Appendix

This is the Appendix for "HeartcareGPT: A Unified Multimodal ECG Suite for Dual Signal–Image Modeling and Understanding".

This Appendix is organized as follows:

- Section A provides the details of our proposed multimodal data engine—**HeartAgent**.
- Section B provides the details of the experimental implementation, the training process of **HeartcareGPT**, the construction details of **Heartcare-400K**, and the specific information of **Heartcare-Bench**.
- Section C shows our detailed ablation experimental results of **HeartcareGPT** and **Beat**.
- Section D shows typical data examples in **Heartcare-400K**.
- Section E lists the broader impact and limitations of this paper.

A. Multimodal Data Engine Details

To efficiently construct the Heartcare-400K dataset, we design **HeartAgent**, an automated multimodal data engine that transforms multi-source ECG data into high-quality, structured QA pairs. As illustrated in Figure 2, the system consists of four core modules that work collaboratively to complete the full pipeline from raw data parsing to task QA generation.

Multimodal Feature Converter. The Converter preprocesses hospital PDF reports into standardized inputs for downstream modules. It extracts patient metadata (e.g., age and gender) and diagnostic details (diagnosis, waveform, rhythm and other features) from the PDF via fitz (PyMuPDF) [16] and regular expressions. Subsequently, diagnostic text is mapped to structured English labels in accordance with the SCP-ECG semantic standard [39].

Noise Filtering and Quality Optimizer. The Optimizer implements a parsing mechanism to unify heterogeneous ECG inputs into standardized 12-lead, 250 Hz digital signals. Structured digital inputs from public datasets are processed via the WFDB toolkit [40]. To correct noise, drift, and missing-segment issues, the Optimizer applies a three-stage enhancement pipeline: (i) resampling all signals to 250 Hz, (ii) applying clean functions in NueroKit2 [28] for lead-level denoising and baseline correction, and (iii) using NeuroKit2's quality scores to extract high-quality segments while discarding low-quality regions.

Diversed Image Generator. The Image Generator processes clear waveform images through a dual-channel pipeline. For high-quality signals originated from the Optimizer, the Generator uses Matplotlib [15] in the final stage to render clean digital signals into images. For hospital reports, it processes PDF-based reports to produce cropped images of the complete 12-lead tracings, while ensuring all data is de-identified.

Multi-Task QA Builder. To enhance model generalization and training consistency across multi-level ECG QA scenarios, we use GPT-4 [33] to directly generate structured multimodal ECG QA samples, constructing the training dataset. Each generated sample consists of four key components: **(i) Context Description**, providing background information such as patient demographics, signal snippets, or clinical report summaries; **(ii) Task Instruction**, specifying the required operation type, including *Closed-QA*, *Open-QA*, *Comparison-QA*, *Report Generation* and *Signal Prediction*; **(iii) Auxiliary Labels**, providing additional structured supervisory information to enhance training quality; and **(iv) Output Format**, standardizing expression of answers to ensure consistency across tasks.

B. Implementation Details

B.1. Model Details

HeartcareGPT employs an architecture design that aligns ECG signals and images with textual inputs in latent space. For signal and image modalities, We use two MLP projectors for cross-modal feature fusion. Notably, we implement LoRA [13] for parameter-efficient fine-tuning, preserving pretrained knowledge while enabling domain-specific adaptation for ECG tasks. This design achieves an optimal balance between model capacity and computational efficiency, establishing a scalable architectural foundation for multimodal ECG understanding.

HeartcareGPT offers two versions: **HeartcareGPT-3.8B** and **HeartcareGPT-7B**, which are based on Phi-3-mini-4k-Instruct [1] and Qwen2.5-7B-Instruct [6] as pre-trained LLMs, respectively. We employs SigLip-So400M-patch14-384 [55] and the proposed Beat as dual-form feature encoders, and extend the model's native textual tokenizer by adding four special tokens: ⟨PRED⟩, ⟨sig⟩, ⟨img⟩ and ⟨text⟩. ⟨PRED⟩ indicates that the model is expected to perform signal-prediction tasks handled by the integrated beat module. ⟨sig⟩ marks the starting boundaries of sequential features, ⟨img⟩ marks the starting boundaries of visual features, and ⟨text⟩ marks the starting boundaries of textual features, enabling the model to more reliably recognize and process these modality-specific representations. Table 5 shows the details.

In addition, for baseline models that support only a single input modality, we apply a unified preprocessing strategy to align modalities, ensuring fairness within the same evaluation framework. These baseline models typically support only image inputs rather than raw signals. Therefore, we convert the ECG signals into images using the Diverse Image Generator module of HeartAgent (presented in Appendix A), and feed the generated images together with the textual ques-

Table 5. Overview of the components of HeartcareGPT.

Model	ViT	Signal Projector	Image Projector	LLM	Params	Vocab Size	LoRA Rank
HeartcareGPT-3.8B	SigLIP-So400M-patch14-384	1-layer MLP	2-layer MLP	Phi-3-mini-4k-Instruct	3.8B	32,069	64
HeartcareGPT-7B	SigLIP-So400M-patch14-384	1-layer MLP	2-layer MLP	Qwen2.5-7B-Instruct	7B	151,851	64

tions into the baseline models to obtain their responses.

B.2. Training Details

We follow a three-stage training paradigm: (i) training Beat to extract high-fidelity ECG embeddings, (ii) warming up the visual and signal projectors to stabilize feature alignment, and (iii) performing joint instruction fine-tuning on Heartcare-400K for end-to-end modeling. This paradigm achieves decoupled feature learning and semantic alignment across stages, enabling the model to maintain signal fidelity while acquiring advanced clinical reasoning capabilities.

Training Beat. Beat is first trained on PTB-XL [46] dataset. We use a joint supervision strategy to optimize reconstruction and prediction losses simultaneously. This stage focuses on learning robust ECG signal representations through DVQ structure.

Warming Up. We warm up the visual and signal projectors using paired ECG images and signals, allowing the model to align heterogeneous features in a controlled manner. By isolating projector optimization, the model avoids early-stage instability and catastrophic feature distortion. This step establishes a coherent multimodal latent space for unified processing.

Joint Fine-Tuning. We perform full-model instruction fine-tuning on Heartcare-400K, enabling the model to associate ECG inputs with diagnostic reasoning, structured reporting, and question answering. The model learns both high-level clinical knowledge and fine-grained ECG interpretation skills. This stage integrates all modalities end-to-end, yielding a clinically aligned and diagnostically capable ECG foundation model.

Hyperparameter configurations for each training stage are detailed in Table 6.

B.3. Construction details of Heartcare-400K

Data Source Details. In the data collection phase, we gather ECG report data with two modalities—digitized raw signals and clinical report images.

PTB-XL [46] is one of the largest publicly available electrocardiogram datasets, comprising 21,799 clinical 12-lead ECG recordings that cover a diverse range of cardiac pathologies as well as healthy control data. Each recording has a duration of 10 seconds with a sampling rate of 500 Hz, accompanied by standardized diagnostic annotations and detailed patient metadata, such as gender and age. We utilize PTB-XL as a high-quality structured data source to enhance the diversity and accuracy of Heartcare-400K in the digital modality.

In contrast, ECG image modality data has long been constrained by acquisition challenges, annotation costs, and privacy concerns, resulting in scarce and outdated publicly available image datasets. To address this issue, we establish collaborations with two top-tier hospitals and collect a total of 12,170 recent ECG report forms through rigorous anonymization and professional physician annotations. Each report is in a standardized PDF format, containing basic patient information, physiological parameters, physician diagnoses, and approximately 5-second 12-lead image recordings, significantly improving the timeliness and clinical usability of the image modality.

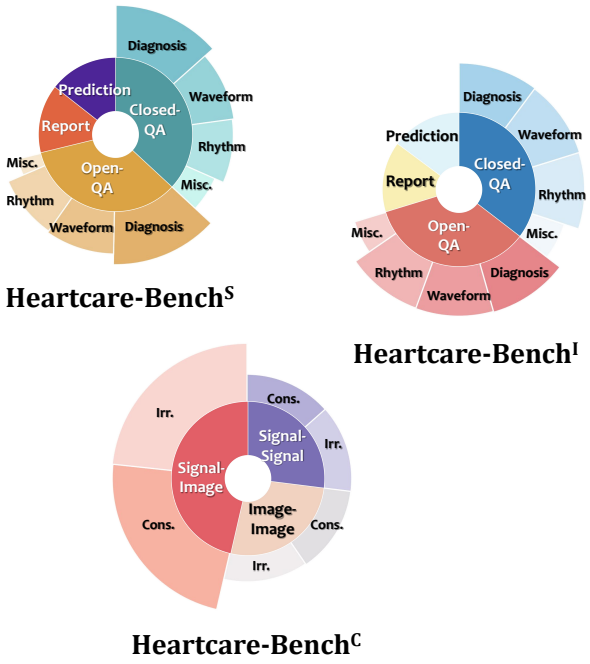


Figure 6. Hierarchical distribution of Heartcare-400K dataset across benchmarks and subtask types.

To provide a comprehensive overview of diagnostic coverage and clinical relevance of Heartcare-400K, Table 10 presents a detailed listing of ECG features organized across four clinically critical dimensions: (i) *Diagnosis* (e.g., Inferior MI, AV block), (ii) *Waveform abnormalities* (e.g., T abnormalities, ST depression), (iii) *Rhythm patterns* (e.g., sinus rhythm, atrial fibrillation), and (iv) *Miscellaneous fea-*

Table 6. Overview of hyperparameter configurations.

Hyperparameter	HeartcareGPT-3.8B			HeartcareGPT-7B		
	Training Beat	Warming up	Joint Fine-Tuning	Training Beat	Warming up	Joint Fine-Tuning
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Adapter LR	/	/	2e-5	/	/	2e-5
Learning Rate	1e-4	2e-4	2e-5	1e-4	2e-4	2e-5
Global Batch Size	32	16	16	32	16	16
Weight Decay	0	0.01	0.01	0	0.01	0.01
Dropout Rate	0	0	0.05	0	0	0.05
LR Scheduler	Cosine Annealing	Linear	Linear	Cosine Annealing	Linear	Linear

tures (e.g., noise, heart axis). This illustrates the dataset's comprehensive and balanced representation across these clinically relevant dimensions.

To visually summarize the dataset composition and task-specific coverage of Heartcare-400K, we provide a hierarchical radial visualization in Figure 6. This depict the distribution of QA subtasks within three benchmarks, Heartcare-Bench^{S}, Heartcare-Bench^{I} and Heartcare-Bench^{C}, highlighting both the relative scale and fine-grained breakdown of different data types.

QA Templates. For datasets that only contain classification or grading labels, we analyze the data characteristics of their labels and design different QA templates for each. This allow us to transform the original data into QA pairs. We additionally provide the prompt used to guide GPT-4 [33] in generating QA templates for diagnosis subtasks in *Closed-QA* data construction pipeline of HeartAgent, as shown in Figure 7. Examples of the QA templates are shown in the Table 9.

B.4. Construction details of Heartcare-Bench

To comprehensively assess model performance across different task types, we design a multi-dimensional evaluation framework tailored to the specific objectives of each module. The evaluation criteria are carefully selected to reflect the core competencies required by each task—ranging from answer correctness and semantic understanding to clinical accuracy and waveform forecasting fidelity.

Closed-QA. We measure model discrimination performance by standard *Accuracy*.

Open-QA. We adopt a dual-track evaluation comprising *F1-Bio* [38] to assess semantic alignment, and *ROUGE-L* [24] to quantify linguistic fluency and contextual style fidelity.

Comparision-QA. We measure model discrimination performance by standard *Accuracy*.

Report Generation. (i) Beyond ROUGE-L, we use *F1-RadGraph* [17] to evaluate the precision of entity and relation extraction in the report structure. (ii) We implement an 100-point, multi-dimensional criteria that evaluates diagnostic completeness, waveform accuracy, rhythm accuracy, report logic, and descriptive norms. Using GPT-4 [33], we tally error types and severity according to this criteria (see Table 7) for weighted penalties. The template used for GPT-4's evaluation is shown in Figure 8. According to the evaluation

Table 7. Evaluation dimensions and weighted criteria for ECG diagnostic reports.

Dimension	Evaluation Criteria	Weight
Diagnosis Completeness	Completeness of abnormal features mentioned	10
	Completeness of key diagnoses included	10
	Absence of critical diagnostic errors	10
	Whether the report describes severity or likelihood of the findings	8
	Whether the report includes suspected diagnoses	7
Waveform Accuracy	Correct identification of anatomical regions (e.g., P/QRS/T waves)	8
	Correct recognition of waveform abnormalities (e.g., ST elevation/depression)	7
Rhythm Accuracy	Correct classification of baseline rhythm (e.g., sinus or ectopic)	4
	Correct classification of arrhythmias (e.g., tachycardia or bradycardia)	4
	Correct interpretation of conduction abnormalities (e.g., location and degree of block)	4
	Accurate detection of pacing signals	3
Report Logic	Report is well-structured and logically organized	5
	Findings are explained in a point-wise or categorized manner	4
	Includes relevant auxiliary information (e.g., age, gender, etc.)	3
	Patient privacy is protected via anonymization	3
Descriptive Norms	Terminology complies with SCP-ECG standards (e.g., use "complete right bundle branch block" instead of "RBBB")	5
	Language avoids inappropriate certainty (e.g., avoids overconfident conclusions)	5
Total Score		**100**

criteria, we grade the reports as follows:

- **Excellent Report (90–100):** Nearly no error with complete diagnostic information, clear structure, and no clinically significant mistakes. Ready for immediate clinical use.
- **Acceptable Report (80–89):** Contains minor errors but maintains diagnostic accuracy and logical flow. Requires minimal editing before clinical application.
- **Review Required Report (60–79):** Has notable errors, incomplete information, or unclear structure. Needs expert verification before use.
- **Unusable Report (< 60):** Contains critical errors, major missing information, or serious diagnostic inaccuracies. Unsafe for clinical decision-making.

Signal Prediction. We delimit the predicted segment the with special token ⟨PRED⟩ and compute the *Mean Squared Error (MSE)* between the forecasted waveform and the true continuation. Lower MSE indicates superior prediction accuracy.

Heartcare-Bench adopts strict patient-level partitioning to ensure that all ECG records from the same individual—across different formats or tasks—reside in a single split. To further eliminate potential leakage, we perform a lightweight yet comprehensive overlap check that includes examination-level matching using study IDs and timestamps, duplicate detection of ECG waveforms via normalized hashing, and visual and textual duplication inspection for rendered ECG images and associated reports. These checks jointly guarantee that no record or derived annotation appears across splits, ensuring fair and leakage-free evaluation.

C. Supplemental Experimental Results

C.1. Ablation Study Results of HeartcareGPT

For completeness, we present in this appendix the full numerical results of all ablation settings of HeartcareGPT-3.8B discussed in Section 6.3. While the main paper visualizes the trends through figures, the detailed data tables are provided here to enable transparent comparison across all variants of the training pipeline (shown in Table 11, 12, 13 and 14) and multimodal integration strategies (shown in Table 15, 16, 17 and 18).

Across all settings, the ablation results consistently validate the effectiveness of our design choices. Removing either Beat training stage or warming up stage leads to substantial degradation in diagnostic accuracy and multimodal alignment. Likewise, replacing our tri-modal fusion with single-modality variants or removing 12-lead sub-image segmentation yields clear declines in performance. These trends reinforce the necessity of ECG-aware representation modeling, stabilized multimodal projection, and fine-grained visual decomposition for enhancing clinical reasoning.

Overall, experimental results further substantiate our claims that each architectural and training design contributes meaningfully to HeartcareGPT's superior diagnostic capability and its robustness across diverse ECG modalities.

C.2. Ablation Study Results of Beat

To validate the performance of our model, we conduct comprehensive experiments based on our ECG tokenizer, Beat. We evaluate its capabilities in both signal reconstruction and prediction tasks under various structural configurations, including the use of DVQ structure, codebook size, and input sequence length.

In Section 5.1, we have introduced the joint supervision strategy adopted in Beat. To facilitate evaluation in experiments, we also consider this total loss as a quantitative metric. Specifically, the overall training objective of Beat $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{pred}} + \lambda_3 \mathcal{L}_{\text{VQ}}, \tag{12}$$

where we set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$ and $\lambda_3 = 0.25$ in our experiments, balancing the contributions of reconstruction, prediction, and vector-quantization objectives. By monitoring the overall training objective as well as its individual components, we can more thoroughly assess how each architectural modification affects representation quality and optimization stability.

For a more comprehensive evaluation of Beat's performance, we additionally employ the codebook utilization as Code%, which quantifies the proportion of codebook entries actually used during encoding.

Table 8. Ablation study for Beat under different configurations.

DVQ	Codebook Size	Input Length	Code%	$\mathcal{L}_{\text{total}}$
-	512	1000	93.75	1.00
✓	1024	1000	79.50	<u>0.79</u>
✓	256	1000	**99.87**	0.81
✓	512	1500	91.46	1.10
✓	512	500	<u>98.20</u>	0.83
✓	512	1000	96.22	**0.76**

As shown in Table 8, the results show that the DVQ structure with a codebook size of 512 and an input length of 1000 strikes a favorable balance between compression efficiency and semantic completeness. Notably, while smaller codebooks or shorter input sequences can achieve higher codebook utilization, their overall loss is higher than our chosen configuration, highlighting that code utilization alone is not sufficient to evaluate tokenizer effectiveness.

We summarize the following observations: (i) The DVQ structure captures global rhythm patterns via the core codebook and refines local variations via the residual codebook, thereby enhancing the clinical semantic integrity of the

discrete representation while maintaining a compact token space. (ii) Enlarging the codebook increases representational granularity but leads to codebook collapse and lower utilization, whereas a smaller codebook fails to capture the complex pathological semantics of ECG signals. (iii) Excessively long or short input sequences degrade codebook utilization and introduce instability in reconstruction and prediction, likely due to imbalanced temporal context or fragmented signal structure. Overall, Beat achieves an effective global-local modeling trade-off through structural and parametric design, significantly improving the quality of ECG tokenization and enabling end-to-end training of ECG and text modalities within Med-MLLMs.

Furthermore, Figure 9 provides a comprehensive visualization of Beat's reconstruction and prediction performance, demonstrating the model's capability to accurately recover input patterns while generating high-fidelity future predictions.

D. Case Study

In this section, we compare generated answers of our proposed HeartcareGPT-3.8B with those of a generalist model (Claude-3.5) and a medical model (MedVLM-R1-2B). Since *Closed-QA* and *Comparison-QA* primarily contain selection-based questions, which are less illustrative for qualitative comparison, we focus on *Open-QA* and *Report Generation* cases where the model's clinical reasoning capabilities can be more clearly observed. Figures 10 and 11 illustrate the performance of these three models on *Open-QA* and *Report Generation* tasks. We highlight statements that match the ground truth in green, and indicate discrepancies in red.

As shown in Figure 10, HeartcareGPT's response closely matches the reference answer, showing its precise understanding of detailed diagnostic questions.

A similar observation can be made from Figure 11. HeartcareGPT produces reports that not only capture clinically relevant findings with high accuracy, but also maintain clear structure. Moreover, the reports show correct usage of ECG-specific phrasing and logical flow.

E. Limitations

Heartcare Suite advances multimodal ECG understanding with potential benefits for clinical diagnosis, medical artificial intelligence (AI) research, and patient care. By integrating raw ECG signals and structured reports, it enables accurate, automated cardiac analysis, particularly valuable in resource-limited settings. The release of Heartcare-400K and Heartcare-Bench fosters transparency and progress in medical AI. However, limitations include dataset biases (e.g., underrepresentation of rare conditions), potential signal fidelity loss in tokenization, and untested real-time monitoring capabilities. Computational costs and regulatory hurdles for clinical deployment remain challenges. Future work should expand data diversity, optimize real-time processing, and validate clinical utility through trials.

Table 9. Sample QA templates in Heartcare-400K.

	Questions
Closed-QA	1. Please assign the most suitable shape and structure classification with a detailed examination of the provided ECG sequence of this subject. A. Non-diagnostic T abnormalities; B. Ventricular premature complex; C. Low QRS voltage in limb leads; D. Non-specific ST elevation.
	2. Investigate the patient's ECG reading and diagnose its classification based on its features. A. Normal ECG; B. Incomplete left bundle branch block; C. Long QTc-interval; D. Complete right bundle branch block.
	3. By conducting a detailed evaluation of the ECG trace of the person, output the correct rate and regularity it should be classified under. A. Bigeminal pattern; B. Sinus tachycardia; C. Sinus rhythm; D. Normal functioning artificial pacemaker.
	4. What would you determine the pattern and timing of this ECG reading to be? A. Atrial fibrillation; B. Atrial flutter; C. Normal functioning artificial pacemaker; D. No abnormality detected.
	5. With precision and attention to detail, work through the subject's ECG reading and give the most appropriate rhythm based on its characteristics. A. Sinus bradycardia; B. Atrial flutter; C. Paroxysmal supraventricular tachycardia; D. Atrial fibrillation.
Open-QA	1. Given the ECG finding, please work through its features and classify the right shape and structure.
	2. Assign the waveform associated with the ECG characteristic.
	3. What pattern and timing does ECG interpretation exhibit?
	4. Through meticulous examination of the patient's ECG sequence, please accurately determine the diagnosis that best defines it.
	5. What rhythm does the given ECG characteristic from the patient exhibit?
Comparison-QA	1. Has non-diagnostic T abnormalities been eliminated in the recent tracing in comparison to the previous one?
	2. Does the second ECG still indicate the presence of ST/T change as compared to the first ECG?
	3. Is atrial flutter still not detected in the recent tracing when compared to the previous one?
	4. Does the second ECG still show the absence of non-specific T-wave changes when compared to the first ECG?
	5. Is the RR interval still considered normal in the latest tracing when compared to the previous tracing?
Report Generation	1. Please provide a comprehensive ECG interpretation for this 69.0-year-old male individual.
	2. Produce a detailed ECG analysis for this 55.0-year-old female case.
	3. Generate a standardized ECG assessment report for this 36.0-year-old male patient.
	4. Based on the ECG signal from this 75.0-year-old male patient, generate a structured clinical 12-lead ECG report.
	5. Analyze the ECG tracing and generate a clinical report for this 77.0-year-old male individual.
	Answers
Positive Condition	1. Based on the ECG pattern, after thorough examination, the form is classified as {condition}.
	2. The diagnostic classification observed in the given ECG observation suggests a evident link to suggestive of {condition}.
	3. After systematic analysis, the ECG evaluation is classified as {condition}.
	4. Clinical findings from this ECG assessment reinforce the presence of {condition} as a evident outcome.
	5. The ECG signal shows evidence of {condition}.
Negative Condition	1. All leads demonstrate physiological waveforms, and the overall conclusion is a normal ECG.
	2. Standard diagnostic criteria confirm that the signal is entirely normal, with no pathological findings.
	3. No evidence of ST-segment elevation, depression, or T-wave inversions.
	4. Healthy cardiac activity.
	5. Heart rate is regular, with consistent P-P and R-R intervals.

Table 10. Systematic Categorization of ECG Features in Heartcare-400K.

Diagnosis	
ischemic in inferior leads	non-specific ischemic
septal hypertrophy	non-specific intraventricular conduction block
subendocardial injury in anterolateral leads	left anterior fascicular block
ischemic in anterior leads	non-diagnostic T abnormalities
anterolateral myocardial infarction	incomplete right bundle branch block
non-specific ST changes	third degree AV block
right ventricular hypertrophy	ischemic in lateral leads
incomplete left bundle branch block	long QTc-interval
first degree AV block	inferoposterior myocardial infarction
right atrial hypertrophy	Wolf-Parkinson-White syndrome
subendocardial injury in anteroseptal leads	inferolateral myocardial infarction
inferior myocardial infarction	posterior myocardial infarction
right atrial overload/enlargement	long QT-interval
complete left bundle branch block	left posterior fascicular block
intraventricular conduction block	complete right bundle branch block
inferoposterolateral myocardial infarction	ST-T changes compatible with ventricular aneurysm
left atrial overload/enlargement	ischemic in inferolateral leads
digitalis-effect	anterior myocardial infarction
ischemic in anterolateral leads	subendocardial injury in lateral leads
subendocardial injury in inferolateral leads	subendocardial injury in inferior leads
biatrial hypertrophy	second degree AV block
left ventricular hypertrophy	lateral myocardial infarction
ischemic in anteroseptal leads	anteroseptal myocardial infarction
electrolytic disturbance or drug	ST-T changes
Waveform	
non-specific ST changes	low QRS voltage in chest leads
low QRS voltage in limb leads	Q waves present
low QRS voltage in left chest leads	long QTc-interval
long QT-interval	digitalis-effect
non-specific ST elevation	high QRS voltage in left ventricular
low QRS voltage in chest and limb leads	non-specific ST depression
low amplitude T-waves	ventricular premature complex
non-diagnostic T abnormalities	inverted T-waves
short PR interval	atrial premature complex
Rhythm	
sinus bradycardia	supraventricular tachycardia
sinus rhythm	sinus tachycardia
atrial bradycardia	atrial fibrillation
sinus arrhythmia	normal functioning artificial pacemaker
atrial flutter	trigeminal pattern
paroxysmal supraventricular tachycardia	bigeminal pattern
Miscellaneous	
burst noise	old stage of myocardial infarction
static noise	early stage of myocardial infarction
baseline drift	middle stage of myocardial infarction
electrodes problems	left axis deviation
ventricular extrasystoles	right axis deviation
supraventricular extrasystoles	

Table 11. Ablation study on training stages for HeartcareGPT-3.8B on *Closed-QA* tasks from Heartcare-Bench^{S} and Heartcare-Bench^{I}.

Training Beat	Warming up	Heartcare-Bench^{S}				Heartcare-Bench^{I}				Avg.
		Diagnosis	Waveform	Rhythm	Misc.	Diagnosis	Waveform	Rhythm	Misc.	
-	-	37.70	53.06	45.71	63.22	43.17	49.64	40.41	34.54	45.93
-	✔	45.82	64.37	57.28	72.25	51.86	60.91	50.19	44.69	55.92
✔	-	74.13	87.10	75.86	71.90	82.03	83.74	71.62	65.38	76.47
✔	✔	**81.95**	**95.94**	**82.79**	**79.84**	**87.85**	**92.21**	**79.25**	**67.80**	**83.33**

Table 12. Ablation study on training stages for HeartcareGPT-3.8B on *Open-QA* tasks from Heartcare-Bench^{S} and Heartcare-Bench^{I}.

Training Beat	Warming up	Heartcare-Bench^{S}								Heartcare-Bench^{I}							
		Diagnosis		Waveform		Rhythm		Misc.		Diagnosis		Waveform		Rhythm		Misc.	
		F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L
-	-	50.32	23.15	57.84	28.97	49.35	25.42	60.27	24.32	53.81	27.19	55.04	25.96	50.91	25.30	56.67	26.18
-	✔	61.18	31.84	61.23	**36.95**	66.59	33.98	58.72	30.21	52.84	28.12	53.47	32.87	61.12	32.45	55.78	**31.55**
✔	-	66.84	28.21	69.42	32.89	75.23	33.12	63.45	28.34	60.71	**31.59**	69.12	29.75	73.67	31.34	**64.27**	28.09
✔	✔	**68.53**	**32.27**	**72.74**	35.38	**78.63**	**36.42**	**65.84**	**30.97**	**63.17**	29.83	**70.92**	**33.57**	**75.36**	**34.69**	61.53	28.24

Table 13. Ablation study on training stages for HeartcareGPT-3.8B on *Comparison-QA* tasks from Heartcare-Bench^{C}.

Training Beat	Warming up	S–S		I–I		S–I		Avg.
		Cons.	Irr.	Cons.	Irr.	Cons.	Irr.	
-	-	44.61	45.18	47.52	51.44	49.68	52.12	48.43
-	✔	48.55	50.22	53.18	55.01	48.87	51.23	51.18
✔	-	63.77	**70.44**	68.45	69.38	72.14	77.29	70.25
✔	✔	**66.40**	67.47	**69.88**	**75.19**	**78.71**	**78.83**	**72.74**

Table 14. Ablation study on training stages for HeartcareGPT-3.8B on *Report Generation* tasks from Heartcare-Bench^{S} and Heartcare-Bench^{I}.

Training Beat	Warming up	Heartcare-Bench^{S}			Heartcare-Bench^{I}		
		Score^{GPT}	F1-Rad	Rouge-L	Score^{GPT}	F1-Rad	Rouge-L
-	-	55.12	17.11	26.05	60.32	19.57	30.27
-	✔	**65.21**	20.11	29.64	68.02	21.14	30.89
✔	-	64.33	22.12	32.20	66.29	22.78	35.43
✔	✔	61.29	**26.84**	**34.39**	**78.50**	**23.10**	**38.68**

Table 15. Ablation study on multimodal integration for HeartcareGPT-3.8B on *Closed-QA* tasks from proposed Heartcare-Bench^{S} and Heartcare-Bench^{I}. *Seg.* = Image segmentation.

Signal	Image	Seg.	Heartcare-Bench^{S}				Heartcare-Bench^{I}				Avg.
			Diagnosis	Waveform	Rhythm	Misc.	Diagnosis	Waveform	Rhythm	Misc.	
-	✓	✓	74.83	87.12	76.46	73.42	80.91	88.05	69.43	53.36	75.45
✓	-	-	72.61	85.37	74.84	63.92	75.46	86.19	71.43	56.18	73.25
✓	✓	-	79.12	93.26	81.64	76.43	84.98	87.65	77.43	62.84	80.42
✓	✓	✓	**81.95**	**95.94**	**82.79**	**79.84**	**87.85**	**92.21**	**79.25**	**67.80**	**83.33**

Table 16. Ablation study on multimodal integration for HeartcareGPT-3.8B on *Open-QA* tasks from our proposed Heartcare-Bench^{S} and Heartcare-Bench^{I}.

Signal	Image	Seg.	Heartcare-Bench^{S}								Heartcare-Bench^{I}							
			Diagnosis		Waveform		Rhythm		Misc.		Diagnosis		Waveform		Rhythm		Misc.	
			F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L	F1-Bio	Rouge-L
-	✓	✓	36.26	19.91	59.33	31.04	62.47	29.30	49.72	24.16	55.39	25.47	58.84	27.69	59.65	28.94	47.42	22.92
✓	-	-	38.74	21.56	51.62	26.83	64.29	30.84	36.67	23.05	43.68	22.19	48.56	24.71	50.97	25.64	32.12	20.14
✓	✓	-	58.18	29.47	70.79	33.56	75.91	29.28	63.87	30.29	62.43	**29.96**	69.71	32.65	73.48	32.61	60.28	**31.34**
✓	✓	✓	**68.53**	**32.27**	**72.74**	**35.38**	**78.63**	**36.42**	**65.84**	**30.97**	**63.17**	29.83	**70.92**	**33.57**	**75.36**	**34.69**	**61.53**	28.24

Table 17. Ablation study on multimodal integration for HeartcareGPT-3.8B on *Comparison-QA* tasks from Heartcare-Bench^{C}.

Signal	Image	Seg.	S–S		I–I		S–I		Avg.
			Cons.	Irr.	Cons.	Irr.	Cons.	Irr.	
-	✓	✓	48.23	57.79	56.91	54.74	60.12	53.87	55.28
✓	-	-	51.46	61.32	53.65	52.83	64.25	67.01	58.42
✓	✓	-	59.64	63.72	62.58	65.84	67.61	71.51	65.15
✓	✓	✓	**66.40**	**67.47**	**69.88**	**75.19**	**78.71**	**78.83**	**72.74**

Table 18. Ablation study on multimodal integration for HeartcareGPT-3.8B on *Report Generation* tasks from Heartcare-Bench^{S} and Heartcare-Bench^{I}.

Signal	Image	Seg.	Heartcare-Bench^{S}			Heartcare-Bench^{I}		
			Score^{GPT}	F1-Rad	Rouge-L	Score^{GPT}	F1-Rad	Rouge-L
-	✓	✓	58.46	16.83	33.29	66.28	20.07	35.79
✓	-	-	61.92	19.34	30.18	67.74	21.99	32.32
✓	✓	-	**64.37**	23.76	33.84	70.13	21.84	34.49
✓	✓	✓	61.29	**26.84**	**34.39**	**78.50**	**23.10**	**38.68**

Closed-QA Generation Prompt of Diagnosis

System Prompt:
You will be given structured ECG-related variables (age, gender, and a diagnosis-probability JSON). Your task is to generate exactly one JSON object that represents a single **Diagnosis Closed-QA** item. The output must contain the fields "type", "question", and "answer", and must follow all rules specified below. No additional commentary, explanations, or extra text is allowed.

Instruction:

- **Age:** {Integer}
- **Gender:** {Male / Female / Other}
- **Diagnosis Type**: {DIAGNOSIS_TYPE_JSON}
  Each key of the diagnosis type JSON is a possible diagnosis and value is a probability (0–100). For example:

```
{
    "Inferior myocardial infarction": 95.0,
    "Atrial fibrillation": 30.0
}
```

  If no diagnoses are present, it is an empty JSON.
- **Rules for Using Probabilities:**
  1. If there is at least one diagnosis **>= 60**, choose the highest-probability diagnosis as the correct option.
  2. If all diagnosis **< 60**, and the diagnosis type JSON is only composed of low-probability values or empty, then the correct answer must be one "Normal" expression.
  3. A "Normal" option may be included when it is not the correct answer, and its wording can follow any standard clinical phrasing indicating normal findings (e.g., expressions such as *"Normal ECG"*, "*Within normal limits"*, or equivalent variants).
  4. Distractors must be clinically plausible but must not overlap with any actual diagnosis in the diagnosis type JSON.
- **Requirements:**
  1. The **"question"** field must be a single string and must mention both the patient's age and gender. The phrasing should vary naturally. For example:
     - *"Based on the ECG signal of a 65-year-old male patient, ..."*
     - *"For the ECG of a 70-year-old female, ..."*
     - "*This ECG from a 55-year-old male patient indicates ..."*
     - *"Considering the ECG tracing of a 60-year-old patient (female), ..."*
  2. The question text must include ONLY four options labeled "A: ...; B: ...; C: ...; D: ..." (separated by semicolons), e.g., *"Based on the ECG signal of a 65-year-old male patient, Which diagnosis is most likely? A: Atrial fibrillation; B: No abnormality detected; C: Inferior myocardial infarction; D: Complete right bundle branch block"*.
  3. The **"answer"** field must be exactly the option label plus content, e.g., *"A: Atrial fibrillation"*.
  4. The correct answer must not consistently appear in the same option position; its placement among A–D should be randomized.
  5. The output must be must be in the form of JSON:

```
{
    "type": "DiagnosisClosedQA",
    "question": ...,
    "answer": ...,
}
```

Figure 7. Prompt for QA pairs generation in *Closed-QA* tasks, diagnosis dimension.

Evaluation Prompt

System Prompt:
You are a professional cardial expert. The diagnostic accuracy of the generated report was judged according to the reference report. There are 17 evaluation indicators, and the calculation method and examples of each indicator are given below. Please compare the generated report with the reference report and score strictly according to the evaluation criteria.

Instruction:

- **Reference Report:** {REFERENCE_REPORT}
- **Generated Report:** {GENERATED_REPORT}
- **Evaluation Criteria:**
  1. Completeness of abnormal features mentioned (higher=more complete): **10**,
  2. Completeness of key diagnoses included (higher=more complete): **10**,
  3. Absence of critical diagnostic errors (higher=better): **8**,

  ...
  17. Whether wording is appropriate, avoiding absolute expressions: **5**
- **Requirements:**
  1. Score each item in the criteria above from 0 to 100 based on comparison with the reference report.
     - A score **from 90 to 100** indicates full compliance with the description;
     - A score **from 80 to 89** indicates substantial compliance with the description;
     - A score **from 60 to 79** indicates partial non-compliance with certain aspects;
     - A score **below 60** indicates complete non-compliance.
  2. Calculate weighted dimension scores: `score_i × weight_i`.
  3. The final total score is the sum of all weighted dimension scores:
     `total_score = sum(score_i × weight_i) / sum(weight_i))`.
  4. The output must be must be in the form of JSON:

```
{
    "item_scores": {
        "1": score_1, "2": score_2, ..., "17": score_17
    },
    "total_score": total_score
}
```

Figure 8. Prompt for report evaluation.

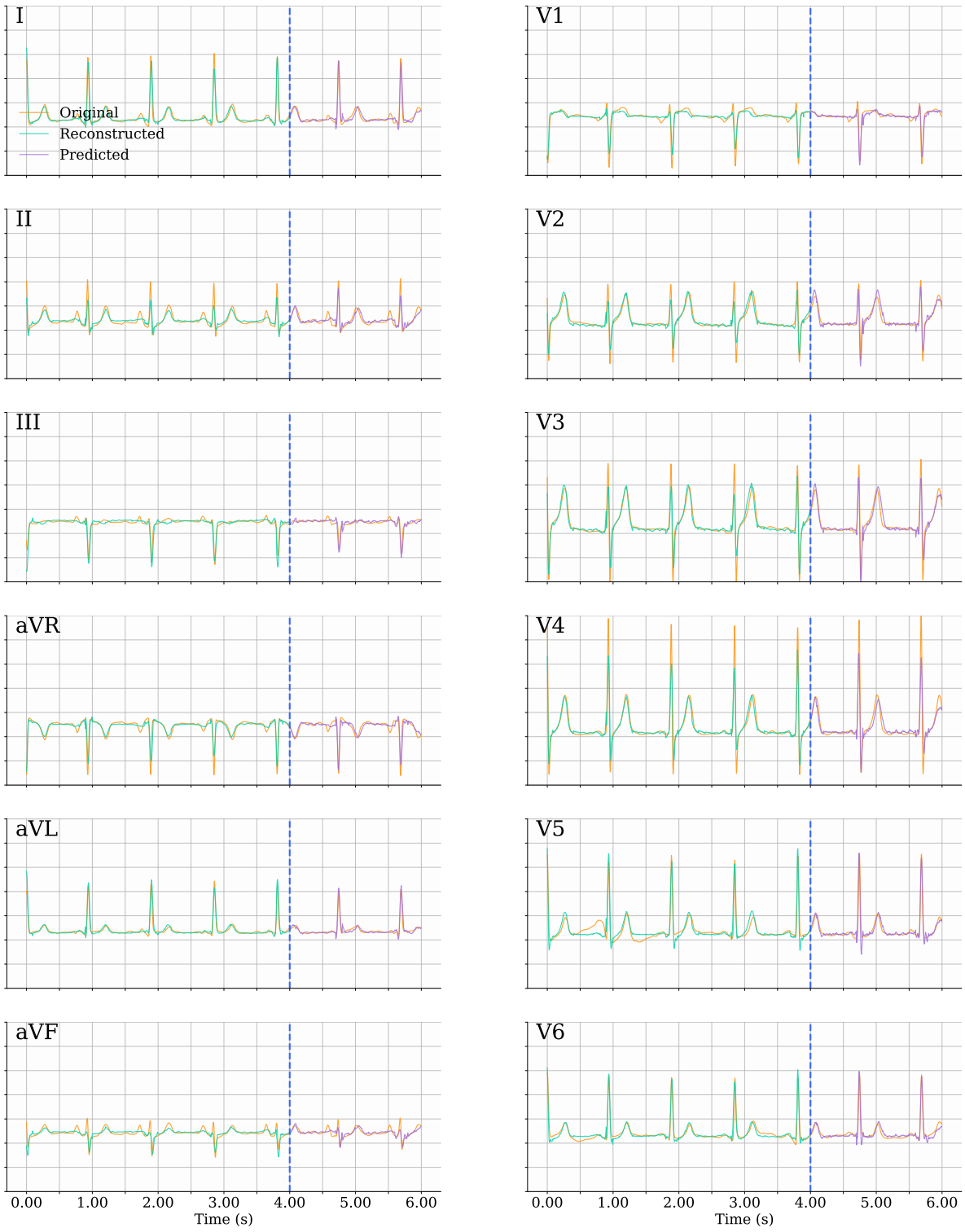


Figure 9. ECG signal reconstruction and prediction with Beat.

Question

Considering the ECG results of a 67-year-old female patient, what is the primary diagnosis?

Ground Truth

Anteroseptal myocardial infarction

HeartcareGPT

Ventricular premature complex and myocardial infarction.

MedVLM-R1-2B

The primary diagnosis for this patient is a 12-lead ECG showing ST-segment elevation in the inferior wall of the heart.

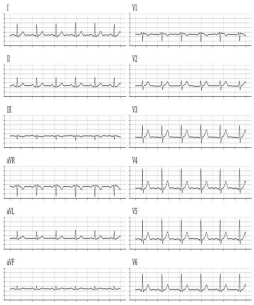
Claude-3.5

The ECG indicates Atrial Fibrillation with a rapid ventricular response.

Figure 10. A case of *open-QA* tasks.

Question

Please provide a comprehensive ECG interpretation for this 69.0-year-old male individual.



Ground Truth

First Degree AV Block

HeartcareGPT

Clinical 12-lead ECG Report

Patient Information:
- Age: 69 years
- Gender: Male

ECG Findings:
Diagnosis:
- First Degree AV Block: Highly likely (100.0% probability)
- Non-specific Ischemic Changes: Less likely (15.0% probability)

Rhythm:
- Baseline rhythm: Normal sinus rhythm
- Conduction issues: Presence of first degree AV block

Waveform:
- No significant waveform abnormalities detected

Summary:
This ECG indicates a first degree AV block, consistent with delayed AV node conduction. There is a low likelihood of non-specific ischemic changes. Overall, the minor conduction anomaly observed does not suggest severe cardiac pathology at present.

MedVLM-R1-2B

The ECG of this 69.0-year-old male individual shows a sinus rhythm with a right bundle branch block (RBBB) pattern. The rhythm is regular, and the waveform is normal. The RBBB pattern indicates that the electrical signals in the right bundle branch are delayed or blocked, which can affect the heart's ability to pump blood effectively. However, the overall ECG findings are normal, suggesting that the patient's heart function is within the expected range for his age.

Claude-3.5

ECG Report

Rhythm Description:
- Regular sinus rhythm
- Heart rate approximately 75-80 beats per minute
- Consistent P-QRS-T wave morphology across all leads

Key Waveform Findings:
- Normal P wave morphology and axis
- QRS complexes: Normal duration (0.08-0.10 seconds)
- ST segments: Isoelectric
- T waves: Upright and symmetrical in most leads
- No significant Q waves
- No ST segment elevation or depression
- No evidence of acute ischemia or infarction

Interpretation/Summary:
Normal 12-lead ECG demonstrating a stable sinus rhythm with no significant pathological findings. The electrical conduction system appears intact, with normal wave morphology and intervals. No acute cardiac abnormalities are detected in this tracing for this 69-year-old male patient.

Recommendations:
- Routine clinical correlation
- No urgent cardiac interventions suggested based on this ECG

Figure 11. A case of *Report Generation* tasks.