GEM:
Empowering MLLM for Grounded ECG Understanding with Time Series and Images

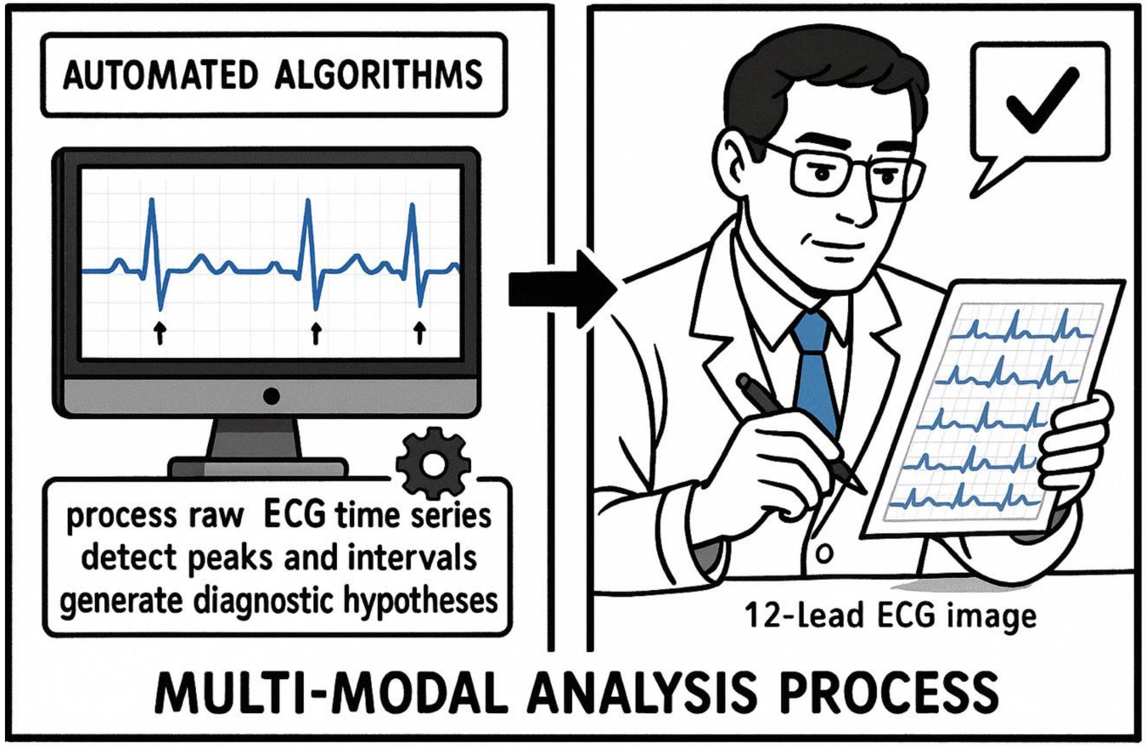
Xiang Lan, PhD

NeurIPS 2025

Motivation

Real-world ECG interpretation is naturally a multi-modal analysis process

1. Machines use algorithms to process raw ECG time series and generate diagnostic hypotheses
2. Cardiologists validate these findings by analyzing 12-lead ECG images


AUTOMATED ALGORITHMS
process raw ECG time series
detect peaks and intervals
generate diagnostic hypotheses
12-Lead ECG image
MULTI-MODAL ANALYSIS PROCESS

Salerno, Stephen M., Patrick C. Alguire, and Herbert S. Waxman. "Training and competency evaluation for interpretation of 12-lead electrocardiograms: recommendations from the American College of Physicians." *Annals of internal medicine* 138.9 (2003): 747-750.

Motivation

Limitation of current ECG models

Single Modality

- Often just time series
- Overlook the benefits of combining all available modalities

No Language Capability

- Inconvenient for both doctors and patients to use

Lack Explainability

- Fail to clearly link predictions to specific waveform features or articulate their reasoning process.

Objectives

1. Language-based ECG interpretation model
    - User friendly (for both clinicians and patients)

2. Make use of different modalities of ECG data
    - Time Series: high resolution
    - Image: easy to use

3. Provide high-granularity ECG interpretation
    - Feature-Grounded Analysis
    - Evidence-Driven Diagnosis
    - Realistic Interpretation Process

High-Granularity ECG Interpretation

- Feature-Grounded Analysis:

  Findings are explicitly tied to detailed ECG features like waveforms and intervals

- Evidence-Driven Diagnosis:

  Conclusions are supported by clear and logical reasoning directly linked to ECG findings

- Realistic Interpretation Process:

  Simulate how a clinician analyzes ECGs and arrive at a diagnosis

Challenges

- Data:

  No existing high-granularity ECG instruction data for LLM

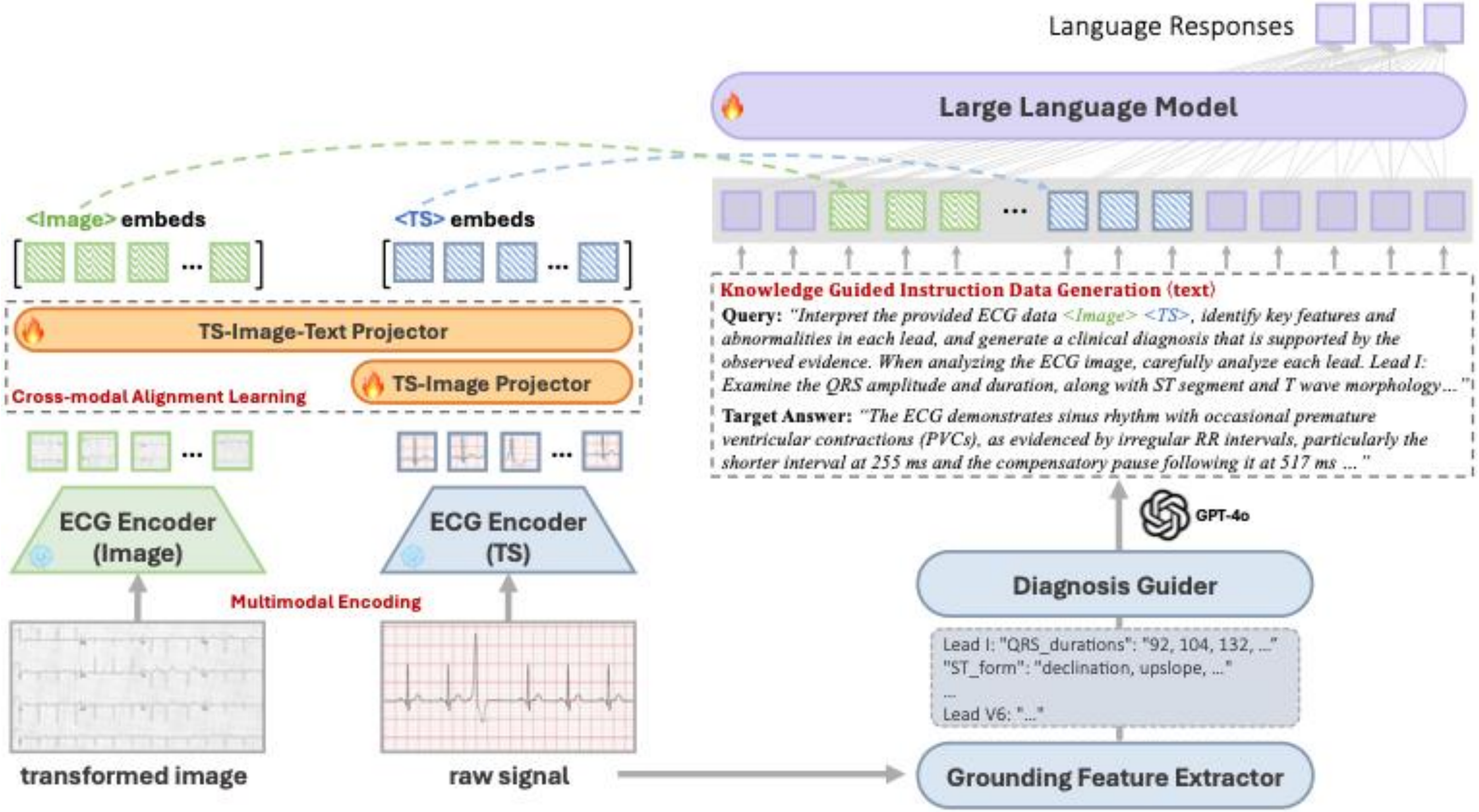
- Model:

  Need integrates ECG signal, ECG image, and text simultaneously

Methods Overview

Data: curate large-scale high-granularity instruction data.

Model: encode image and time series into the same language space.



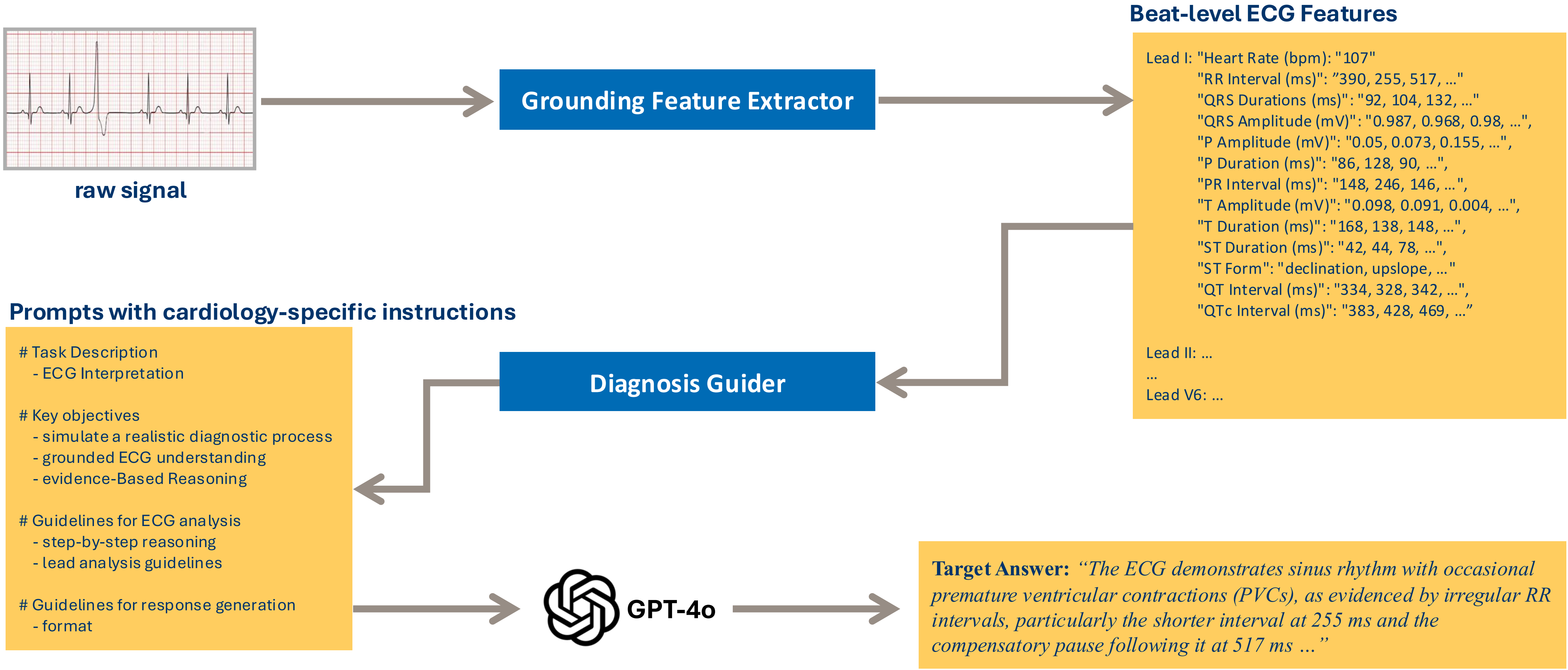
Knowledge Guided Instruction Data Generation (text)

Query: *"Interpret the provided ECG data <Image> <TS>, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence. When analyzing the ECG image, carefully analyze each lead. Lead I: Examine the QRS amplitude and duration, along with ST segment and T wave morphology…"*

Target Answer: *"The ECG demonstrates sinus rhythm with occasional premature ventricular contractions (PVCs), as evidenced by irregular RR intervals, particularly the shorter interval at 255 ms and the compensatory pause following it at 517 ms …"*

Lead I: "QRS_durations": "92, 104, 132, …"
"ST_form": "declination, upslope, …"
…
Lead V6: "…"

Knowledge-Guided Instruction Data Generation



raw signal

Grounding Feature Extractor

Beat-level ECG Features

```
Lead I: "Heart Rate (bpm): "107"
        "RR Interval (ms)": ”390, 255, 517, …"
        "QRS Durations (ms)": "92, 104, 132, …"
        "QRS Amplitude (mV)": "0.987, 0.968, 0.98, …",
        "P Amplitude (mV)": "0.05, 0.073, 0.155, …",
        "P Duration (ms)": "86, 128, 90, …",
        "PR Interval (ms)": "148, 246, 146, …",
        "T Amplitude (mV)": "0.098, 0.091, 0.004, …",
        "T Duration (ms)": "168, 138, 148, …",
        "ST Duration (ms)": "42, 44, 78, …",
        "ST Form": "declination, upslope, …"
        "QT Interval (ms)": "334, 328, 342, …",
        "QTc Interval (ms)": "383, 428, 469, …”

Lead II: …
…
Lead V6: …
```

Diagnosis Guider

Prompts with cardiology-specific instructions

# Task Description
  - ECG Interpretation

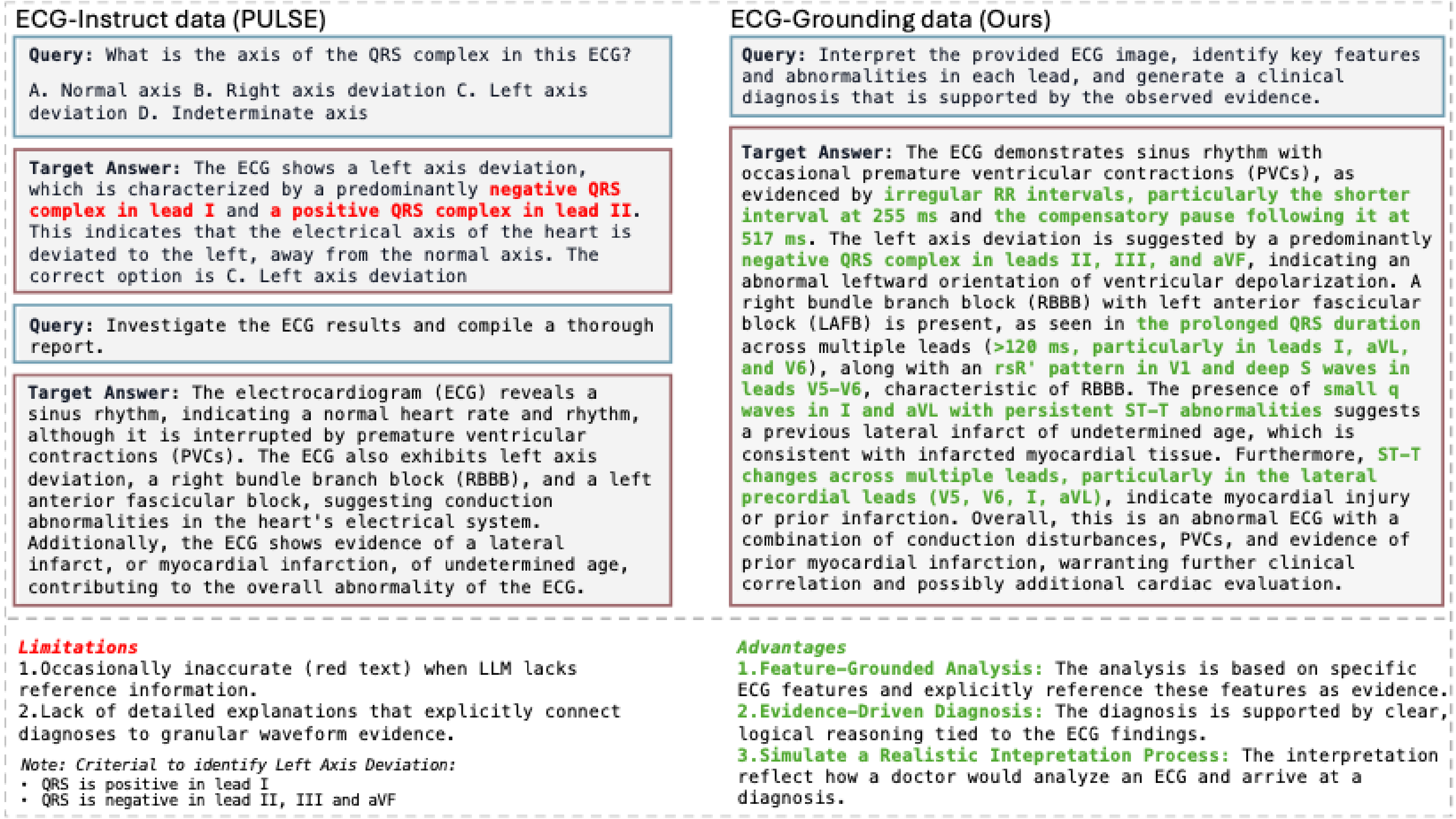
# Key objectives
  - simulate a realistic diagnostic process
  - grounded ECG understanding
  - evidence-Based Reasoning

# Guidelines for ECG analysis
  - step-by-step reasoning
  - lead analysis guidelines

# Guidelines for response generation
  - format

Target Answer: *“The ECG demonstrates sinus rhythm with occasional premature ventricular contractions (PVCs), as evidenced by irregular RR intervals, particularly the shorter interval at 255 ms and the compensatory pause following it at 517 ms …”*

Knowledge-Guided Instruction Data Generation



ECG-Instruct data (PULSE)

Query: What is the axis of the QRS complex in this ECG?

A. Normal axis B. Right axis deviation C. Left axis deviation D. Indeterminate axis

Target Answer: The ECG shows a left axis deviation, which is characterized by a predominantly **negative QRS complex in lead I** and **a positive QRS complex in lead II**. This indicates that the electrical axis of the heart is deviated to the left, away from the normal axis. The correct option is C. Left axis deviation

Query: Investigate the ECG results and compile a thorough report.

Target Answer: The electrocardiogram (ECG) reveals a sinus rhythm, indicating a normal heart rate and rhythm, although it is interrupted by premature ventricular contractions (PVCs). The ECG also exhibits left axis deviation, a right bundle branch block (RBBB), and a left anterior fascicular block, suggesting conduction abnormalities in the heart's electrical system. Additionally, the ECG shows evidence of a lateral infarct, or myocardial infarction, of undetermined age, contributing to the overall abnormality of the ECG.

Limitations

1.Occasionally inaccurate (red text) when LLM lacks reference information.
2.Lack of detailed explanations that explicitly connect diagnoses to granular waveform evidence.

Note: Criterial to identify Left Axis Deviation:

- QRS is positive in lead I
- QRS is negative in lead II, III and aVF

ECG-Grounding data (Ours)

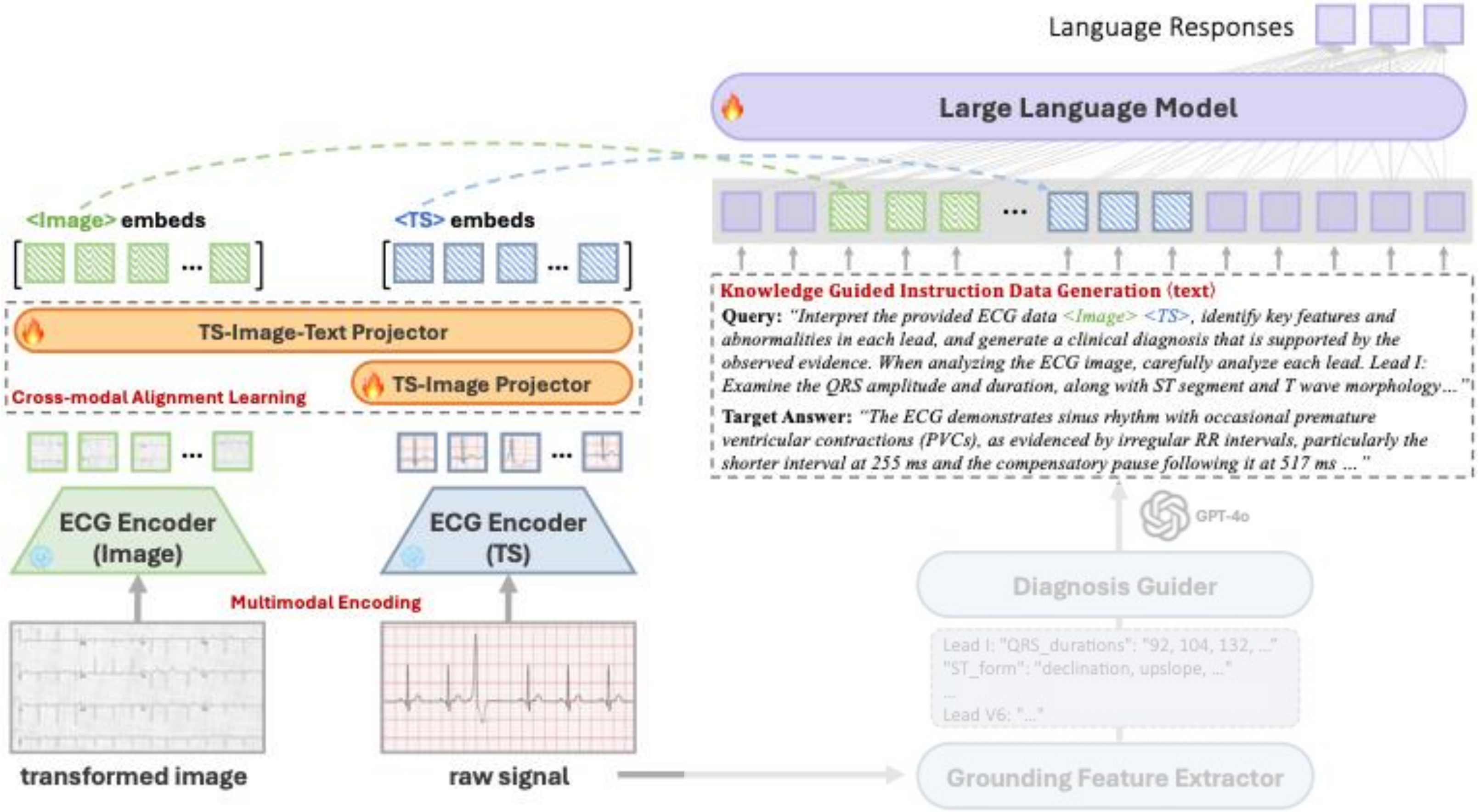
Query: Interpret the provided ECG image, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence.

Target Answer: The ECG demonstrates sinus rhythm with occasional premature ventricular contractions (PVCs), as evidenced by **irregular RR intervals, particularly the shorter interval at 255 ms** and **the compensatory pause following it at 517 ms**. The left axis deviation is suggested by a predominantly **negative QRS complex in leads II, III, and aVF**, indicating an abnormal leftward orientation of ventricular depolarization. A right bundle branch block (RBBB) with left anterior fascicular block (LAFB) is present, as seen in **the prolonged QRS duration** across multiple leads (**>120 ms, particularly in leads I, aVL, and V6**), along with an **rsR' pattern in V1 and deep S waves in leads V5-V6**, characteristic of RBBB. The presence of **small q waves in I and aVL with persistent ST-T abnormalities** suggests a previous lateral infarct of undetermined age, which is consistent with infarcted myocardial tissue. Furthermore, **ST-T changes across multiple leads, particularly in the lateral precordial leads (V5, V6, I, aVL)**, indicate myocardial injury or prior infarction. Overall, this is an abnormal ECG with a combination of conduction disturbances, PVCs, and evidence of prior myocardial infarction, warranting further clinical correlation and possibly additional cardiac evaluation.

Advantages

1.Feature-Grounded Analysis: The analysis is based on specific ECG features and explicitly reference these features as evidence.
2.Evidence-Driven Diagnosis: The diagnosis is supported by clear, logical reasoning tied to the ECG findings.
3.Simulate a Realistic Intepretation Process: The interpretation reflect how a doctor would analyze an ECG and arrive at a diagnosis.

Multimodal Encoding & Cross-modal Learning

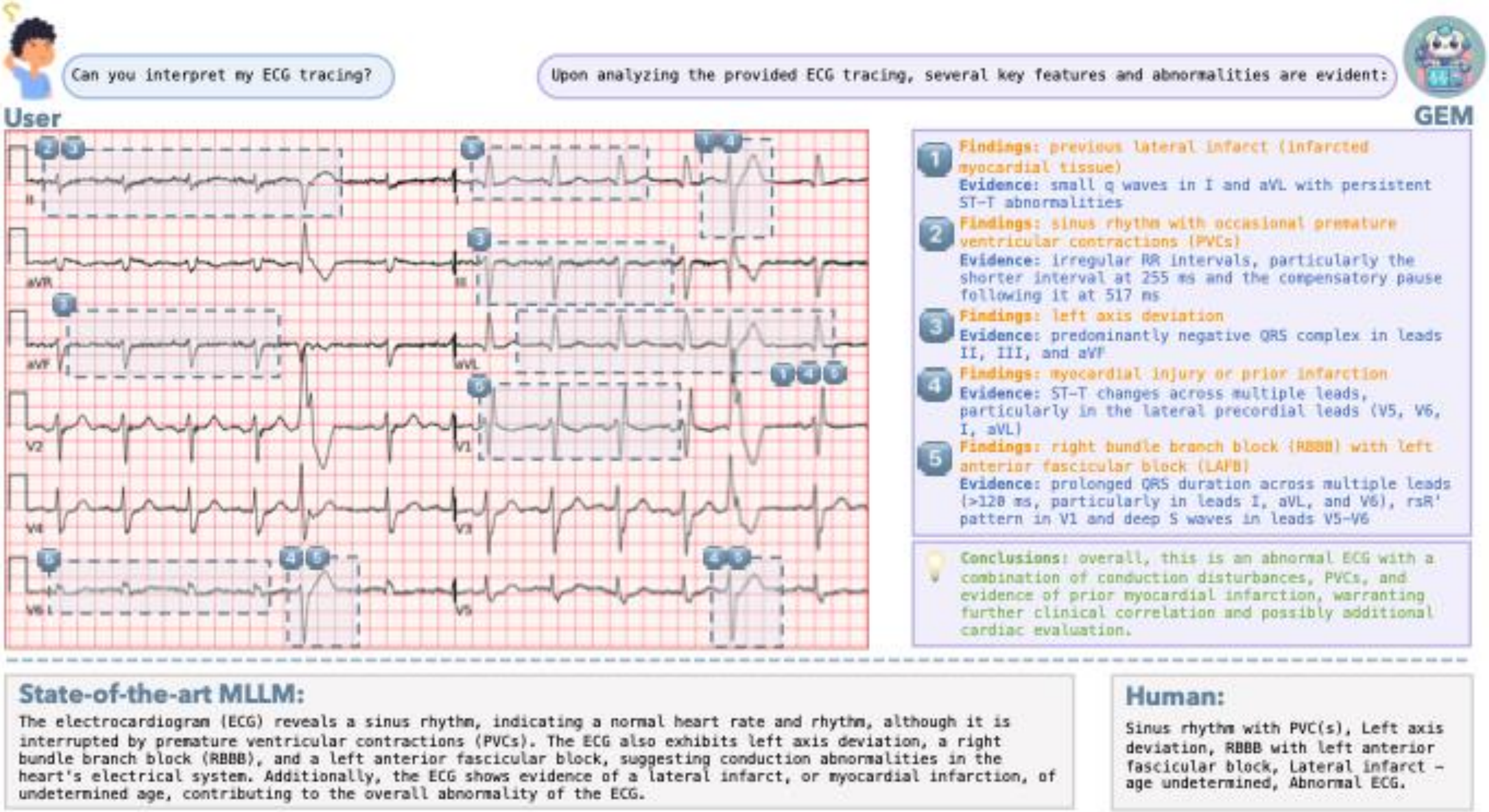


Knowledge Guided Instruction Data Generation (text)

Query: *"Interpret the provided ECG data <Image> <TS>, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence. When analyzing the ECG image, carefully analyze each lead. Lead I: Examine the QRS amplitude and duration, along with ST segment and T wave morphology ..."*

Target Answer: *"The ECG demonstrates sinus rhythm with occasional premature ventricular contractions (PVCs), as evidenced by irregular RR intervals, particularly the shorter interval at 255 ms and the compensatory pause following it at 517 ms ..."*

GEM



Can you interpret my ECG tracing?

Upon analyzing the provided ECG tracing, several key features and abnormalities are evident:

1. **Findings:** previous lateral infarct (infarcted myocardial tissue)
   Evidence: small q waves in I and aVL with persistent ST-T abnormalities
2. **Findings:** sinus rhythm with occasional premature ventricular contractions (PVCs)
   Evidence: irregular RR intervals, particularly the shorter interval at 255 ms and the compensatory pause following it at 517 ms
3. **Findings:** left axis deviation
   Evidence: predominantly negative QRS complex in leads II, III, and aVF
4. **Findings:** myocardial injury or prior infarction
   Evidence: ST-T changes across multiple leads, particularly in the lateral precordial leads (V5, V6, I, aVL)
5. **Findings:** right bundle branch block (RBBB) with left anterior fascicular block (LAFB)
   Evidence: prolonged QRS duration across multiple leads (>120 ms, particularly in leads I, aVL, and V6), rsR' pattern in V1 and deep S waves in leads V5-V6

Conclusions: overall, this is an abnormal ECG with a combination of conduction disturbances, PVCs, and evidence of prior myocardial infarction, warranting further clinical correlation and possibly additional cardiac evaluation.

State-of-the-art MLLM:

The electrocardiogram (ECG) reveals a sinus rhythm, indicating a normal heart rate and rhythm, although it is interrupted by premature ventricular contractions (PVCs). The ECG also exhibits left axis deviation, a right bundle branch block (RBBB), and a left anterior fascicular block, suggesting conduction abnormalities in the heart's electrical system. Additionally, the ECG shows evidence of a lateral infarct, or myocardial infarction, of undetermined age, contributing to the overall abnormality of the ECG.

Human:

Sinus rhythm with PVC(s), Left axis deviation, RBBB with left anterior fascicular block, Lateral infarct – age undetermined, Abnormal ECG.

Training

Some implantation details:

- Around 6,000 input token & 300 output token per sample in generating ECG-Grounding
- 8 A100 80G GPUs, 20 hours per epoch for GEM's training
- Base LLM: Vicuna-7B
- Finetuning mode: full parameter finetune

Training data:

ECG-Instruct	**ECG-Grounding (ours)**
▪ 1,156,110 instruction-response pairs	▪ 30,000 instruction-response pairs
▪ MIMIC-IV-ECG, PTB-XL, ECG-QA, CODE-15%	▪ MIMIC-IV-ECG
▪ Build basic language capabilities related to ECG tasks	▪ **Build grounded ECG understanding capabilities**

Evaluation

Grounded ECG understanding test

- Diagnosis accuracy
- Analysis completeness
- Analysis relevance
- Lead assessment coverage
- Lead assessment accuracy
- ECG feature grounding
- Evidence-based reasoning
- Clinical diagnostic fidelity

Abnormality detection

- AUC
- F1
- Hamming Loss

Report generation and QA

- Report score from GPT-4o
- QA accuracy

Human evaluation

- Reliability Metrics
  - Analytical Relevance (1-5)
  - Analytical Accuracy (1-5)
  - Analytical Completeness (1-5)
- Usefulness Metrics
  - Reasoning Quality (1-5)
  - Findings Novelty (1-5)
  - Clinical Value (1-5)
  - Overall Satisfaction (1-5)

Grounded ECG Understanding Test

Comprehensively evaluate whether the model achieves clinically grounded ECG understanding capabilities comparable to cardiologists.

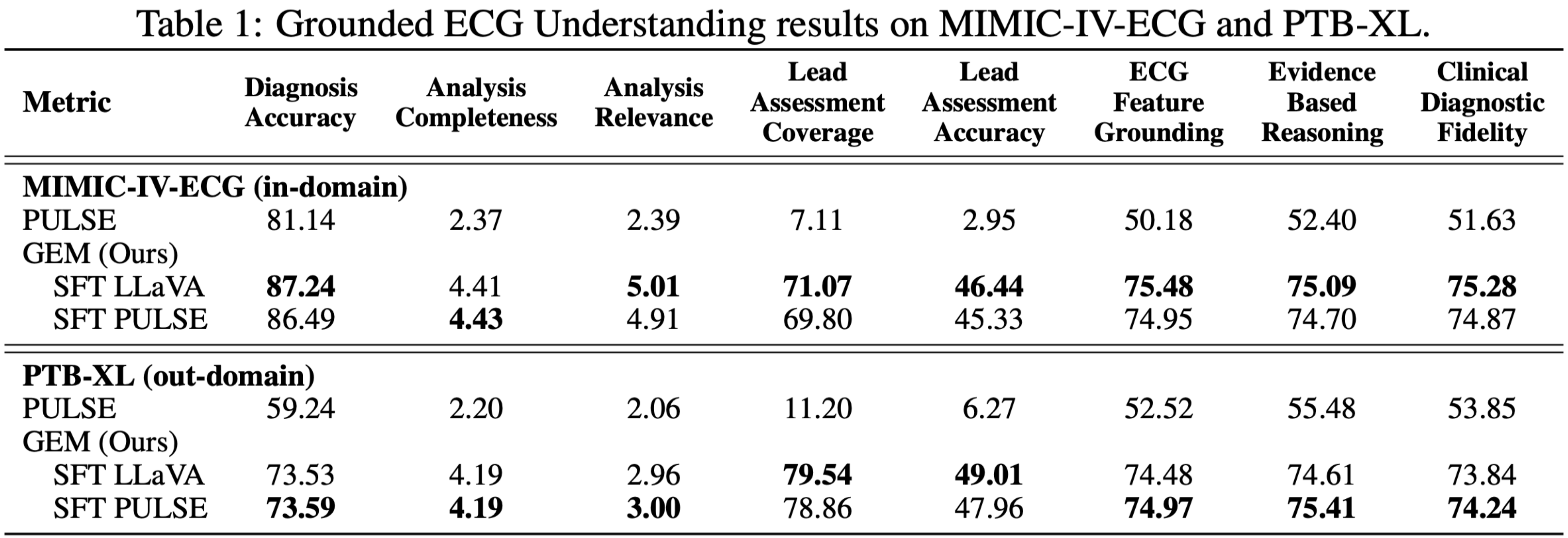


Table 1: Grounded ECG Understanding results on MIMIC-IV-ECG and PTB-XL.

Metric	**Diagnosis Accuracy**	**Analysis Completeness**	**Analysis Relevance**	**Lead Assessment Coverage**	**Lead Assessment Accuracy**	**ECG Feature Grounding**	**Evidence Based Reasoning**	**Clinical Diagnostic Fidelity**
MIMIC-IV-ECG (in-domain)								
PULSE	81.14	2.37	2.39	7.11	2.95	50.18	52.40	51.63
GEM (Ours)								
SFT LLaVA	**87.24**	4.41	**5.01**	**71.07**	**46.44**	**75.48**	**75.09**	**75.28**
SFT PULSE	86.49	**4.43**	4.91	69.80	45.33	74.95	74.70	74.87
PTB-XL (out-domain)								
PULSE	59.24	2.20	2.06	11.20	6.27	52.52	55.48	53.85
GEM (Ours)								
SFT LLaVA	73.53	4.19	2.96	**79.54**	**49.01**	74.48	74.61	73.84
SFT PULSE	**73.59**	**4.19**	**3.00**	78.86	47.96	**74.97**	**75.41**	**74.24**

PULSE: state-of-the-art large language model trained for ECG
SFT LLaVA: large language model has not trained with ECG related tasks
SFT PULSE: large language model has trained with ECG related tasks

MIMIC-IV-ECG: 2,381 samples
PTB-XL: 2,041 samples

ECG-Bench

Assess model's capability in cardiac abnormality detection and report generation.

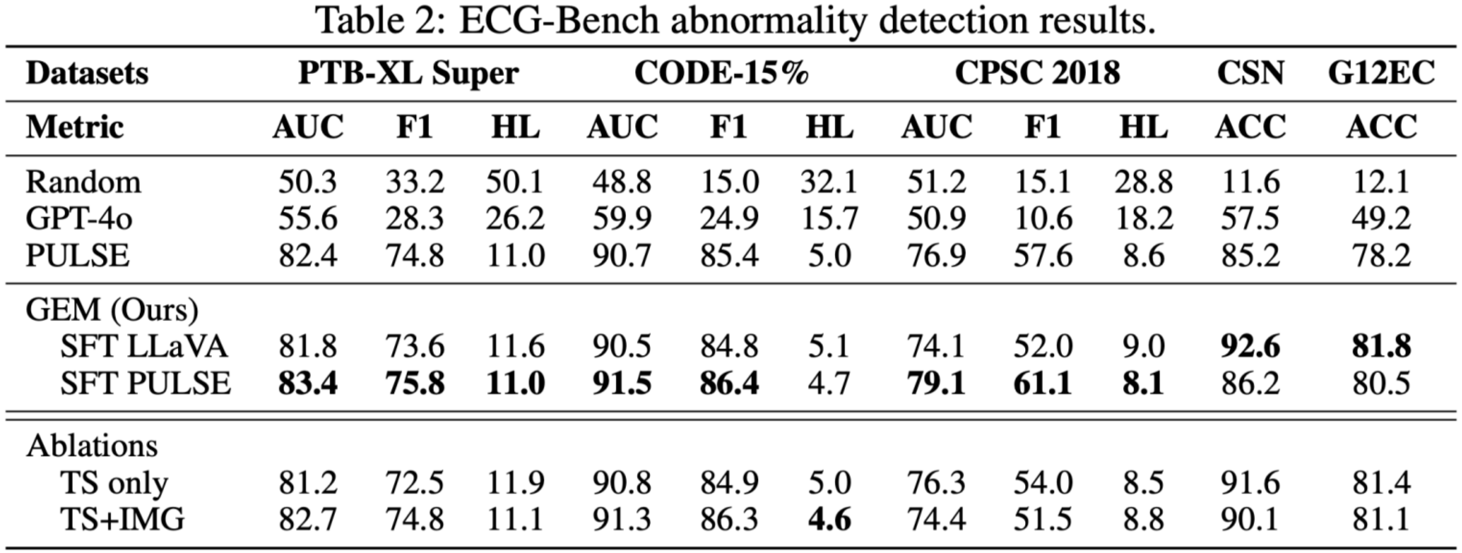


Table 2: ECG-Bench abnormality detection results.

Datasets	**PTB-XL Super**			**CODE-15%**			**CPSC 2018**			**CSN**	**G12EC**
Metric	**AUC**	**F1**	**HL**	**AUC**	**F1**	**HL**	**AUC**	**F1**	**HL**	**ACC**	**ACC**
Random	50.3	33.2	50.1	48.8	15.0	32.1	51.2	15.1	28.8	11.6	12.1
GPT-4o	55.6	28.3	26.2	59.9	24.9	15.7	50.9	10.6	18.2	57.5	49.2
PULSE	82.4	74.8	11.0	90.7	85.4	5.0	76.9	57.6	8.6	85.2	78.2
GEM (Ours)											
SFT LLaVA	81.8	73.6	11.6	90.5	84.8	5.1	74.1	52.0	9.0	**92.6**	**81.8**
SFT PULSE	**83.4**	**75.8**	**11.0**	**91.5**	**86.4**	4.7	**79.1**	**61.1**	**8.1**	86.2	80.5
Ablations											
TS only	81.2	72.5	11.9	90.8	84.9	5.0	76.3	54.0	8.5	91.6	81.4
TS+IMG	82.7	74.8	11.1	91.3	86.3	**4.6**	74.4	51.5	8.8	90.1	81.1

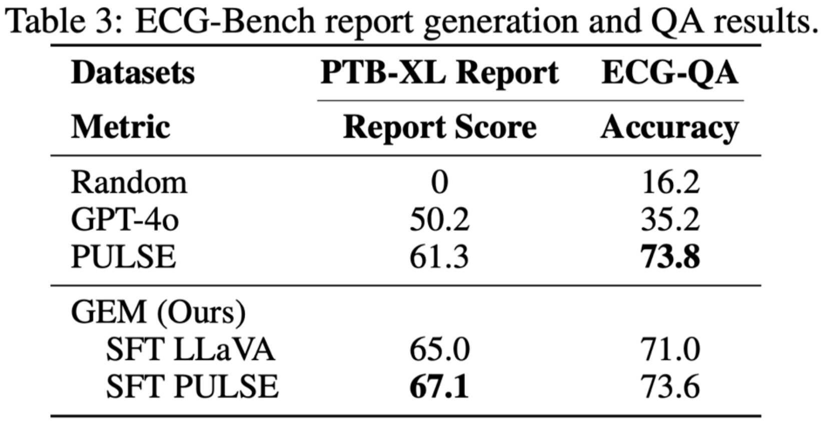


Table 3: ECG-Bench report generation and QA results.

Datasets	**PTB-XL Report**	**ECG-QA**
Metric	**Report Score**	**Accuracy**
Random	0	16.2
GPT-4o	50.2	35.2
PULSE	61.3	**73.8**
GEM (Ours)		
SFT LLaVA	65.0	71.0
SFT PULSE	**67.1**	73.6

Human Expert Evaluation

8 board-certified cardiologists:

Data Quality Evaluation:
- ***GPT-4o generated data (200 cases)***
- ***Deepseek-R1 generated data (200 cases)***

Model Evaluation:
- ***GEM generated interpretations (200 cases)***

1) Assess the quality of GPT-4o generated training data

2) Test the effectiveness of open-source substitutes

3) Validate the clinical utility of the GEM model

Scoring criteria (1-5)

Reliability

Relevance: ***Do the model's analyses closely support the diagnosis, and is there corresponding ECG evidence?***
5: Every analysis point is highly relevant to the diagnosis, with clear supporting evidence.
4: Most analyses are strongly relevant, with minor insufficiencies.
Accuracy: ***Are there any medical factual errors in the model's output?***
5: Completely accurate
4: Mostly accurate
Completeness: ***Does the model comprehensively discuss key ECG components relevant to the diagnosis, including rhythm, intervals, and wavefoms?***
5: All relevant ECG features (rhythm, PR, QRS, ST, T waves, intervals, etc.) are accurately discussed.
4: Most key ECG features are covered, with minor omissions.

Usefulness

Reasoning: ***Does the model provide a clear, evidence-based reasoning process similar to that of a clinician, logically deriving the diagnosis from ECG features?***
5: Clear and coherent reasoning structure, explaining each step from ECG to diagnosis causally.
4: Overall reasonable reasoning, but some steps lack detail.
Novelty: Does the model provide insights or findings not noticed by the clinician?
4: Novel and somewhat insightful content.
3: Some new findings, but of limited value.
Clinical Value: Does the model output help in clinical decision-making?
5: Direct and significant support for clinical judgment; content is clear and reliable.
4: Most content is helpful and practically useful.
Satisfaction: Subjective rating of the overall quality of this analysis.
5: Very satisfied.
4: Satisfied

Human Expert Evaluation

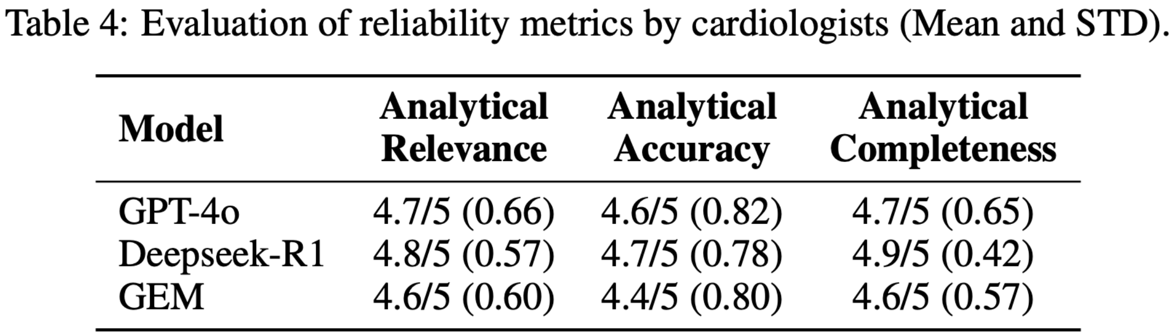


Table 4: Evaluation of reliability metrics by cardiologists (Mean and STD).

Model	**Analytical Relevance**	**Analytical Accuracy**	**Analytical Completeness**
GPT-4o	4.7/5 (0.66)	4.6/5 (0.82)	4.7/5 (0.65)
Deepseek-R1	4.8/5 (0.57)	4.7/5 (0.78)	4.9/5 (0.42)
GEM	4.6/5 (0.60)	4.4/5 (0.80)	4.6/5 (0.57)

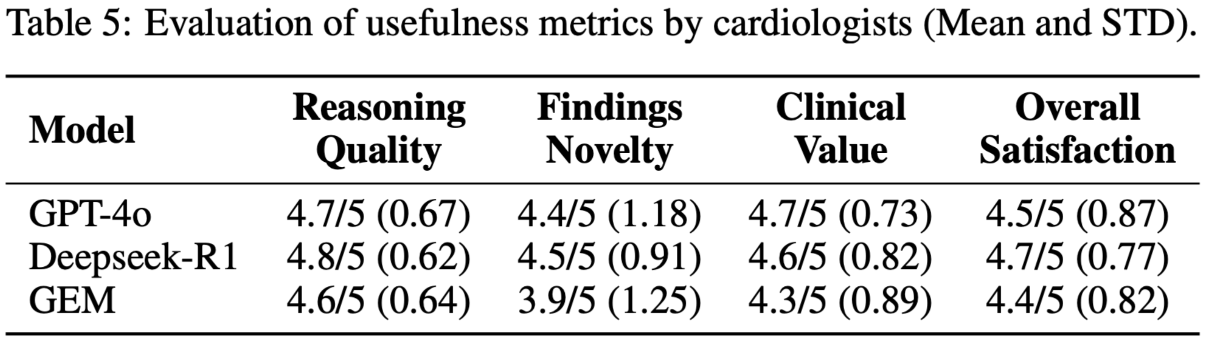


Table 5: Evaluation of usefulness metrics by cardiologists (Mean and STD).

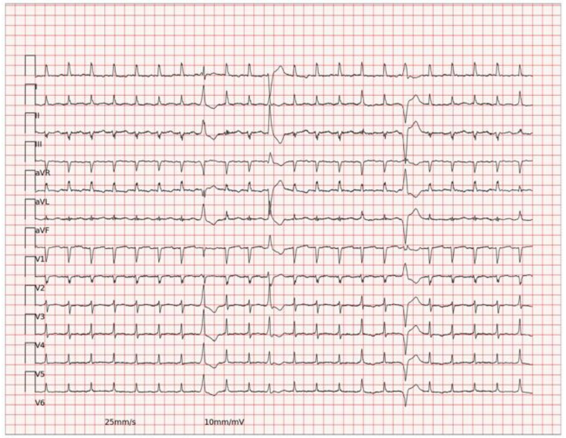
Model	**Reasoning Quality**	**Findings Novelty**	**Clinical Value**	**Overall Satisfaction**
GPT-4o	4.7/5 (0.67)	4.4/5 (1.18)	4.7/5 (0.73)	4.5/5 (0.87)
Deepseek-R1	4.8/5 (0.62)	4.5/5 (0.91)	4.6/5 (0.82)	4.7/5 (0.77)
GEM	4.6/5 (0.64)	3.9/5 (1.25)	4.3/5 (0.89)	4.4/5 (0.82)

1) GPT-4o can generate ECG interpretations that are both clinically reliable and practically valuable.

2) DeepSeek R1 even better.

3) GEM can deliver clinically meaningful and accurate interpretations that align well with cardiologists' expectations.

Human Expert Evaluation



GEM Interpretation: Upon analyzing the provided ECG image and computed measurements, several key features and abnormalities are evident. The heart rate is elevated at 121 bpm, indicating a rapid ventricular response. The rhythm appears irregular, consistent with atrial fibrillation, as seen by the absence of distinct P waves and the presence of irregular RR intervals. In the precordial leads, particularly V1 through V4, there are low QRS voltages, which could suggest conditions such as pericardial effusion or obesity. The QRS complexes in these leads are notably diminished, which aligns with the low voltage finding. Furthermore, the presence of premature ventricular contractions (PVCs) is suggested by the irregular RR intervals and the occasional absence of preceding P waves. The ST-T changes observed in the anterolateral leads, particularly V2 to V6, are nonspecific but could indicate ischemic changes or other cardiac conditions. The QTc intervals are prolonged in several leads, which may suggest a risk for arrhythmias. The combination of these findings, including the rapid ventricular response, irregular rhythm, low QRS voltages, and nonspecific ST-T changes, supports a diagnosis of atrial fibrillation with rapid ventricular response and PVCs. The low QRS voltages in the precordial leads add another layer of complexity, potentially indicating underlying cardiac conditions. The ECG is classified as abnormal due to these findings, which warrant further clinical correlation and possibly additional diagnostic testing to determine the underlying cause of these abnormalities.

Ground-truth Report: Atrial fibrillation with rapid ventricular response with frequent multifocal PVCs or aberrant ventricular conduction; Possible anterior infarct; Inferior/lateral ST-T changes are nonspecific; Low QRS voltages in precordial leads; Abnormal ECG.

PULSE Interpretation: The electrocardiogram (ECG) reveals atrial fibrillation with a rapid ventricular response, accompanied by premature ventricular contractions (PVCs) or aberrant ventricular conduction, indicating irregular heart rhythm. Additionally, the ECG suggests possible anterior myocardial infarction, although the age of the infarct cannot be determined. The inferior and lateral ST-T changes are nonspecific, and the QRS voltages are low in the precordial leads, contributing to an overall abnormal ECG pattern.

Discussion and Future Works

1 LLM-as-a-Judge may introduce potential bias

- Conduct a comparative analysis using multiple independent LLMs (e.g., Deepseek-R1, LLaMA) to evaluate the same test outputs.
- Scale human expert evaluation.

2 Emotional impact of model outputs

- better account for the emotional or psychological impact of model outputs.

Summary

- We introduced GEM, the first multimodal large language model integrating ECG signals, ECG plots, and text for grounded ECG understanding.

- We created ECG-Grounding, the first high-granularity ECG grounding dataset comprising 30,000 instruction pairs annotated with heartbeat-level physiological features.

- We proposed the Grounded ECG Understanding task, a clinically motivated benchmark designed to comprehensively assess a model's ECG interpretation capability.

Data&Model

Code

Thank You