REISA: An Electrocardiogram Diagnostic Model Based on Retrieval-Augmented Generation and Instruction-Driven Spatial Addressing

Mingrui Li[1[0009-0008-9137-6439]],Yinling Wang[1*],Hangzhou Wang[1],Ziyang He[1]

[1] School of Cyberspace Security, Zhengzhou University, Zhengzhou, China
yinlingwang@zzu.edu.cn

Abstract. Current ECG diagnostic models effectively identify cardiac pathologies in open domains but struggle to establish granular semantic-morphological rela-tionships. Existing paradigms rely heavily on Large Language Models for di-agnostic synthesis, which carries inherent risks of medical hallucinations, and often lack spatial awareness in multi-lead signal processing for location-sensitive diseases. To address these challenges, we propose REISA, an evi-dence-based ECG diagnostic framework with enhanced spatial perception. REISA integrates Retrieval-Augmented Generation (RAG) to fuse original clinical reports with authoritative morphological criteria, thereby anchoring di-agnostic logic to clinical evidence rather than the internal parametric memory of LLMs. Furthermore, we introduce a Morphological Semantic Dehydration Strategy (MSDS) to reconstruct hybrid reports into high-signal-to-noise ratio morphological descriptors by filtering out non-pathological narrative noise via strict physical constraint protocols. Central to our framework is the Instruction-driven Spatial Addressing (ISSA) mechanism, which utilizes the instruction-following capabilities of text encoders to transform descriptors into dynamic spatial query vectors, enabling precise addressing of lead-specific regions with-in signal feature maps. Systematic cross-domain zero-shot evaluations across multiple clinical datasets demonstrate that REISA effectively corrects the logi-cal biases of conventional models, achieving substantial improvements in di-agnostic accuracy and robustness. Ultimately, REISA provides a robust ap-proach for constructing evidence-based ECG diagnostic systems, demonstrat-ing significant clinical utility for automated ECG diagnosis in diverse practical environments.

Keywords: Electrocardiogram, Retrieval-augmented generation, Morphological semantic dehydration, Instruction-driven spatial addressing.

1 Introduction

1.1 A Subsection Sample

Cardiovascular diseases (CVDs) are the leading cause of death globally and a primary cause of physical disability and reduced quality of life , with their incidence showing a continuous upward trend [1]. As a vital standard for clinical cardiac assessment and early diagnostic screening, electrocardiogram (ECG) interpretation depends heavily on professional expertise and is susceptible to inter-observer variability [2]. Furthermore, in remote or medical resource-deficient areas, the scarcity of professional cardiologists often leads to diagnostic delays, further highlighting the urgent need for robust, automated ECG interpretation systems.

Driven by the development of deep learning, data-driven automated methods have achieved significant breakthroughs in tasks such as arrhythmia recognition[3]. Early studies utilized 1-D convolutional neural networks (1D-CNN), recurrent neural networks (RNN), and their variants to improve classification accuracy, or incorporated entropy features and attention mechanisms to enhance spatio-temporal modeling[4]. To address the issue of high annotation costs, self-supervised learning strategies have also been employed to optimize performance in small-sample scenarios[5]. Nevertheless, traditional deep learning still faces significant limitations: first, these models lack an understanding of medical knowledge, diagnostic logic, and textual semantics, resulting in weak interpretability and a clear divergence from clinical reasoning; second, most models focus exclusively on signal waveform learning and struggle to integrate textual information such as medical history and symptoms, which limits their comprehensive judgment capability for complex ECG abnormalities.

In recent years, the introduction of Large Language Models (LLMs) has propelled a transformation in ECG diagnosis from simple pattern recognition toward knowledge-integrated intelligent reasoning, aligning diagnostic systems more closely with the interpretation logic of clinicians [6]. Nevertheless, LLMs face significant limitations. On one hand, due to the lack of domain-specific professional knowledge constraints, LLMs that rely heavily on general corpora are prone to medical hallucinations, leading to label distortion and degraded generalization performance, which makes it difficult to meet the stringent requirements of actual diagnostic scenarios [7]. On the other hand, since ECG signals possess inherent spatio-temporal attributes—where specific pathological changes are typically anchored to specific lead channels—traditional encoders lack instruction-following capabilities. Consequently, they cannot precisely address lead coordinates within feature maps through the content of text labels, which may result in excessively low diagnostic accuracy for location-sensitive diseases [8]. To address these issues, researchers have developed domain-specific ECG diagnostic models, achieving alignment between waveform features and clinical semantics through large-scale ECG-text paired pre-training. For instance, Yu et al. implemented zero-shot diagnosis by transforming ECG features into prompts, breaking through the limitations of small-sample scenarios [9]. The CVDLLM framework proposed by Qiu et al. enhances reasoning capability and interpretability by fusing semantic embeddings with graph attention features [10]. Furthermore, some studies have begun to introduce Retrieval-Augmented Generation (RAG) methods, embedding authoritative professional knowledge to mitigate the inherent structural deficiencies of LLMs in this field [11].

To address these challenges, this study proposes REISA, an ECG zero-shot diagnostic model based on Retrieval-Augmented Generation (RAG). This model utilizes RAG technology to generate augmented diagnostic reports, introduces a "morphological bottleneck" strategy to filter textual noise, and employs the ISSA mechanism to achieve targeted addressing and feature extraction within 12-lead spatio-temporal feature maps. Our main contributions are summarized as follows:

1. We utilize MedCPT to retrieve information from a knowledge base constructed based on AHA/ACC guidelines [12]. By performing semantic fusion to reconstruct morphological descriptors with clinical certainty, we alleviate the medical hallucinations caused by the over-reliance of LLMs on internal parametric knowledge.

2. By formulating morphological constraint protocols, we reconstruct original reports and evidence-based evidence into high-signal-to-noise ratio (SNR) descriptors, significantly reducing textual redundancy and enhancing the capture precision of the diagnostic model [13] for fine-grained electrophysiological features.

3. Leveraging the instruction-understanding capability of encoders [14], we transform lead position information within the text into proactive query instructions. By intervening in the allocation of cross-attention weights among the 12 channels, we achieve precise observation of specific lead regions, effectively alleviating the defect of vague localization in previous models during multi-lead analysis.

2 Methods

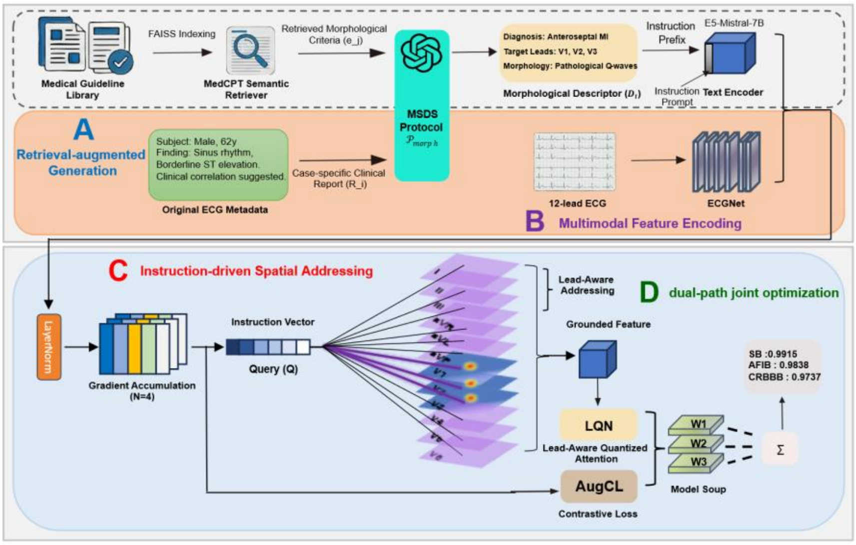

Fig. 1. Overall Architecture of the REISA Framework

An overall overview of the proposed REISA framework is shown in Figure 1, which comprises four core modules: Retrieval-Augmented Generation (RAG) [11], multimodal encoding, spatial-aware alignment, and dual-path joint optimization. First, in the retrieval-augmented generation stage, an evidence-based knowledge base is constructed based on authoritative AHA/ACC/HR guidelines, and the MedCPT retriever is utilized [12] to retrieve morphological criteria corresponding to the labels. Subsequently, the Morphological Semantic Dehydration (MSDS) protocol guides GPT-4 to fuse original reports with the retrieved evidence, utilizing an information bottle-neck strategy [13] to eliminate clinical noise and generate high-signal-to-noise ratio (SNR) morphological descriptors. In the encoding and calibration phase, E5-Mistral-7B [15] and ECGNet transform the descriptors and signals into high-dimensional embeddings and spatio-temporal feature maps, respectively. A LayerNorm module is in-troduced to alleviate the logit collapse problem caused by the asymmetry in encoder scales. Building upon this, the instruction-driven spatial-aware alignment mecha-nism maps descriptors into dynamic query vectors carrying lead constraints, guiding the attention to focus precisely on pathology-related channels within the feature maps, thereby solving the problem of vague localization in multi-lead analysis. Fi-nally, the system combines the global alignment of AugCL with the local multi-label prediction of LQN. Through the collaborative optimization of dual parallel loss functions and an ensemble strategy, the fusion of global semantics and local ad-dressing is achieved to construct the final ECG diagnostic model.

2.1 Diagnostic Description Generation for Original Reports Based on Retrieval-Augmented Generation

To address the medical hallucination and redundancy issues arising from the direct generation of descriptions by LLMs, this study proposes a Retrieval-Augmented Generation (RAG) pipeline [11] constrained by authoritative guidelines and constructs an offline knowledge base retrieval pipeline based on FAISS. First, morphological determination criteria for 105 categories of ECG diagnosis were extracted from authoritative guidelines such as AHA/ACC to construct an expert knowledge base $\mathcal{K}_{\mathrm{ref}}$. Subsequently, each criterion entry e_j was mapped into a 768-dimensional semantic vector using the MedCPT Article Encoder (Φ_{art}) [12], and a vector index based on inner products was constructed using FAISS to achieve efficient retrieval of authoritative knowledge. During the data processing phase, the system extracts the original diagnostic label l_i for each data entry and encodes it into a query vector via the MedCPT Query Encoder (Φ_{que}). The morphological criterion e_i^* with the highest correlation is then retrieved by calculating the cosine similarity. The retrieval operator is defined as:

$$e_i^* = \mathrm{argmax}_{e_j \in \mathcal{K}_{ref}} \left(\frac{\Phi_q(l_i) \cdot \Phi_a(e_j)}{\| \Phi_q(l_i) \| \| \Phi_a(e_j) \|} \right) \tag{1}$$

Where e_i^* represents the retrieved evidence-based medical grounds that best match the current case. Subsequently, e_i^* is concatenated with the original report R_i containing case-specific information to construct a hybrid context $C_i = [R_i \oplus e_i^*]$. Then, guided by the morphological semantic dehydration protocol $\mathcal{P}_{morph}$ [13], the hybrid context C_i is reconstructed using the logical reasoning capability of GPT-4. Redundant noise is eliminated, and non-standard expressions in the report are physically aligned. The final generated morphological descriptor D_i is defined as:

$$D_i = \mathrm{Generator}_{GPT4}(C_i, \mathcal{P}_{morph}) \tag{2}$$

This procedure ensures that the synthesized pre-training text retains the individualized characteristics of the case while adhering to the rigorous standards of medical guidelines.

2.2 Morphological Semantic Dehydration Strategy

Guided by the Information Bottleneck (IB) principle [13], we propose the Morphological Semantic Dehydration (MSDS) strategy to address the issue of gradient scattering and hindered signal feature capture caused by "semantic water" (e.g., narrative noise and treatment suggestions) in clinical reports. To this end, we designed a strict morphological constraint protocol P$morph$. During the RAG generation phase, GPT-4 is guided to act as a knowledge distiller [6] to perform three-dimensional semantic reconstruction: (1) Identifying colloquial findings in the original report R_i and correcting them into standardized morphological terminology with reference to the retrieved authoritative criteria e_i^* (e.g., correcting "extremely irregular heartbeat" to "Irregularly irregular R-R intervals," and "low voltage" to "QRS amplitude < 0.5mV in limb leads"); (2) Eliminating all non-electrophysiological descriptions such as clinical suggestions and diagnostic conjectures to ensure that every term can be directly mapped to the physical features of ECG signals, thereby enhancing the certainty of the supervision signals; and (3) Explicitly labeling the leads where pathological features are located in the generated D_i (e.g., "Focus on Leads V1-V3"), transforming static diagnostic conclusions into dynamic addressing instructions to provide a high-resolution query operator [14] for the subsequent Instruction-driven Spatial Addressing (ISSA) module in Section 3.4.

This strategy operates during the generation phase of diagnostic descriptions for pre-training data based on Retrieval-Augmented Generation. We utilize the retrieved original evidence C_i as the input context for GPT-4, directing the model to perform controlled generation operations under the strict morphological constraint protocol P$morph$.

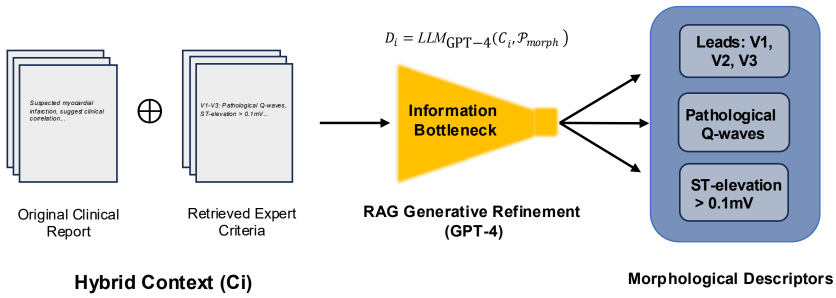

Fig. 2. Information Bottleneck Mechanism in the Morphological Semantic Dehydration Strategy (MSDS)

MSDS reconciles the academic conflict between "standardization" and "individualization" [5]. Relying solely on knowledge bases would render the text modality for similar diseases too monotonous, thus losing the diversity of individual waveforms; conversely, the exclusive use of original reports introduces significant noise. MSDS-guided RAG generation ensures that the morphological descriptors D_i encompass both the "medical invariants" from guidelines and case-specific pathological details (e.g., specific values of ST-segment elevation), thereby enhancing the model's sensitivity to subtle pathological distortions (such as slight LPR interval prolongation) and its cross-domain robustness. Unlike the KED model [9], which aligns with macro-level diagnostic conclusions, REISA aligns signals with pure physical waveform indicators upon which the pathological conclusions depend [8]. This de-contextualization effectively suppresses the semantic shift caused by stylistic variations across different medical institutions, supporting the deployment of automated diagnosis in resource-deficient regions and enhancing the equity and credibility of global computer-aided diagnosis and treatment.

2.3 Instruction-driven Spatial Addressing Mechanism

ECG pathological features are highly lead-specific [4]; however, general-purpose text encoders often lack instruction-following capabilities, treating lead identifiers (e.g., "V1-V3") as isolated semantic symbols rather than spatial constraints [8]. This "correct recognition but shifted localization" phenomenon causes models to perform near-randomly in the zero-shot diagnosis of location-sensitive conditions like LPR. To resolve this, we designed the ISSA mechanism, leveraging the instruction-following E5-Mistral-7B [15] to transform descriptors D_i into dynamic query vectors with spatial guidance capabilities. Unlike static semantic embeddings, this encoder recognizes the instruction attributes of lead identifiers [14], allowing the resulting vector $\mathbf{v}_{inst}$ to carry spatial directional bias alongside pathological semantics, thereby guiding the model to proactively address specific signal channels within the feature maps.

The ISSA mechanism is implemented within the cross-attention framework of the Label Query Network (LQN) [10]. We define the instruction vector v_{inst} as the query (Q), while the spatio-temporal feature map $M_{sig} \in \mathbb{R}^{12\times T\times d}$ output by the signal encoder serves as the key/value pairs (K/V). Learnable matrices W_Q and W_K project both modalities into a unified metric space, where the resulting attention weights α_i reflect the spatial activation degree of the textual instructions across different lead channels. The final grounded feature $z_{grounded}$ is an instruction-filtered representation with high clinical credibility, providing the model with an evidence-based basis for spatial decision-making. The calculation process for the attention weights is as follows:

$$e_i = \frac{(\mathbf{W}_Q\mathbf{v}_{inst})(\mathbf{W}_K\mathbf{M}_{sig,i})^\top}{\sqrt{d}}, i \in \{1, \dots, 12\} \tag{3}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{12}\exp(e_j)} \tag{4}$$

$$\mathbf{z}_{grounded} = \sum_{i=1}^{12}\alpha_i\,\mathbf{v}_i \tag{5}$$

By explicitly embedding lead constraints into the Query vector, this mechanism forces the attention weights α_i to concentrate on the pathology-related lead channels in $\mathbf{M}_{sig}$ (such as the Septal Leads shown in Figure 3). This design achieves a paradigm shift from "average alignment of global features" to "targeted alignment of local evidence." The final generated Grounded Feature $\mathbf{z}_{grounded}$ provides the model with a basis for spatial decision-making with evidence-based support, fundamentally correcting the model's recognition logic for weak pathological signals.

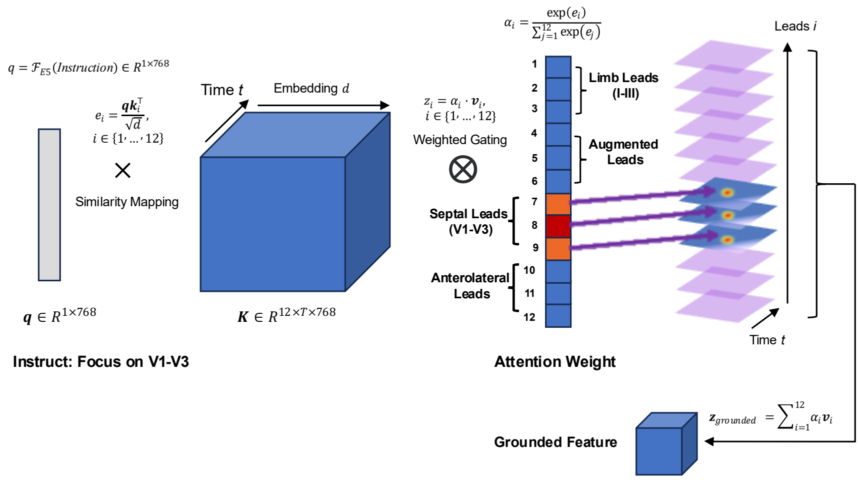

Fig. 3. Mathematical and Geometric Interpretation of the Instruction-driven Spatial Addressing Mech-anism

3 Experiments

3.1 Datasets and Data Preprocessing

To evaluate the robustness of the REISA framework, we conducted experiments on three large-scale ECG datasets covering diverse ethnic groups and environments:

1.**MIMIC-IV-ECG:** Contains 772,155 12-lead samples synchronized with clinical reports, covering 105 diagnostic categories anchored to the knowledge base via MedCPT [12]. This dataset is used for pre-training to establish mapping relationships between signal features and morphological descriptors.

2.**PTB-XL:** Includes 21,837 10-second records following SCP-ECG standards [16], covering rhythm, morphology, and diagnosis. It is used to evaluate the model's diagnostic accuracy and interpretability regarding fine-grained morphological features (e.g., LPR interval).

3.**CPSC2018:** Comprises 6,877 variable-length Asian population records (6–60 seconds) from heterogeneous devices. It is used to assess cross-domain robustness against racial physiological differences and signal interference [17].

This multi-faceted data composition establishes a systematic evaluation framework, designed to comprehensively validate the zero-shot transferability and generalization resilience [3] of the REISA framework when confronted with complex heterogeneous clinical environments, diverse ethnic physiological characteristics, and varying recording protocols.

Table 1. Statistical overview and characteristics of the clinical ECG datasets used for pre-training and evaluation.

Dataset	Records	Leads	Hz	Classes	Region
MIMIC-IV	800000	12	100/500	105	US
PTB-XL	21837	12	100/500	71	Germany
CPSC2018	6877	12	500	9	China

Raw signals were resampled to 100 Hz and truncated via a sliding window (fixed at 10 seconds, 1000 sampling points), followed by lead-wise Z-score normalization to eliminate voltage variations among heterogeneous devices. Model training was conducted on an NVIDIA RTX 4090, utilizing a 4-step gradient accumulation to achieve an equivalent batch size of 1024, with the AdamW optimizer (learning rate: 1×10^{-4}, weight decay: 1×10^{-5}) and a warm-up learning rate scheduler configured for a total of 15 epochs. Performance was evaluated using AUROC, Macro F1, and the Matthews Correlation Coefficient (MCC), with all metrics reported alongside 95% confidence intervals (CIs) estimated through 1000 bootstrap resamples. Finally, decision thresholds were determined through a class-wise Max-MCC search on the validation set.

3.2 Experimental Results

We compared the zero-shot performance of REISA with existing baseline models, including the fully supervised model xResNet1d [16], the contrastive learning-based CLOCS [17], and the multimodal model KED [9], as shown in Table 2.

Table 2. Comparative analysis of zero-shot diagnostic performance

Dataset	Method	Learing Pardigm	Knowledge Paradigm	AUROC	F1-SCORE	MCC
PTB-XL	CLOCS	SSL	N/A	0.6980	0.4210	0.3150
	KED	Foundation	Generative Narrative	0.7440	0.4800	-
	ESI	Foundation	Augmented Caption	0.7570	0.5210	-
	REISA(Ours)	Foundation	Evidence-Anchored Instruction	0.7626	0.5501	0.2644

CPSC2018	CLOCS	SSL	N/A	0.7120	0.4560	0.3620
	KED	Foundation	Generative Narrative	0.8790	0.5120	-
	CVDLLM	Foundation	Case-based Graph	0.7820	0.4950	-
	REISA(Ours)	Foundation	Evidence-Anchored Instruction	0.8232	0.4573	0.4005

Experimental results are shown in Table 2, where REISA demonstrates significant performance advantages in zero-shot diagnostic tasks. On the PTB-XL dataset, it achieved an average AUROC of 0.7626, not only outperforming the signal self-supervised model CLOCS but also surpassing existing multimodal foundation models KED (0.7440) and ESI (0.7570) [8]. Particularly in the F1-SCORE (0.5501) and MCC (0.2644) metrics, which represent decision quality, REISA constructed a more discriminative decision boundary. This is attributed to the MSDS strategy effectively reducing semantic ambiguity during the alignment process. Conversely, on the CPSC2018 dataset, although KED exhibited strong ranking capability in terms of AUROC, REISA demonstrated superior diagnostic consistency in the MCC (0.4005) metric, proving the effectiveness of the ISSA mechanism. Integrating the performance across all datasets, REISA maintained a high degree of logical robustness in data testing spanning different geographical regions and heterogeneous devices. This powerfully demonstrates that anchoring foundation models to "evidence-based medical knowledge" rather than simple "textual narratives" is the core pathway [6] for enhancing ECG zero-shot transfer capability, providing empirical support for the construction of globally applicable automated diagnostic systems.

Table 3. Multi-dimensional evaluation details for core clinical conditions.

Category	Model Method	AUROC	Sens.	Spec.	LR+	LR-	MCC
AFIB	KED	0.9650	0.7650	0.9730	28.3	0.24	-
	REISA	0.9838	0.9211	0.9824	52.47	0.08	0.8446
CRBBB	KED	0.9720	0.9040	0.9940	150.9	0.09	-
	REISA	0.9737	0.7778	0.9581	18.57	0.23	0.6910
ASMI	KED	0.3100	0.2370	0.9650	6.77	0.79	-
	REISA	0.6068	0.2330	0.8948	2.21	0.85	0.2819
1AVB	REISA	0.8706	0.2000	0.9896	19.30	0.81	0.2725
NORM	REISA	0.7451	0.9107	0.5014	1.82	0.18	0.3002

Table 3 provides a detailed demonstration of the multi-dimensional diagnostic efficacy of REISA across typical clinical conditions. In atrial fibrillation (AFIB), REISA achieved an AUROC of 0.9838 and a high Positive Likelihood Ratio (LR+) of 52.47, significantly outperforming KED (0.7650). This demonstrates that the MSDS strategy enhances the model's capability to capture key morphological invariants, such as P-wave disappearance and irregular rhythm. For ASMI, a location-sensitive condition, KED exhibited logical failure below the random level; in contrast, REISA successfully improved the metrics to a 0.6068 AUROC and a 0.2819 MCC by introducing explicit lead-space constraints through the ISSA mechanism, effectively alleviating the spatial awareness deficiency of existing models when processing lesions in specific leads. Overall results indicate that REISA can effectively model complex pathological morphology and spatial associations through evidence-based instruction alignment, demonstrating superior decision-making robustness compared to narrative-based models.

3.3 Ablation Study

To quantify the contribution of each methodological innovation within the REISA framework to model performance, we performed systematic progressive ablation experiments on the PTB-XL dataset [16]. All experimental groups were conducted in a unified experimental environment to ensure an accurate assessment of the effectiveness of the paradigm shift from traditional "semantic alignment" to the "instruction addressing" proposed in this paper. The experimental results are summarized in Table 4.

Table 4. Hierarchical ablation study of REISA core components on the PTB-XL dataset

Group	Paradigm	MSDS	ISSA	AUROC	LPR	F1
w/o 1	LLM Memory	Raw Text	None	0.6942	0.5550	0.4521
2	Retrieval	Raw RAG	None	0.7215	0.6480	0.4862
3	Retrieval	MSDS	None	0.7394	0.6912	0.5284
4	REISA	MSDS	ISGA	0.7626	0.8417	0.5501

Table 4 illustrates the performance gains of each core component. A comparison between Group 1 and Group 2 shows that the introduction of external evidence-based knowledge improved the average AUROC from 0.6942 to 0.7215, verifying the necessity of anchoring to authoritative guidelines rather than the parametric memory of LLMs [7]. The comparison between Group 2 and Group 3 indicates that the MSDS strategy increased the F1 score to 0.5284 by filtering out non-pathological noise, demonstrating the promotional effect of high-SNR text on cross-modal alignment [13]. After introducing the ISSA mechanism in Group 4, the AUROC for location-sensitive conditions (e.g., LPR) rose from 0.6912 to 0.8417, confirming that spatial addressing vectors can resolve the "lack of spatial awareness" issue in multi-lead signals. In summary, each stage—knowledge retrieval, semantic reconstruction, and spatial grounding—makes a substantial contribution to the zero-shot diagnostic robustness of REISA.

3.4 Interpretability Analysis: Saliency Analysis and Lead Addressing Capability Verification

Table 4 shows that the AUROC for LPR significantly improved from 0.5550 to 0.8417, demonstrating that REISA effectively alleviates the lack of spatial awareness. Grad-CAM audit reveals a fundamental shift in decision logic: although LPR diagnosis relies on the measurement of the P-QRS interval, the baseline model (AUC 0.5550) exhibits distinct "attention dispersion," with its focus disorderly distributed across limb leads and ST-T segments. In the absence of instruction addressing, cross-attention is biased by signal amplitude, causing gradient energy to erroneously concentrate on high-power components such as the R-wave peak in lead V5. Consequently, traditional models treat text labels as global descriptors, making it difficult to isolate subtle interval features from background morphological noise.

Fig. 4. Comparison of diagnostic saliency for LPR based on Grad-CAM. (a) The BERT-based base-line model exhibits scattered attention on non-pathological segments. (b) Guided by textual evidence, the instruction-driven REISA precisely grounds its attention on the P-R segment.

In contrast, REISA exhibits a marked "instruction-guided focusing" effect. Under the guidance of specific lead instructions (e.g., measuring PR duration), the ISSA mechanism precisely locks attention onto the PR segment and the isoelectric interval, effectively bypassing QRS peak interference and supporting a robust AUC of 0.8417. Visual evidence demonstrates that the text encoder, leveraging its instruction-following capability [14], achieves dynamic, lead-specific addressing, marking a shift in the model from relying on "global waveform correlation" to "clinical-logic-driven identification of localized segments." This feature extraction mechanism based on clinical logic avoids interference from high-amplitude components. This is of great significance for enhancing the transparency and safety of ECG diagnostic models in clinical cardiology applications.

4 Conclusion

The REISA framework proposed in this study achieves a paradigm shift in multimodal ECG diagnosis from "probabilistic generation" to "evidence-based alignment." By anchoring diagnostic logic to authoritative guidelines via the RAG mechanism, REISA demonstrates the decisive role of supervision signal accuracy in rigorous medical scenarios. Methodologically, the MSDS strategy effectively resolves the gradient scattering problem by constructing a physical feature bottleneck, significantly enhancing the model's precision in capturing subtle electrophysiological distortions. Simultaneously, the ISSA mechanism formalizes lead localization as an instruction-driven addressing task, successfully fixing the "spatial blind spot" of foundation models and validating its logical correctness in managing location-sensitive diseases. REISA demonstrates a robust AUROC of 0.7626 and superior cross-domain generalization capability on datasets such as PTB-XL, demonstrating the effectiveness of evidence-based anchoring and instruction-following encoders in enhancing model interpretability and accuracy. Despite remaining limitations in identifying rare pathologies, this work provides a new paradigm for constructing high-confidence ECG diagnostic models. Future research will further extend this instruction-addressing paradigm to long-term dynamic ECG monitoring and other multi-channel physiological signal domains.

References

1. Liu, M.B., et al.: 2023 Report on Cardiovascular Health and Diseases in China. Chinese Circulation Journal 39(1), 20–38 (2025)
2. Saab, K., et al.: Capabilities of multi-modal large language models in medicine. Nature Medicine 30(3), 756–765 (2024)
3. Tian, Y., et al.: Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. Cell Reports Medicine 5(12), 101875 (2024)
4. Zhang, J., et al.: Mlbf-net: A multi-lead-branch fusion network for multi-class arrhythmia classification. IEEE Journal of Translational Engineering in Health and Medicine 12, 1–11 (2024)
5. Tu, T., et al.: Towards Generalist Biomedical AI. NEJM AI 1(3), 1–12 (2024)
6. Moor, M., et al.: Foundation models for generalist medical artificial intelligence. Nature 616(7956), 259–265 (2023)
7. Zakka, C., et al.: Almanac: Retrieval-augmented language models for clinical medicine. Nature Medicine 30(4), 1096–1103 (2024)
8. Yu, H., et al.: ECG Semantic Integrator (ESI): A Foundation ECG Model Pretrained with LLM-Enhanced Cardiological Text. arXiv:2405.19366 (2024)
9. Yu, H., et al.: Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation. In: Proc. ML4H, pp. 650–663 (2023)
10. Qiu, X., et al.: CVDLLM: Automated Cardiovascular Disease Diagnosis with LLM-Assisted Graph Attention. Proc. IEEE TAI (In Press, 2025)
11. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Proc. NeurIPS, vol. 33, pp. 9459–9474 (2020)
12. Jin, Q., et al.: MedCPT: Contrastive Learning of Medical Terminology Representations. Bioinformatics 39(11), btad651 (2023)
13. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. Proc. IEEE ITW, pp. 1–5 (2015)
14. Wang, L., et al.: Improving text embeddings with large language models. arXiv:2401.00368 (2024)
15. Jiang, A.Q., et al.: Mistral 7B. arXiv:2310.06825 (2023)
16. Wagner, P., et al.: PTB-XL, a large publicly available electrocardiography dataset. Sci. Data 7(1), 154 (2020)
17. Murawwat, S., et al.: Denoising and classification of arrhythmia using MEMD and ANN. Alex. Eng. J. 61(4), 2807–2823 (2022)