EchoGPT: An Interactive Cardiac Function Assessment Model for Echocardiogram Videos

Bo Xu[1] , **Quanhao Zhu**[1] , **Qingchen Zhang**[2] , **Mengmeng Wang**[3] , **Liang Zhao**[1*] , **Hongfei Lin**[1] , **Jing Ren**[4] and **Feng Xia**[4]
[1]Dalian University of Technology, China
[2]Hainan University, China
[3]Zhejiang University of Technology, China
[4]RMIT University, Australia
{boxu, liangzhao, hflin}@dlut.edu.cn, zhuqh19@gmail.com, zhangqingchen@hainanu.edu.cn, mengmengwang@zju.edu.cn, jing.ren@ieee.org , f.xia@ieee.org

Abstract

With the development of wearable cardiac ultrasound devices, it is no longer sufficient to solely rely on doctors for diagnosing long-term echocardiogram videos. Automated diagnosis of echocardiogram videos has now become a research hotspot. Existing studies only analyze echocardiogram video through discriminative models, which have limited question-answering capabilities. Therefore, this study innovatively proposes a large language model with cardiac ultrasound diagnostic capabilities—EchoGPT. EchoGPT integrates the robust communication and comprehension capabilities of large language models (LLMs) with the diagnostic prowess of traditional medical models, empowering patients to obtain accurate medical indicator data and comprehend their health conditions through interactive questioning with the model. The model is capable of local deployment on personal computers, effectively safe guarding user privacy. EchoGPT operates through three main components: left ventricle segmentation, left ventricular ejection fraction (LV_{EF}) prediction, and finetuning of video-text LLMs. Experimental results demonstrate EchoGPT's superior accuracy in predicting LV_{EF} compared to other models, and positive feedback from professional physicians through questionnaire surveys, validating its potential in practical applications. The demo is available at https://github.com/zhuqh19/EchoGPT.

1 Introduction

In the realm of modern medicine, the continuous development of new materials and the proliferation of wearable devices have led to an increasing demand for portable physiological data monitoring devices. Particularly in the field of echocardiogram video monitoring, advancements in Bioadhesive Ultrasound (BAUS) [Liu *et al.*, 2024b; Wang *et al.*, 2022] technology have opened new possibilities for cardiac health monitoring. However, given that the video data collected by wearable devices tend to be of considerable duration (usually exceeding one hour), it is extremely challenging for physicians to manually diagnose echocardiogram videos, and the accuracy of such diagnoses cannot be guaranteed. Automated diagnosis of echocardiogram videos has now become a research hotspot.

In order to address the aforementioned issue, it is necessary to develop a multimodal model capable of diagnosing based on echocardiogram videos. Multimodal models like CLIP, including BioMedCLIP, PubMedCLIP and EchoCLIP [Radford *et al.*, 2021; Zhang *et al.*, 2024b; Eslami *et al.*, 2023; Christensen *et al.*, 2024], excel in understanding tasks and offer preliminary diagnostics. However, they often overlook patients' comprehension, hindering effective communication. Additionally, their accuracy in medical metrics and general question-answering capabilities are limited compared to larger language models, restricting their interactive and precise utility. With the development of large language models [OpenAI *et al.*, 2024; Sun *et al.*, 2025], especially multimodal large language models, a new avenue has been paved for patients to gain immediate insights into their health conditions. Existing video-text multimodal large models [Wang *et al.*, 2024; Ataallah *et al.*, 2024a; Zhang *et al.*, 2024c; Cheng *et al.*, 2024], trained mostly on general datasets, struggle to accurately capture the features of medical data, resulting in inaccurate responses. In some medical imaging fields, relevant multimodal image-text large language models (LLMs) have emerged. For instance, SkinGPT [Zhou *et al.*, 2024] is a multimodal LLM designed for the diagnosis of dermatological images. However, there remains a dearth of corresponding multimodal LLMs in the domain of medical video diagnosis, especially echocardiogram video.

In light of these challenges, as shown in Figure 1, this study introduces an innovative medical diagnostic large language model tailored for echocardiogram videos—EchoGPT. This model combines the powerful communication and comprehension capabilities of large language models with the di-

*Corresponding author.

agnostic capabilities of traditional medical models, enabling patients to receive accurate medical indicator data and understand their health conditions through interactive questioning with the model. This not only enhances patients' self-monitoring capabilities and alleviates the workload of physicians but also supports local deployment on personal computers, effectively safeguarding users' personal privacy.

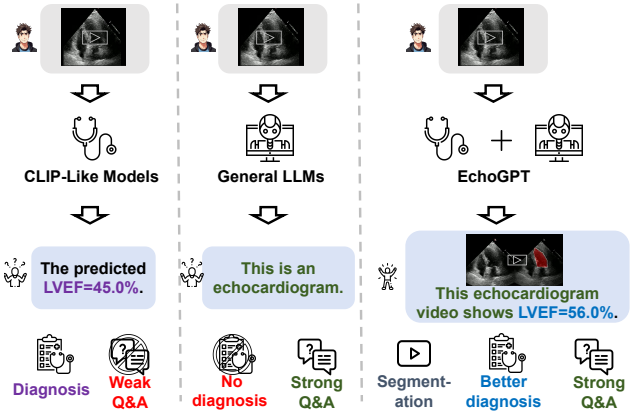

Figure 1: Challenge and our solution.

The primary contributions of our paper are as follows:

- **Pioneering Application of LLM in Echocardiogram Diagnosis**: To the best of our knowledge, EchoGPT is the first multimodal echocardiogram diagnosis large language model (LLM). EchoGPT represents the first instance of leveraging LLM in the field of medical video modeling, especially echocardiogram video diagnosis, marking a significant advancement in the application of AI technology for cardiac health monitoring.
- **Local Deployment for Enhanced Privacy and Accessibility**: EchoGPT's capability for local deployment on personal computers positions it at the forefront of privacy protection in medical diagnostics. As advanced sensors become more ubiquitous, the model's advantages in safeguarding user privacy and facilitating convenient access to healthcare insights will become increasingly evident.
- **State-of-the-Art Performance in Diagnostic Accuracy and Interactivity**: EchoGPT achieves cutting-edge performance in both diagnostic accuracy and interactivity. The model's sophisticated understanding of echocardiogram videos and its ability to engage in interactive questioning with patients set new standards for AI-assisted medical diagnostics.

2 Related Work

Medical Vision-Language Models. The integration of visual and linguistic data in medicine has led to advanced models that boost understanding and analysis. These models handle both visual and textual medical data, aiding in diagnosis, research, and patient care. This development is a major step in medical AI, offering tools to enhance healthcare professionals' expertise and patient outcomes. While CLIP [Radford *et al.*, 2021] has been influential, its general training has limitations in medical contexts. PubMedCLIP [Eslami *et al.*, 2023], fine-tuned on radiographic data, shows better performance in radiology tasks. BioMedCLIP [Zhang *et al.*, 2024b], pre-trained on scientific data, outperforms general models in biomedical tasks. EchoCLIP [Christensen *et al.*, 2024], tailored for echocardiogram videos, improves performance on relevant datasets. However, these models still face challenges in user interactivity and data accuracy due to CLIP's inherent constraints.

Video-Text Multimodal LLMs. In the general video description domain, multimodal LLMs such as Alibaba's Qwen2-VL [Bai *et al.*, 2023; Wang *et al.*, 2024] with visual enhancement, Peking University's LLava-Video [Zhang *et al.*, 2024c] with two-stage training, and Video-Llama2 [Zhang *et al.*, 2023; Cheng *et al.*, 2024] from the Video-LLaMA series are advancing rapidly. MiniGPT4-Video [Ataallah *et al.*, 2024a; Ataallah *et al.*, 2024b; Chen *et al.*, 2023; Zhu *et al.*, 2023] is notable for translating visual features into LLM space. However, these models perform poorly in medical data comprehension despite excelling in general video understanding and interactivity. These AI models integrate visual and linguistic processing for innovative healthcare applications. In medical 2D imaging, numerous multimodal image-text LLMs have emerged, such as SkinGPT for dermatology, CheXGPT for chest X-rays, Dia-LLaMA for CT report generation, ECG-LLM for electrocardiography, and miniGPT-Med for general medical visual question answering [Zhou *et al.*, 2024; Gu *et al.*, 2024; Chen *et al.*, 2024; Yu *et al.*, 2024; Alkhaldi *et al.*, 2024]. Nevertheless, there are still no relevant multimodal LLMs applied to echocardiogram video diagnosis.

3 Method

3.1 Problem Formulation

The model framework proposed in this study aims to achieve automated assessment of cardiac function through the integrated analysis of echocardiogram videos and user-submitted textual queries. The inputs to the model include an echocardiogram video x and a user query text y. The format of the video data x is defined as $x \in \mathbb{R}^{H\times W\times C\times T}$, where H, W, C, and T represent the height, width, number of color channels, and the number of frames of the video, respectively. For the user query text y, its format is defined as $y \in \mathbb{R}^{N\times M}$, where N denotes the number of words, and M denotes the vector dimension of each word. The objective of this research is to ultimately output an echocardiogram video x_{seg} with a left ventricular segmentation mask to provide a visual result to aid in clinical diagnosis. Furthermore, we aspire to generate a medical diagnostic report that includes the specific left ventricular ejection fraction (LV_{EF}) and enable the model to provide corresponding responses to user queries y.

3.2 Model Framework of EchoGPT

Our model framework, as depicted in Figure 2, comprises the following three parts:

Part 1: Left Ventricle Segmentation. The objective of this module is to precisely segment the left ventricle region from echocardiogram videos, providing visual reference and guidance for both physicians and users. The input video x is first divided into a sequence of frames, then deep visual features are extracted through an Atrous Convolution network, followed by a linear layer that maps these features to a segmentation mask space. The final output is a video with a left ventricle segmentation mask, offering physicians assisting visual diagnostic results and aiding users in more intuitively understanding the diagnostic outcomes.

Part 2: $\mathbf{LV_{EF}}$ Prediction. In this part, the model aims to predict the left ventricular ejection fraction (LV_{EF}) based on echocardiogram videos. Visual features are extracted from the video data x through 3D CNNs and spatiotemporal convolutional networks, followed by a linear layer that maps these features to the ejection fraction space to obtain the predicted LV_{EF} value. This LV_{EF} value serves as prompting information, transmitted alongside the user's query to the Text Tokenizer to obtain corresponding textual features, which are then mapped into the context space of the large language model LLM.

Part 3: Finetuning Video-Text LLM. The inputs for this part include the echocardiogram video x, the user query text y, and the LV_{EF} value obtained from Part 2. The video x is divided into a sequence of frames and visual features are extracted through a Vision Transformer (ViT). To accommodate the context window limitations of the LLM, we perform a concatenate operation on the visual tokens output by ViT to reduce the number of tokens, achieving information compression. The compressed visual features are then mapped to the context space of the LLM through a linear layer. After integrating the aforementioned context features, the LLM generates corresponding diagnostic reports or responses based on the user's query.

Algorithm 1: Forward Propagation of EchoGPT

Input : Echocardiogram video $x \in \mathbb{R}^{H\times W\times C\times T}$, User question text $y \in \mathbb{R}^{N\times M}$
Output: Segmented video x_{seg}, LV_{EF} value, Responses

Part 1: Left Ventricle Segmentation
for *each frame in video* x **do**
  extract frame features using Atrous Convolution
  project features to segmentation mask space using a linear layer
end
combine segmentation masks to get x_{seg}
Part 2: $\mathbf{LV_{EF}}$ Prediction
extract visual features from x using 3D CNN and Spatiotemporal convolutions
map features to LV_{EF} space using a linear layer to get LV_{EF}
Part 3: Finetuning Video-Text LLM
for *each frame in video* x **do**
  extract frame features using Vision Transformer
end
concatenate ViT output to reduce token numbers
map visual features to LLM context space using a linear layer
generate responses based on y, LV_{EF}, and context using LLM

3.3 Left Ventricle Segmentation Model

This model employs a semantic segmentation architecture based on DeepLabV3-ResNet50 [Chen *et al.*, 2017], incorporating Atrous Convolution technology, dedicated to the segmentation task of the left ventricle in echocardiogram video sequences. Given that the echocardiogram video sequences in the EchoNet-Dynamic dataset are only annotated with expert segmentation masks at two key frames, end-systole and end-diastole, this research fully leverages this sparsely annotated characteristic and devises corresponding weakly supervised training strategies. The core features of the adopted DeepLabV3-ResNet50 model are reflected in: multi-scale feature extraction, capturing contextual information through Atrous Convolutions with varying dilation rates; spatial pyramid pooling, integrating multi-scale feature representations; and an encoder-decoder structure, ensuring the precision of segmentation boundaries. After processing the input data through Atrous Convolution, informative feature maps are obtained, which are ultimately mapped to binary segmentation masks via a linear layer to distinguish between left ventricular and non-left ventricular regions. In terms of training strategy, this study opts to train the model from scratch, rather than directly applying pre-trained weights, to better adapt to the unique characteristics of echocardiogram data. After training for 50 epochs, the segmentation model is capable of accurately segmenting the left ventricle from echocardiogram video sequences, with visualization results that can assist physicians in diagnosis and help users gain a deeper understanding of diagnostic outcomes.

3.4 $\mathbf{LV_{EF}}$ Prediction Model

This model employs a three-dimensional convolutional neural network architecture based on R2Plus1D-18 [Tran *et al.*, 2014; Tran *et al.*, 2017], aimed at predicting the ejection fraction of echocardiogram videos through spatiotemporal feature extraction. The method fully leverages the advantages of transfer learning from pre-trained models and has been specifically optimized for the characteristics of echocardiogram images. Specifically, the model utilizes the pre-trained R2Plus1D-18 as the backbone network and incorporates the following key improvements to adapt to the ejection fraction prediction task: firstly, replacing the original classification layer with a single-output linear layer for regression prediction; secondly, retaining the feature extraction capability of spatiotemporal convolutional layers to capture key spatiotemporal features in the videos; and lastly, integrating residual connections to ensure the effective propagation of deep features. With these enhancements, the prediction model can effectively predict the corresponding left ventricular ejection fraction from echocardiogram videos. The obtained precise ejection fraction indices will be combined with user queries as prompts, transmitted to the large language

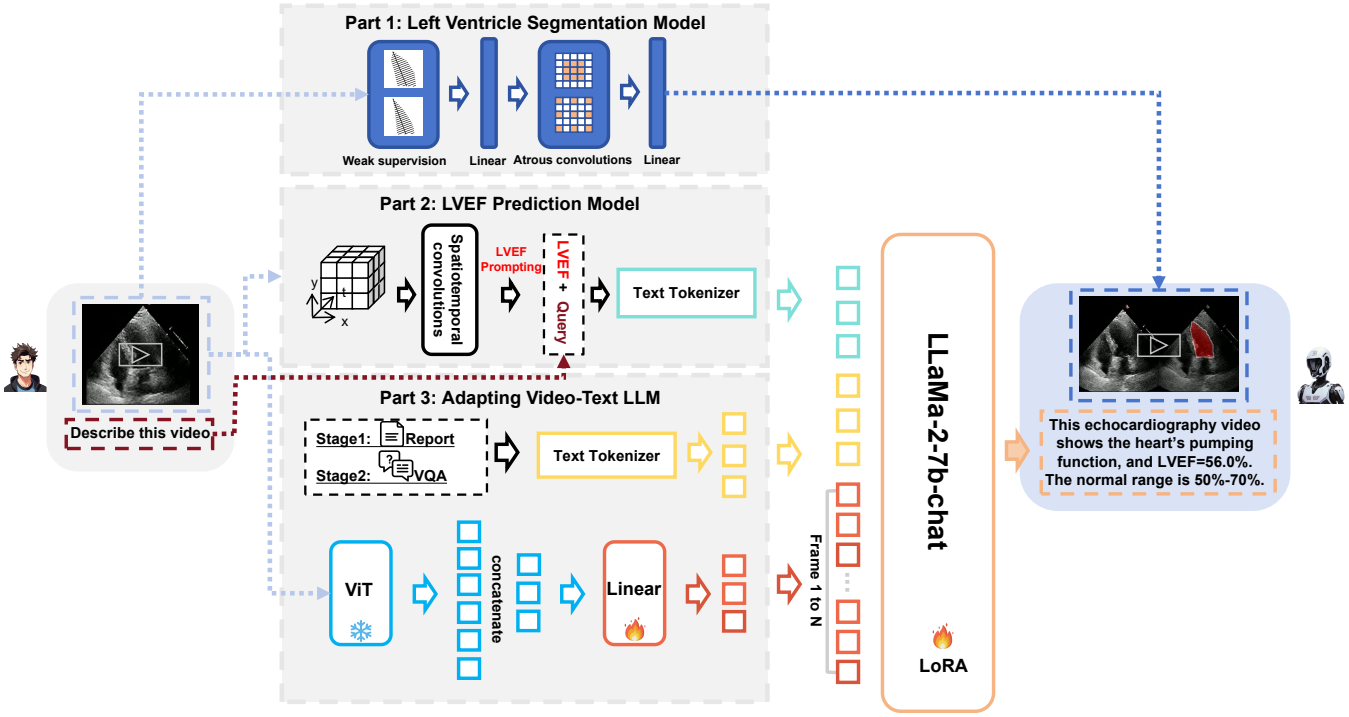

Figure 2: The framework of EchoGPT.

model, to achieve more accurate interaction and analysis.

3.5 Finetuning Video-Text Large Language Model

The training process in this part is divided into two stages:

Stage1 : Echocardiogram video-Medical Report pair pretraining. In this stage, the model processes multiple video frames to comprehend the video content. Specifically, a maximum of N frames are sampled from each video, with N, equal to 45, determined by the context window size of the language model. We employed predefined prompts with the following template: $< s > [INST] < Img >< FrameFeature_1 > ... < Img >< FrameFeature_N >< Instruction >< /INST >$. A linear layer was trained to project the visual features encoded by the visual encoder [Dosovitskiy *et al.*, 2021; Li *et al.*, 2023; Sun *et al.*, 2023; Chen *et al.*, 2020] (e.g., ViT) into the text space of the LLM [Touvron *et al.*, 2023; Zhang *et al.*, 2024a], utilizing image description loss (captioning loss). The encoded video frames were input into the model alongside the corresponding medical reports; $< FrameFeature >$ in the prompt was replaced with the encoded sampled video frames, and $< Instruction >$ was replaced with randomly selected instructions from a predefined set, which included the left ventricular ejection fraction (LV_{EF}) corresponding to the video, such as "Generate a report according to this video whose LV_{EF} is 56%." The EchoNet-Dynamic dataset, annotated with corresponding medical report data, was used for large-scale medical report generation training.

Stage 2 : Echocardiogram Video question answering instruction finetuning. The same training strategy as the first stage was employed in this stage, but with a focus on using high-quality medical visual question-and-answer datasets for instruction fine-tuning [Liu *et al.*, 2024a]. This helps enhance the model's ability to interpret input videos and generate precise answers to corresponding questions. The template was the same as in the second phase, except that $< Instruction >$ was replaced with general questions mentioned in the question-and-answer dataset. The EchoNet-Dynamic dataset, annotated with relevant medical visual question-and-answer data, was used for medical visual question-and-answer training.

Regarding training details, both stages maintained a batch size of 2 and utilized the AdamW optimizer with a cosine learning rate scheduler, setting the learning rate to 1e-4. The visual backbone was the Vision Transformer with frozen weights. The linear projection layer was trained from scratch, and the LoRA (Low-rank Adaptation) method [Hu *et al.*, 2021] was used for efficient fine-tuning of the large language model Llama2-7B-Chat. Specifically, the W_q and W_v components were fine-tuned, with the rank (r) set to 64 and the LoRA-alpha value equal to 16. The entire model maintained a consistent image resolution of 224×224 pixels throughout all stages to ensure uniformity.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets

Table 1 presents a detailed overview of the publicly available echocardiogram video dataset EchoNet-Dynamic, our corresponding diagnostic report annotation dataset, and the associated medical visual question-answering dataset.

Dataset	Content	Type	Example	Amount
EchoNet-Dynamic	Echocardiogram videos from Stanford University Medical Center	-	-	10030
Diagnosis Report	Medical text report annotation of Echocardiogram videos	-	This is the report... Examination area: Heart...resulting in an LVEF of 49.7%...The left ventricular systolic and diastolic functions of the heart are reduced, and further diagnosis and treatment are recommended.	7465
Visual Question Answering	Echocardiogram video related Question Answering	Descriptive	Q:What kind of video is this? A:It is an echocardiogram video.	18596
		Explanatory	Q:What is EF (Ejection Fraction)? A:EF (Ejection Fraction) refers to the proportion of blood volume in the left ventricle that is pumped out during contraction.	21577
		Advisory	Q:My LVEF is below the normal range, is it necessary to quit smoking? A:Yes, smoking is extremely detrimental to heart health...	12320

Table 1: Introduction of Datasets.

EchoNet-Dynamic Dataset. The EchoNet-Dynamic dataset [Ouyang *et al.*, 2020; Ouyang *et al.*, 2019] includes over 10,000 cardiac ultrasound videos from Stanford University Medical Center, featuring diverse patients. Each video, recorded from the apical four-chamber view, comes with clinical data like EF, ESV, and EDV, and includes left ventricular mask annotations by doctors at end-systolic and end-diastolic phases.

Diagnosis Report Annotation of Echocardiogram Videos. We invited an experienced echocardiogram physician to annotate each video with detailed medical reports. These reports describe the videos and assess left ventricular pump function using LV_{EF}, aiming to support cardiac function analysis and clinical decisions.

Construction of Video Question Answering (VQA) Dataset. Our VQA dataset includes three question types: descriptive, explanatory, and advisory. Descriptive questions help users understand echocardiogram video features. Explanatory questions provide insights into cardiac ultrasound data for better heart health comprehension. Advisory questions offer professional advice on treatment and lifestyle changes, particularly for those at risk of heart pump insufficiency or failure, along with early intervention suggestions.

Evaluation Metrics

Left ventricular ejection fraction (LV_{EF}) is a critical cardiac performance metric that quantifies the percentage of blood pumped out of the left ventricle with each contraction, reflecting the heart's ability to efficiently circulate blood. To systematically assess the performance of the model proposed in this study compared to other existing models in predicting the accuracy of LV_{EF}, a series of quantitative metrics were adopted for experimental evaluation:

- **AUC (40%)**: Measures the model's ability to distinguish between LV_{EF} values below and above 40%.
- **AUC (50%)**: Assesses the model's capability to differentiate LV_{EF} values below and above 50%, important for identifying cardiac insufficiency.
- **AUC (70%)**: Evaluates the model's performance in extreme cases.
- **R^2**: Reflects the proportion of LV_{EF} variation explained by the model.
- **MAE (Mean Absolute Error)**: Measures the average difference between predicted and actual LV_{EF} values.

In summary, these metrics collectively form a multidimensional assessment framework for comprehensively evaluating and comparing the performance of different models in predicting LV_{EF}.

4.2 Experimental Implementation

Quantitative Experiment. In the quantitative segment of this study, a comparative experiment was conducted to evaluate the performance of various models in predicting the accuracy of LV_{EF}. Specifically, we selected models such as EchoCLIP, CLIP, BioMedCLIP, and PubMedCLIP [Christensen *et al.*, 2024; Radford *et al.*, 2021; Zhang *et al.*, 2024b; Eslami *et al.*, 2023], and applied them to over 1200 test videos from the EchoNet-Dynamic dataset to predict left ventricular ejection fraction. By comparing the predicted outcomes of these models with actual clinical data, we quantified their accuracy and performed statistical analysis. Concurrently, in order to empirically validate the efficacy of EchoGPT in practical applications, we have solicited feedback from a cohort of seasoned physicians specializing in echocardiogram through a structured questionnaire survey. This methodological approach was employed to gather expert opinions and assess the real-world utility of the model within the domain of cardiac ultrasound imaging.

Qualitative Experiment. In the qualitative experimental phase, we focused on assessing the capabilities of large-scale visual-text multimodal language models in generating diagnostic reports and answering medical-related questions. To this end, we selected models including MiniGPT4-Video, Qwen2-VL, Video-Llama2, and Llava-Video [Ataallah *et al.*, 2024a; Wang *et al.*, 2024; Cheng *et al.*, 2024; Zhang *et al.*, 2024c], and employed a case study approach to demonstrate the effectiveness of these models in practical applications by presenting specific cases to reflect the differences in their performance. In conjunction with this, to substantiate the efficacy of our two-stage training strategy, we conducted ablation studies to compare the actual performance of EchoGPT, which underwent the complete two-stage training regimen, with that of EchoGPT variants trained solely in stage 1 and stage 2, respectively. This comparative analysis was designed to elucidate the contributions of each training phase to the overall performance of the model.

All experiments were conducted in a high-performance computing environment equipped with four RTX-4090 GPUs and an Intel(R) Xeon(R) Platinum 8336C CPU, ensuring the sufficiency of computational resources and the reliability of the experimental results.

4.3 Quantitative Results

Comparison of Predictive Accuracy for Left Ventricular Ejection Fraction ($\mathbf{LV_{EF}}$). In our study, we compared the performance of various models in predicting left ventricular ejection fraction (LV_{EF}). The experimental results in Table 2 demonstrated that our proposed EchoGPT model outperformed all other models across all evaluated metrics.

Specifically, EchoGPT achieved AUC values of 0.9519, 0.9221, and 0.8475 for the thresholds of 40%, 50%, and 70%, respectively, which are significantly higher than the performance of other models. This indicates that EchoGPT possesses an exceedingly high degree of accuracy in identifying samples with ejection fractions below 40%, below 50%, and above 70%. In terms of the R^2 score, EchoGPT reached 0.6848, implying that our model accounts for approximately 68.48% of the variability in ejection fraction, the highest proportion among all compared models. Furthermore, EchoGPT's mean absolute error (MAE) was 7.5934, the lowest among all models, which further corroborates the high precision of EchoGPT in predicting left ventricular ejection fraction. In comparison, other models such as EchoCLIP, BioMedCLIP, PubMedCLIP, and CLIP did not perform as well as EchoGPT.

In summary, the EchoGPT model demonstrated superior performance in predicting left ventricular ejection fraction, both in terms of distinguishing the severity of ejection fraction and overall predictive accuracy, significantly outperforming other models under comparison. These results substantiate the potential application of EchoGPT in the field of cardiac function assessment, offering a robust tool for clinical decision-making.

Method	Interactable	EchoNet-Dynamic				
		AUC (40%) ↑	AUC (50%)↑	AUC (70%)↑	R^2↑	MAE↓
BioMedCLIP	✗	0.4973	0.5003	0.4641	0.0001	13.8036
PubMedCLIP	✗	0.5799	0.5805	0.5199	0.0096	24.0842
CLIP	✗	0.5815	0.5658	0.5641	0.0171	8.8818
EchoCLIP	✗	0.8393	0.7920	0.7707	0.4067	10.3734
EchoGPT (ours)	✓	**0.9519**	**0.9221**	**0.8475**	**0.6848**	**7.5934**

Table 2: Comparison of Predictive Accuracy for Left Ventricular Ejection Fraction (LV_{EF}).

Questionnaire-based Evaluation of EchoGPT. To assess the diagnostic accuracy of EchoGPT and its practical utility to users, we engaged a panel of five physicians specializing in echocardiogram to evaluate the system's performance. For this evaluation, we randomly selected 100 videos from the test set of the EchoNet-Dynamic dataset and employed a questionnaire-based assessment methodology.We provided EchoGPT with the following prompts to elicit responses:

1. Could you describe this echocardiogram video for me?
2. My LV_{EF} is not in the normal range, what should I do for this case?

Based on the responses provided by EchoGPT, physicians were queried on the accuracy and relevance of its diagnosis, the informativeness of its answers, the utility of its suggestions, its capacity to assist in medical diagnosis, its ability to enhance users' understanding of their health conditions, the privacy protection afforded by local deployment, and their willingness to utilize EchoGPT. The results, as depicted in the table, indicate that in the majority of cases, physicians concurred with EchoGPT's responses, often expressing agreement or strong agreement. Figure 3 indicates that EchoGPT demonstrates commendable performance in both the accuracy of its diagnoses and the practicality of its interactive capabilities.

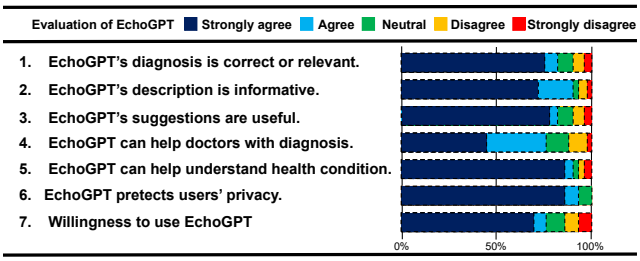

Figure 3: Questionnaire-based evaluation of EchoGPT.

4.4 Qualitative Results

Ablation Study. To ensure that EchoGPT provides accurate diagnostic reports while maintaining a high level of interactivity, we adopted the two-stage training strategy as mentioned in Section 3.5. To validate the substantial benefits of this training strategy in enhancing the model's actual performance, we conducted ablation experiments to confirm the necessity of both training stages. As illustrated in Figure 4, the first image depicts a communication dialogue between a physician and a patient, while the second image shows an interactive dialogue between EchoGPT and a user. It can be observed that EchoGPT, after undergoing the two-stage training, is capable of accurately providing correct diagnostic reports and responding to user-related inquiries with concise language. The third image demonstrates the responses of EchoGPT trained only in the first stage; although it can provide correct diagnostic reports, it fails to respond precisely and concisely to other user queries, often including redundant information. The fourth image displays the performance of EchoGPT trained solely in the second stage, which evidently cannot provide accurate diagnostic reports. In summary, the results of the ablation experiments substantiate the effectiveness and necessity of the two-stage training strategy. EchoGPT, after being trained in both stages, can achieve diagnostic results comparable to those of physicians and can accurately and concisely answer user-related questions.

Comparative Experiment. To verify the superior performance of EchoGPT in terms of interactivity and diagnostic report generation capabilities, we also compared it with other Video-Text Large Language Models (LLMs). As shown in the four lower images of Figure 4, these illustrate the actual conversational effects of four different Video-Text LLMs with users. MiniGPT4-Video and Video-Llama2 were both unable to correctly identify the characteristics of echocardiogram videos, and even failed to provide correct descriptions of the echocardiogram videos. While Qwen2-VL and Llava-Video could correctly recognize echocardiogram videos and provide some descriptive language, none of the aforementioned models were able to calculate the left ventricular ejection fraction, and thus could not provide actual diagnostic results. In contrast, the second image at the top demonstrates the conversational effect of EchoGPT, which not only possesses the strong interactivity of large language models but also can provide accurate diagnostic results, effectively assisting users in understanding their health conditions.

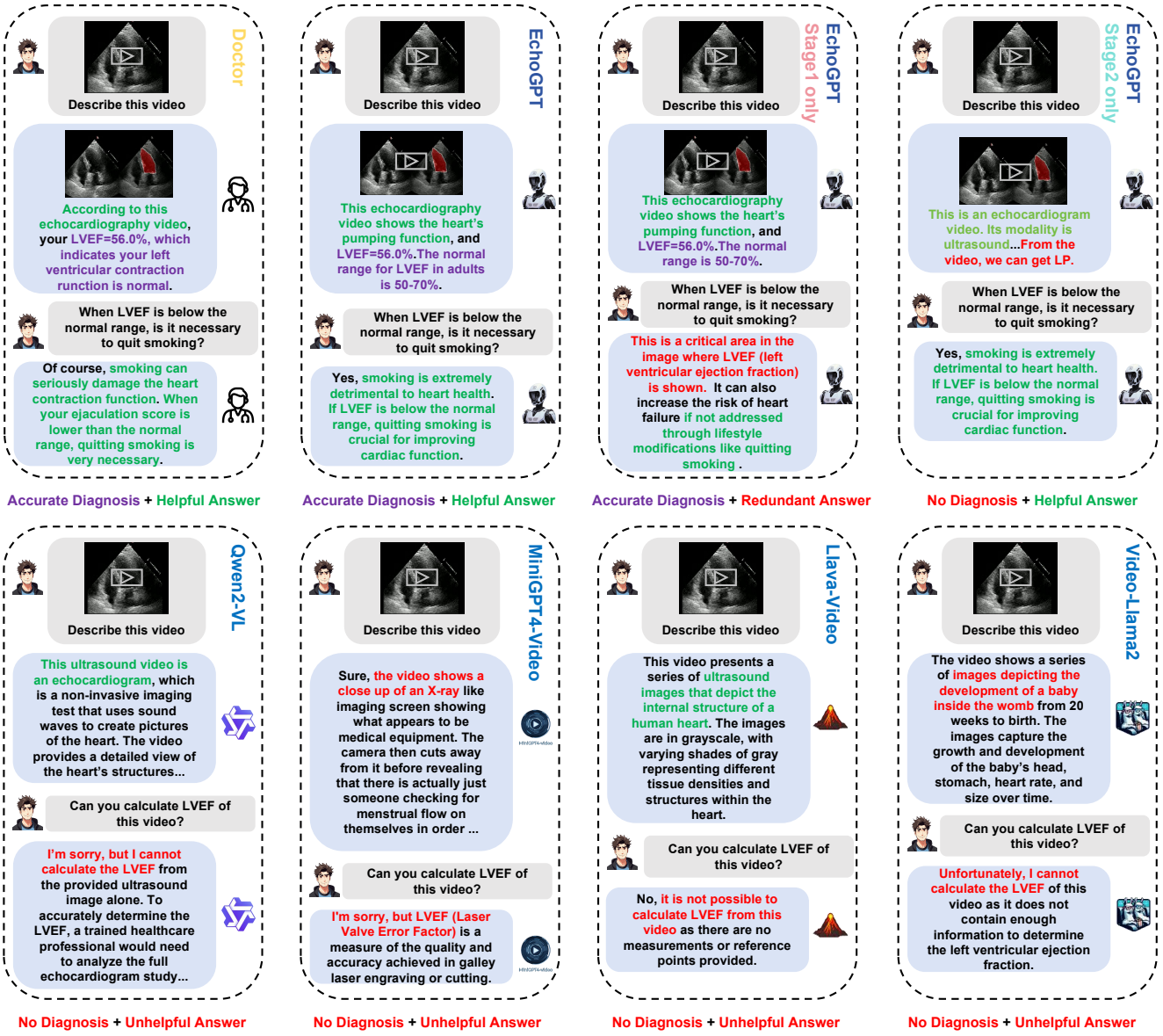

Figure 4: Qualitative Results of Ablation Study and Comparative Experiment.

5 Conclusion

This study successfully developed and validated EchoGPT, a large language model for medical diagnosis tailored to echocardiogram videos. EchoGPT not only enhances patients' ability for self-monitoring and alleviates the workload of physicians but also protects users' personal privacy through local deployment. EchoGPT demonstrates exceptional performance in diagnostic accuracy and user interactivity, setting a new standard for AI-assisted medical diagnostics. Employing a two-stage training strategy, EchoGPT is capable of providing accurate diagnostic reports and responding to user inquiries with concise language. Furthermore, EchoGPT outperforms existing models in predicting left ventricular ejection fraction, garnering recognition from clinical physicians. Future work will explore the local deployment of EchoGPT on mobile devices and further optimize its conversational capabilities.

Acknowledgments

This work is supported by the Science and Technology Project of Liaoning Province (2023JH2/101700363, 2024JH2/102600027), in part by the National Natural Science Foundation of China under Grant 62072073, 62106034, in part by the Fundamental Research Funds for the Central Universities under Grant DUT24ZD124, in part by the Dalian Innovation Fund 2021JJ12GX016, and the Science and Technology Project of Dalian City (2024JJ12GX025, 2023JJ12SN029 and 2023JJ11CG005).

References

[Alkhaldi *et al.*, 2024] Asma Alkhaldi, Raneem Alnajim, Layan Alabdullatef, Rawan Alyahya, Jun Chen, Deyao Zhu, Ahmed Alsinan, and Mohamed Elhoseiny. Minigpt-med: Large language model as a general interface for radiology diagnosis, 2024.

[Ataallah *et al.*, 2024a] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.

[Ataallah *et al.*, 2024b] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos, 2024.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.

[Chen *et al.*, 2023] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[Chen *et al.*, 2024] Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. Dia-llama: Towards large language model-driven ct report generation, 2024.

[Cheng *et al.*, 2024] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[Christensen *et al.*, 2024] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, 2024.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[Eslami *et al.*, 2023] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163, 2023.

[Gu *et al.*, 2024] Jawook Gu, Kihyun You, Han-Cheol Cho, Jiho Kim, Eun Kyoung Hong, and Byungseok Roh. Chexgpt: Harnessing large language models for enhanced chest x-ray report labeling, 2024.

[Hu *et al.*, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024.

[Liu *et al.*, 2024b] Hsiao-Chuan Liu, Yushun Zeng, Chen Gong, Xiaoyu Chen, Piotr Kijanka, Junhang Zhang, Yuri Genyk, Hisham Tchelepi, Chonghe Wang, Qifa Zhou, and Xuanhe Zhao. Wearable bioadhesive ultrasound shear wave elastography. *Science Advances*, 10(6):eadk8426, 2024.

[OpenAI *et al.*, 2024] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and et al. Gpt-4 technical report, 2024.

[Ouyang *et al.*, 2019] David Ouyang, Bryan He, Amirata Ghorbani, Matthew P. Lungren, Euan A. Ashley, David H. Liang, and James Y. Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. 2019.

[Ouyang *et al.*, 2020] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023.

[Sun *et al.*, 2025] Weilin Sun, Xinran Li, Manyi Li, Kai Xu, Xiangxu Meng, and Lei Meng. Hierarchically-structured open-vocabulary indoor scene synthesis with pre-trained large language model, 2025.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[Tran *et al.*, 2014] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[Tran *et al.*, 2017] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017.

[Wang *et al.*, 2022] Chonghe Wang, Xiaoyu Chen, Liu Wang, Mitsutoshi Makihata, Hsiao-Chuan Liu, Tao Zhou, and Xuanhe Zhao. Bioadhesive ultrasound for long-term continuous imaging of diverse organs. *Science*, 377(6605):517–523, 2022.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Yu *et al.*, 2024] Han Yu, Peikun Guo, and Akane Sano. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text, 2024.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[Zhang *et al.*, 2024a] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2024.

[Zhang *et al.*, 2024b] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024.

[Zhang *et al.*, 2024c] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024.

[Zhou *et al.*, 2024] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and Xin Gao. Pretrained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.