

anyECG-chat: A Generalist ECG-MLLM for Flexible ECG Input and Multi-Task Understanding

Haitao Li^{1,2}, Ziyu Li¹, Yiheng Mao¹, Ziyi Liu³, Zhoujian Sun⁴, Zhengxing Huang¹

¹Zhejiang University

²Shanghai Innovation Institute

³Transtek Medical Electronics Co., Ltd.

⁴Ant Group

lihaitao@zju.edu.cn, zhengxinghuang@zju.edu.cn

Abstract

The advent of multimodal large language models (MLLMs) has sparked interest in their application to electrocardiogram (ECG) analysis. However, existing ECG-focused MLLMs primarily focus on report generation tasks, often limited to single 12-lead, short-duration (10s) ECG inputs, thereby underutilizing the potential of MLLMs. To this end, we aim to develop a MLLM for ECG analysis that supports a broader range of tasks and more flexible ECG inputs. However, existing ECG-QA datasets are often monotonous. To address this gap, we first constructed the anyECG dataset, which encompasses a wide variety of tasks, including report generation, abnormal waveform localization, and open-ended question answering. In addition to standard hospital ECGs, we introduced long-duration reduced-lead ECGs for home environments and multiple ECG comparison scenarios commonly encountered in clinical practice. Furthermore, we propose the anyECG-chat model, which supports dynamic-length ECG inputs and multiple ECG inputs. We trained the model using a three-stage curriculum training recipe with the anyECG dataset. A comprehensive evaluation was conducted, demonstrating that anyECG-chat is capable of supporting various practical application scenarios, including not only common report generation tasks but also abnormal waveform localization for long-duration reduced-lead ECGs in home environments and comprehensive comparative analysis of multiple ECGs.

Code — <https://github.com/CuCl-2/anyECG-chat>

Introduction

Electrocardiograms (ECGs) play a pivotal role in clinical practice, serving as a cornerstone for diagnosing and monitoring various cardiac conditions (Sahoo et al. 2020; Rath et al. 2021; Ayano et al. 2022). Traditional methods for ECG analysis often rely on task-specific models, which are typically constrained to single tasks such as arrhythmia detection or report generation (Wang et al. 2023; Eldele et al. 2021; Zhang et al. 2022; Hu, Chen, and Zhou 2023; Na et al. 2024). These approaches lack the flexibility to handle diverse tasks and do not support interactive human-machine communication. The emergence of multimodal large language models (MLLMs) (Liu et al. 2023c; Wang et al. 2024;

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.




ReportGen
<p>ECG (10s, 12leads):</p>  <p>Q: Please provide the report for the following ECG. A: Report: Sinus rhythm, Poor R wave progression - probable normal variant, Borderline ECG</p>
Localization
<p>ECG (15s, 2leads):</p>  <p>Q: Examine this ECG and point out where the Premature ventricular contraction is located. A: Duration: 1.9s-3.1s, 6.8s-8.1s, 14.3s-15.0s</p>
MultiECG
<p>Three ECGs (10s, 12leads):</p>  <p>Q: How do the ECGs collected 0 days ago, 1323 days ago, and 1924 days ago compare in terms of rhythm and overall assessment? A: The ECGs collected at these times show sinus rhythm as the primary rhythm, with one instance of sinus bradycardia, and all are assessed as normal or normal except for rate, indicating a stable cardiac condition over time.</p>

Table 1: anyECG dataset examples.

Alayrac et al. 2022; Chen et al. 2024) has introduced a transformative paradigm for ECG analysis. By leveraging their powerful natural language reasoning capabilities and multimodal perception abilities, MLLMs can perform a wide range of tasks while enabling natural language-based human interaction. Consequently, ECG-focused MLLMs have garnered significant attention as a promising research direction.

Previous research on ECG MLLMs has explored various approaches. A straightforward method involves using an external classifier or feature extractor to convert ECG signals into a series of textual labels, which are then passed to the LLM (Liu et al. 2023a; Oh et al. 2023; Yu, Guo, and Sano 2023). However, this approach often results in significant in-

formation loss. Another line of work focuses on native ECG MLLMs, where ECG representations are directly fed into the LLM. For instance, MEIT (Wan et al. 2024) employs a projection mechanism to align ECG embeddings with the semantic space of the LLM, enabling the generation of ECG reports. Similarly, PULSE (Liu et al. 2024b) addresses real-world scenarios involving ECG images by constructing an ECG image-based MLLM that supports tasks such as report generation, waveform classification, and rhythm analysis. Building on these advancements, (Zhao et al. 2024) introduces multi-turn dialogue capabilities, enabling iterative interactions. Additionally, it leverages retrieval-augmented generation (RAG) (Lewis et al. 2020) to enrich the LLM with ECG-specific knowledge.

Despite the advancements in ECG-MLLMs, their current applications are predominantly limited to single-task scenarios such as report generation or label classification (Wan et al. 2024; Liu et al. 2024b; Zhao et al. 2024; Li et al. 2024a). In essence, ECG reports are composed of a series of labels related to rhythm, morphology, and diagnosis, making report generation and label classification fundamentally the same task (Gow et al. 2023; Wagner et al. 2020). However, the core objective of MLLMs is to address diverse, multi-task challenges rather than being limited to a single task (Wang et al. 2024). Consequently, existing ECG-MLLMs fail to fully harness the potential of MLLMs. Moreover, these models are typically restricted to processing single, 12-lead, 10-second ECG inputs (Wan et al. 2024; Liu et al. 2024b; Zhao et al. 2024), which are inadequate for modern use cases. For instance, they cannot effectively handle the long-duration, reduced-lead ECGs commonly generated in home environments (Gu et al. 2024) or the multi-ECG comparison scenarios frequently encountered in clinical practice. To bridge this gap, there is a pressing need for a more versatile ECG-MLLM capable of supporting a broader range of tasks, particularly fine-grained localization tasks, and accommodating more flexible ECG inputs, including long-duration ECGs, reduced-lead ECGs, and multiple ECGs.

However, existing ECG question-answering datasets (Oh et al. 2023; Wan et al. 2024; Liu et al. 2024b) are often overly simplistic and fail to meet the requirements for diverse tasks and flexible input scenarios. To address these limitations, we developed a novel dataset named anyECG, which comprises three subsets: ReportGen, Localization, and Multi-ECG. These subsets encompass a wide range of tasks, including report generation, abnormal waveform localization, and open-ended question answering. Additionally, we introduced long-duration ECGs, reduced-lead ECGs, and multi-ECG inputs to better align with modern clinical and home-monitoring scenarios.

To support these diverse tasks and flexible input formats, we propose the anyECG-chat Model which uses dynamic ECG input mechanism to support dynamic-length ECG inputs and multiple ECG inputs seamlessly. We employed a three-stage curriculum learning (Gong et al. 2023; Wang et al. 2024) approach to train the model, enabling it to evolve from coarse perception to fine-grained understanding, and ultimately to instruction-following and multi-ECG comparison tasks.

We evaluated our model on three tasks. In the ReportGen task, out-of-domain testing on six unseen ECG datasets showed superior generalization compared to existing ECG-MLLMs. For the Localization task, using a reserved test set, our model outperformed traditional segmentation models and other ECG-MLLMs by enabling fine-grained, second-level abnormality localization and handling dynamic-length ECG inputs. It also showed strong zero-shot performance in unseen single-lead scenarios. In the MultiECG task, our model consistently led on the MIMIC Multi-ECG QA and ECG-QA datasets. Furthermore, our model demonstrated robust multi-turn dialogue capabilities.

Our contributions can be summarized as follows:

- We introduce the anyECG dataset, which moves beyond traditional ECG report generation to fine-grained waveform localization and open-ended question answering. It also accommodates a wider variety of ECG input formats, including multi-ECG comparisons and long-duration, reduced-lead recordings.
- We proposed the anyECG-chat architecture, which is specifically designed to handle dynamic ECG inputs, enabling it to address the diverse scenarios presented by the anyECG dataset.
- We employed a three-stage curriculum learning approach, consisting of pre-training, fine-grained pre-training, and instruction tuning. The resulting anyECG-chat model demonstrates strong performance across various tasks, including report generation, waveform localization, and multi-ECG comparison.

Related Work

ECG Understanding: In recent years, the paradigm of ECG understanding has gradually shifted from traditional supervised learning (Ribeiro et al. 2020) to self-supervised learning (Chen et al. 2020; Grill et al. 2020; Chen and He 2021), which leverages large amounts of unlabeled data for pre-training. Self-supervised ECG learning can be broadly categorized into contrastive self-supervised learning (Wang et al. 2023; Eldele et al. 2021) and generative self-supervised learning (Zhang et al. 2022; Hu, Chen, and Zhou 2023; Na et al. 2024). Both approaches, however, require fine-tuning on downstream task data and are not inherently suited for zero-shot scenarios. Inspired by CLIP (Radford et al. 2021), several multimodal contrastive learning methods for ECG-report pairs (Li et al. 2024b; Liu et al. 2024a; Yu, Guo, and Sano 2024; Li et al. 2025) have emerged. However, these models lack a decoder and are therefore limited to discriminative tasks. They are ill-suited for diverse generative applications and cannot accommodate multiple tasks within a single model. In contrast, this paper introduces anyECG-chat, a generative MLLM capable of performing a wide range of tasks guided by textual instructions. This approach unlocks the potential for diverse and flexible applications in ECG understanding.

ECG-MLLMs Inspired by advancements in large vision-language models (Liu et al. 2023c; Wang et al. 2024; Alayrac et al. 2022; Chen et al. 2024), ECG-MLLMs have emerged as a promising direction for ECG analysis. A

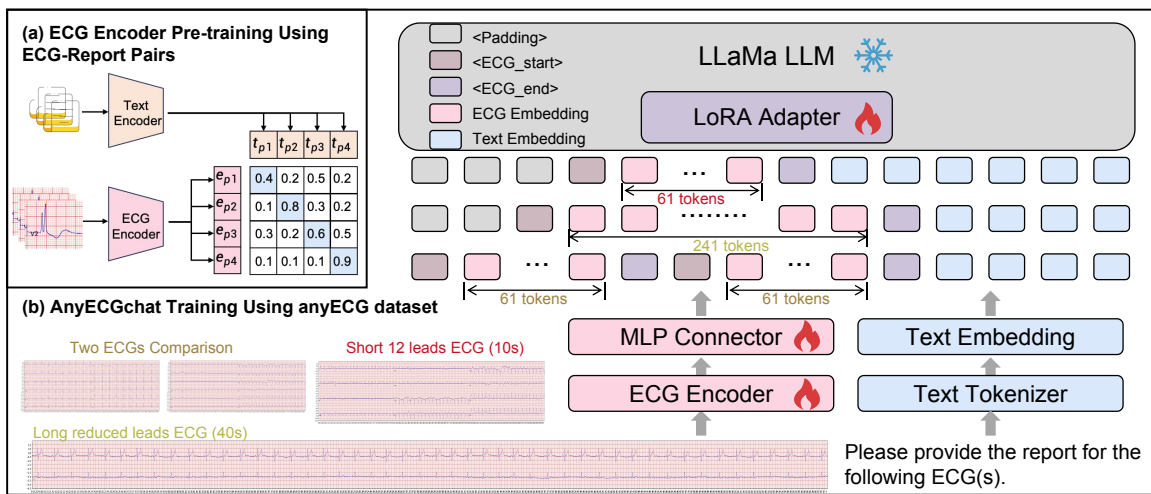


Figure 1: The overview of anyECG-chat architecture.

straightforward approach involves using external classifiers or feature extractors to convert ECG signals into a series of textual labels, which are then fed into the LLM (Liu et al. 2023a; Oh et al. 2023; Yu, Guo, and Sano 2023). However, this method often results in significant information loss. An alternative approach like MEIT (Wan et al. 2024) focuses on building native ECG MLLMs, where ECG embeddings are directly input into the LLM. PULSE (Liu et al. 2024b) addresses real-world scenarios involving ECG images by constructing an ECG image-based MLLM capable of supporting tasks such as report generation, waveform classification, and rhythm analysis. Building on these advancements, (Zhao et al. 2024) introduces multi-turn dialogue capabilities, allowing for iterative interactions. Additionally, it incorporates RAG (Lewis et al. 2020) to enhance the LLM with ECG-specific knowledge.

Despite these successes, existing ECG-MLLMs are often limited to report generation applications or label classification (Wan et al. 2024; Liu et al. 2024b; Zhao et al. 2024) and are typically restricted to processing single, 12-lead, 10-second ECG inputs. This limitation makes them inadequate for modern scenarios. This paper aims to develop a more versatile ECG-MLLM that supports a broader range of tasks and accommodates more flexible ECG inputs.

AnyECG Dataset

The existing ECG-QA datasets (Oh et al. 2023; Wan et al. 2024; Liu et al. 2024b) are relatively monotonous and fail to meet the requirements for supporting multi-task and flexible ECG inputs. To address this limitation, we constructed a novel dataset named anyECG. In terms of tasks, prior ECG-QA datasets are often restricted to report generation or label classification, which significantly underestimates the potential of MLLMs. Therefore, we introduced a broader range of tasks, including not only report generation but also more complex tasks requiring fine-grained perception, such as waveform localization, as well as diverse open-ended question answering facilitated by LLMs. Regard-

Dataset	Source	Duration	Leads	ECGs per QA	QA Pairs
ReportGen					
MIMIC-ECG ReportGen	MIMIC	10s	12	1	773,268
Localization					
ESTD Loc	ESTD	10s	2	1	39,110
ESTD Loc Long	ESTD	10-60s	2	1	19,555
MIT-ST Loc	MIT-ST	10s	2	1	6,500
MIT-ST Loc Long	MIT-ST	10-60s	2	1	3,250
MIT-Arr Loc	MIT-Arr	10s	2	1	54,440
MIT-Arr Loc Long	MIT-Arr	10-60s	2	1	27,220
MultiECG					
MIMIC Multi-ECG QA	MIMIC	10s	12	2-6	135,094
ECG-QA(10%)	PTB-XL	10s	12	1-2	33,220

Table 2: Overview of the anyECG Dataset. ESTD: European ST-T Database, MIT-ST: MIT-BIH ST Change Database, MIT-Arr: MIT-BIT Arrhythmia Database.

ing ECG signals, previous datasets typically utilize single, short-duration (10s), 12-lead ECG. This setup is inadequate for modern scenarios, such as the large volume of long-duration, reduced-lead ECGs generated in home environments, and the multi-ECG comparison scenarios commonly encountered in clinical practice. To this end, we incorporated long-duration ECGs, reduced-lead ECGs, and multi-ECG inputs into our dataset. Specifically, anyECG consists of three components: anyECG-ReportGen, anyECG-Localization, and anyECG-MultiECG. We provide an example from each component in the Table 1 and summarize the dataset statistics in Table 2. Notably, we standardized the sampling frequency of all ECG datasets to 100 Hz and normalized the ECG signals to a range of -1 to 1.

Our anyECG dataset is constructed by reorganizing existing datasets or with the assistance of LLM-generated data. Specifically, anyECG-ReportGen is constructed by reorganizing the MIMIC-ECG (Gow et al. 2023) dataset, anyECG-Localization is built by reorganizing three long-duration, 2-lead ECG datasets: the European ST-T Database (Taddei et al. 1992), the MIT-BIH ST Change Database (Albrecht 1983), and the MIT-BIT Arrhythmia Database (Moody and

Mark 2001). anyECG-MultiECG is constructed based on the MIMIC-ECG (Gow et al. 2023) and PTB-XL (Wagner et al. 2020) datasets. The detailed construction process is described in Appendix AnyECG Dataset Construction.

AnyECG-Chat Architecture

The architecture of our model is illustrated in Figure 1. It consists of an ECG encoder, a large language model (LLM), a modality alignment module, and LoRA adapters. Previous ECG MLLMs were often limited to single 12-lead, short-duration (10s) ECG inputs. To enable our model to handle the diverse scenarios and flexible ECG inputs in the anyECG dataset, we introduced a Dynamic ECG Input mechanism. We will elaborate on each component and the Dynamic ECG Input mechanism in detail below.

ECG Encoder

The performance of multimodal large language models (MLLMs) in question answering (QA) tasks heavily relies on the perceptual capabilities of the ECG encoder. Instead of training the ECG encoder from scratch, we opted to pre-train it using contrastive learning (Li et al. 2024b) on the MIMIC-ECG (Gow et al. 2023) dataset, which contains 800,000 ECGs and their corresponding reports.

We employed a ViT-base (Dosovitskiy et al. 2020) architecture as the ECG encoder. However, since ViT-base is originally designed for image data, we redefined the patching mechanism to accommodate the temporal and multi-lead nature of ECG signals; and introduced lead embeddings (Na et al. 2024), adjusted positional embeddings to capture the spatiotemporal structure of ECG data.

Specifically, let an ECG signal be represented as $X \in \mathbb{R}^{L \times T}$, where L is the number of leads and T is the signal length. First, we standardized the sampling frequency to 100 Hz and normalized each lead to the range $[-1, 1]$ to mitigate measurement biases from different devices and enhance generalization. To adapt the patching mechanism, we applied spatio-temporal patchifying with a patch size of $(1, 200)$. For example, given a preprocessed ECG from MIMIC-ECG $X \in \mathbb{R}^{12 \times 1000}$, each lead is divided into 5 patches, resulting in a total of 60 patches across all leads. Additionally, we introduced a [CLS] token to capture global features.

Traditional ViT models rely solely on positional embeddings, which are insufficient for capturing inter-lead relationships in ECG data. To address this limitation, we introduced lead embeddings, denoted as E_{lead} , to encode the spatial relationships between leads. Patches from the same lead share the same lead embedding, while patches from different leads at the same time share the same positional embedding. The final input embedding for each patch is computed as:

$$E = E_{\text{signal}} + E_{\text{pos}} + E_{\text{lead}},$$

where E_{pos} represents the positional embedding, and E_{signal} is the patch embedding derived from the ECG signal.

Large Language Model

In this paper, we utilize the Meta-Llama-3-8B-Instruct (Grattafiori et al. 2024) as our LLM. To prevent overfitting

and catastrophic forgetting, which could significantly degrade the model’s ability to respond to general queries, we opted for Low-Rank Adaptation (LoRA) (Hu et al. 2021) instead of full parameter fine-tuning. In anyECG-chat, we inject LoRA adapters (rank=8 and $\alpha = 16$) to the projection layers for query and key in all self-attention layers of the LLaMA model.

Modality Connector

Various modality connectors have been explored in prior research on vision-language models (VLMs), including cross-attention mechanisms (Alayrac et al. 2022), Q-formers (Li et al. 2023b), and simple linear projections (Liu et al. 2023c). In this work, to balance effectiveness and efficiency, we adopt a two-layer MLP with GELU activation as the modality connector, inspired by LLaVA 1.5 (Liu et al. 2023b).

Dynamic ECG Input

The Dynamic ECG Input mechanism is designed to empower anyECG-chat with the ability to handle diverse scenarios and flexible ECG inputs, including varying-length ECGs, reduced-lead ECGs, and multi-ECG inputs. To achieve this, two key challenges must be addressed: (1) embedding dynamic-length and reduced-lead ECGs effectively, and (2) ensuring that multiple ECG embeddings can be input into the LLM while maintaining clear distinctions between different ECGs.

For the first challenge, since our ECG encoder is pre-trained on the MIMIC-ECG dataset using 10-second, 100 Hz, 12-lead ECGs, we adopt the following strategies: For ECGs shorter than 10 seconds, zero-padding is applied to match the required length. For ECGs longer than 10 seconds, they are first padded to the nearest multiple of 10 seconds and then segmented into 10-second clips. These clips are individually processed by the ECG encoder, and the resulting embeddings are concatenated to get the final ECG embedding sequences. The [CLS] tokens from each segment are averaged to produce the final [CLS] embedding for the long-duration ECG. For reduced-lead ECGs, missing leads are similarly zero-padded to ensure compatibility with the encoder. As mentioned above, since our ECG encoder incorporates lead embeddings, it can capture the relationships between leads even for missing leads.

To address the second challenge, and to ensure the LLM can distinguish between multiple ECG inputs without conflating them into a single long-duration ECG, we introduce special tokens $\langle \text{ECG_start} \rangle$ and $\langle \text{ECG_end} \rangle$. These tokens are added before and after each ECG embedding, enabling the LLM to clearly identify and differentiate between individual ECG inputs.

Training Recipe

We designed a three-stage curriculum learning approach tailored to the varying complexity of tasks in the anyECG dataset. Inspired by (Gong et al. 2023; Wang et al. 2024), this approach comprises pretraining, fine-grained pretraining, and open-ended instruction tuning. Notably, the ECG

Stage	Trained Params	Training Task	Samples	LR	Batch Size	Epochs
1	Connector + ECG encoder	ReportGen	773,268	1×10^{-4}	256	2
2	Connector + ECG encoder + LoRA	ReportGen + Localization	923,343	1×10^{-4}	64	2
3	Connector + LoRA	ReportGen + Localization + MultiECG	1,091,657	1×10^{-4}	64	1

Table 3: Overview of the Training Recipe

encoder was pre-trained on the MIMIC-ECG dataset using contrastive learning prior to these three stages.

In Stage 1, the model was trained on the anyECG-ReportGen dataset with a frozen LLM; only the ECG encoder and Connector were updated to align ECG and LLM embeddings. Stage 2 added the more demanding anyECG-Localization dataset, requiring fine-grained waveform localization. Here, we jointly trained the ECG encoder, Connector, and fine-tuned the LLM using LoRA to enhance localization performance. In Stage 3, we introduced open-ended QA tasks using the full anyECG dataset, freezing the ECG encoder. This phase emphasized instruction-following and incorporated multi-ECG inputs for comparative reasoning. Notably, the dataset for each stage includes the dataset from the previous stage, preventing the model from forgetting. This progressive three-stage training strategy allowed the model to evolve from coarse perception to fine-grained understanding, and finally to instruction-following and multi-ECG comparison tasks. By gradually increasing task complexity, the approach mitigates the risk of the model relying excessively on textual reasoning, which could lead to hallucinations, especially when its ECG perception capabilities are underdeveloped. The training recipe and detailed hyperparameters are summarized in Table 3.

Experiments

We evaluated the performance of anyECG-chat across three tasks. For Report Generation task, since MIMIC-ECG dataset was used for contrastive pretraining and Stage 1 training, we performed out-of-domain testing on six unseen ECG datasets to ensure fairness. For Localization, we used the reserved test set from anyECG-Localization to evaluate the model’s performance and further assessed its zero-shot capability in single-lead scenarios. For Multi-ECG, we evaluated the model on the reserved test set of the MIMIC Multi-ECG QA dataset and the ECG-QA dataset, which contains multi-turn question-answering tasks. We also conducted a qualitative analysis of multi-turn instruction-following capabilities. An overview of the evaluation datasets is provided in Appendix Evaluation Dataset Overview.

Report Generation

As discussed above, we used six OOD ECG classification datasets to evaluate the generalization capability of anyECG-chat. Notably, ECG reports are essentially composed of labels, and using classification metrics to evaluate the model provides a more accurate measure of its understanding of ECGs compared to traditional text similarity metrics like BLEU or ROUGE. An example is that prediction ‘sinus tachycardia’ and ground truth ‘sinus bradycardia’ would yield a classification score of 0, while semantic

similarity metrics might still assign a non-zero score.

To compare anyECG-chat with existing models, we prompted the anyECG-chat with the query, “Please provide the report for the following ECG.” The reports generated by anyECG-chat and the dataset label names were then encoded using a text encoder (BioBERT (Deka, Jurek-Loughrey et al. 2022)). Finally, the cosine similarity between the text embeddings of anyECG-chat’s output and each label was computed to derive the prediction scores.

We compared anyECG-chat against several supervised methods (Wang et al. 2023; Na et al. 2024), discriminative zero-shot methods (Liu et al. 2024a), and other generative zero-shot methods (Li et al. 2023a; Liu et al. 2024b; Wan et al. 2024) using AUC as the evaluation metric. The results, presented in Table 4, demonstrate that anyECG-chat achieved the best performance among generative zero-shot methods like PULSE (Liu et al. 2024b) and MEIT (Wan et al. 2024). Although our method does not outperform discriminative zero-shot methods, such a comparison is actually unfair because MERL (Liu et al. 2024a) uses the labels of each dataset as prior knowledge, whereas our model does not require any label information. Our method directly generates labels without relying on prior knowledge. Despite this unfair comparison, our model still achieves comparable performance on PTBXL-Rhythm, PTBXL-Sub, and CSN. We also present the results using semantic similarity metrics as a reference in Table 5. The reported metrics are averaged across the six datasets.

Localization Task

Results For the localization task, we used the reserved test set from anyECG-Localization to evaluate the model’s performance using the Intersection over Union (IoU) metric. We also compared its performance against other supervised methods (Moskalenko, Zolotych, and Osipov 2020) and existing ECG-MLLMs. As expected, other ECG-MLLMs lacked the fine-grained temporal perception required for second-level localization. When asked to identify the location of abnormal waveforms, they could only provide lead-level answers (see Appendix Case Study for detail). Detailed results are presented in Figure 2.

For short-duration ECGs, the results demonstrate that although Unet (Moskalenko, Zolotych, and Osipov 2020) is a dedicated model for segmentation tasks, anyECG-chat significantly outperformed Unet on the European ST-T and MIT-BIH ST Change datasets and achieved comparable performance on the MIT-BIH Arrhythmia dataset. Interestingly, we observed that Unet exhibited consistent performance across different datasets, whereas anyECG-chat showed varying performance. This discrepancy may be attributed to the diverse training data used for anyECG-chat,

macro-AUC	PTBXL Super	PTBXL Sub	PTBXL Form	PTBXL Rhythm	CPSC	CSN
Supervised: <i>dedicated model tailored for each dataset</i>						
ASTCL	81.02	76.51	66.99	76.05	79.51	75.79
ST-MEM	71.36	63.59	66.07	74.85	70.39	71.36
Discriminative Zero-Shot: <i>requires pre-defined labels</i>						
MERL	74.20	75.70	65.90	78.50	82.80	74.40
Generative Zero-Shot: <i>directly outputs labels</i>						
LLaVa-Med	51.21	58.33	69.12	75.77	56.07	60.54
MEIT	62.34	57.91	61.12	70.45	62.38	62.73
PULSE	66.61	61.32	63.82	73.91	66.15	64.18
anyECG-chat	68.95	73.10	64.55	77.60	71.05	71.29

Table 4: Results of Classification.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
LLaVa-Med	0.47	0.44	0.40	0.38	0.62	0.57	0.52
MEIT	0.47	0.43	0.40	0.37	0.64	0.60	0.54
PULSE	0.50	0.46	0.43	0.40	0.68	0.64	0.58
anyECG-chat	0.53	0.51	0.47	0.44	0.72	0.68	0.60

Table 5: Results of Report Generation

beyond the anyECG-localization dataset, which likely enhanced its ability to perceive different types of abnormalities across datasets. For long-duration ECGs, Unet was unable to handle dynamic-length ECGs due to architectural limitations, whereas anyECG-chat successfully processed these inputs, further showcasing its flexibility and robustness.

Zero-Shot Single Lead Localization Though anyECG-localization dataset only includes 2-lead ECGs, we also evaluated the model’s zero-shot capability in single-lead scenarios. Three single-lead cases were tested: masking the first lead, masking the second lead, and masking a random lead, with the masked lead values set to zero. The results, shown in Figure 2, indicate that anyECG-chat achieves comparable performance in both short-duration and long-duration ECGs when the first lead is masked in the European ST-T dataset and when the second lead is masked in the MIT-BIH ST Change and MIT-BIH Arrhythmia datasets. This demonstrates the model’s zero-shot capability in single-lead scenarios. However, performance drops significantly when the other lead is masked, likely because the queried abnormal waveform features are present only in the masked lead.

Multi-ECG Comparison

For the multi-ECG comparison task, we evaluated our model using two datasets: MIMIC Multi-ECG QA and ECG-QA. The former includes scenarios involving comparisons of 2 to 6 ECGs, while the latter focuses solely on comparisons between 2 ECGs. As previously mentioned, since the answers in ECG-QA are relatively concise, we limited the training data to 10% of the original dataset to prevent the model from overfitting to short responses.

MIMIC Multi-ECG QA Since the MIMIC Multi-ECG QA dataset is constructed using Llama-3.3-70B-Instruct (Touvron et al. 2023; Grattafiori et al. 2024) for open-ended QA tasks, it lacks explicit metrics for direct evaluation. To address this, we employed QwQ-32B (Team 2025; Yang et al. 2024), as the evaluation model. To ensure fairness,

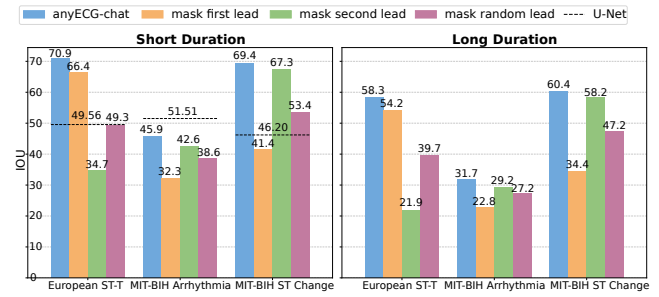


Figure 2: Results of Localization and Zero-Shot Single Lead ECG Localization. Since LLaVa-Med, MEIT and PULSE failed to provide second-level localization, scoring 0, they are omitted from the figure.

we did not use the answer generated by Llama-3.3-70B-Instruct as the gold standard for QwQ’s evaluation. Instead, we provided QwQ with the questions and the corresponding reports for each ECG, allowing it to assess the quality of the outputs without bias. The evaluation scores ranged from 0 to 5. The detailed prompt is provided in Appendix Case Study. Furthermore, to further validate the accuracy of LLM-based evaluation, we sampled 120 data points for human scoring, as detailed in Appendix Human Scoring, which demonstrates a strong correlation between human scores and QwQ’s scores.

We compared the outputs of anyECG-chat and other ECG-MLLMs. It is worth noting that since LLaVa-Med, MEIT and PULSE were not trained to handle multi-ECG inputs, we adapted their usage to support multi-ECG comparison tasks while maintaining consistency with their training setup. Specifically, we first processed each ECG individually to generate its corresponding report. These reports were then concatenated, along with an image combining all the ECGs, and provided as input to ECG-MLLMs to answer multi-ECG comparison questions. The score distributions for these models are shown in Figure 3. Notably, anyECG-chat achieved significantly higher scores compared to the other two models. Additionally, we analyzed the average scores of each model across different numbers of ECG inputs, as well as the number of times each model achieved the highest score among the three models. The results, as shown in Table 6, indicate that anyECG-chat maintains notable robustness as the number of input ECGs increases. Furthermore, anyECG-chat secured the highest score in 816 out of 1,152 questions, demonstrating a substantial performance advantage over the other two models.

ECG-QA For the ECG-QA dataset, we compared anyECG-chat with several discriminative models (Chen et al. 2022; Moon et al. 2022) and other ECG-MLLMs (Liu et al. 2024b; Wan et al. 2024). As mentioned earlier, the answers in ECG-QA are relatively concise, often limited to a few short phrases. Consequently, discriminative methods model the QA task as a multi-label classification problem, which requires predefined possible labels as prior knowledge. In contrast, ECG-MLLMs, as generative methods, directly produce answers without relying on predefined labels.

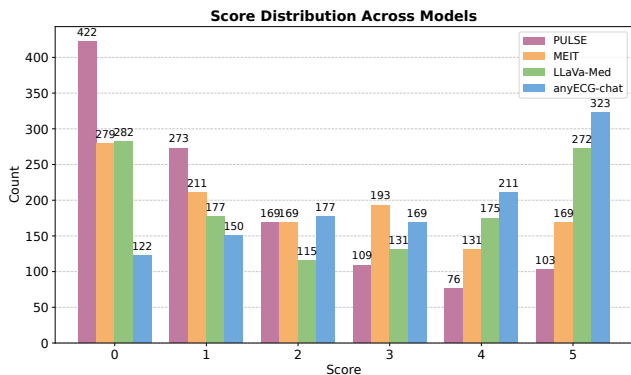


Figure 3: The Score Distribution on MIMIC Multi-ECG QA.

# ECGs	2	3	4	5	6	All	Highest (#)
PULSE	1.66	1.40	1.13	1.46	1.84	1.53	264/1152
MEIT	2.10	1.85	1.75	1.80	2.00	1.90	415/1152
LLaVa-Med	2.75	2.19	2.34	2.09	2.24	2.48	553/1152
anyECG-chat	3.30	2.81	2.59	2.63	2.98	3.02	816/1152

Table 6: Average Scores for MIMIC Multi-ECG QA Across Different Numbers of ECG Inputs.

We used exact match accuracy as the evaluation metric, and the results are presented in Table 7. Although anyECG-chat does not outperform discriminative models that leverage predefined labels, it achieves the best performance among generative ECG-MLLMs, even when trained on only 10% of the training data. Notably, it excels in CI-Verify and CC-Verify tasks, achieving accuracies of 70.1% and 67.9%, respectively, demonstrating its strong capability in multi-ECG comparison tasks.

Ablation Study and Analysis

Model Architecture We performed a series of ablation studies, the results are summarized in Table 8. The reported metrics represent the average performance on the three dimensions. The ablation studies reveal several critical insights. First, initializing the ECG encoder with random weights instead of using multimodal contrastive pre-training significantly degrades performance across all three tasks, underscoring the importance of pre-trained representations in capturing rich ECG features. Second, removing lead embedding results in slight performance drops, particularly in tasks requiring complex spatial relationships, such as localization and MultiECG. Third, omitting the dynamic input mechanism restricts the model to processing fixed 10-second ECG inputs. While short ECG classification tasks remain largely unaffected, localization and MultiECG performance suffer due to the inability to handle long-duration or multi-segment ECGs.

Training Strategy Fully fine-tuning all parameters instead of using LoRA substantially increases computational cost without noticeable performance improvements, highlighting the efficiency of LoRA-based adaptation. When experience replay was removed, where earlier data and tasks

EM Acc.	S	S	S	CC	CC	CI	CI
	Verify	Choose	Query	Verify	Query	Verify	Query
Discriminative Model: <i>requires possible labels</i>							
M ³ AE	74.6	57.1	41.0	75.5	20.1	75.3	4.2
MedViLL	73.9	54.1	40.4	74.3	22.0	77.5	3.5
Fusion Transformer	72.1	46.4	37.4	71.9	18.4	68.1	2.2
Generative ECG-MLLM: <i>directly outputs answers</i>							
LLaVa-Med(0%)	34.7	0	0	11.9	0	36.8	0
MEIT(0%)	42.2	2.5	0.6	28.3	0.4	40.5	0
PULSE(100%)	64.6	56.1	2.4	52.9	3.9	57.1	0
anyECG-chat(10%)	69.6	50.1	20.1	68.0	8.6	72.1	1.2

S: Single, CC: Comparison-Consecutive, CI: Comparison-Irrelevant.

Table 7: Performance Comparison on ECG-QA.

Configuration	Classification	Localization	MultiECG
Default	71.09	56.10	3.02
w/o Contrastive Pre-training	68.24	53.69	2.82
w/o Lead Embedding	70.10	55.53	2.93
w/o Dynamic Input Mechanism	70.55	52.66	2.74
Full Parameter Tuning	71.02	56.15	3.04
w/o Experience Replay	66.32	53.88	3.04
w/o Curriculum Training	61.21	50.62	2.95

Table 8: Ablation Study Results.

were excluded from subsequent training stages. This approach led to noticeable forgetting of previously learned information (Rolnick et al. 2019; Scialom, Chakrabarty, and Muresan 2022). Additionally, when curriculum training was eliminated, the three-stage training pipeline was replaced with a single-stage training approach. This resulted in a substantial decline in performance across both Classification and Localization tasks. The drop can be attributed to the inherent complexity of the MultiECG task, which involves handling open-ended question answering. Directly starting training on such high-difficulty data caused the model to rely heavily on its language modeling capabilities, leading to increased hallucination.

Multi-Turn QA Although anyECG dataset contains only single-turn QA, we hypothesize that anyECG-chat can handle multi-turn QA due to LoRA-based fine-tuning, which preserves the LLM’s pre-trained abilities. As illustrated in Appendix Multi-Turn QA, the model shows strong multi-turn instruction-following behavior, indicating its potential as a teaching aid for physicians despite the lack of quantitative evaluation.

Conclusion

In this paper, we introduced anyECG-chat, a MLLM designed for diverse ECG analysis tasks. By leveraging the novel anyECG dataset and a three-stage curriculum training strategy, anyECG-chat demonstrated strong performance across report generation, waveform localization, and multi-ECG comparison tasks. The proposed Dynamic ECG Input mechanism further enhanced the model’s flexibility, enabling it to handle varying-length, reduced-lead, and multi-ECG inputs seamlessly. Experimental results showed that anyECG-chat outperformed existing ECG-MLLMs in multiple scenarios and exhibited robust zero-shot capabilities.

Acknowledgments

This work was partially supported by the Technical Innovation key project of Zhejiang Province (2024C03023), and the National Nature Science Foundation of China under Grant No. 82272129 and 82572380.

References

- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.
- Albrecht, P. 1983. *ST Segment Characterization for Long Term Automated ECG Analysis*. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.
- Ayano, Y. M.; Schwenker, F.; Dufera, B. D.; and Debelee, T. G. 2022. Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review. *Diagnostics*, 13(1): 111.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.
- Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.
- Chen, Z.; Du, Y.; Hu, J.; Liu, Y.; Li, G.; Wan, X.; and Chang, T.-H. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 679–689. Springer.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; 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*.
- Deka, P.; Jurek-Loughrey, A.; et al. 2022. Evidence Extraction to Validate Medical Claims in Fake News Detection. In *International Conference on Health Information Science*, 3–15. Springer.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwok, C. K.; Li, X.; and Guan, C. 2021. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.
- Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; and Glass, J. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.
- Gow, B.; Pollard, T.; Nathanson, L. A.; Johnson, A.; Moody, B.; Fernandes, C.; Greenbaum, N.; Waks, J. W.; Eslami, P.; Carbonati, T.; et al. 2023. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. *Type: dataset*, 6: 13–14.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent: a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.
- Gu, H. Y.; Huang, J.; Liu, X.; Qiao, S. Q.; and Cao, X. 2024. Effectiveness of single-lead ECG devices for detecting atrial fibrillation: An overview of systematic reviews. *Worldviews on Evidence-Based Nursing*, 21(1): 79–86.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hu, R.; Chen, J.; and Zhou, L. 2023. Spatiotemporal self-supervised representation learning from multi-lead ECG signals. *Biomedical Signal Processing and Control*, 84: 104772.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.
- Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llavamed: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.
- Li, H.; Li, Z.; Mao, Y.; Liu, Z.; Sun, Z.; and Huang, Z. 2024a. De-biased Multimodal Electrocardiogram Analysis. *arXiv preprint arXiv:2411.14795*.
- Li, H.; Liu, C.; Ding, Z.; Liu, Z.; and Huang, Z. 2025. Fine-Grained ECG-Text Contrastive Learning via Waveform Understanding Enhancement. *arXiv preprint arXiv:2505.11939*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, J.; Liu, C.; Cheng, S.; Arcucci, R.; and Hong, S. 2024b. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, 402–415. PMLR.
- Liu, C.; Ma, Y.; Kothur, K.; Nikpour, A.; and Kavehei, O. 2023a. BioSignal Copilot: Leveraging the power of LLMs in drafting reports for biomedical signals. *medRxiv*, 2023–06.
- Liu, C.; Wan, Z.; Ouyang, C.; Shah, A.; Bai, W.; and Arcucci, R. 2024a. Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement. *arXiv preprint arXiv:2403.06659*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023b. Improved Baselines with Visual Instruction Tuning.

- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual Instruction Tuning.
- Liu, R.; Bai, Y.; Yue, X.; and Zhang, P. 2024b. Teach Multi-modal LLMs to Comprehend Electrocardiographic Images. *arXiv preprint arXiv:2410.19008*.
- Moody, G. B.; and Mark, R. G. 2001. The impact of the MIT-BIH arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3): 45–50.
- Moon, J. H.; Lee, H.; Shin, W.; Kim, Y.-H.; and Choi, E. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12): 6070–6080.
- Moskalenko, V.; Zolotykh, N.; and Osipov, G. 2020. Deep learning for ECG segmentation. In *Advances in neural computation, machine learning, and cognitive research III: selected papers from the XXI international conference on neuroinformatics, October 7-11, 2019, Dolgoprudny, Moscow Region, Russia*, 246–254. Springer.
- Na, Y.; Park, M.; Tae, Y.; and Joo, S. 2024. Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram. *arXiv preprint arXiv:2402.09450*.
- Oh, J.; Bae, S.; Lee, G.; Kwon, J.-m.; and Choi, E. 2023. ECG-QA: A Comprehensive Question Answering Dataset Combined With Electrocardiogram. *arXiv preprint arXiv:2306.15681*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Rath, A.; Mishra, D.; Panda, G.; and Satapathy, S. C. 2021. Heart disease detection using deep learning methods from imbalanced ECG samples. *Biomedical Signal Processing and Control*, 68: 102820.
- Ribeiro, A. H.; Ribeiro, M. H.; Paixão, G. M.; Oliveira, D. M.; Gomes, P. R.; Canazart, J. A.; Ferreira, M. P.; Andersson, C. R.; Macfarlane, P. W.; Meira Jr, W.; et al. 2020. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature communications*, 11(1): 1760.
- Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.
- Sahoo, S.; Dash, M.; Behera, S.; and Sabut, S. 2020. Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey. *Irbm*, 41(4): 185–194.
- Scialom, T.; Chakrabarty, T.; and Muresan, S. 2022. Fine-tuned language models are continual learners. *arXiv preprint arXiv:2205.12393*.
- Taddei, A.; Distanti, G.; Emdin, M.; Pisani, P.; Moody, G.; Zeelenberg, C.; and Marchesi, C. 1992. The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. *European heart journal*, 13(9): 1164–1172.
- Team, Q. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Kreiseler, D.; Lunze, F. I.; Samek, W.; and Schaeffter, T. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific data*, 7(1): 154.
- Wan, Z.; Liu, C.; Wang, X.; Tao, C.; Shen, H.; Peng, Z.; Fu, J.; Arcucci, R.; Yao, H.; and Zhang, M. 2024. MEIT: Multi-modal electrocardiogram instruction tuning on large language models for report generation. *arXiv preprint arXiv:2403.04945*.
- Wang, N.; Feng, P.; Ge, Z.; Zhou, Y.; Zhou, B.; and Wang, Z. 2023. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Yu, H.; Guo, P.; and Sano, A. 2023. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In *Machine learning for health (ML4H)*, 650–663. PMLR.
- Yu, H.; Guo, P.; and Sano, A. 2024. ECG Semantic Integrator (ESI): A Foundation ECG Model Pretrained with LLM-Enhanced Cardiological Text. *arXiv preprint arXiv:2405.19366*.
- Zhang, H.; Liu, W.; Shi, J.; Chang, S.; Wang, H.; He, J.; and Huang, Q. 2022. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–15.
- Zhao, Y.; Zhang, T.; Wang, X.; Han, P.; Chen, T.; Huang, L.; Jin, Y.; and Kang, J. 2024. ECG-Chat: A Large ECG-Language Model for Cardiac Disease Diagnosis. *arXiv preprint arXiv:2408.08849*.