

# Teach Multimodal LLMs to Comprehend Electrocardiographic Images

Anonymous ACL submission

## Abstract

Electrocardiograms (ECGs) are essential, non-invasive diagnostic tools for assessing cardiac conditions. Existing methods suffer from limited generalizability, focusing on a narrow range of conditions, and typically depend on raw physiological signals, which may not be available in resource-limited settings where only printed or digital ECG images are accessible. Recent advancements in multimodal large language models (MLLMs) present promising opportunities for addressing these challenges. However, the application of MLLMs to ECG image interpretation remains challenging due to the lack of instruction-tuning data and well-established ECG image benchmarks for quantitative evaluation. To address these challenges, we introduce ECGInstruct, the first ECG image instruction-tuning dataset with over one million samples, covering a wide range of ECG-related tasks from diverse data sources. We develop PULSE, a fully open-source MLLM for ECG image interpretation trained on ECGInstruct. We curate ECGBench, a human expert-curated benchmark covering four key ECG image interpretation tasks across nine different datasets. Our experiments show that PULSE sets a new state-of-the-art, outperforming general MLLMs with an average accuracy improvement of 21% to 33%. This work highlights the potential of PULSE to enhance ECG interpretation in clinical practice.<sup>1</sup>

## 1 Introduction

Electrocardiograms (ECGs) are essential, non-invasive tools for diagnosing cardiovascular diseases. Despite the availability of automated ECG diagnosis models (Hannun et al., 2019; Ribeiro et al., 2020; Hughes et al., 2021), their clinical adoption remains challenging. Many models can only classify a restricted set of conditions (Ribeiro et al., 2020), limiting their ability to detect previously unseen abnormalities. Additionally, they rely on time-series physiological signals, which may be unavailable in resource-limited settings (Siontis

et al., 2021) where ECGs are stored only as *printed or digital images* (Sangha et al., 2022, 2023).

Recent advances in multimodal large language models (MLLMs) (OpenAI, 2023; Li et al., 2024a,b; Liu et al., 2024c) have demonstrated exceptional capabilities in vision-language tasks, opening new possibilities for ECG interpretation directly from ECG images, the primary format used by clinicians (Cuevas-González et al., 2022).

However, adapting MLLMs for ECG image analysis presents several obstacles. First, there are no large-scale ECG image datasets, as most existing ECG datasets contain only raw signal data, necessitating the creation of ECG images. Second, instruction-tuning datasets specific to ECG images are lacking, requiring the development of high-quality instruction-response pairs tailored to ECG interpretation. Lastly, the absence of a standardized benchmark for evaluating MLLM performance on ECG images makes it difficult to quantify progress and identify areas for improvement.

To address these challenges, we introduce a comprehensive suite of resources aimed at advancing ECG image interpretation: 1) ECGInstruct, the first large-scale ECG image instruction-tuning dataset with over one million ECG image-text samples; 2) PULSE, a fully open-source 7B MLLM trained on ECGInstruct that achieves state-of-the-art performance in ECG diagnosis; and 3) ECGBench, a human expert-curated benchmark covering four ECG image interpretation tasks across nine different datasets.

Specifically, **ECGInstruct** integrates diverse ECG-related tasks informed by clinical expertise, ensuring real-world applicability. To enhance robustness, we introduce synthetic ECG images with common artifacts, helping the model generalize to noisy clinical data. We collect ECG samples from multiple geographic regions to promote adaptability across different populations and healthcare systems. Additionally, we employ a large-scale synthetic data generation pipeline that leverages LLMs for instruction-tuning, with strict quality control through expert validation and automated

<sup>1</sup>All code, data and models are available at [anonymous.4open.science/r/PULSE-4ECD](https://anonymous.4open.science/r/PULSE-4ECD)

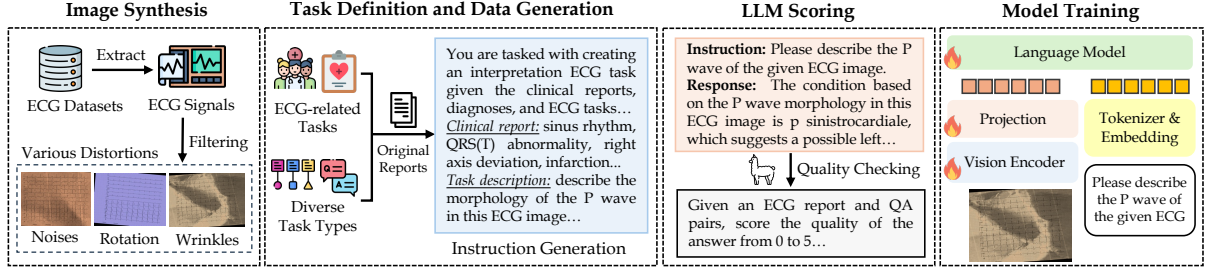


Figure 1: Overview of ECG image instruction-tuning data generation and MLLM training. (1) ECG images are synthesized from raw signals with realistic distortions. (2) ECG instruction-tuning data (ECGInstruct) is curated from clinician-defined tasks, diagnoses, and clinical reports. (3) An independent LLM scores the data for quality control. (4) PULSE, an MLLM trained on ECGInstruct, enables advanced ECG image interpretation.

assessment. Using the LLaVA architecture (Liu et al., 2024c), we train **PULSE-7B** on ECGInstruct, demonstrating that this straightforward training recipe significantly enhances ECG interpretation across various tasks.

For evaluation, we develop **ECGBench**, a benchmark designed with clinical experts to comprehensively assess ECG image interpretation. ECGBench includes four key tasks: (1) Abnormality detection, aligning diagnostic labels across heterogeneous datasets; (2) Report generation, standardizing expert-validated clinical reports; (3) MMMU ECG, a high-quality multi-choice question set with strict quality control; and (4) ECG Arena, a multi-turn interaction task that simulates real-world reasoning. Our evaluation strategy combines traditional metrics like AUC and F1 with GPT-4o-based scoring for open-ended tasks, ensuring both objectivity and clinical relevance.

Evaluated on ECGBench, our fully open-source PULSE sets a new state-of-the-art, significantly outperforming proprietary MLLMs across all benchmarks with an average accuracy gain of 21% to 33% compared to GPT-4o on out-of-domain datasets. Ablation experiments demonstrate the importance of incorporating diverse data sources and ECG instruction tasks into the training data. A case study and discussion further illustrate the model’s effectiveness in ECG image interpretation.

## 2 Teach MLLMs to Read ECG Images

In this section, we introduce ECGInstruct, the first large and comprehensive ECG image instruction-tuning data generation (Sec. 2.1) and PULSE, a fully open 7B MLLM trained on ECGInstruct for ECG image interpretation (Sec. 2.2). The overview of instruction-tuning data generation and MLLM training is shown in Fig. 1.

### 2.1 ECGInstruct

Existing ECG datasets lack ECG images for training MLLMs. To address this gap, we curate ECGInstruct, an instruction-tuning dataset with 1) realistic image synthesis resembling artifacts in paper ECGs, 2) diverse expert-curated tasks, and 3) data from varied geographical regions. We show the data summary in Table 1.

#### ECG image synthesis with various distortions

To enhance the robustness and real-world applicability of our model, we synthesize ECG images mimicking common artifacts found in paper ECGs. We adopt an ECG image synthesis tool (Shivashankara et al., 2024) that provides various imperfections such as grid line interference, creases, wrinkles, paper rotations, etc. By including these synthesized artifacts, we aim to train models that can effectively interpret ECGs in less-than-ideal conditions, as often encountered in clinical settings. More details are provided in Appendix C.

#### ECG-related tasks with clinical experts’ insights

To construct a comprehensive set of ECG-related tasks, we consulted domain experts to curate diverse and clinically relevant tasks covering four different categories. Each category is designed to address specific aspects of ECG interpretation and analysis, including (1) basic feature recognition (see examples in Appendix Fig. A1), (2) heart rhythm analysis (see examples in Appendix Fig. A2), (3) morphology and pathology identification (see examples in Appendix Fig. A3) and (4) clinical report generation (see examples in Appendix Fig. A4). Basic feature recognition (e.g., interval or segment, etc.) forms the foundation of ECG interpretation, enabling the model to grasp essential cardiac parameters. Heart rhythm analysis (e.g., arrhythmias, conduction abnormalities,





































Source Dataset	Task	Type	N
PTB-XL (Wagner et al., 2020)	Feature	   	30K
	Rhythm	   	36K
	Morphology	   	67K
	Report		16K
ECG-QA (Oh et al., 2024)	Feature		40K
	Rhythm		9K
	Morphology		90K
MIMIC ECG (Gow et al., 2023)	Feature	   	29K
	Rhythm	   	115K
	Morphology	   	169K
	Report		487K
CODE-15% (Ribeiro et al., 2021)	Feature		22K
	Rhythm		14K
	Morphology		31K
Total (ECGInstruct)			1.2M

Table 1: Summary of ECGInstruct. Feature: basic feature recognition, Rhythm: heart rhythm analysis, Morphology: morphology and pathology identification, Report: clinical report generation. : close-ended QA, : open-ended QA, : fill-in-the-blank, : multi-choice QA (See Appendix Table A1 for more details).

etc.) and morphology and pathology identification (e.g., wave shape, pathological conditions, etc.) are more advanced and critical aspects of ECG analysis, ensuring that the model can detect and classify complex conditions accurately. Lastly, clinical report generation mirrors the process of healthcare professionals synthesizing a comprehensive interpretation of an ECG. By incorporating clinical experts’ insights, we encourage the model to learn the practical skills required in a clinical context.

**Diverse types of tasks and data sources** Based on the original diagnoses and clinical reports from the existing ECG datasets, we curate diverse types of tasks including multi-choice questions, fill-in-the-blank, close-ended QA, and open-ended QA. This variety of task types not only enhances the model’s versatility but also mimics the diverse cognitive processes involved in real-world ECG interpretation. By incorporating these varied task types, we aim to develop a more robust and adaptable model capable of handling a wide spectrum of ECG-related queries and analyses.

To ensure broad applicability and generalizability, we collect publicly available ECG data from four different sources across geographically distinct regions: 1) PTB-XL (Wagner et al., 2020): a large Germany-based ECG dataset; (2) MIMIC-IV-ECG (Gow et al., 2023): a large set of ECGs

for patients who appear in the MIMIC-IV Clinical Database from Beth Israel Deaconess Medical Center in Boston (Johnson et al., 2023); 3) CODE-15% (Ribeiro et al., 2021): an ECG dataset from a central ECG repository from Minas Gerais, Brazil under the clinical outcomes in digital electrocardiology (CODE) study (Ribeiro et al., 2019); 4) ECG-QA (Oh et al., 2024), a question answering dataset for ECGs that is constructed based on PTB-XL (Wagner et al., 2020). This diverse geographical representation enhances the model’s ability to generalize across different populations and healthcare systems, accounting for potential variations in ECG patterns and interpretations across regions.

**Quality control** To guarantee the quality of generated instructions and corresponding responses, we apply an independent LLM as a judge to evaluate and score the content. This process involves several steps: 1) initial generation: instructions and responses are first generated using our primary model; 2) evaluation criteria: we establish a set of evaluation criteria including the instruction relevance, clarity, answerability of the responses, etc; 3) LLM judge and scoring: an independent LLM (Llama 3 (Meta, 2024)) is used as a judge to assess each instruction-response pair against established criteria and assign scores (see prompt in Appendix Fig. A8); 4) feedback loop: low-scoring items are flagged for human expert review and potential revision or removal; 5) iterative refinement: based on the scoring patterns and human expert input, we continually refine our instruction generation process. By combining automated LLM evaluation with human expert oversight, we create a robust system for maintaining and improving the quality of our instruction-response pairs.

**Data synthesizing at scale** Since large-scale annotation of ECG features is extremely expensive and time-consuming, we develop an automatic data synthesizing pipeline to address this data scarcity issue. We utilized clinical reports from PTB-XL and MIMIC-IV-ECG as initial seed data and leveraged an advanced LLM (i.e., Llama-3-70B-Instruct) for data synthesis. Building upon the expert-in-the-loop process and diverse data resources described in the previous sections, we synthesized a substantial volume of ECG-related instructions and corresponding responses. These were based on expert-provided examples and real-world scenarios, with the specific prompts used in this process detailed in the Appendix E. For datasets lacking comprehen-



sive reports, such as CODE-15%, we manually constructed diverse templates to transform the existing data into an instruction-response format. Formally, we have the data synthesis process as follows,

$$D_s = F_s(\text{Prompt}_s(D); \theta) \quad (1)$$

where  $\theta$  is the teacher LLM (i.e., Llama3-70B-Instruct),  $D$  is the initial seed data (e.g., clinical reports from PTB-XL),  $\text{Prompt}_s$  is the text prompt used to guide the generation,  $F_s$  is the quality control function used to shepherd the synthetic data.

## 2.2 PULSE: Model Training

We develop PULSE, a fully open-source 7B MLLM trained on ECGInstruct for ECG image interpretation. Our model architecture closely follows that of LLaVA (Liu et al., 2024b,c), adapting it for ECG image analysis. We use a vision encoder to process ECG images and a large language model as the text decoder, connected via a projection layer. We organize the data into three components: the image, the instructions, and the outputs. The instruction is query or task related to the ECG image and the output is the expected response or prediction base on the image and instruction. We place the image at the beginning of each conversation, serving as the visual grounding for the entire dialogue. During training, we update the parameters of the vision encoder ( $\theta_{\text{enc}}$ ), projection layer ( $\theta_{\text{proj}}$ ), and the language model ( $\theta_{\text{llm}}$ ) using an autoregressive training objective. Specifically, given an instruction  $\mathbf{Q}$ , a reference answer  $\mathbf{A}$ , and an image  $\mathbf{I}$ , all represented as sequences of tokenized inputs, we train the model by maximizing the likelihood of each token in  $\mathbf{A}$ , indexed by  $i = 1, \dots, L$ :

$$\mathcal{L} = - \sum_{i=1}^L \log p(\mathbf{A}_i \mid \mathbf{A}_{1:i-1}, \mathbf{Q}, \mathbf{I}; \theta_{\text{enc}}, \theta_{\text{proj}}, \theta_{\text{llm}}) \quad (2)$$

This objective guides the model to generate responses that are both accurate and contextually appropriate by leveraging information from the instruction, prior generated tokens, and the visual features extracted from the ECG image.

## 3 ECGBench

In this section, we present ECGBench, a comprehensive benchmark for evaluating MLLMs on ECG image interpretation. Our benchmark contains both repurposed tasks from existing datasets and newly created tasks from external resources. We provide the data curation process in Fig. 2 and details of each evaluation dataset in Appendix Table A2.

## 3.1 Evaluation Task Curation

**Abnormality detection** This task focuses on detecting cardiac abnormalities using ECG images. We curate this task by repurposing six existing ECG datasets: three in-domain datasets: PTB-XL (Supper) (Wagner et al., 2020), CODE-15% (Ribeiro et al., 2021), ECG-QA (Oh et al., 2024), and three out-of-domain datasets: CPSC 2018 (Liu et al., 2018), CSN (Zheng et al., 2020a,b) and G12EC (Liu et al., 2018). For all datasets, we first synthesize images using raw signals and then curate queries based on the original diagnosis and reports. For datasets with fewer than 10 diagnostic labels, we curate close-ended questions. For those with more labels, we construct multi-choice questions with 8 options, including the original diagnosis and randomly sampled negative labels.

**Report generation** This task involves generating detailed reports for given ECG images. We benchmark using 500 randomly selected reports from the test set of PTB-XL, which contains high-quality ECG reports written and validated by cardiologists. Similarly, the ECG images are synthesized from the raw signals.

**MMMU ECG** Inspired by MMMU (Yue et al., 2024), a widely adopted evaluation benchmark for MLLMs, we manually curated an ECG version with 200 multi-choice questions with the help of medical school students. The curation process involved three key steps: (1) **Resource selection:** We gathered ECG materials from diverse and reliable sources such as ECG textbooks, clinical case reports from medical journals, and widely used online ECG learning materials. This ensures the comprehensiveness and quality of collected ECG examples and interpretations. (2) **Question creation and collection:** Five medical school students with basic knowledge of ECG were recruited for this task. They extracted existing questions from the collected resources. For ECG images accompanied only by clinical interpretations, the annotators created questions based on these interpretations. Additionally, they formulated new questions drawing from their expertise, ensuring a balance between various ECG interpretation aspects (e.g., rhythm analysis, morphology assessment, clinical interpretation). (3) **Quality control:** To maintain high standards, we implemented a quality control process. In particular, each question was reviewed by at least two other annotators, checking for ac-

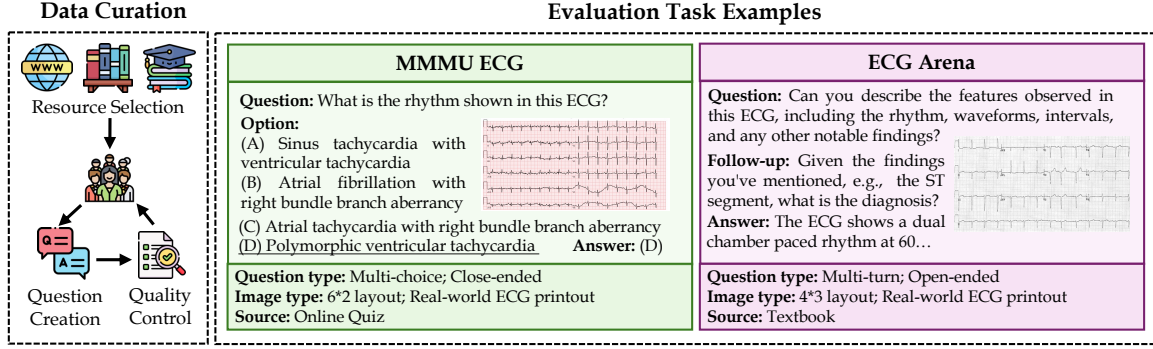


Figure 2: The evaluation data curation and example questions of MMMU ECG and ECG Arena in ECGBench. More details for abnormality detection and report generation are provided in Appendix Fig. A11

curacy and clarity. An independent reviewer cross-checked the final images, questions, and answers against the original sources to ensure fidelity to the source material. Any discrepancies or ambiguities were resolved during this process.

**ECG Arena** To assess the model’s instruction-following ability in ECG comprehension, we developed ECG Arena, inspired by MT-Bench (Zheng et al., 2024) and Arena-hard (Chiang et al., 2024) used in general LLM chat evaluations. We manually curated 50 multi-turn ECG-related questions, focusing on open-ended interactions. The data curation process for ECG Arena, like MMMU ECG, involves three main steps: resource selection, question creation, and quality control. The key distinction is that MMMU ECG focuses on multiple-choice questions, whereas ECG Arena involves more complex, flexible multi-turn, open-ended questions. Each follow-up question is contingent on the initial question and its response, making the process more challenging and reflective of real-world applications. Since multi-turn conversations are rare in existing sources, this posed significant challenges during data curation. To address this, annotators created such conversations by referencing original clinical interpretations and ECG images. The questions are designed to feel natural and simulate a real clinical setting (e.g., the first question may ask about basic findings from the image, followed by a question about potential clinical causes or diagnoses based on those findings).

### 3.2 Evaluation Metrics

**Abnormality detection:** We use macro AUC, macro F1, and hamming loss (HL) for multi-label datasets, and accuracy for others. **Report generation:** We employ GPT-4o as a judge, evaluating reports based on rhythms, waveform, and diagnosis,

with a maximum score of 100 points (see evaluation prompt in Appendix Fig. A9). **MMMU ECG:** We use accuracy as the primary metric, with systematic, rule-based evaluation pipelines to ensure consistent scoring. **ECG Arena:** GPT-4o assesses model performance by comparing generated responses with ground truth answers, considering accuracy, completeness, and instruction adherence, with a maximum score of 100 points (see evaluation prompt in Appendix Fig. A10). More evaluation details are provided in the Appendix G.

## 4 Experiments

### 4.1 Methods for Comparison

In order to evaluate the performance of our proposed model, we compare it against a set of established methods including domain-specific methods and state-of-the-art MLLMs.

• **Domain-specific methods.** We consider six domain-specific methods including five signal-based methods: METS (Li et al., 2024c), MERL (Liu et al., 2024a), ST-MEM (Na et al., 2023), MMCL (Turgut et al., 2025), MOMENT (Goswami et al., 2024); one image-based method: ECG-GPT (Khunte et al., 2024).

• **Proprietary MLLMs.** We consider three proprietary MLLMs: GPT-4o, GPT-4o mini (OpenAI, 2024), Gemini 1.5 Pro (Reid et al., 2024), and Claude 3.5 Sonnet (Anthropic, 2024).

• **Open-source MLLMs.** We select various open-source models to ensure coverage across different model sizes and visual components, including the general models LLaVA-1.5 (Liu et al., 2024d,b), LLaVA-1.6 (Liu et al., 2024c), Phi-3-Vision (Abdin et al., 2024), Idefics2-8B (Laurençon et al., 2024), DeepSeek-VI-7B (Lu et al., 2024a), Mantis-8B-siglip-Llama3 (Jiang et al., 2024), MiniCPM-V-2.6 (Yao et al., 2024), InternVL2 (Chen et al.,

Datasets	PTB-XL Super			PTB-XL Report	CODE-15%			ECG-QA
Metric	AUC	F1	HL	Report Score	AUC	F1	HL	Accuracy
Random	50.3	33.2	50.1	0	48.8	15.0	32.1	16.2
Domain-specific Methods								
METS	-	65.7 <sup>†</sup>	-	N/A	-	-	-	N/A
MERL	74.2 <sup>†</sup>	-	-	N/A	-	-	-	N/A
ST-MEM	71.4 <sup>†</sup>	-	-	N/A	-	-	-	N/A
ECG-GPT	69.5*	53.9*	20.1*	47.8*	68.9*	40.1*	17.4*	N/A
Proprietary MLLMs								
GPT-4o	<u>55.6</u>	<u>28.3</u>	<u>26.2</u>	<u>50.2</u>	<u>59.9</u>	<u>24.9</u>	15.7	<u>35.2</u>
GPT-4o mini	52.0	20.4	31.7	37.1	57.5	22.0	<u>15.1</u>	14.9
Gemini 1.5 Pro	50.7	15.3	27.9	35.9	56.7	20.0	15.9	33.2
Claude 3.5 Sonnet	54.0	27.5	29.6	43.7	58.3	20.3	17.8	34.2
Open-source MLLMs								
LLaVA-Med	50.0	12.3	28.1	24.3	69.2	27.0	33.4	<u>29.5</u>
LLaVA-1.6-34B	50.2	19.9	36.0	17.0	57.2	12.8	16.6	<u>22.4</u>
LLaVA-OneVision-7B	49.8	11.4	34.5	30.0	58.7	17.0	20.6	20.4
LLaVA-OneVision-72B	50.6	29.6	50.4	40.6	52.3	7.0	<u>13.1</u>	25.0
Deepseek-VL-Chat-7B	50.9	15.7	27.9	15.6	63.7	<u>27.5</u>	22.4	21.1
MiniCPM-V-2.6	49.0	<u>37.7</u>	63.8	15.4	56.6	25.3	22.0	20.8
Phi-3-Vision-128k-Instruct	50.0	29.6	48.4	20.2	<u>69.6</u>	22.6	38.8	28.4
Qwen2-VL-72B	<u>54.0</u>	28.3	30.2	<u>48.9</u>	60.6	23.6	16.1	23.7
InternVL2-8B	50.6	14.3	<u>27.8</u>	38.1	55.8	16.1	17.7	22.3
InternVL2-40B	51.2	18.7	34.6	41.8	56.7	16.2	17.4	18.2
PULSE-7B (Ours)	<b>82.9</b>	<b>76.9</b>	<b>10.2</b>	<b>65.4</b>	<b>91.7</b>	<b>87.0</b>	<b>4.6</b>	<b>71.6</b>
Δ over best proprietary MLLM	+27	+49	+16	+15	+32	+62	+11	+36
Δ over best open-source MLLM	+29	+39	+18	+17	+22	+60	+9	+42

Table 2: **In-domain evaluation results.** <sup>†</sup> indicates results from original papers, \* denotes results obtained using the provided online software, N/A indicates methods not applicable or not designed for certain tasks, and - indicates unreported scores in original papers. Results on all baselines are provided in Appendix Table A4.

2023, 2024) and state-of-the-art multimodal models LLaVA-OneVision (Li et al., 2024a), Qwen2-VL (Wang et al., 2024), and the domain-specific models LLaVA-Med (Li et al., 2024b).

## 4.2 Implementation Details

We follow the architecture of LLaVA-v1.6-Vicuna-7B, which includes three core components: a vision encoder, a large language model, and a projector to align image and text modalities. We format all datasets into a chatbot-style multi-turn dialogue format and use the special token “<image>” to represent image features within the text data. We utilize anyres to support the model’s ability to recognize ECG images of various sizes that may appear in real-world scenarios. We freeze the parameters of the vision encoder and fine-tune all parameters of the projector and LLM. We use a learning rate of 2e-5, set the batch size to 128, and employ a cosine scheduler with a 5% warm-up period for three epochs. The model is trained on 8 H100 GPUs, each running for 40 hours, totaling 320 GPU hours

of computation. Detailed implementation details are provided in Appendix H.

## 4.3 Main Results

We show in-domain and out-of-domain results in Table 2 and Table 3 respectively. Overall, we observe that PULSE achieves state-of-the-art performance on different datasets and tasks.

**Results on in-domain datasets** As shown in Table 2, PULSE demonstrates significant improvements over both proprietary and open-source MLLMs across all in-domain datasets. Specifically, PULSE surpasses the best proprietary model (GPT-4o) with a 27% improvement in AUC, a 15-point gain in report score, and a 36% increase in accuracy on the PTB-XL Super, PTB-XL Report, and ECG-QA tasks, respectively. Moreover, PULSE achieves notable gains over the best open-source model, with a 29% improvement in AUC, a 17-point gain in report score, and a 42% increase in accuracy on the same tasks.

Datasets	CPSC 2018			CSN	G12EC	MMMU ECG	ECG Arena
Metric	AUC	F1	HL	Accuracy	Accuracy	Accuracy	Arena Score
Random	51.2	15.1	28.8	11.6	12.1	24.2	0
Domain-specific Methods							
METS	-	-	-	N/A	N/A	N/A	N/A
MERL	<b>82.8<sup>†</sup></b>	-	-	N/A	N/A	N/A	N/A
ST-MEM	70.4 <sup>†</sup>	-	-	N/A	N/A	N/A	N/A
ECG-GPT	69.3*	44.0*	9.9*	N/A	N/A	N/A	N/A
Proprietary MLLMs							
GPT-4o	50.9	10.6	18.2	<u>57.5</u>	49.2	43.5	33.5
GPT-4o mini	49.2	11.0	25.5	32.1	33.2	39.5	30.1
Gemini-1.5-Pro	50.1	7.4	20.5	50.5	36.0	40.0	31.2
Claude 3.5 Sonnet	<u>52.8</u>	<u>11.5</u>	18.9	51.5	<u>51.4</u>	42.0	<u>37.1</u>
Open-source MLLMs							
LLaVA-Med	50.0	2.5	20.2	13.8	14.1	27.0	15.9
LLaVA-1.6-34B	49.6	<u>19.3</u>	62.8	44.3	<u>45.9</u>	31.0	17.5
LLaVA-OneVision-7B	49.6	8.0	28.3	23.3	25.7	26.0	22.5
LLaVA-OneVision-72B	51.5	12.8	29.4	44.0	42.6	35.0	15.5
Deepseek-VL-Chat-7B	50.7	6.0	20.0	35.7	32.9	34.5	15.3
MiniCPM-2.6	50.0	18.0	48.4	12.7	19.6	34.5	20.4
Phi-3-Vision-128k-Instruct	50.6	19.0	70.2	14.8	18.4	31.0	11.3
Qwen2-VL-72B	50.7	9.8	<u>18.9</u>	35.5	42.9	35.0	10.3
InternVL2-8B	52.1	8.2	22.2	<u>47.7</u>	37.5	30.0	22.9
InternVL2-40B	<u>52.4</u>	8.2	21.4	41.0	45.0	30.5	<u>28.0</u>
PULSE-7B (Ours)	80.7	<b>65.4</b>	<b>6.8</b>	<b>87.9</b>	<b>81.4</b>	<b>64.0</b>	<b>41.5</b>
$\Delta$ over best proprietary MLLM	+28	+54	+11	+30	+30	+21	+4
$\Delta$ over best open-source MLLM	+28	+46	+12	+40	+36	+26	+14

Table 3: **Out-of-domain (OOD) evaluation results.** <sup>†</sup> indicates results from original papers, \* denotes results obtained using the provided online software, N/A indicates methods not applicable or not designed for certain tasks, and - indicates unreported scores in original papers. Results on all baselines are provided in Appendix Table A5.

These results highlight the complexity of ECG image interpretation, a task where even the best proprietary models perform near randomly. By fine-tuning on ECGInstruct, PULSE achieves substantial performance improvements, demonstrating the importance of high-quality and task-related instruction-tuning. Moreover, while certain domain-specific methods (e.g., MERL) achieve comparable performance on specific datasets, their specialized designs limit their generalization to other diverse tasks, restricting their broader applicability in real-world, complex healthcare scenarios.

**Results on out-of-domain datasets** Table 3 presents the comparison results on out-of-domain datasets, where PULSE consistently delivers outstanding performance. Notably, it achieves a significant 21% improvement in accuracy on the MMMU ECG benchmark compared to GPT-4o. This substantial improvement indicates the PULSE’s robustness and ability to generalize to unseen data.

The ECG Arena benchmark presents a more

challenging task for all models, characterized by a multi-turn, open-ended question-answering format, which closely simulates real clinical scenarios. Despite these challenges, PULSE still surpasses the best proprietary model by 4 points and outperforms the leading open-source model by an impressive 14 points in terms of arena score. These results highlight PULSE’s relative strength in handling complex, clinically-oriented ECG interpretation and analysis. Additionally, the performance gap across models on this challenging benchmark indicates considerable room for future improvements in this task.

#### 4.4 Ablation Study

**Effect of training data source** Given that ECGInstruct is compiled from diverse datasets, it is crucial to examine how each dataset contributes to the model’s overall performance. Table 4 presents a comparative analysis of models trained on various dataset combinations. The model trained exclusively on PTB-XL (P) exhibits



Training Data	PTB-XL Super	PTB-XL Report	CSN	CODE 15%	ECG QA	CPSC	G12	MMMU ECG	ECG Arena	AVG
P	70.3	60.8	85.5	33.1	29.6	31.2	68.6	46.0	31.0	50.7
P + M	76.2	66.5	91.4	50.1	33.6	60.2	82.0	64.5	39.6	62.7
P + M + C	76.2	67.9	90.2	87.4	41.2	58.8	78.7	61.5	42.0	67.1
P + M + C + E	76.9	65.4	87.9	87.0	71.6	65.4	81.4	64.0	41.5	<b>71.2</b>

Table 4: Performance of different training dataset combinations. P: PTB-XL, M: MIMIC-IV-ECG, C: CODE-15%, E: ECG-QA. F1 for PTB-XL Super, CODE-15%, and CPSC; Accuracy for CSN, ECG-QA, G12, and MMMU ECG; Report Scores for PTB-XL Report; Arena Scores for ECG Arena. **AVG** denotes the average across all metrics.

the lowest performance across all datasets, indicating the limitations of relying on a single data source for effective generalization. As we progressively incorporate additional datasets into the training set, the model’s performance consistently improves. These results highlight the importance of curating diverse training data, as expanding beyond a single source enhances the model’s capacity to generalize across datasets and tasks. We provide the ablation study on different instruction tasks in Appendix Table A8.

**Comparison between signal encoder and image encoder** We present the comparison results between the image-based encoder (ours) and the signal-based encoder in Fig. 3. Both models were trained using the exact same data and architectural framework, differing only in the approach used to encode ECGs (i.e., treating ECGs as time-series or digital images). The results show that the image-based encoder consistently outperforms the signal-based encoder across different evaluation tasks, with particularly significant improvements observed in out-of-domain datasets. These findings highlight that encoding ECGs as images not only aligns with the goal of enabling broader applicability of automated ECG diagnosis, especially in resource-constrained or remote settings (where only printed or digital ECGs are available) but also empirically surpasses the performance of signal-based encoder model.

#### 4.5 Case Study

We present some examples from our benchmark, comparing the outputs of our model with GPT-4o for ECG report generation (Appendix Figs. A13-A15) and ECG Arena (Appendix Fig. A16). While GPT-4o is capable of generating reports and answering questions by following instructions, it often produces responses that, although well-structured and seemingly relevant, contain sig-

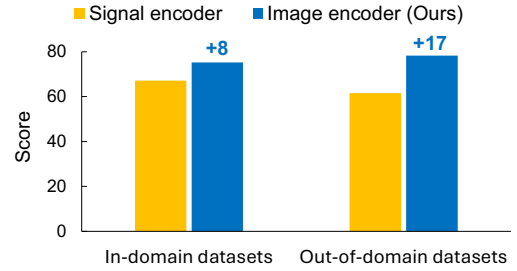


Figure 3: Performance comparison between signal-based encoder and image-based encoder.

nificant inaccuracies in interpretation. In contrast, PULSE consistently provides more accurate responses that align closely with the ground truths. Additionally, we observed that GPT-4o tends to over-rely on its OCR capabilities when textual information (e.g., printed axis labels, numerical values like heart rate or QRS duration) is present in images, leading to superficial reasoning based on text rather than a deep analysis of visual data. As shown in Appendix Fig. A15, GPT-4o identifies a left axis deviation based on the printed QRS axis degree, without analyzing the visual waveform patterns. If such axis information were absent, the model would likely fail to identify the deviation.

## 5 Conclusion

In this paper, we study the problem of ECG image interpretation, which is a crucial task in assessing cardiac conditions. We develop PULSE, a fully open-source MLLM trained on ECGInstruct with over one million samples across a diverse range of ECG-related tasks. Evaluated on the proposed benchmark, ECGBench, our model shows state-of-the-art performance, surpassing both proprietary and open-source MLLMs across multiple in-domain and out-of-domain evaluation datasets. This work demonstrates the potential of using MLLMs for enhancing ECG image analysis and interpretation in clinical applications.



## Limitations

While PULSE demonstrates superior performance across various evaluation datasets, there remains room for improvement in handling more complex and open-ended tasks, such as report generation and multi-turn conversations. These tasks demand advanced reasoning abilities and strong instruction-following capabilities, which are crucial for clinical applications.

To address these challenges, future work can focus on two key areas: (1) incorporating a more diverse set of instruction-following data to enhance the model’s generalizability, and (2) scaling up high-quality chain-of-thought (CoT) and multi-turn training data, leveraging expert insights, structured knowledge databases (e.g., SNOMED CT (Stearns et al., 2001)), medical literature, and clinical guidelines. This will guide the model through intermediate reasoning steps, including identifying key ECG features, mapping them to diagnoses, and generating well-grounded clinical rationales.

Our data ablation studies (Tables 4 and A8) suggest that increasing and diversifying training data can further enhance model performance. Based on these findings, we plan to explore these directions in future research to bridge the identified gaps.

## References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anthropic. 2024. [Claude 3.5 sonnet](#). Accessed: September 24, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Daniel Cuevas-González, Juan Pablo García-Vázquez, Miguel Bravo-Zanoguera, Roberto López-Avitia, Marco A Reyna, Nestor Alexander Zermeño-Campos, and María Luisa González-Ramírez. 2022. Ecg standards and formats for interoperability between mhealth and healthcare information systems: a scoping review. *International Journal of Environmental Research and Public Health*, 19(19):11941.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tio, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [InstructBLIP: Towards general-purpose vision-language models with instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*.

B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernandes, N. Greenbaum, J. W. Waks, P. Eslami, T. Carbonati, A. Chaudhari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, and S. Horng. 2023. [Mimic-iv-ecg: Diagnostic electrocardiogram matched subset](#).

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69.

J Weston Hughes, Jeffrey E Olgin, Robert Avram, Sean A Abreau, Taylor Sittler, Kaahan Radia, Henry Hsia, Tomos Walters, Byron Lee, Joseph E Gonzalez, et al. 2021. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA cardiology*, 6(11):1285–1295.

Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhui Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Akshay Khunte, Veer Sangha, Evangelos K Oikonomou, Lovedeep S Dhillon, Arya Aminoroaya, Andreas Coppi, Sumukh Vasisht Shankar, Bobak J Mortazavi, Deepak L Bhatt, Harlan M Krumholz, et al. 2024. Automated diagnostic reports from images of electrocardiograms at the point-of-care. *medRxiv*.

670	Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? <i>arXiv preprint arXiv:2405.02246</i> .	726
671		727
672		728
673		729
674	Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. <i>arXiv preprint arXiv:2408.03326</i> .	730
675		
676		
677		
678		
679	Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. <i>Advances in Neural Information Processing Systems</i> , 36.	
680		
681		
682		
683		
684		
685	Jun Li, Che Liu, Sibao Cheng, Rossella Arcucci, and Shenda Hong. 2024c. Frozen language model helps ecg zero-shot learning. In <i>Medical Imaging with Deep Learning</i> , pages 402–415. PMLR.	731
686		732
687		733
688		734
689	Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024a. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. <i>arXiv preprint arXiv:2403.06659</i> .	735
690		
691		
692		
693		
694	Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. 2018. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. <i>Journal of Medical Imaging and Health Informatics</i> , 8(7):1368–1373.	736
695		737
696		
697		
698		
699		
700		
701	Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 26296–26306.	738
702		
703		
704		
705		
706	Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. <a href="#">Llava-next: Improved reasoning, ocr, and world knowledge</a> .	
707		
708		
709	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. <i>Advances in neural information processing systems</i> , 36.	
710		
711		
712	Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. <a href="#">Deepseek-vl: Towards real-world vision-language understanding</a> . <i>Preprint</i> , arXiv:2403.05525.	
713		
714		
715		
716		
717		
718	Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahron Kim, Dimitra Pouli, Ankush Patel, et al. 2024b. A multimodal generative ai copilot for human pathology. <i>Nature</i> , pages 1–3.	
719		
720		
721		
722		
723	Meta. 2024. <a href="#">Introducing meta llama 3: The most capable openly available llm to date</a> . Accessed: 2024-10-01.	
724		
725		
	Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2023. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In <i>The Twelfth International Conference on Learning Representations</i> .	740
		741
	Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon-myung Kwon, and Edward Choi. 2024. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. <i>Advances in Neural Information Processing Systems</i> , 36.	742
		743
	OpenAI. 2023. <a href="#">Gpt-4v(ision) technical work and authors</a> .	744
		745
	OpenAI. 2024. <a href="#">Gpt-4o contributions</a> .	746
		747
	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	748
		749
	Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. <i>arXiv preprint arXiv:2403.05530</i> .	750
		751
	Antônio H Ribeiro, GM Paixao, Emilly M Lima, Manoel Horta Ribeiro, Marcelo M Pinto Filho, Paulo R Gomes, Derick M Oliveira, Wagner Meira Jr, Thomas B Schon, and Antonio Luiz P Ribeiro. 2021. Code-15%: A large scale annotated dataset of 12-lead ecgs. <i>Zenodo</i> , Jun, 9.	752
		753
		754
	Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. 2020. Automatic diagnosis of the 12-lead ecg using a deep neural network. <i>Nature communications</i> , 11(1):1760.	755
		756
		757
		758
		759
		760
	Antonio Luiz P Ribeiro, Gabriela MM Paixao, Paulo R Gomes, Manoel Horta Ribeiro, Antonio H Ribeiro, Jessica A Canazart, Derick M Oliveira, Milton P Ferreira, Emilly M Lima, Jermana Lopes de Moraes, et al. 2019. Tele-electrocardiography and bigdata: the code (clinical outcomes in digital electrocardiography) study. <i>Journal of electrocardiology</i> , 57:S75–S78.	761
		762
		763
		764
		765
		766
		767
		768
		769
	Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. <i>arXiv preprint arXiv:2404.18416</i> .	770
		771
		772
		773
		774
	Veer Sangha, Bobak J Mortazavi, Adrian D Haimovich, Antônio H Ribeiro, Cynthia A Brandt, Daniel L Jacoby, Wade L Schulz, Harlan M Krumholz, Antonio Luiz P Ribeiro, and Rohan Khera. 2022. Automated multilabel diagnosis on electrocardiographic images and signals. <i>Nature communications</i> , 13(1):1583.	775
		776
		777
		778
		779
		780

781	Veer Sangha, Arash A Nargesi, Lovedeep S Dhingra, Akshay Khunte, Bobak J Mortazavi, Antônio H Ribeiro, Evgeniya Banina, Oluwaseun Adedola, Nadish Garg, Cynthia A Brandt, et al. 2023. Detection of left ventricular systolic dysfunction from electrocardiographic images. <i>Circulation</i> , 148(9):765–777.	837
782		838
783		839
784		840
785		841
786		842
787		843
		844
788	Kshama Kodthalu Shivashankara, Deepanshi, Afagh Mehri Shervedani, Gari D Clifford, Matthew A Reyna, and Reza Sameni. 2024. Ecg-image-kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. <i>Physiological Measurement</i> .	845
789		846
790		847
791		848
792		
793		
794	Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. <i>Nature</i> , 620(7972):172–180.	849
795		850
796		851
797		852
798		853
799	Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. <i>arXiv preprint arXiv:2305.09617</i> .	854
800		855
801		856
802		857
803		858
804		
805	Konstantinos C Siontis, Peter A Noseworthy, Zach I Attia, and Paul A Friedman. 2021. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. <i>Nature Reviews Cardiology</i> , 18(7):465–478.	859
806		860
807		861
808		862
809		863
810		864
811		865
812	Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In <i>Proceedings of the AMIA Symposium</i> , page 662. American Medical Informatics Association.	866
813		867
814		868
815	Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip J Aston, Ashish Sundar, Claus Graff, Jørgen K Kanters, Wilhelm Haverkamp, Olaf Dössel, Axel Loewe, et al. 2023. Ptb-xl+, a comprehensive electrocardiographic feature dataset. <i>Scientific data</i> , 10(1):279.	869
816		870
817		871
818		
819		
820		
821	Özgün Turgut, Philip Müller, Paul Hager, Suprosanna Shit, Sophie Starck, Martin J Menten, Eimo Martens, and Daniel Rueckert. 2025. Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging. <i>Medical Image Analysis</i> , page 103451.	872
822		873
823		874
824		875
825		876
826		
827	Patrick Wagner, Nils Strodthoff, Ralf-Dieter Boussetjot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. Ptb-xl, a large publicly available electrocardiography dataset. <i>Scientific data</i> , 7(1):1–15.	877
828		878
829		879
830		880
831		881
832		882
833	Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Zhenwu Peng, Jie Fu, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. 2024. Electrocardiogram instruction tuning for report generation. <i>arXiv preprint arXiv:2403.04945</i> .	883
834		884
835		885
836		886
	Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. <i>arXiv preprint arXiv:2409.12191</i> .	887
		888
		889
		890
		891
		892
		893
		894
		895
		896
		897
		898
		899
		900
		901
		902
		903
		904
		905
		906
		907
		908
		909
		910
		911
		912
		913
		914
		915
		916
		917
		918
		919
		920
		921
		922
		923
		924
		925
		926
		927
		928
		929
		930
		931
		932
		933
		934
		935
		936
		937
		938
		939
		940
		941
		942
		943
		944
		945
		946
		947
		948
		949
		950
		951
		952
		953
		954
		955
		956
		957
		958
		959
		960
		961
		962
		963
		964
		965
		966
		967
		968
		969
		970
		971
		972
		973
		974
		975
		976
		977
		978
		979
		980
		981
		982
		983
		984
		985
		986
		987
		988
		989
		990
		991
		992
		993
		994
		995
		996
		997
		998
		999
		1000

## Table of Contents in Appendix

<b>A</b>	<b>Related Work</b>	<b>13</b>
<b>B</b>	<b>Preliminary on 12-lead ECG</b>	<b>13</b>
<b>C</b>	<b>Details of ECG Image Synthesis</b>	<b>14</b>
<b>D</b>	<b>Details of Instruction-tuning Datasets</b>	<b>14</b>
<b>E</b>	<b>Prompts</b>	<b>20</b>
<b>F</b>	<b>Details of Evaluation Datasets</b>	<b>26</b>
<b>G</b>	<b>Details of Evaluation Metrics</b>	<b>27</b>
<b>H</b>	<b>Implementation Details</b>	<b>27</b>
<b>I</b>	<b>Main Results</b>	<b>29</b>
<b>J</b>	<b>Ablation Study</b>	<b>32</b>
<b>K</b>	<b>Human Expert Performance</b>	<b>33</b>
<b>L</b>	<b>Human Evaluation</b>	<b>33</b>
<b>M</b>	<b>Case Study</b>	<b>34</b>



## A Related Work

**Domain-specific Models for ECG.** Many domain-specific models have been proposed to enhance automatic ECG diagnosis (Hannun et al., 2019; Ribeiro et al., 2020; Hughes et al., 2021). For example, Ribeiro et al. (2020) applied convolutional neural networks (CNNs) to encode ECG signals for diagnosing 6 types of abnormalities. To reduce dependence on high-quality labeled data, recent studies (Li et al., 2024c; Liu et al., 2024a; Na et al., 2023) have further explored self-supervised learning approaches using unlabeled ECG training data. For example, Liu et al. (2024a) proposed an ECG representation learning framework by integrating the ECG signals and clinical reports, showing improved performance in zero-shot ECG classification tasks. Despite these successes, most approaches treat ECG data as temporal physiological signals, which could be limiting in certain resource-constrained or remote settings where only printed or digital images are available. Recently, a few methods (Sangha et al., 2022, 2023; Khunte et al., 2024) have been proposed for ECG diagnosis using ECG images. For example, Khunte et al. (2024) developed a diagnostic report generation framework for ECG images, which is built upon a BEiT (Bao et al., 2021) vision transformer encoder and a GPT-2 (Radford et al., 2019) decoder. However, their model is only capable of the clinical report generation task, without generalizability to other diverse tasks. In contrast, our study investigates the capabilities of MLLMs for ECG image interpretation. We curate a diverse range of instruction-tuning datasets to fine-tune the model, thus improving model generalizability.

**MLLMs in Healthcare** Recent advancements in MLLMs have shown promising results in various healthcare domains. General medical multimodal models such as LLaVA-Med (Li et al., 2024a), Med-PaLM (Singhal et al., 2023a,b), and Med-Gemini (Saab et al., 2024) have demonstrated capabilities in processing diverse medical data types. Additionally, domain-specific multimodal models have been developed for specialized fields like pathology (Lu et al., 2024b; Xu et al., 2024) and radiology (Wu et al., 2023). These models have shown potential in integrating visual and textual information to support clinical decision-making and medical education. However, despite the importance of ECG data in cardiac diagnosis and monitoring, current MLLMs often struggle to process ECG images effectively. This limitation highlights a significant gap in the application of MLLMs to cardiology, where the ability to interpret both visual ECG representations and accompanying clinical information is crucial.

**Multimodal Instruction-tuning.** Instruction-tuning has proven effective in the multimodal domain, particularly in vision-language models like LLaVA (Liu et al., 2024d), MiniGPT-4 (Zhu et al., 2023) and InstructBLIP (Dai et al., 2023). These models demonstrate impressive generalizability across various visual understanding and reasoning tasks. While multimodal instruction-tuning has been applied to general medical imaging tasks (Li et al., 2024b; Singhal et al., 2023a), its application to ECG images remains largely unexplored. A recent work (Wan et al., 2024) introduced a targeted instruction-tuning framework and fine-tuned existing open-source LLMs for ECG report generation. However, their 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 also treats ECG data as temporal signals, whereas our paper focuses on encoding ECG images with MLLMs, which is more applicable to real scenarios where only printed or digital ECG images are available.

## B Preliminary on 12-lead ECG

ECG is a vital diagnostic tool that measures the electrical activity of the heart over time, providing insights into both spatial and temporal aspects of cardiac function. Typically, an ECG recording is presented as a 12-lead multivariate time series, where each lead offers a unique perspective on heart activity. The six limb leads (I, II, III, aVR, aVL, and aVF) assess the electrical movements across the arms and legs, giving views from the frontal plane. Simultaneously, the six precordial leads (V1, V2, V3, V4, V5, and V6) monitor the chest, offering horizontal plane views. In this paper, we focus on ECG images that are synthesized from raw signals.

## C Details of ECG Image Synthesis

We employ the ECG-image-kit (Shivashankara et al., 2024) framework to synthesize diverse ECG images from raw signal data. This toolkit allows for the generation of ECG images under various conditions by introducing a range of distortions and noises to better simulate real-world clinical data.

Specifically, in addition to generating standard 12-lead ECG images—characterized by black waveforms on a white background, red grid lines, and a 4x3 layout—we introduce a variety of perturbations to the images. These modifications include the addition of wrinkles and creases, simulating the physical wear and tear commonly observed in paper-printed ECGs. Our image synthesis process includes various augmentation methods to simulate physical distortions, image quality variations, and layout alterations. We introduce wrinkles and creases to mimic wear and tear commonly observed in paper-printed ECGs, and apply random rotations at varying angles to simulate misaligned scans or prints. To account for different acquisition systems and scanning qualities, we vary image resolutions and introduce random background colors, such as slight yellowing to represent aging or poor scanning quality. We also add noise to the images to simulate imperfections in the scanning or printing process. Furthermore, we experiment with different aspect ratios, overall image sizes, and ECG plot positions within the image to reflect the heterogeneity of ECG printouts across different systems and formats. In some cases (with a 0.02 probability), we randomly remove grid lines to account for variations in ECG presentation.

To further enrich the synthetic images, we randomly insert meta-information into the image header to simulate the annotations typically seen in clinical ECG reports. For the PTB-XL dataset, we extract patient demographics (e.g., age, gender) and basic ECG features (e.g., heart rate, axis deviations) from the associated PTB-XL feature annotation dataset, PTB-XL+ (Strodthoff et al., 2023). This extracted data is used to impute realistic meta-information, which is then randomly printed on the synthesized image. This random insertion of meta-data not only increases the visual variety of the images but also provides additional context, simulating real-world ECG prints that include patient and diagnostic information. To further increase diversity, we adopt alternative lead configurations beyond the standard 4x3 layout, such as 12x1 (single row of leads), 6x2 (two rows of six leads), and other commonly used clinical formats. These variations ensure that our model is exposed to a wide range of ECG presentation styles.

The augmentation process is designed to balance the dataset, with an approximate ratio of 1:1 between augmented and standard ECG images. This balance ensures that the model is exposed to both clean and distorted images, aiding in its generalization to real-world clinical scenarios.

## D Details of Instruction-tuning Datasets

The detailed statistics of curated instruction-tuning datasets are provided in Table A1, with training examples of Figs. A1, A2, A3, A4.

Source Dataset	Task	Type	# Samples
PTB-XL	Basic Feature Recognition	Close-ended QA	22,759
		Open-ended QA	906
		Fill-in-blank	449
		Multi-choice QA	5,716
	Heart Rhythm Analysis	Close-ended QA	19,550
		Open-ended QA	201
		Fill-in-blank	436
		Multi-choice QA	16,179
	Morphology and Pathology Identification	Close-ended QA	50,166
		Open-ended QA	2,649
		Fill-in-blank	680
		Multi-choice QA	13,432
	Clinical Report	Open-ended QA	16,302
PTB-XL Total			149,425
ECG-QA	Basic Feature Recognition	Close-ended QA	40,154
	Heart Rhythm Analysis	Close-ended QA	8,993
	Morphology and Pathology Identification	Close-ended QA	90,211
ECG-QA Total			139,358
MIMIC-ECG	Basic Feature Recognition	Close-ended QA	759
		Open-ended QA	4,759
		Fill-in-blank	6,470
		Multi-choice QA	17,186
	Heart Rhythm Analysis	Close-ended QA	48,625
		Open-ended QA	5,262
		Fill-in-blank	11,487
		Multi-choice QA	49,352
	Morphology and Pathology Identification	Close-ended QA	8,241
		Open-ended QA	81,080
		Fill-in-blank	18,264
		Multi-choice QA	61,456
	Clinical	Open-ended QA	486,746
MIMIC-ECG Total			799,687
CODE-15%	Basic Feature Recognition	Close-ended QA	22,177
	Heart Rhythm Analysis	Close-ended QA	13,893
	Morphology and Pathology Identification	Close-ended QA	31,570
CODE-15% Total			67,640
ECGInstruct Total			1,156,110

Table A1: Detailed data statistics of ECGInstruct.

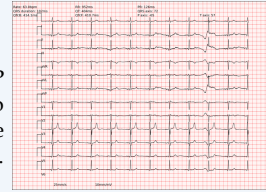
## Training Data Examples: Basic Feature Recognition

### Instruction:

What are the main ECG features I'm seeing here?

### Response:

Let's break it down. We have a rhythm that appears to be irregular, with varying P-P intervals. The heart rate is not explicitly stated, but based on the ECG tracing, it seems to be within the normal range. The cardiac axis appears to be normal as well. One notable feature is the presence of an rSr' pattern in lead V1, which is likely a normal variant. Additionally, we can see some nonspecific T wave changes in the septal leads.



### Follow-up Instruction:

Can you explain the significance of the rSr' pattern in lead V1? Is it a cause for concern?

### Response:

The rSr' pattern in lead V1 is often seen in healthy individuals, especially in younger people. It's considered a normal variant because it doesn't necessarily indicate any underlying cardiac pathology. The "r" represents a small initial positive deflection, followed by a larger negative deflection (the "S" wave), and then a secondary positive deflection (the "r'"). This pattern is thought to be related to the normal activation sequence of the right ventricle. Since it's isolated to lead V1 and doesn't appear to be associated with any other abnormalities, it's unlikely to be a cause for concern in this case.

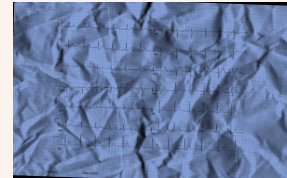
**Sub-task:** Basic Waveform Understanding; **Question type:** Open-ended; **Source dataset:** MIMIC-IV-ECG

### Instruction:

What can be inferred about the heart's electrical conduction system based on this ECG?

Options:

- A. There is a block in the electrical conduction pathway
- B. There is an abnormality in the heart's electrical conduction system
- C. The heart's electrical conduction system is functioning normally
- D. There is a rhythm disturbance



### Response:

This image indicates that the heart's electrical conduction system is functioning normally, as there are no apparent abnormalities or disturbances in the electrical activity. The presence of a sinus rhythm and a normal axis suggests that the heart's electrical conduction system is operating within normal limits, without any significant defects or anomalies. This ECG image is consistent with a normal heart, and there is no evidence of any conduction system abnormalities.

We can conclude that the answer is C. The heart's electrical conduction system is functioning normally

**Sub-task:** Conduction System Interpretation; **Question type:** Multi-choice QA; **Source dataset:** PTB-XL

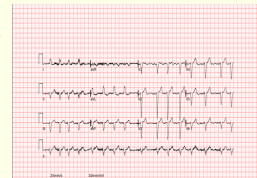
### Instruction:

You are reviewing an ECG tracing in a clinical setting. Based on the ECG morphology, fill in the blanks:

This ECG tracing is most consistent with a rhythm of \_\_\_\_\_, with a cardiac axis of \_\_\_\_\_, and a bundle branch block of the \_\_\_\_\_ bundle.

### Response:

This ECG tracing is most consistent with a rhythm of **atrial fibrillation**, with a cardiac axis of **left axis deviation**, and a bundle branch block of the **left bundle**.



**Sub-task:** Axis and Rhythm Determination; **Question type:** Fill-in-the-blank; **Source dataset:** MIMIC-IV-ECG

### Instruction:

Given this ECG image, please answer the following question. From the provided options, select all that apply. List your selections, separated by slashes.

Question:

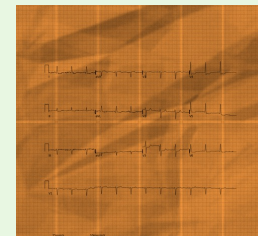
Which ECG leads are affected by baseline drift?

Options:

lead III, lead V2, none, lead V4, lead aVL, lead V5, lead V6, lead aVF, lead V3, lead aVR, lead V1, lead I, lead II. Provide the accurate response without additional details.

### Response:

lead aVF/lead V3



**Sub-task:** Lead Placement Determination; **Question type:** Close-ended QA; **Source dataset:** ECG-QA

Figure A1: The Examples of basic feature recognition instructions for finetuning PULSE.



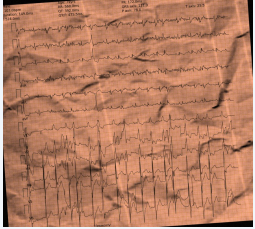
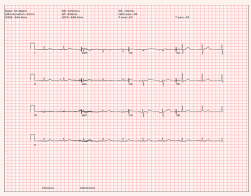
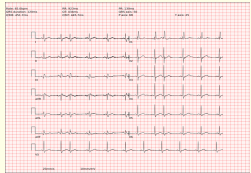
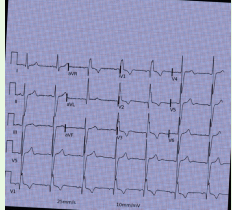
Training Data Examples: Heart Rhythm Analysis	
<p><b>Instruction:</b> Analyze the ECG image and identify the combination of cardiac rhythm abnormalities that can be diagnosed simultaneously in this patient. Rank the abnormalities in order of their potential impact on the patient's hemodynamic stability, with 1 being the most critical. Use a semicolon-separated list with the rhythm abnormality followed by its severity ranking</p> <p><b>Response:</b> Sinus tachycardia; 1; Right bundle branch block; 3; Premature atrial contraction; 2; Premature ventricular contraction; 2. Justification: Sinus tachycardia is ranked as the most critical abnormality since it can lead to decreased cardiac output, increased myocardial oxygen demand, and potentially exacerbate underlying cardiac conditions. Premature atrial contractions and premature ventricular contractions are ranked equally as they can both disrupt normal cardiac function, although their individual impact might be less severe than sinus tachycardia. Right bundle branch block, while still a significant abnormality, is ranked last as it may not have an immediate impact on hemodynamic stability, but rather serves as an indicator of underlying cardiac pathology.</p>	
<b>Sub-task:</b> Rhythm Analysis; <b>Question Type:</b> Open-ended; <b>Source Dataset:</b> PTB-XL	
<p><b>Instruction:</b> Imagine you're an ER doctor reviewing this ECG, what would be your primary concern regarding the patient's heart rhythm? Directly output one of the following options without explanation:</p> <p><b>Options:</b> A) Tachycardia; B) Ventricular Fibrillation; C) Atrial Fibrillation; D) Bradycardia; E) Long QT Syndrome; F) First-Degree AV Block; G) Second-Degree AV Block; H) Normal Sinus Rhythm</p> <p><b>Response:</b> D) Bradycardia</p>	
<b>Sub-task:</b> Arrhythmia Identification; <b>Question Type:</b> Multi-choice QA; <b>Source Dataset:</b> MIMIC-IV-ECG	
<p><b>Instruction:</b> If you were to categorize the ECG rhythm of this patient into one of the following options: Normal Sinus Rhythm, Atrial Fibrillation, Ventricular Tachycardia, or Sinus Rhythm with Premature Atrial Contractions, which one would you choose? _____</p> <p><b>Response:</b> Sinus Rhythm with Premature Atrial Contractions</p>	
<b>Sub-task:</b> Rhythm Classification; <b>Question Type:</b> Fill-in-the-blank; <b>Source Dataset:</b> PTB-XL	
<p><b>Instruction:</b> As part of a heart health assessment, your task is to examine the ECG, identify all fitting options from the provided list, and enumerate them using semicolon: 1dAVb(1st degree av block), RBBB(right bundle branch block), LBBB(left bundle branch block), SB(sinus bradycardia), ST(sinus tachycardia), AF(atrial fibrillation). Only answer based on the given options without any explanation.</p> <p><b>Response:</b> RBBB(right bundle branch block)</p>	
<b>Sub-task:</b> Conduction Abnormality Detection; <b>Question Type:</b> Close-ended QA; <b>Source Dataset:</b> CODE-15%	

Figure A2: The Examples of heart rhythm analysis instructions for finetuning PULSE.

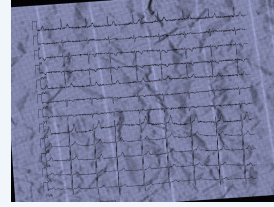
## Training Data Examples: Morphology and Pathological Condition Identification

### Instruction:

What specific features of the ECG tracing suggest the presence of a right bundle branch block, and how do these features relate to the overall cardiac conduction pattern, considering the morphology of the QRS complexes in leads V1 and V2? Provide a detailed explanation in 2-3 sentences.

### Response:

The ECG tracing exhibits a right bundle branch block (RBBB) due to the characteristic wide S waves in leads I and V6, accompanied by QS complexes in leads V2, indicative of delayed right ventricular depolarization. The QRS complexes in lead V1 display a prolonged duration, exceeding 140 ms, with a characteristic "M" or "W" shape, further supporting the diagnosis of RBBB. This conduction pattern is a result of a block in the right bundle branch, leading to an altered sequence of ventricular activation, causing the delayed and abnormal depolarization of the right ventricle.



**Sub-task:** Wave Morphology Analysis; **Question Type:** Open-ended; **Source Dataset:** PTB-XL

### Instruction:

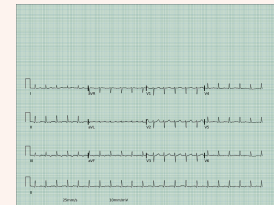
Answer the following multiple choice question. Think step by step before answering. What is the likely location of a previous myocardial infarction (MI) based on this ECG?

### Options:

A. Anterior wall B. Lateral wall C. Inferior wall D. Septal wall

### Response:

The ECG shows a pattern consistent with a previous MI, as evidenced by the abnormal Q-waves and ST-segment changes. The location of the MI can be inferred by the leads involved. We can observe that the inferior leads (II, III, and aVF) show the most prominent changes, indicating that the MI likely occurred in the inferior wall. Based on this, we select C. Inferior wall



**Sub-task:** Interpretation of Ischemia or Infarction; **Question Type:** Multi-choice QA; **Source Dataset:** MIMIC-IV-ECG

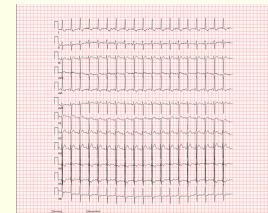
### Instruction:

Compare the QRS complex morphology in leads V1-V4 to that of a normal sinus rhythm ECG. Explain the significance of the observed differences, considering the context of intraventricular conduction disorder and delayed R progression.

Then, fill in the blanks: The QRS complex morphology in leads V1-V4 is most indicative of \_\_\_\_\_. (Provide a detailed explanation in 2-3 sentences before filling in the blank.)

### Response:

The QRS complex morphology in leads V1-V4 exhibits a notable difference from that of a normal sinus rhythm ECG. Specifically, the delayed R progression and intraventricular conduction disorder in V4 suggest a disruption in the normal ventricular depolarization sequence, resulting in altered QRS complex morphology. This deviation is significant because it indicates a pathological EKG pattern, which may be associated with underlying cardiac conditions such as left ventricular hypertrophy. The QRS complex morphology in leads V1-V4 is most indicative of **pathological left ventricular conduction**.



**Sub-task:** Pathological Conduction Identification; **Question Type:** Fill-in-the-blank; **Source Dataset:** PTB-XL

### Instruction:

Imagine you are an expert cardiologist reviewing this ECG image. Considering the patient's history and this ECG, what could have contributed to the development of the prolonged QT interval? Rank your top three possibilities in order of likelihood, using a semicolon-separated list.

### Response:

Hypokalemia; certain medications or drugs; genetic mutations or channelopathies.



**Sub-task:** Electrolyte imbalances Interpretation; **Question Type:** Open-ended QA; **Source Dataset:** PTB-XL

Figure A3: The Examples of morphology and pathological condition identification instructions for finetuning PULSE.

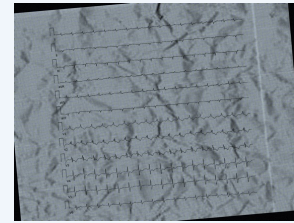
## Training Data Examples: Clinical Reporting

### Instruction:

Review this electrocardiogram image and provide your expert diagnosis.

### Response:

The electrocardiogram (ECG) reveals a sinus tachycardia, characterized by a heart rate exceeding 100 beats per minute, indicating an abnormal acceleration of the heart rhythm. This is further accentuated by the presence of an exaggerated left type, suggesting an increased electrical activity in the left ventricle. Additionally, peripheral hypotension is noted, indicating lower-than-normal blood pressure in the peripheral arteries, which may be a concern as it can lead to inadequate perfusion of vital organs. The ECG also demonstrates an incomplete right bundle branch block (RBBB), which suggests a delay or obstruction in the electrical conduction pathway of the right ventricle. Furthermore, the QRS complex is abnormal, with an altered morphology and duration, indicating a disruption in the ventricular depolarization process. The presence of an old inferior infarct is likely, suggesting a previous myocardial infarction (MI) in the inferior wall of the heart. Moreover, the ECG shows signs of old ST and T wave abnormalities, indicating possible previous episodes of ischemia or injury to the myocardium. The ECG also suggests likely anterior ischemia or left ventricular strain, which may indicate ongoing or previous damage to the anterior wall of the heart.



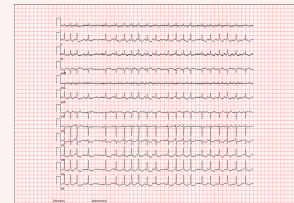
**Question Type:** Open-ended; **Source Dataset:** PTB-XL

### Instruction:

Examine the ECG signal image and produce a thorough report based on your findings, ending with the final diagnosis.

### Response:

The electrocardiogram (ECG) reveals atrial fibrillation with a rapid ventricular response, indicating a chaotic and rapid heart rhythm originating from the atria. Additionally, there are signs of a possible anterior myocardial infarction, although the age of the infarct cannot be determined from this ECG. Furthermore, ST-T changes in the inferior and lateral leads suggest myocardial ischemia, which may be partly attributed to the rapid heart rate and rhythm, but also raises concerns for underlying coronary artery disease. Final Diagnosis: Atrial fibrillation with rapid ventricular response, possible anterior myocardial infarction, and signs of myocardial ischemia.



**Question Type:** Open-ended; **Source Dataset:** MIMIC-ECG

Figure A4: The Examples of clinical reporting instructions for finetuning PULSE.

## **E Prompts**

The prompt used to synthesize instruction tasks (Fig. A5). The prompt used to synthesize ECG multi-turn dialogue (Fig. A6). The prompt used to revise (and translate) original reports (Fig. A7). The prompt used to score and filter generated instruction data (Fig. A8). The prompt used to evaluate the generated report (Fig. A9). The prompt used to evaluate the ECG Arena (Fig. A10).



Prompt: Multi-task Data Synthesizing
<p>Your task: Create a complex ECG visual task based on the given report and target task type:</p> <p>Guidelines for task creation:</p> <ol style="list-style-type: none"> <li>Design a concise yet challenging graduate-level task that requires deep reasoning.</li> <li>Frame the task as interacting with an actual ECG image, without mentioning the report. Make the task visually centric, assuming direct ECG image analysis.</li> <li>Strictly base all information on the given ECG report only. Avoid tasks and answers that are inconsistent with the report.</li> <li>Avoid restating the report or using phrases like "As described in the report."</li> <li>Generate one task from a diverse range of task types, including but not limited to: <ul style="list-style-type: none"> <li>Direct questions (e.g. "What is the heart rhythm?")</li> <li>Hypothetical scenarios (e.g. "Imagine you're an ER doctor reviewing this ECG...")</li> <li>Comparative tasks (e.g. "How does this ECG differ from a normal sinus rhythm?")</li> <li>Explanation requests (e.g. "Explain the significance of the QS complexes seen in V2.")</li> <li>Problem-solving scenarios (e.g. "Given these ECG findings, what further tests might you order?")</li> <li>Educational prompts (e.g. "Teach a medical student about the key features of this ECG.")</li> <li>Role-playing scenarios (e.g. "You're consulting with a cardiologist about this ECG. What do you tell them?")</li> <li>Decision-making tasks (e.g. "Based on this ECG, would you clear this patient for surgery? Why or why not?")</li> </ul> </li> <li>Specify a clear, appropriate output format within the task instructions (free-form, "think-step-by-step", direct output the short answer (in one phrase or one sentence), JSON format, table, list, different delimiters (such as commas, semicolons, numeric order), etc.). Do not limited to the given task type and format, you have the freedom to design any type of task you deem appropriate.</li> <li>Focus the task on one or more of the following ECG analysis aspects: <ol style="list-style-type: none"> <li>Basic ECG feature interpretation (e.g. heart rate, rhythm, cardiac axis)</li> <li>Diagnosis and classification (e.g. diagnosis identification, waveform classification, rhythm classification)</li> <li>Waveform and interval analysis (e.g. P wave morphology, PR interval, QT interval, QRS complexes, T wave morphology)</li> </ol> </li> <li>Ensure the task complexity aligns with the given report's information.</li> </ol> <p>After creating the task:</p> <ol style="list-style-type: none"> <li>Provide a detailed, accurate answer to your own task.</li> <li>Ensure your answer is comprehensive and strictly based on the report.</li> <li>Strictly follow the output format and requirements specified in your task instructions.</li> </ol> <p>ECG Report: {report}</p> <p>Target Task Type: {target}</p> <p>Present your work in this format:</p> <p>Task: [Concise content of the ECG tasks, including required output format. Do not include phrases like "Output format:..." or like "[Insert image here]", but in more natural expression. ]</p> <p>Response: [Comprehensive response following the task's requirements, strictly based on the report]</p> <p>Do not include any content outside of the Task and Response sections.</p>

Figure A5: The prompt used to synthesize ECG instruction tasks based on clinical reports.

Prompt: Multi-turn Dialogue Synthesizing
<p>Your task: Create a 2-4 turn dialogue between a medical professional and an AI assistant analyzing an ECG, based on the given report:</p> <p>Guidelines for dialogue creation:</p> <ol style="list-style-type: none"> <li>Design a series of questions and answers that progressively explore the ECG findings in depth, suitable for graduate-level medical professionals.</li> <li>Frame the dialogue as if the medical professional is directly analyzing an actual ECG image, without mentioning the report. Make the conversation visually centric, assuming direct ECG image analysis.</li> <li>Strictly base all information on the given ECG report only. Avoid including details inconsistent with the report.</li> <li>Do not use phrases like "As described in the report," "The report mentions," or "The term..." The dialogue should not appear to reference an external report.</li> <li>Begin with direct questions about basic ECG features, then progress to more complex interpretations and clinical implications.</li> <li>Include a mix of question types, with an emphasis on direct questions: <ul style="list-style-type: none"> <li>- Direct questions (e.g., "What are the main ECG features?", "What is the heart rhythm?")</li> <li>- Requests for explanations (e.g., "Can you explain the significance of the QS complexes?", "What the cause of these features?")</li> <li>- Clinical reasoning questions (e.g., "Given these findings, what's your diagnosis?")</li> <li>- Hypothetical scenarios (e.g., "How would you manage a patient presenting with this ECG?")</li> </ul> </li> <li>Focus the dialogue on one or more of the following ECG analysis aspects: <ol style="list-style-type: none"> <li>Basic ECG feature interpretation (e.g., heart rate, rhythm, cardiac axis)</li> <li>Diagnosis and classification (e.g. diagnosis identification, waveform classification, rhythm classification)</li> <li>Waveform and interval analysis (e.g. P wave morphology, PR interval, QT interval, QRS complexes, T wave morphology)</li> <li>Clinical implications and management</li> </ol> </li> <li>Ensure the dialogue complexity aligns with the given report's information.</li> </ol> <p>After creating the dialogue:</p> <ol style="list-style-type: none"> <li>Provide extremely comprehensive and detailed answers from the AI assistant's perspective. Each response should thoroughly cover all relevant aspects of the question asked.</li> <li>Ensure all answers are comprehensive and strictly based on the report, without explicitly referencing it.</li> <li>Make the dialogue flow naturally, as if a real user is progressively exploring the ECG findings.</li> <li>Structure the AI assistant's responses to be highly readable: <ul style="list-style-type: none"> <li>- Break down complex information into digestible parts.</li> <li>- Use bullet points or numbered lists to organize information</li> <li>- Include brief explanations of medical terms or concepts when necessary</li> <li>- Provide context for why certain findings are significant</li> </ul> </li> </ol> <p>Aim for a balance between depth of information and clarity of presentation in each response.</p> <p>ECG Report: {report}</p> <p>Present your work in this format:</p> <p>Human: [First question about the ECG]</p> <p>Assistant: [Comprehensive response based strictly on the report]</p> <p>Human: [Follow-up question delving deeper into the ECG analysis]</p> <p>Assistant: [Detailed answer providing further insights]</p> <p>[Continue the dialogue for up to 2 more turns if necessary, ensuring a natural progression of inquiry]</p> <p>Do not include any content outside of the dialogue format. Ensure that the entire conversation appears to be about analyzing an actual ECG image, without any indication that the information comes from a written report.</p>

Figure A6: The prompt used to synthesize ECG multi-turn dialogue as instruction-tuning data.

Prompt: Report Revision
<p>I will provide you with an ECG report. Please expand the report into a comprehensive and detailed version, considering all aspects mentioned in the original report. The expanded version should be at least 4 sentences long. Ensure that you elaborate on each point from the original report, providing more context and explanation where possible. Do not add any new content, interpretations, or conclusions beyond what is explicitly stated in the original report. Avoid using phrases like "Here is the revised report" or similar introductions. Simply begin with the expanded content.</p> <p>Original Report: {report} Expanded Report:</p>

Figure A7: The prompt used to revise (and translate) original reports.

Prompt: Instruction Data Scoring
<p>Task: Given an ECG report and a corresponding question-answer pair, score the quality of the answer based on the guidelines provided. The score should range from 0 to 5, where 0 represents poor quality and 5 represents excellent quality. You should be strict when giving the final assessment if some of the criteria are not satisfied. Please consider the following criteria for scoring:</p> <ol style="list-style-type: none"> <li><b>1. Relevance:</b> Does the answer directly address the question asked?</li> <li><b>2. Accuracy:</b> Is the information in the answer accurate and consistent with the ECG report?</li> <li><b>3. Usefulness:</b> Does the answer provide helpful information that would aid understanding or decision-making based on the ECG report?</li> <li><b>4. Constructed Information:</b> Does the answer invent details not present in the ECG report?</li> <li><b>5. Presence of Direct Report Quotation:</b> A good answer does not simply quote or directly replicate phrases from the ECG report. It should assume that the questioner does not know the report's specific content. The presence of direct report quotations is not allowed in the answer, otherwise, the overall scores should be at most 2.</li> </ol> <p>Output format: Please first output a single line containing a comprehensive explanation of your evaluation, avoiding any potential bias. In the subsequent line, please provide the value indicating the scores in the format: "Score: [your rating score]"</p> <p>Please apply the above scoring guide to the following ECG report and question-answer pair:</p> <p>ECG Report: {report} Question: {question} Answer: {answer}</p>

Figure A8: The prompt used to score and filter generated instruction data.

Prompt: Evaluation of Report Generation
<p>Evaluate the alignment and quality of a generated ECG report by comparing it to a ground truth clinician's report. The evaluation will focus on three key aspects: Diagnosis, Waveform, and Rhythm. Use specific criteria for each aspect and be precise in comparing medical terminologies. Only focus on information present in the ground truth report, identifying any mistakes. Remain objective and do not let the response length affect your evaluation.</p> <p>Evaluation Criteria:</p> <p><b>1. Diagnosis (0-10):</b>  Assess how well the generated ECG report matches the clinical diagnoses in the ground truth report. Focus on conditions like conduction disturbances, ischemia, hypertrophy, and other abnormalities as presented in the ground truth report.</p> <ul style="list-style-type: none"> <li>- 10: All key diagnoses are correctly identified with no errors or omissions.</li> <li>- 5: Partially accurate, with some diagnoses identified correctly but key conditions missing or incorrect.</li> <li>- 0: Fails to identify key diagnoses, with multiple critical errors.</li> </ul> <p><b>2. Waveform (0-10):</b>  Evaluate the accuracy and quality of the ECG waveform morphology in the generated report compared to the ground truth. Focus on abnormalities in P-wave, QRS complex, ST changes, T-wave, and intervals (PR, QT), ensuring waveform morphology is consistent with the ground truth.</p> <ul style="list-style-type: none"> <li>- 10: All waveform abnormalities are correctly identified without errors.</li> <li>- 5: Some waveform abnormalities are identified, but key issues are missed or misinterpreted.</li> <li>- 0: Fails to identify key waveform abnormalities, with multiple critical errors.</li> </ul> <p><b>3. Rhythm (0-10):</b>  Assess the accuracy and clarity of rhythm interpretation in the generated report. Focus on identifying and describing normal and abnormal rhythms (e.g., sinus rhythm, atrial fibrillation, ventricular tachycardia) as presented in the ground truth report.</p> <ul style="list-style-type: none"> <li>- 10: Rhythm interpretation is fully accurate and clearly described.</li> <li>- 5: Rhythm interpretation is partially accurate but contains notable errors or omissions.</li> <li>- 0: Rhythm interpretation is largely incorrect, with critical errors.</li> </ul> <p>Please organize your output in a JSON format of diagnosis, form and rhythm, with a brief explanation of each aspect. For example: {Diagnosis: {Score: \$SCORE\$, Explanation: \$EXPLANATION\$}}</p> <p>[The Start of Ground Truth Report]  {ground_truth_report}  [The End of Ground Truth Report]  [The Start of Generated Report]  {generated_report}  [The End of Generated Report]</p>

Figure A9: The prompt used to evaluate the generated report.



Prompt: Evaluation of ECG Arena
<p>Evaluate the quality of a model's response to an ECG-related question by comparing it with a given ground truth answer. Focus on three aspects: accuracy, completeness, and instruction adherence. Be precise and objective, especially when identifying errors in medical terminology. Do not let the response length affect your evaluation.</p> <p>Evaluation Criteria:</p> <p><b>1. Accuracy (0-10):</b>  How well does the model's response match the ground truth, particularly in ECG interpretation and diagnosis? This score emphasizes whether the key information is correct, such as the correct identification of waveforms, intervals, and clinical diagnoses.  - 10: Fully accurate, with correct ECG interpretation, terminology, and diagnosis.  - 5: Partially accurate, with some correct information but notable errors or omissions.  - 0: Largely inaccurate or misleading.</p> <p><b>2. Completeness (0-10):</b>  Does the response cover essential aspects of ECG interpretation (e.g., rhythm, axis, waveforms, clinical causes) mentioned in the ground truth? This score focuses on whether the answer is comprehensive and includes as much essential information as possible.  - 10: Comprehensive, covering all key details.  - 5: Partially complete, with important points missing.  - 0: Incomplete, lacking critical information.</p> <p><b>3. Instruction Adherence (0-10):</b>  Does the model follow the specific instructions in the question (e.g., listing features, suggesting a diagnosis)? This score focuses on how well the model follows the task instructions, regardless of the correctness of the answer.  - 10: Fully follows instructions.  - 5: Partially follows instructions, with some deviations.  - 0: Fails to follow instructions or provides an irrelevant response.</p> <p>Please organize your output in a JSON format of accuracy, completeness, and instruction adherence, with a brief explanation of each aspect. For example: {Accuracy: {Score: \$SCORE\$, Explanation: \$EXPLANATION\$}}</p> <p>[The Start of Ground Truth Answer]  {ground_truth_answer}  [The End of Ground Truth Answer]</p> <p>[The Start of Model's Response]  {model_response}  [The End of Model's Response]</p>

Figure A10: The prompt used to evaluate the ECG Arena.

## F Details of Evaluation Datasets

We provide the data curation process in Fig. A11 and details of each evaluation dataset in Table A2.

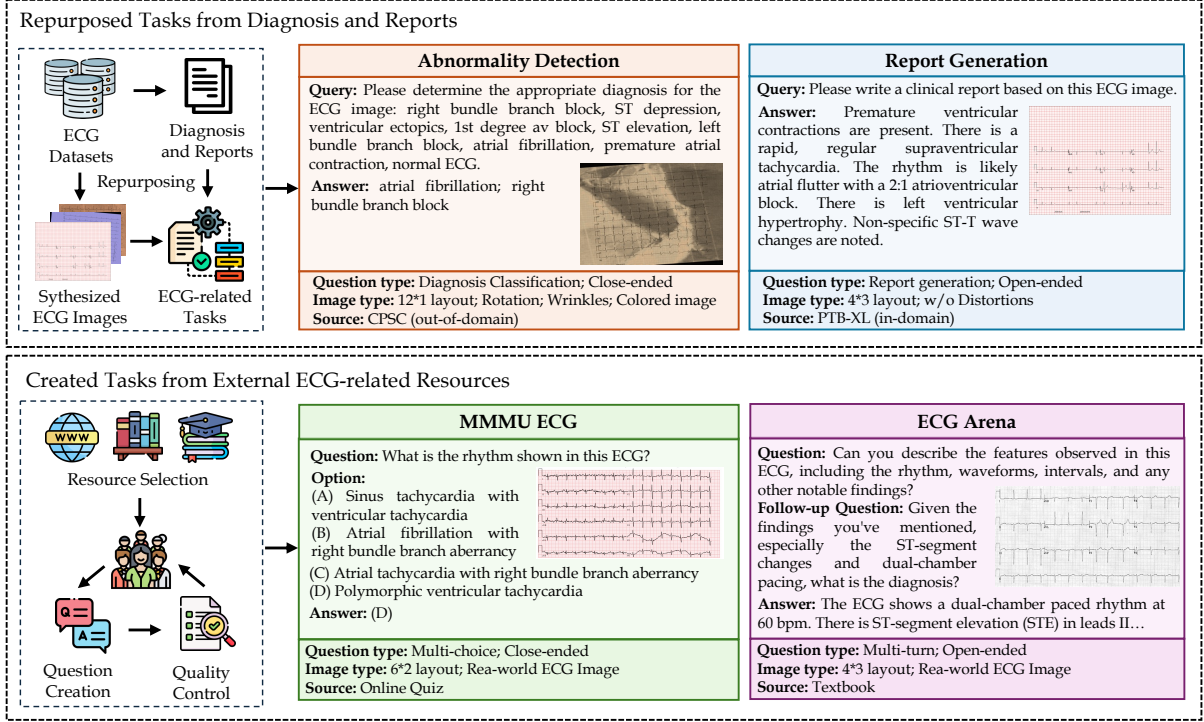


Figure A11: The data curation process for ECGBench. There are four key tasks involved: (1) two repurposed tasks (abnormality detection and report generation) derived from existing ECG datasets, where ECG images are synthesized from raw signals, and queries/answers are extracted based on diagnostic and clinical reports; (2) Two newly developed tasks using external resources, where ECG images and associated questions and answers are collected and generated from real-world sources.

Evaluation Dataset	Task	Type	# Samples	In-Domain?
PTB-XL Super	Abnormality Detection	Close-ended	2,082	YES
PTB-XL Report	Report Generation	Open-ended	500	YES
CODE-15%	Abnormality Detection	Close-ended	1,400	YES
ECG-QA	Abnormality Detection	Close-ended	1,317	YES
CPSC 2018	Abnormality Detection	Close-ended	2,061	NO
CSN	Abnormality Detection	MCQ (8-option)	1,611	NO
G12EC	Abnormality Detection	MCQ (8-option)	2,026	NO
MMMU ECG	Multimodal Understanding	MCQ (4-option)	200	NO
ECG Arena	Multi-turn Conversation	Open-ended	50	NO

Table A2: Overview of evaluation datasets in ECGBench. This collection contains both in-domain and out-of-domain problems across four key tasks with diverse answer types.

## G Details of Evaluation Metrics

**Abnormality Detection.** we utilize multi-label classification metrics, including Macro AUC, Macro F1, and Hamming Loss, to evaluate the datasets PTB-XL Super, CODE-15%, and CPSC 2018, where multiple correct labels may exist. For the ECG-QA, CSN, and G12EC datasets, we adopt accuracy as the evaluation metric.

**Report Generation.** Rather than relying on traditional text generation metrics, we leverage strong LLMs as evaluators, following the approach of [Zheng et al. \(2024\)](#). This method provides a more nuanced evaluation by focusing on key aspects of the reports. Specifically, we use GPT-4o to compare the model-generated reports against those written by cardiologists. We introduce a “Report Perfect Score”, which is based on three critical components of a generated report: (1) Rhythms (0 to 10 points), (2) Waveform Morphology (0 to 10 points), and (3) Diagnosis (0 to 10 points). The final score is the average of these three components, scaled to a maximum of 100 points. The prompt used to query GPT-4o for evaluating the report score is provided in Fig. A9.

**MMMU ECG.** We adopt accuracy as the primary metric. We have designed systematic, rule-based evaluation pipelines to ensure robust and consistent scoring. To mitigate the potential influence of any intermediate generations (e.g., reasoning steps) in long responses, we employ robust regular expressions and develop response-processing workflows. These are used to extract answer options from the long responses for accurate answer matching. In cases where no valid answer can be extracted from the model’s response, we perform random selection to assign a score.

**ECG Arena.** We also employ a strong judge model, GPT-4o, to assess model performance by comparing generated responses with ground truth answers. The evaluation considers three perspectives, each scored on a scale of 0-10: Accuracy (how closely the model’s response matches the ground truth), Completeness (whether the model provides a comprehensive answer covering all aspects of ECG interpretation), and Instruction Adherence (how well the model follows the specific instructions in the question). We calculate the final score by averaging these three aspects and scaling to a maximum of 100 points. The specific prompt used for GPT-4 evaluation is provided in Fig. A10.

## H Implementation Details

We follow the model architecture of LLaVA, which includes three core components: a vision encoder, a large language model, and a projector to align image and text modalities. Table A3 summarizes all the model parameters. Specifically, for the LLM, we utilize Vicuna-1.5-7B, while the vision encoder is based on CLIP-ViT-Large-Patch14-336. We employ a 2-layer MLP as a projector to map the visual features from the CLIP encoder onto the tokens used by the LLM. These features are mapped onto predefined image tokens, which encapsulate the features of ECG images. The tokens representing ECG features are then concatenated as an image context preceding the dialogue.

We format all datasets into a chatbot-style multi-turn dialogue format (same as Vicuna-1.5-7B) and use the special token <image> to represent image features within the text data. For example, a sample data instance is: “Human: <image> Describe this ECG image. \n Assistant: This image ...”. To enhance the model’s ability to handle ECG images of various sizes encountered in real-world scenarios, we employ Anyres. Anyres divides high-resolution images into multiple sub-images of size 336x336. The features of these sub-images are then concatenated with the global features of the original image to form the final image representation.

We fine-tune all parameters of the vision encoder, projector, and LLM. The training process uses a learning rate of  $2e-5$ , a batch size of 128, and a cosine scheduler with a 5% warm-up period over three epochs. The loss is calculated using the cross-entropy loss function, focusing on the response portion of the dialogue.

<b>Model Parameters</b>	
Total	7.06B
Vision Encoder(clip-vit-large-patch14-336)	303.5M
Connector	21M
LLM(Vicuna-1.5-7B)	6.74B
<b>Training Parameters</b>	
Learning Rate	2e-5
Weight Decay	0.0
Warmup Ratio	0.03
Learning Rate Scheduler	Cosine
Batch Size	128
<b>Vision Encoder Arch</b>	
Hidden Size	1024
Input Resolution	336
ViT Layer	24
ViT Heads	16
Patch Size	14
<b>LLM Arch</b>	
Hidden Size	4096
Max Context Length	4096
Attention Heads	32
Hidden Layers	32
KV Heads	32

Table A3: Summary of model parameters.

I Main Results

We provide the overall results in Fig. A12, results on all baselines in Table A4 for in-domain evaluation and Table A5 for out-of-domain evaluation.

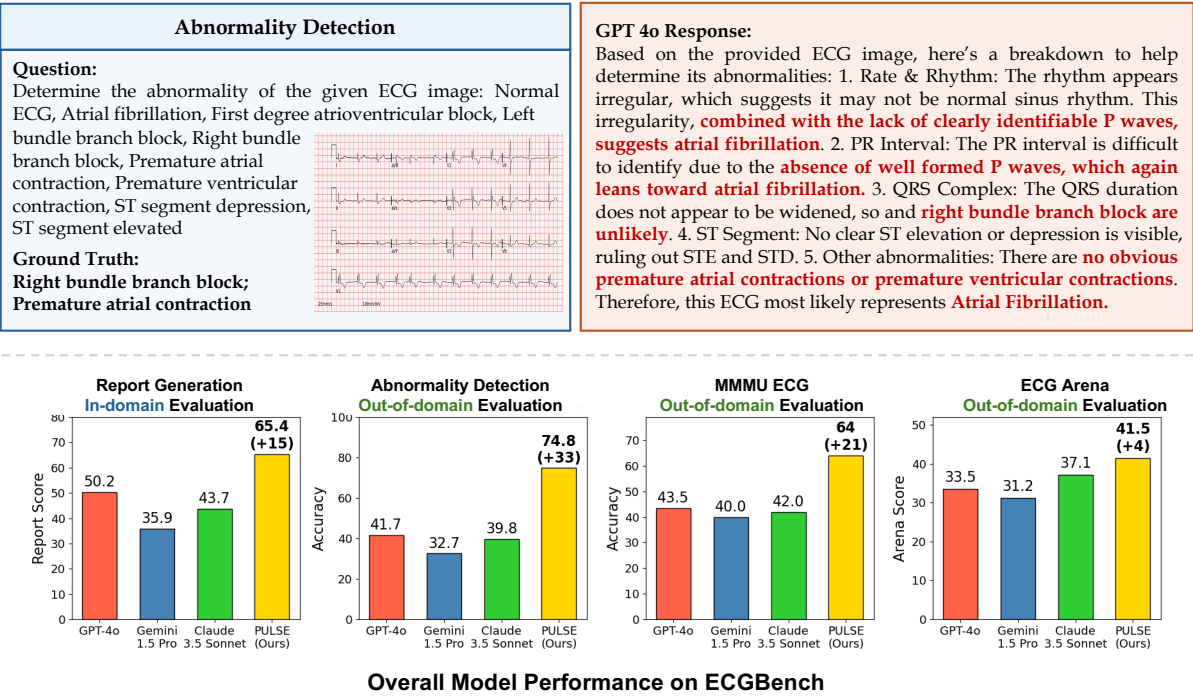


Figure A12: The proposed PULSE demonstrates superior performance across multiple in-domain and out-of-domain datasets on our constructed ECGBench compared with advanced proprietary MLLMs (e.g., GPT-4o). Notably, the proprietary MLLMs often fail to accurately interpret ECG images, generating well-structured and contextually relevant responses but ultimately incorrect (with errors highlighted in red) compared to the ground truth diagnosis.



Datasets	PTB-XL Super			PTB-XL Report	CODE-15%			ECG-QA
Metric	AUC	F1	HL	Report Score	AUC	F1	HL	Accuracy
Random	50.3	33.2	50.1	0	48.8	15.0	32.1	16.2
Domain-specific Methods								
METS	-	65.7 <sup>†</sup>	-	N/A	-	-	-	N/A
MERL	74.2 <sup>†</sup>	-	-	N/A	-	-	-	N/A
ST-MEM	71.4 <sup>†</sup>	-	-	N/A	-	-	-	N/A
MMCL	81.6	-	-	N/A	-	-	-	N/A
MOMENT	<b>83.3</b>	-	-	N/A	-	-	-	N/A
ECG-GPT	69.5*	53.9*	20.1*	47.8*	68.9*	40.1*	17.4*	N/A
Proprietary MLLMs								
GPT-4o	<u>55.6</u>	<u>28.3</u>	<u>26.2</u>	<u>50.2</u>	<u>59.9</u>	<u>24.9</u>	15.7	<u>35.2</u>
GPT-4o mini	52.0	20.4	31.7	37.1	57.5	22.0	<u>15.1</u>	14.9
Gemini 1.5 Pro	50.7	15.3	27.9	35.9	56.7	20.0	15.9	33.2
Claude 3.5 Sonnet	54.0	27.5	29.6	43.7	58.3	20.3	17.8	34.2
Open-source MLLMs								
LLaVA-Med	50.0	12.3	28.1	24.3	69.2	27.0	33.4	<u>29.5</u>
LLaVA-1.5-7B	50.0	12.3	28.1	27.2	63.9	19.2	25.3	25.2
LLaVA-1.5-13B	50.0	35.2	48.4	20.7	53.9	13.1	13.6	21.2
LLaVA-1.6-Vicuna-7B	50.0	15.8	29.4	16.5	50.1	1.0	13.6	13.3
LLaVA-1.6-Vicuna-13B	50.0	20.1	38.3	5.9	53.0	3.6	16.6	22.0
LLaVA-1.6-34B	50.2	19.9	36.0	17.0	57.2	12.8	16.6	22.4
LLaVA-OneVision-7B	49.8	11.4	34.5	30.0	58.7	17.0	20.6	20.4
LLaVA-OneVision-72B	50.6	29.6	50.4	40.6	52.3	7.0	<u>13.1</u>	25.0
Deepseek-VL-Chat-7B	50.9	15.7	27.9	15.6	63.7	<u>27.5</u>	22.4	21.1
Idefics2-8B	50.7	21.9	31.2	10.6	49.0	17.9	47.9	26.1
Mantis-8B-siglip-Llama3	50.6	20.4	30.0	16.0	57.5	17.9	15.7	23.8
MiniCPM-V-2.6	49.0	<u>37.7</u>	63.8	15.4	56.6	25.3	22.0	20.8
Phi-3-Vision-128k-Instruct	50.0	29.6	48.4	20.2	<u>69.6</u>	22.6	38.8	28.4
Qwen2-VL-7B	51.3	22.4	30.8	43.0	60.7	24.8	20.5	20.4
Qwen2-VL-72B	<u>54.0</u>	28.3	30.2	<u>48.9</u>	60.6	23.6	16.1	23.7
InternVL2-8B	50.6	14.3	<u>27.8</u>	38.1	55.8	16.1	17.7	22.3
InternVL2-40B	51.2	18.7	34.6	41.8	56.7	16.2	17.4	18.2
InternVL2-Llama3-76B	50.4	9.4	35.6	41.4	59.0	20.2	20.5	21.8
PULSE-7B (Ours)	<b>82.9</b>	<b>76.9</b>	<b>10.2</b>	<b>65.4</b>	<b>91.7</b>	<b>87.0</b>	<b>4.6</b>	<b>71.6</b>
Δ over best proprietary MLLM	+27	+49	+16	+15	+32	+62	+11	+36
Δ over best open-source MLLM	+29	+39	+18	+17	+22	+60	+9	+42

Table A4: In-domain evaluation results on representative baselines. <sup>†</sup> indicates results from original papers, \* denotes results obtained using the provided online software, N/A indicates methods not applicable or not designed for certain tasks, and - indicates unreported scores in original papers.

Datasets	CPSC 2018			CSN	G12EC	MMMU ECG	ECG Arena
Metric	AUC	F1	HL	Accuracy	Accuracy	Accuracy	Arena Score
Random	51.2	15.1	28.8	11.6	12.1	24.2	0
Domain-specific Methods							
METS	-	-	-	N/A	N/A	N/A	N/A
MERL	<b>82.8<sup>†</sup></b>	-	-	N/A	N/A	N/A	N/A
ST-MEM	70.4 <sup>†</sup>	-	-	N/A	N/A	N/A	N/A
MMCL	52.7	-	-	N/A	N/A	N/A	N/A
MOMENT	50.5	-	-	N/A	N/A	N/A	N/A
ECG-GPT	69.3*	44.0*	9.9*	N/A	N/A	N/A	N/A
Proprietary MLLMs							
GPT-4o	50.9	10.6	<u>18.2</u>	<u>57.5</u>	49.2	<u>43.5</u>	33.5
GPT-4o mini	49.2	11.0	25.5	32.1	33.2	39.5	30.1
Gemini-1.5-Pro	50.1	7.4	20.5	50.5	36.0	40.0	31.2
Claude 3.5 Sonnet	<u>52.8</u>	<u>11.5</u>	18.9	51.5	<u>51.4</u>	42.0	<u>37.1</u>
Open-source MLLMs							
LLaVA-Med	50.0	2.5	20.2	13.8	14.1	27.0	15.9
LLaVA-1.5-7B	50.0	2.5	20.0	32.1	25.4	33.0	12.7
LLaVA-1.5-13B	50.4	13.3	30.1	30.7	30.7	35.0	13.1
LLaVA-1.6-Vicuna-7B	50.5	<u>19.7</u>	66.0	23.7	23.3	28.0	16.0
LLaVA-1.6-Vicuna-13B	50.0	19.3	62.8	31.4	35.0	38.0	17.9
LLaVA-1.6-34B	49.6	19.3	62.8	44.3	<u>45.9</u>	31.0	17.5
LLaVA-OneVision-7B	49.6	8.0	28.3	23.3	25.7	26.0	22.5
LLaVA-OneVision-72B	51.5	12.8	29.4	44.0	42.6	35.0	15.5
Deepseek-VL-Chat-7B	50.7	6.0	20.0	35.7	32.9	34.5	15.3
Idefics2-8B	49.0	17.9	47.9	22.8	26.2	36.0	4.9
Mantis-8B-siglip-Llama3	51.3	19.1	48.5	17.6	22.6	<u>38.5</u>	13.6
MiniCPM-2.6	50.0	18.0	48.4	12.7	19.6	34.5	20.4
Phi-3-Vision-128k-Instruct	50.6	19.0	70.2	14.8	18.4	31.0	11.3
Qwen2-VL-7B	49.4	17.5	46.3	25.5	32.9	31.5	8.5
Qwen2-VL-72B	50.7	9.8	<u>18.9</u>	35.5	42.9	35.0	10.3
InternVL2-8B	52.1	8.2	22.2	<u>47.7</u>	37.5	30.0	22.9
InternVL2-40B	<u>52.4</u>	8.2	21.4	41.0	45.0	30.5	<u>28.0</u>
InternVL2-Llama3-76B	51.3	6.5	20.4	26.6	34.7	38.0	22.5
PULSE-7B (Ours)	80.7	<b>65.4</b>	<b>6.8</b>	<b>87.9</b>	<b>81.4</b>	<b>64.0</b>	<b>41.5</b>
Δ over best proprietary MLLM	+28	+54	+11	+30	+30	+21	+4
Δ over best open-source MLLM	+28	+46	+12	+40	+36	+26	+14

Table A5: Out-of-domain evaluation results. <sup>†</sup> indicates results from original papers, \* denotes results obtained using the provided online software, N/A indicates methods not applicable or not designed for certain tasks, and - indicates unreported scores in original papers.

## J Ablation Study

**Effect of Freezing the Vision Encoder.** We performed an ablation study by freezing the vision encoder parameters during training and reported the model’s performance in Table A6. The results indicate a further decline in performance, with the average score dropping from 71.2 to 68.1, compared to the original model with unfrozen vision encoder parameters.

Models	PTB-XL Super	PTB-XL Report	CSN	CODE-15	ECG-QA	CPSC	G12	MMMU ECG	ECG Arena	AVG
Unfrozen ViT	<b>76.9</b>	<b>65.4</b>	<b>87.9</b>	<b>87.0</b>	71.6	<b>65.4</b>	<b>81.4</b>	<b>64.0</b>	<b>41.5</b>	<b>71.2</b>
Frozen ViT	74.8	61.3	85.2	85.4	<b>73.8</b>	57.6	78.2	58.0	38.9	68.1

Table A6: Ablation results on the impact of freezing vision encoder parameters

**Fine-tuning results against other MLLMs** We fine-tune Qwen2-VL-7B using ECGInstruct and show the results in Table A7. The performance of the two backbone models is comparable, with LLaVA slightly outperforming Qwen.

Models	PTB-XL Super	PTB-XL Report	CSN	G12	MMMU ECG
PULSE (Qwen2-VL-7B)	75.1	62.8	85.2	79.2	60.4
PULSE (LLaVA-v1.6 -Vicuna-7B)	<b>76.9</b>	<b>65.4</b>	<b>87.9</b>	<b>81.4</b>	<b>64.0</b>

Table A7: Comparison of different MLLM backbones

**Effect of Instruction Task.** To understand the individual contribution of each ECG-related task to model performance, we analyze combinations of four instruction tasks. As shown in Table A8, adding more tasks progressively improves performance across multiple benchmarks. Models trained solely on basic feature recognition (F) performed poorly across all metrics, highlighting the limitations of a single-task approach. In contrast, the sequential addition of tasks led to substantial performance gains across multiple benchmarks. The model incorporating all four tasks achieved the highest performance, indicating a more comprehensive understanding of ECG images.

Instruction Task	PTB-XL Super	PTB-XL Report	CSN	CODE-15	ECG-QA	CPSC	G12	MMMU ECG	ECG Arena	AVG
F	14.4	40.1	59.3	12.8	52.6	10.3	14.4	40.0	15.0	28.8
F + R	29.0	58.1	86.5	74.9	59.2	38.8	70.5	53.5	27.9	55.4
F + R + M	72.5	61.7	87.9	84.3	66.4	51.6	74.2	58.5	33.0	65.6
F + R + M + C	76.9	65.4	87.9	87.0	71.6	65.4	81.4	64.0	41.5	<b>71.2</b>

Table A8: Performance of different ECG-related instruction task combinations. F: basic feature recognition, R: heart rhythm analysis, M: morphology and pathology identification, C: clinical report generation. F1 for PTB-XL Super, CODE-15%, and CPSC; Accuracy for CSN, ECG-QA, G12, and MMMU ECG; Report Scores for PTB-XL Report; Arena Scores for ECG Arena. **AVG** denotes the average across all metrics.

**K Human Expert Performance**

We engaged three domain experts specializing in ECG to evaluate a sample of 30 questions from the MMMU ECG. The performance comparison is provided in Table A9. Our results highlight a significant performance gap between current MLLMs and human expertise, indicating the need for further improvements in model capabilities for ECG image analysis.

Category	MMMU ECG (%)
Human Expert (Low)	70.0
Human Expert (Medium)	90.0
Human Expert (High)	93.3
Human Expert (Average)	84.4
GPT-4o	43.5
PULSE	64.0

Table A9: Comparison results on MMMU ECG among human experts, GPT-4o, and PULSE

**L Human Evaluation**

We performed a human evaluation on 50 sampled reports from the PTB-XL Report and 30 questions from the ECG Arena. Human evaluation scores and their Pearson correlation with LLM-based scores are presented in Table A10. The results indicate a strong correlation between the two sets of scores.

Models	PTB-XL Report (LLM score)	PTB-XL Report (Human score)	Pearsons Correlation	ECG Arena (LLM score)	ECG Arena (Human score)	Pearsons Correlation
GPT-4o	51.9	50.8	93.4	32.8	35.1	92.0
PULSE	62.8	64.1	91.9	37.4	39.0	91.7

Table A10: Comparison of human evaluation scores and LLM-based scores on report generation and ECG arena.

## **M Case Study**

We provide three examples of ECG report generation (Fig. [A13](#), Fig. [A14](#) and Fig. [A15](#)) and one example of ECG Arena (Fig. [A16](#)).



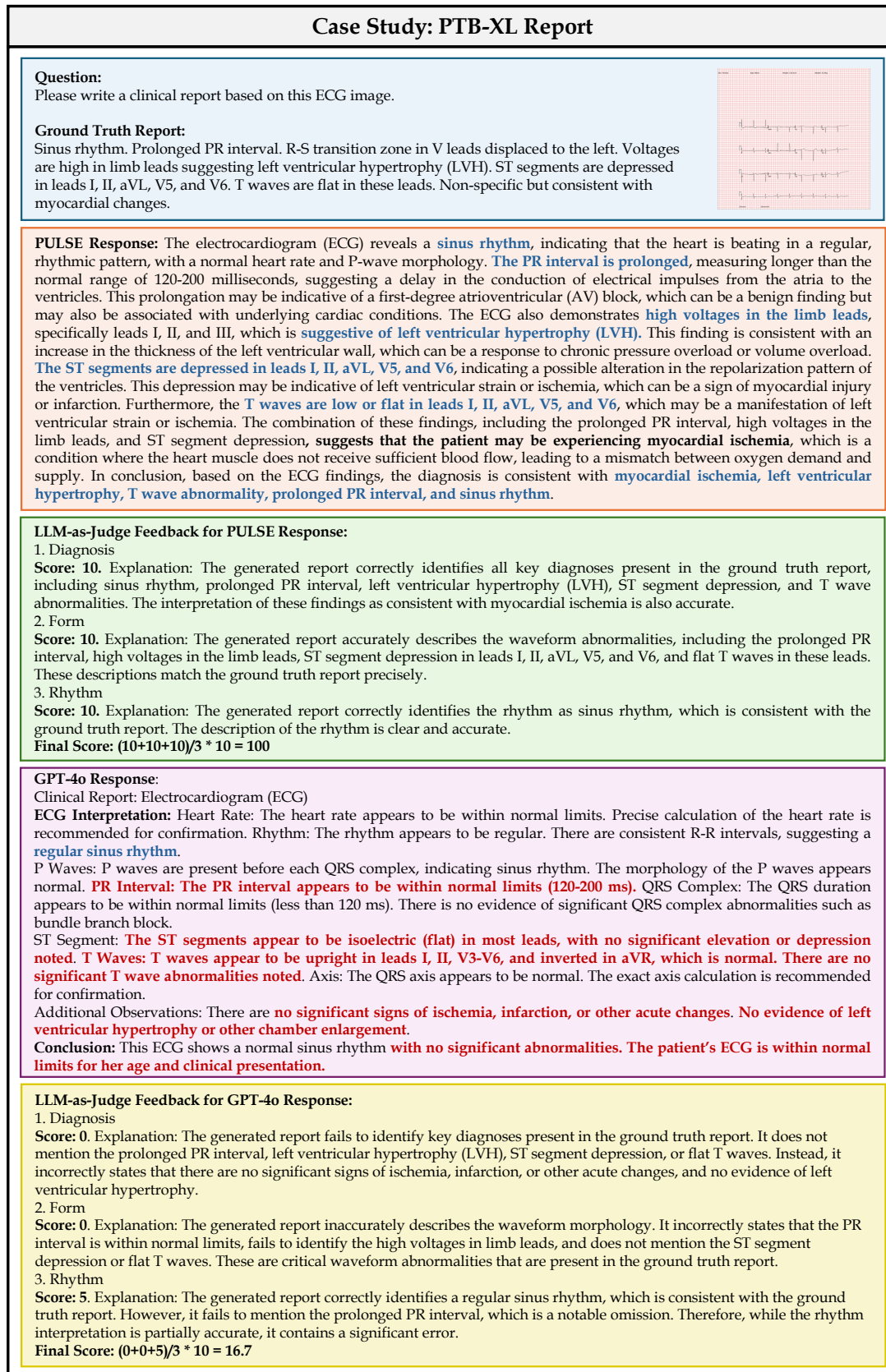


Figure A13: Comparison of model outputs on ECG report generation task (Example 1). Blue indicates correct information, while red highlights errors. Our model's output fully aligns with the ground truth, with a report score of 10. In comparison, GPT-4's report, though structurally sound, contains notable inaccuracies despite its initial appearance of relevance.

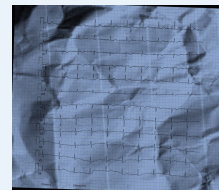
## Case Study: PTB-XL Report

### Question:

Please write a clinical report based on this ECG image.

### Ground Truth Report:

Sinus rhythm. T waves are low or flat in V3, V4, V5 and slightly inverted in V2. Non-specific but consistent with myocardial ischemia. The age of the changes is uncertain.



**PULSE Response:** The electrocardiogram (ECG) reveals a **normal sinus rhythm**, characterized by a regular heartbeat with a rate between 60-100 beats per minute, indicating that the heart's natural pacemaker, the sinoatrial (SA) node, is functioning properly. The axis of the heart's electrical activity is also within normal limits, suggesting that the electrical impulses are being conducted through the heart in a normal and coordinated manner. However, the ST segment and **T waves exhibit abnormalities, which may be indicative of myocardial ischemia or strain**, particularly in the anterior region of the heart. This could be a sign of reduced blood flow to the heart muscle, potentially leading to damage or even necrosis of the myocardial tissue. The presence of these abnormalities in the ST segment and T waves suggests that the heart is under stress, which may be caused by a variety of factors, including coronary artery disease, hypertension, or cardiomyopathy. Overall, the diagnosis based on this ECG report is **sinus rhythm** with ST changes, indicating that the heart is functioning normally in terms of rhythm, but is **exhibiting signs of potential myocardial ischemia or strain**.

### LLM-as-Judge Feedback for PULSE Response:

#### 1. Diagnosis

**Score: 8.** Explanation: The generated report correctly identifies the sinus rhythm and the presence of abnormalities in the ST segment and T waves, which are consistent with myocardial ischemia. However, it does not specifically mention the low or flat T waves in V3, V4, V5, and the slightly inverted T wave in V2, which are key details in the ground truth report.

#### 2. Form

**Score: 7.** Explanation: The generated report accurately identifies the presence of ST segment and T wave abnormalities, which are indicative of myocardial ischemia. However, it lacks the specific details about the T wave morphology in V2, V3, V4, and V5, which are crucial for a precise comparison with the ground truth report.

#### 3. Rhythm

**Score: 10.** Explanation: The generated report correctly identifies the sinus rhythm and provides a clear and accurate description of the heart's rhythm, matching the ground truth report.

**Final Score:  $(8+7+10)/3 * 10 = 83.3$**

### GPT-4o Response:

#### ECG Interpretation:

- Heart Rate: The heart rate is approximately 75 beats per minute, which is within the normal range (60-100 bpm).
- Rhythm: The rhythm appears to be regular with consistent R-R intervals, indicating a **normal sinus rhythm**.
- P Waves: P waves are present before each QRS complex, indicating normal atrial depolarization. The P wave morphology is normal.
- PR Interval: The PR interval is within the normal range (120-200 ms), suggesting normal atrioventricular conduction.
- QRS Complex: The QRS duration is within the normal range (less than 120 ms), indicating normal ventricular depolarization. There are no significant Q waves observed, which rules out previous myocardial infarction.
- ST Segment: The ST segments are isoelectric (flat) in all leads, indicating no signs of acute ischemia or injury.
- T Waves: **T waves are upright in leads I, II, V3-V6, and inverted in aVR, which is a normal finding. There are no signs of T wave inversion or hyperacute T waves.**
- QT Interval: The QT interval appears to be within the normal range (less than 440 ms for males and less than 460 ms for females), indicating no prolonged QT syndrome.
- Axis: The QRS axis is normal, approximately 0 to +90 degrees.
- Additional Observations: No evidence of left or right ventricular hypertrophy. No signs of bundle branch block or other conduction abnormalities.

Conclusion: The ECG findings are within normal limits. There are **no signs of acute ischemia, infarction**, or other significant cardiac abnormalities. The patient has a **normal sinus rhythm** with normal atrioventricular and intraventricular conduction.

### LLM-as-Judge Feedback for GPT-4o Response:

#### 1. Diagnosis

**Score: 0.** Explanation: The generated report fails to identify the key diagnosis of myocardial ischemia indicated by the T wave abnormalities in the ground truth report. It incorrectly concludes that there are no signs of acute ischemia or other significant cardiac abnormalities.

#### 2. Form

**Score: 0.** Explanation: The generated report does not mention the low or flat T waves in V3, V4, V5, and the slightly inverted T waves in V2, which are critical waveform abnormalities noted in the ground truth report. Instead, it incorrectly states that T waves are upright in V3-V6 and does not acknowledge any T wave inversion.

#### 3. Rhythm

**Score: 10.** Explanation: The generated report correctly identifies the rhythm as normal sinus rhythm, which is consistent with the ground truth report.

**Final Score:  $(0+0+10)/3 * 10 = 33.3$**

Figure A14: Comparison of model outputs on ECG report generation task (Example 2). Blue indicates correct information, while red highlights errors. Our model's output mostly aligns with the ground truth, achieving a report score of 83.3. In comparison, GPT-4's output correctly identifies only the ECG rhythm, omitting most other key details.

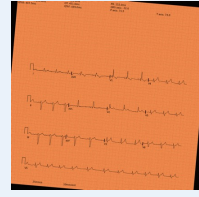
## Case Study: PTB XL Report

### Question:

Please write a clinical report based on this ECG image.

### Ground Truth Report:

sinus arrhythmia, left axis deviation, left anterior hemiblock, right bundle branch block, bifascicular block, abnormal QRS(T), possible lateral myocardial damage



**PULSE Response:** The electrocardiogram (ECG) reveals a **sinus arrhythmia**, characterized by an irregular heart rhythm with varying intervals between beats, originating from the sinoatrial node. This arrhythmia is often benign and can be influenced by factors such as respiration, with heart rate increasing during inhalation and decreasing during exhalation. The ECG also demonstrates an **overdriven left type**, indicating that the left ventricle is dominant in the cardiac conduction system, which can be a normal variant or associated with certain cardiac conditions. Furthermore, the ECG shows evidence of **left anterior hemiblock**, a conduction defect affecting the anterior fascicle of the left bundle branch, resulting in delayed activation of the anterior wall of the left ventricle. This can be a sign of underlying cardiac disease or injury. Additionally, the presence of **right bundle branch block** is noted, indicating a conduction defect affecting the right bundle branch, leading to delayed activation of the right ventricle. The combination of left anterior hemiblock and RBBB is consistent with **bifascicular block**, a condition where two of the three main fascicles of the cardiac conduction system are affected. The ECG also reveals a **left axis deviation**, indicating that the electrical axis of the heart is shifted to the left, which can be a normal variant or associated with certain cardiac conditions such as left ventricular hypertrophy or cardiomyopathy. Overall, the ECG findings are consistent with a diagnosis of **left axis deviation, sinus arrhythmia, left anterior fascicular block, and right bundle branch block**.

### LLM as Judge Feedback for PULSE Response:

#### 1. Diagnosis

**Score: 7.** Explanation: The generated report correctly identifies sinus arrhythmia, left axis deviation, left anterior hemiblock, and right bundle branch block. It also correctly identifies the combination of left anterior hemiblock and RBBB as bifascicular block. However, it fails to mention 'abnormal QRS(T)' and 'possible lateral myocardial damage,' which are key diagnoses.

#### 2. Form

**Score: 5.** Explanation: The generated report accurately describes the conduction defects (left anterior hemiblock, right bundle branch block) and left axis deviation. However, it does not address 'abnormal QRS(T)' and 'possible lateral myocardial damage,' which are critical waveform abnormalities mentioned in the ground truth report.

#### 3. Rhythm

**Score: 10.** Explanation: The generated report provides a detailed and accurate description of sinus arrhythmia, including its characteristics and potential influences. This matches the rhythm interpretation in the ground truth report.

**Final Score:  $(7+5+10)/3 * 10 = 73$**

### GPT 4o Response: Clinical Report: Electrocardiogram (ECG) Analysis:

- Heart Rate and Rhythm: Heart Rate: Approximately 75 beats per minute (bpm). Rhythm: **Sinus rhythm is present**, as indicated by the presence of P waves before each QRS complex and a consistent PR interval.
- Intervals and Durations: PR Interval: 152 ms (within normal range: 120-200 ms). QRS Duration: 125 ms (slightly prolonged; normal range: 80-120 ms). QT Interval: 421 ms (QT interval, QTcB: 447.5 ms, QTcF: 439.0 ms; within normal limits for QTc).
- Axis: QRS Axis: **72.0 degrees (left axis deviation)**. P Axis: 71.5 degrees. T Axis: 73.0 degrees.
- Waveform Analysis: P Waves: Normal morphology, consistent in all leads. QRS Complexes: Slightly widened, indicating a possible intraventricular conduction delay. ST Segment: No significant elevation or depression noted. T Waves: Normal morphology, no inversion or significant abnormalities observed.
- Additional Observations: Lead V1-V6: Normal R wave progression. No evidence of acute ischemia or infarction. No significant arrhythmias detected.

#### Conclusion:

The ECG shows a **sinus rhythm with a heart rate of approximately 75 bpm**. There is a slight prolongation of the QRS duration, suggesting a possible intraventricular conduction delay. The QRS axis is deviated to the left. No acute ischemic changes or significant arrhythmias are noted. The overall findings are within normal limits, with minor deviations that may warrant further clinical correlation and follow up.

### LLM as Judge Feedback for GPT 4o Response:

#### 1. Diagnosis

**Score: 5.** Explanation: The generated report correctly identifies left axis deviation but misses several key diagnoses present in the ground truth report, such as sinus arrhythmia, left anterior hemiblock, right bundle branch block, bifascicular block, abnormal QRS(T), and possible lateral myocardial damage. The generated report mentions a possible intraventricular conduction delay, which is not explicitly stated in the ground truth report.

#### 2. Form

**Score: 5.** Explanation: The generated report correctly identifies the left axis deviation and slightly prolonged QRS duration. However, it fails to identify the abnormal QRS(T) and possible lateral myocardial damage mentioned in the ground truth report. The generated report also does not mention any abnormalities in the P-wave, ST changes, or T-wave, which are critical for a comprehensive waveform analysis.

#### 3. Rhythm

**Score: 5.** Explanation: The generated report identifies a sinus rhythm but misses the sinus arrhythmia mentioned in the ground truth report. The generated report does not identify any significant arrhythmias, which is a notable omission given the ground truth report's mention of sinus arrhythmia.

**Final Score:  $(5+5+5)/3 * 10 = 50$**

Figure A15: Comparison of model outputs on ECG report generation task (Example 3). **Blue** indicates correct information, while **red** highlights errors. Our model's output mostly aligns with the ground truth report, achieving a report score of 73. In comparison, GPT-4's output partially aligns with the ground truth report.

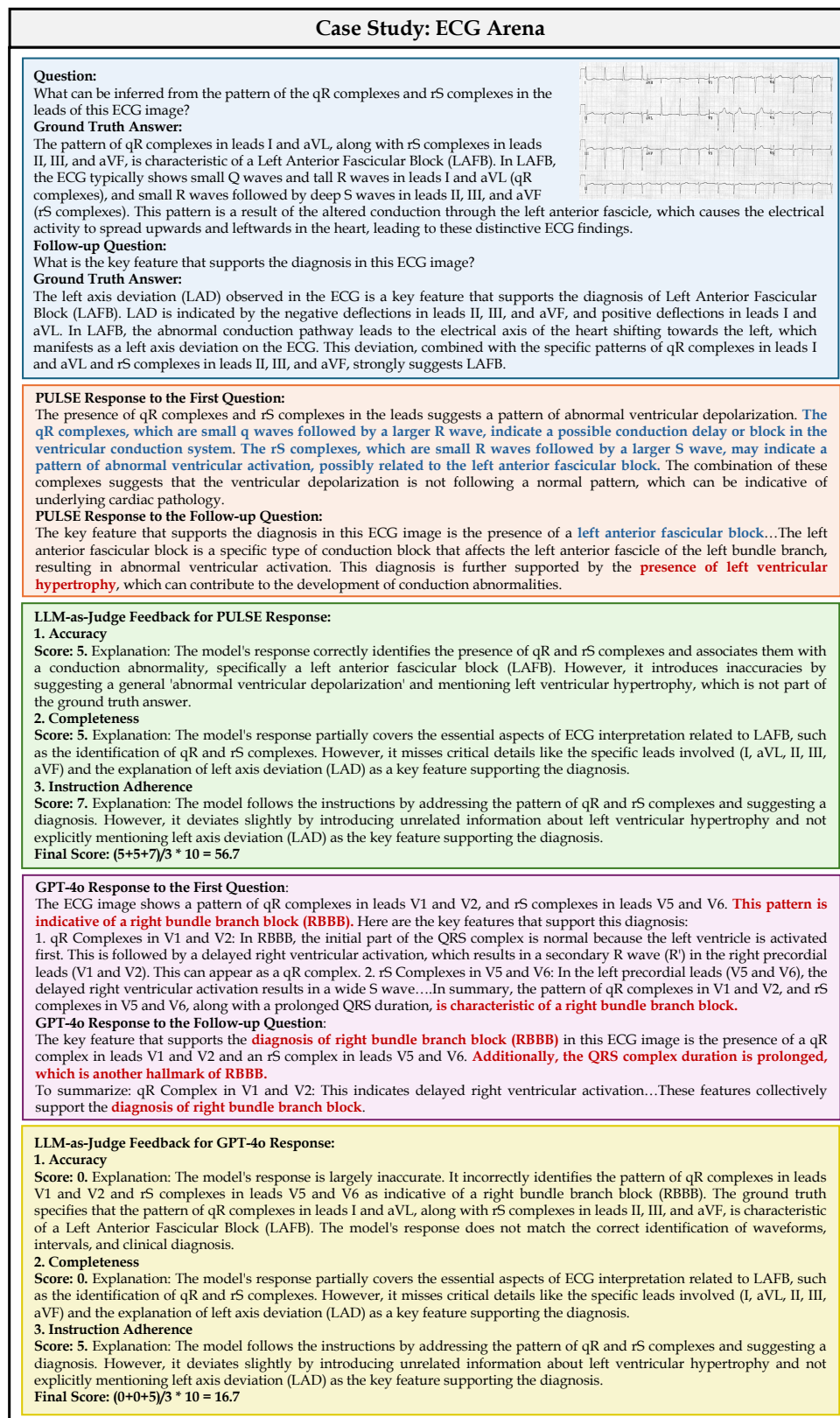


Figure A16: Comparison of model outputs on ECG Arena (Example 1). **Blue** indicates correct information, while **red** highlights errors. Given the challenging nature of this task, our model's output partially aligns with the ground truth, and GPT-4o's output largely deviates from the reference.