieves better results in report generation and multi-turn reasoning. This suggests that the decoupled RAG framework improves interpretability and complex clinical reasoning, while retaining competitive classification performance.

Table 1. Experimental Results on Four Core Tasks.

Category	Datasets	CPSC 2018			ECG-QA	PTB-XL Report	ECG Arena
	Metric	AUC	FI	HL	Accuracy	Report Score	Arena Score
	Random	51.2	15.1	28.8	16.2	0	0
Proprietary MLLMs	GPT-4o	50.9	10.6	18.2	35.2	50.2	33.5
	GPT-4o-mini	49.2	11.0	25.5	14.9	37.1	30.1
	Claude 3.5 Sonnet	52.8	11.5	18.9	34.2	43.7	37.1
Open-source MLLMs	LLaVA-One-Vision-72B	51.5	12.8	<u>29.4</u>	25.0	40.6	15.5
	Deepseek-VL-Chat-7B	50.7	6.0	20.0	21.1	15.6	15.3
	Mantis-8B-siglip-Llama3	51.3	19.1	48.5	23.8	16.0	13.6
ECG-specialized MLLMs	PULSE	76.9	57.6	<u>8.6</u>	**73.8**	61.3	38.9
	GEM	<u>79.1</u>	**61.1**	**8.1**	<u>73.6</u>	<u>67.1</u>	<u>40.0</u>
Ours	NeuroECG-RAG	**79.2**	<u>59.8</u>	8.9	73.4	**67.5**	**42.8**

3.5 Ablation Experiments

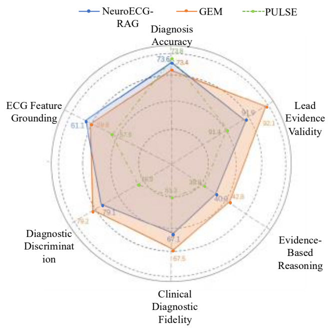

Fig. 4. Comparison of Key Capability Dimensions among NeuroECG-RAG, GEM, and PULSE.

To examine the contribution of each core component, we design five ablation variants, as shown in **Table 2**. Removing the visual expert and using only Bio-RAG with

Qwen3.5 results in a weak text-only baseline, since the system can only retrieve evidence from the user query and cannot reliably activate ECG-specific diagnostic nodes.

The visual Expert + Bio-RAG variant mainly reflects the capability of the visual expert itself. It performs well on basic abnormality recognition, achieving 57.6, 73.8, 61.3, and 38.9 on CPSC2018, ECG-QA, PTB-XL Report, and ECG Arena, respectively, but remains limited in report generation and multi-hop reasoning due to the lack of full evidence integration.

Adding Qwen3.5 to the visual expert improves PTB-XL Report and ECG Arena scores to 62.2 and 39.0, suggesting that Qwen3.5 helps organize long-form explanations. However, performance slightly drops on CPSC2018 and ECG-QA, indicating that additional abstraction may lose fine-grained threshold-sensitive ECG details.

When standard flat vector retrieval is added, the model further improves on PTB-XL Report and ECG Arena, reaching 63.4 and 41.1, but decreases on CPSC2018 and changes little on ECG-QA. This suggests that standard RAG provides useful textual evidence for generation and reasoning, but may introduce noise for label prediction tasks that depend on precise waveform features.

The full model, Visual Expert + Bio-RAG + Qwen3.5, achieves the most balanced performance, with scores of 59.8, 73.4, 67.5, and 42.8 across the four tasks. Compared with standard RAG, its gains come from structured association and multi-hop propagation rather than retrieval alone: the visual expert provides medical anchors, Bio-RAG organizes evidence chains through triple filtering and Personalized PageRank, and Qwen3.5 generates diagnoses under explicit evidence constraints.

Table 2. Ablation Study Results.

Class	CPSC 2018	ECG-QA	PTB-XL Report	ECG Arena
Bio-RAG+Qwen3.5	50.0	64.22	0	0
Visual Expert+Bio-RAG	57.6	73.8	61.3	38.9
Visual Expert+Qwen3.5	57.2	72.9	62.2	39.0
Visual Expert+RAG+Qwen3.5	56.8	73.1	63.4	41.1
Visual Expert+Bio-RAG+Qwen3.5	**59.8**	**73.4**	**67.5**	**42.8**

4 Conclusion

In this study, we propose NeuroECG-RAG, a biomimetic decoupled retrieval-augmented framework for intelligent ECG interpretation. It decomposes ECG understanding into structured perception, knowledge retrieval, and evidence-constrained reasoning. A visual expert first extracts fine-grained waveform features from ECG images, which are then linked to an open knowledge graph for multi-hop evidence retrieval.

The retrieved evidence is finally used to generate diagnoses under explicit constraints. Experiments show that NeuroECG-RAG performs well in report generation, multi-turn question answering, and complex clinical reasoning, while reducing hallucination risks. Overall, this work offers a more interpretable and scalable solution for intelligent ECG interpretation.

References

1. Petmezas, G., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J.A., Katsaggelos, A.K., Maglaveras, N.: State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review. JMIR Medical Informatics 10(8), e38454 (2022)
2. Sangha, V., Mortazavi, B.J., Haimovich, A.D., Ribeiro, A.H., Brandt, C.A., Jacoby, D.L., Schulz, W.L., Krumholz, H.M., Ribeiro, A.L.P., Khera, R.: Automated Multilabel Diagnosis on Electrocardiographic Images and Signals. Nature Communications 13(1), 1583 (2022)
3. Pham, H.M., Tang, J., Saeed, A., Ma, D.: Q-Heart: ECG Question Answering via Knowledge-Informed Multimodal LLMs. arXiv preprint arXiv:2505.06296 (2025)
4. Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P.S., Andersson, C.R., Macfarlane, P.W., Meira Jr, W., et al.: Automatic Diagnosis of the 12-Lead ECG Using a Deep Neural Network. Nature Communications 11(1), 1760 (2020)
5. Liu, R., Bai, Y., Yue, X., Zhang, P.: Teaching Multimodal LLMs to Comprehend 12-Lead Electrocardiographic Images. npj Digital Medicine (2026)
6. Wan, Z., Liu, C., Wang, X., Tao, C., Shen, H., Xiong, J., Arcucci, R., Yao, H., Zhang, M.: MEIT: Multimodal Electrocardiogram Instruction Tuning on Large Language Models for Report Generation. In: Findings of the Association for Computational Linguistics: ACL (Findings of ACL), pp. 14510–14527 (2025)
7. Jin, J., Wang, H., Wu, X., Fang, X., Lan, X., Wang, Z., Zhang, D., Liu, B., Zhang, Y., Wu, X., et al.: ECG-R1: Protocol-Guided and Modality-Agnostic MLLM for Reliable ECG Interpretation. arXiv preprint arXiv:2602.04279 (2026)
8. Lan, X., Wu, F., He, K., Zhao, Q., Hong, S., Feng, M.: GEM: Empowering MLLM for Grounded ECG Understanding with Time Series and Images. arXiv preprint arXiv:2503.06073 (2025)
9. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In: International Conference on Machine Learning (ICML), pp. 19730–19742 (2023)
10. Surís, D., Menon, S., Vondrick, C.: ViperGPT: Visual Inference via Python Execution for Reasoning. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11888–11898 (2023)
11. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval-Augmented Language Model Pre-Training. In: International Conference on Machine Learning (ICML), pp. 3929–3938 (2020)
12. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 9459–9474 (2020)
13. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In: Annual Meeting of the Association for Computational Linguistics (ACL), pp. 10014–10037 (2023)

14. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In: International Conference on Learning Representations (ICLR) (2024)
15. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D.: RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In: International Conference on Learning Representations (ICLR) (2024)
16. Jeong, S., Baek, J., Cho, S., Hwang, S.J., Park, J.C.: Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 7036–7050 (2024)
17. Gutiérrez, B.J., Shu, Y., Gu, Y., Yasunaga, M., Su, Y.: HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In: Advances in Neural In-formation Processing Systems (NeurIPS), pp. 59532–59569 (2024)
18. Hampton, J., Hampton, J., Adlam, D.: The ECG Made Easy. 10th edn. Elsevier (2024)
19. Yu, H., Guo, P., Sano, A.: Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation. In: Machine Learning for Health (ML4H), pp. 650–663 (2023)
20. Yang, S., Kautz, J., Hatamizadeh, A.: Gated Delta Networks: Improving Mamba2 with Delta Rule. arXiv preprint arXiv:2412.06464 (2024)