SilVar-Med: A Speech-Driven Visual Language Model for Explainable Abnormality Detection in Medical Imaging

Tan Hanh Pham[1], *Chris Ngo*[2], *Trong Duong Bui*[3], *Minh Luu Quang*[4],
Tan Huong Pham[5], *Truong Son Hy*[6]

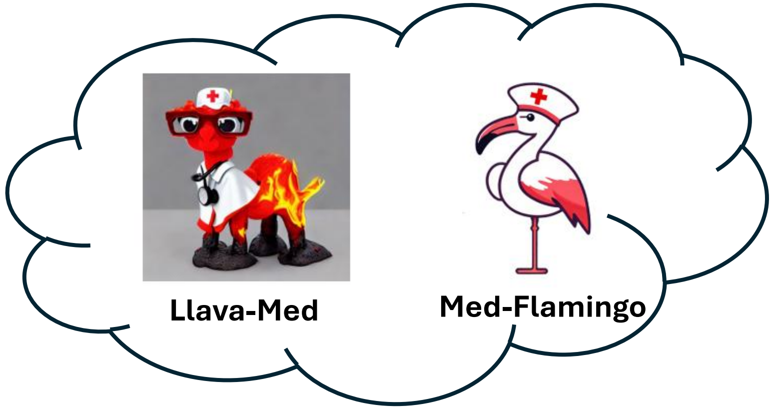
[1]*Harvard Medical School, USA*, [2]*Knovel Engineering Lab, Singapore*, [3]*Vietnam Military Medical University*,
[4]*108 Military Central Hospital, Vietnam*, [5]*Can Tho University of Medicine and Pharmacy*,
[6]*University of Alabama at Birmingham, USA*

Paper ID: MAR-18

MAR
CVPR 2025
Multimodal Algorithmic Reasoning

Presenter: Tan-Hanh Pham

Motivation



Medical vision language models (VLMs)

Existing medical VLMs: Med-Flamingo, Llava-Med, MiniGPT-Med, MedBLIP, ...

Limitation: Most of these model do not provide reasoning behind medical image diagnosis and limited by text+image instruction models.

Potential problem

Figure 1: Surgery (Harvard Medical School)

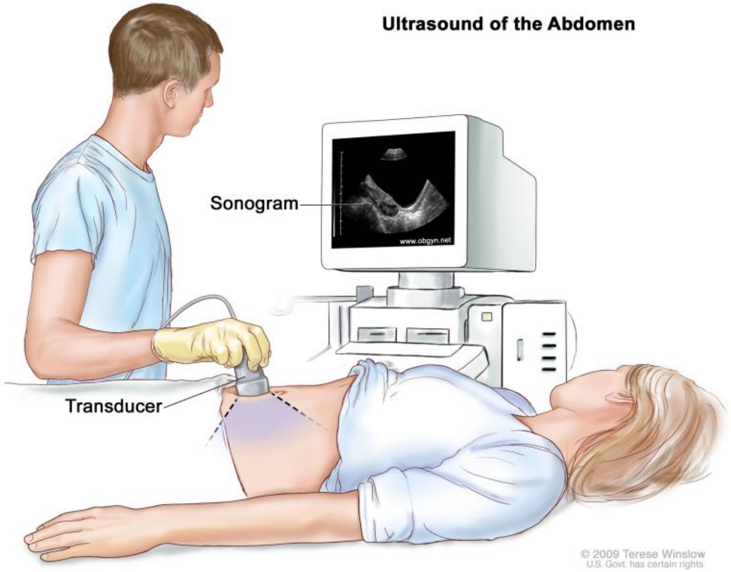


Figure 2: Abdominal ultrasound (National Cancer Institute)

There is a potential for speech-instruction medical VLMs (audio+image) in scenarios such as surgery or text-based interaction is often impractical for physicians.

Contributions

- **Method**: We propose SilVar-Med, a ***speech-driven medical VLM*** that enables intuitive human-machine interaction in healthcare.
- **Dataset benchmark**: a ***reasoning*** dataset for abnormality detection
- **Evaluation metrics**: ***Model's reasoning*** abilities and human evaluation

Method

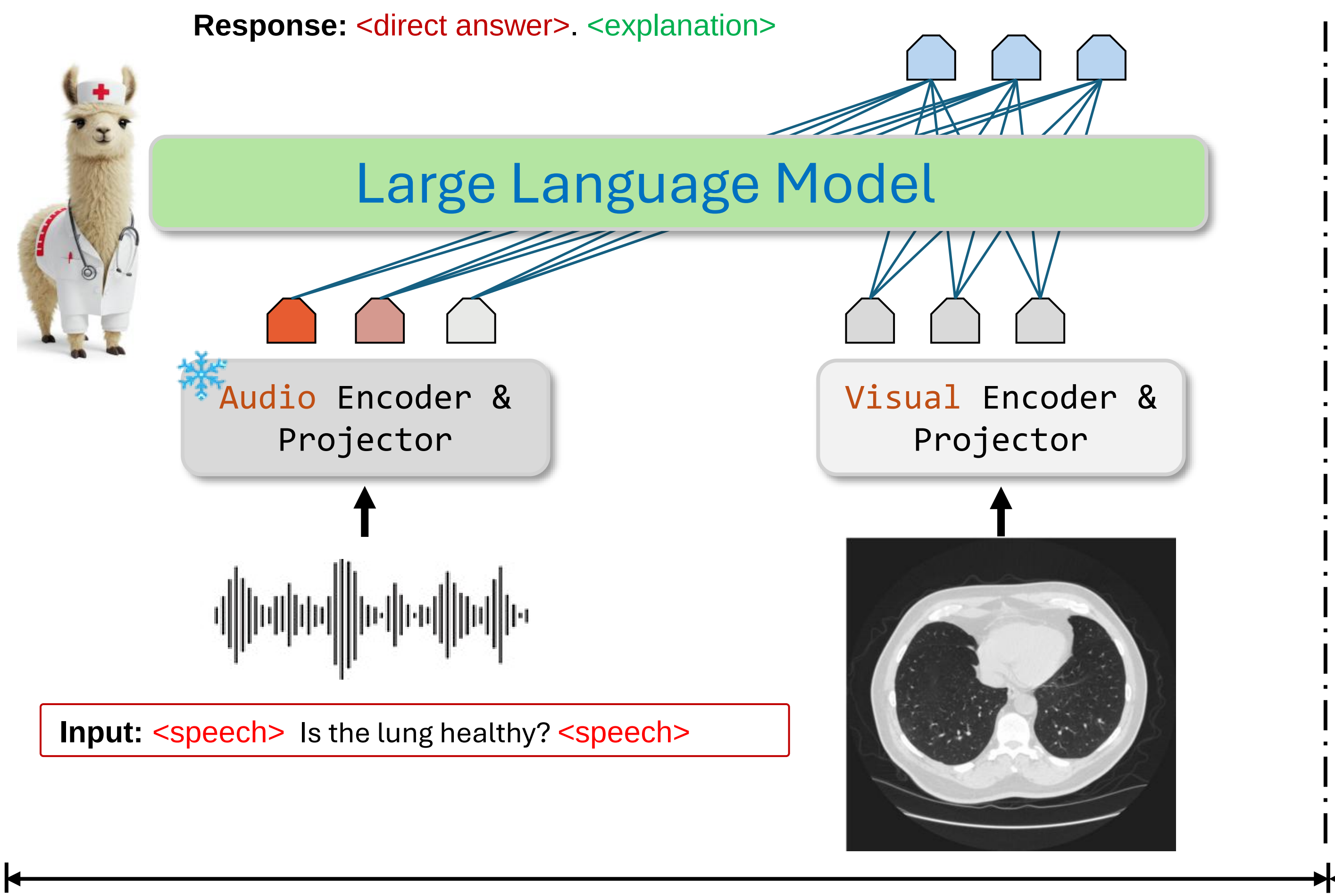


Figure 3: Speech-Driven Medical VLM for Explainable Abnormality Detection in Medical Imaging.

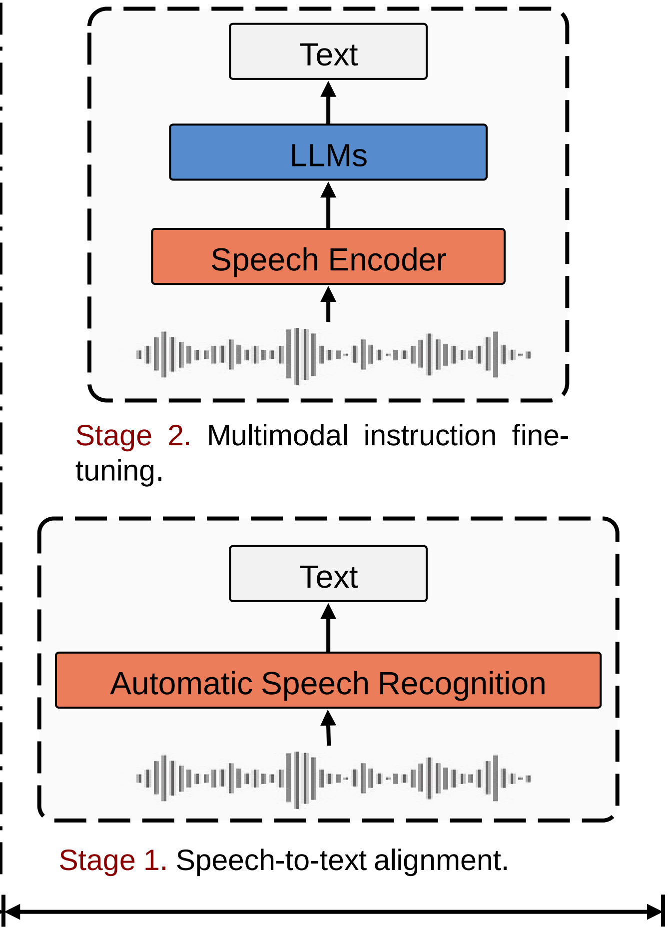


Figure 4: Training pipeline

LLM	Llama 3.1-8B	Audio encoder	Whisper Tiny	Visual Encoder	PubMedCLIP

Datasets

Dataset	Train	Validation	Test
VQA-RAD	1.7k	None	451
SLAKE (English)	4.9k	1k	1k
VQA-Med 2019	12.7k	2k	500
SilVar-Med (ours)	716	-	150

Table 1: Training dataset

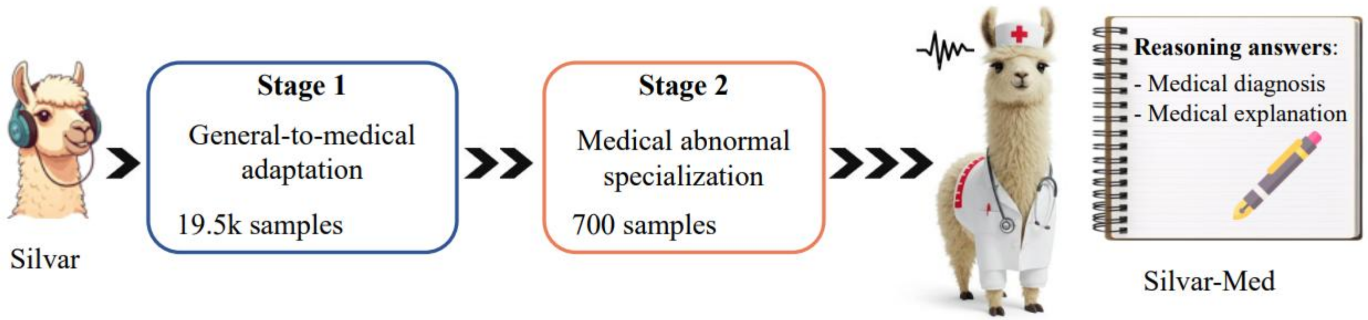


Figure 5: Adaptive training pipeline

Sample 1: No abnormality detection
Question: Is the lung healthy?
Answer: <answer>The lung appears healthy based on the provided imaging.<answer> <explanation> There are no visible signs of abnormalities such as lesions, fluid accumulation, or structural changes. The lung fields are clear, indicating normal function and absence of disease. <explanation>

Sample 2: Abnormality detection
Question: What diseases are included in the picture?
Answer: <answer> The image indicates the presence of lung cancer, characterized by abnormal growths in the lung tissue. <answer> <explanation> The CT scan shows potential lesions or nodules that may suggest malignancy. <explanation>

Experimental Results

Automatic speech recognition

Models	WER		CER	
	train	test	train	test
Whisper Tiny	2.01	2.67	2.01	2.99
Whisper Small	2.02	4.57	1.59	3.50

Table 2: Automatic speech recognition.

Reasoning evaluation

	Human			LLM-as-judge	
Reasoning accuracy	**Exp 1**	**Exp 2**	**Exp 3**	**GPT4o**	**Gemini**
SilVar-Med with the langue module of Llama 3.1 8B					
Completely Incorrect	11	6	13	39.00	22.00
Significantly Incorrect	28	30	33	9.67	23.67
Partially Correct	13	15	28	39.67	54.00
Fully Correct	96	97	74	59.67	48.33

Table 3: Reasoning evaluation (Exp denotes expert).

Output: <direct answer>. <explanation>

0: Completely Incorrect – The prediction fails to answer the question, is off-topic, or entirely unrelated to the ground truth.
1: Significantly Incorrect – The prediction attempts to answer the question but does not match the ground truth in terms of understanding, terminology, or core explanation.
2: Partially Correct – The prediction directly answers the question and provides an explanation. Both the answer and the explanation reflect a reasonable understanding of the main idea, though they contain minor irrelevant or incorrect information.
3: Fully Correct – The prediction completely aligns with the ground truth, providing

Experimental Results

Benchmarking the model on different datasets

Models	Instruction	SLAKE			VQA-RAD		
		Ref	Open	Closed	Ref	Open	Closed
Representatives of existing studies in the literature							
LLaVA [31]	Text		78.18	63.22		50.0	65.07
LLaVA-Med (From LLaVA) [25]	Text		83.08	85.34		61.52	84.19
LLaVA-Med (BioMed CLIP) [25]	Text		87.11	86.78		64.75	83.09
LLaVA-Med++ (w/ Medtrinity) [48]	Text		86.20	89.20		77.10	86.00
LLaVA-Med++ (w/o Medtrinity) [48]	Text		79.30	84.00		64.60	77.00
MMBERT General [22]	Text		-	-		63.10	77.90
MEVF+SAN [35]	Text		-	-		40.70	74.10
CR [52]	Text		-	-		60.00	79.30
Q2ATransformer [32]	Text				79.19		81.20
PubMedCLIP [16]	Text	78.40		82.50	60.10		80.00
BiomedCLIP [54]	Text	82.05		89.7	67.60		79.80
M2I2 [27]	Text	74.70		91.10	66.50		83.50
SilVar-based studies with our own experiment							
SilVar-Med 3.1 8B (Llama 3.1-8B)	Speech		74.08	79.44		55.34	62.56
SilVar-Med 3.1 8B (Llama 3.1-8B)	Text		74.32	80.03		55.21	60.86
Ablation studies of SilVar-Med using different language models for the decoder							
SilVar-Med DR8B (Deepseek R1 Distill-Llama-8B)	Speech		76.50	83.80		58.85	68.35
SilVar-Med DR8B (Deepseek R1 Distill-Llama-8B)	Text		77.12	82.11		60.31	67.98
SilVar-Med 2 7B (Llama 2)	Speech		73.23	76.34		54.75	57.77
SilVar-Med 2 7B (Llama 2)	Text		64.21	75.54		55.65	75.78

Table 4: Benchmarking on the SLAKE and VQA-RAD datasets.

Models	Accuracy	BLEU	Bert-sim
Representatives of existing studies in the literature			
ImageCLEF [7]	62.4 %	64.4 %	-
MedVINT [55]			0.63
Med-Flamingo [33]			0.65
MMBERT [22]	67.2 %	69.0 %	
CGMVQA [40]	60.0 %	61.9 %	
VGG16+BERT [49]	62.4 %	64.4 %	
SilVar-based studies with our own experiment			
SilVar-Med 3.1 8B (speech)	64.99%	62.24 %	0.80
SilVar-Med 3.1 8B (text)	66.10%	63.11%	0.80
Ablation studies			
SilVar-Med DR8B (speech)	68.03%	69.12%	0.80
SilVar-Med DR8B (text)	68.77%	68.34%	0.81
SilVar-Med 2 7B (speech)	66.36%	61.93%	0.79
SilVar-Med 2 7B (text)	67.02%	59.17%	0.80

Table 5: Benchmarking on the VQA 2019 dataset.

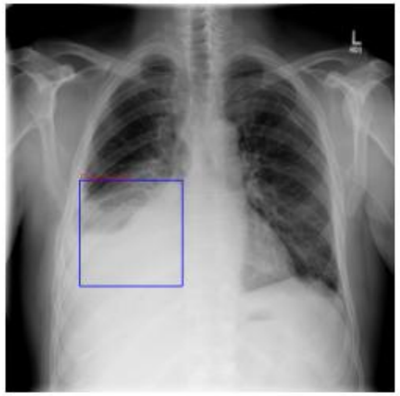
Compared to text-based medical VLMs, SilVar-Med achieves comparable performance despite the challenge of speech-based instructions.

Experimental Results

Examples of predictions

Example 1: Abnormality detection.

Does the lung look abnormal?



Ground truth: Yes, the lung appears abnormal as there are irregularities in the lung fields that may indicate potential pathology. The presence of unusual opacities or changes in the lung structure can suggest conditions such as infection, inflammation, or other lung diseases.

SilVar-Med (Llama 3.1): Yes, the lung appears abnormal due to the presence of irregularities in the lung fields, which may indicate potential pathology such as infection, inflammation, or other lung conditions. The overall lung structure shows signs of asymmetry or density changes that deviate from normal anatomy. These findings warrant further investigation to determine the underlying cause.

GPT-4o mini: The lung appears to be within normal limits based on the provided X-ray image. There are no obvious signs of abnormalities such as masses, infiltrates, or significant opacities. The lung fields are clear, and the vascular markings are normal, indicating healthy lung tissue. However, a definitive assessment should be made by a qualified radiologist for a comprehensive evaluation.

Gemini 1.5 Flash: Based on the provided chest X-ray, the lungs appear normal. There is no visible evidence of infiltrates, masses, or pneumothorax. The lung fields demonstrate normal aeration and the heart size appears within normal limits. Further evaluation may be necessary depending on the clinical presentation.

Figure 6: Model prediction.

Conclusion

- **Model**: We propose SilVar-Med, a **speech-driven medical VLM** that enables intuitive human-machine interaction in healthcare.
- **Dataset**: Reasoning dataset and extensive experiments for abnormality detection.
- **Evaluation metrics**: Model reasoning using LLM-as-judge and human evaluation,.. There is a gap between human and LLM-as-judge evaluation.
- **Application**: Verbal communication with VLM (e.g. surgery, abdominal ultrasound, ...)

Silvar-Med

Thank you!

Time for Q&A