ICCV #2873

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

Anonymous ICCV submission

Paper ID 2873

Abstract

Medical Visual Language Models have shown great potential in various healthcare applications, including medical image captioning and diagnostic assistance. However, most existing models rely on text-based instructions, limiting their usability in real-world clinical environments especially in scenarios such as surgery, text-based interaction is often impractical for physicians. In addition, current medical image analysis models typically lack comprehensive reasoning behind their predictions, which reduces their reliability for clinical decision-making. Given that medical diagnosis errors can have life-changing consequences, there is a critical need for interpretable and rational medical assistance. To address these challenges, we introduce an end-to-end speech-driven medical VLM, SilVar-Med, a multimodal medical image assistant that integrates speech interaction with VLMs, pioneering the task of voice-based communication for medical image analysis. In addition, we focus on the interpretation of the reasoning behind each prediction of medical abnormalities with a proposed reasoning dataset. Through extensive experiments, ***we demonstrate a proof-of-concept study for reasoning-driven medical image interpretation with end-to-end speech interaction****. We believe this work will advance the field of medical AI by fostering more transparent, interactive, and clinically viable diagnostic support systems. Our code and dataset are publicly available at SiVar-Med.*

1. Introduction

Recently, advancements in Visual Language Models (VLMs) have demonstrated the potential of Large Language Models (LLMs) to process both images and text at the same time [2, 4, 26, 30, 39]. In the medical domain, VLMs have gained increasing attention for their ability to facilitate intuitive human-machine interactions such as MedBLIP [11], Med-flamingo [33], Llava-Med [25], improving clinical decision-making and diagnostic assistance. These models are particularly valuable for medical imaging analysis, where they can process complex radiological images — such as X-ray [19, 24], MRI, and CT scans [48] — and generate meaningful textual descriptions. By leveraging deep learning techniques, VLMs can assist professionals in interpreting medical images, identifying abnormalities, and supporting diagnostic workflows.

Despite these advancements, most existing medical VLMs remain limited to text-based interactions, which may not be optimal in time-sensitive clinical settings or for visually impaired users. While some proprietary VLMs, such as GPT-4o [36] and Gemini [43], support speech-driven interactions, they are not open-source, restricting fine-tuning for downstream tasks. Recently, SilVar [37], a speech-driven multimodal model for reasoning-based visual question answering and object localization, has emerged as a pioneering effort in the field. Despite its potential applications in the medical domain, speech-based medical instruction for VLMs remains underexplored in open-source research, and existing models lack the capability to process and reason through spoken queries effectively.

Furthermore, while there are some several benchmarks, have been introduced to evaluate the performance of medical VLMs such as MultiMedEval [41], MultiMedQA [42], OmniMedVQA [18], existing evaluation methods primarily focus on image captioning tasks and are limited in assessing the reasoning behind predictions. In addition, commonly used datasets such as SLAKE [29], VQA-Med [7], VQA-RAD [23], and PathVQA [17] primarily evaluate models using text-based instructions with short-answer responses, often without requiring deeper reasoning or justification. To address this limitation, LLaVA-Med [25] introduced a medical chat assistant capable of answering open-ended research questions, but its functionality is restricted to image-text inputs and limited by the number of supported medical image modalities. OmniMedVQA [18], on the other hand, aggregates multiple available datasets to create a larger benchmark for multiple-choice question-answering tasks and utilizes LLMs as judges for evaluation. However, both approaches lack a structured framework for evaluating reasoning abilities in medical VLMs, particularly in the context of abnormality detection.

To bridge this gap, we propose SilVar-Med, an end-to-end speech-instructed medical VLM that enables users to interact with the model verbally. Our approach not only introduces speech-driven interaction but also concentrates in reasoning abnormality detection by incorporating structured reasoning into predictions. To this end, we introduce a demonstrated dataset, designed for reasoning abnormality detection through speech instructions. Additionally, we propose a novel evaluation metric that leverages LLMs as judges to assess the reasoning capabilities of medical VLMs. Our contribution is summarized as follows:

- We propose SilVar-Med, a speech-driven medical VLM that enables intuitive human-machine interaction in healthcare.
- We focus on investigating the model's reasoning abilities behind abnormality detection, addressing the limitations of predictions without explanations or short predictions.
- We introduce a dataset for speech-instructed medical abnormality detection, enhancing multimodal learning in medical AI.
- We propose a comprehensive reasoning evaluation metric together with LLMs as judges for medical VLMs.

2. Related Work

2.1. Medical Vision Language Models

Over the past five years, there has been a rapid development of LLMs and VLMs such as Gemini and GPT-4 [1, 9], alongside the emergence of open-source models like the Llama family [15, 44, 45], Mistral family [20], Qwen family [6, 50], and Vicuna [56]. These models have significantly advanced natural language understanding, but their capabilities have been further extended by VLMs, which integrate visual and textual modalities [26]. VLMs enable models to process both images and text, enhancing applications such as visual question answering (VQA), medical image interpretation, and image captioning. There are many VLMs including Flamingo [2], BLIP [31], MiniGPT-v2 [10], MiniGPT-4 [57], LLaVA [30], and InternVL [13], which have demonstrated remarkable progress in general-domain visual-language tasks.

Inspired by these advancements, researchers have developed domain-specific VLMs for medical applications [11, 46, 54]. One of the pioneering studies in this field is Med-Flamingo [33], which extends Flamingo to the medical domain by pretraining on multimodal knowledge sources spanning various medical disciplines. Similarly, LLaVA-Med [25] filters image-text pairs from PMC-15M [54] to train a biomedical-specialized VLM leveraging LLaVA-pretrained parameters. In addition to medical image report generation and medical image captioning models [51], MiniGPT-Med [3] and Lite-GPT [24] extend MiniGPTs [10, 57] to generate bounding boxes along with predictions, enabling localized abnormality detection. Furthermore, Merlin [8] is one of the pioneering models for 3D VLMs, capable of processing 3D medical images alongside their corresponding textual radiology reports, along with RadFM [47]. Other notable studies, such as PubMedCLIP [16], BiomedCLIP [54], and BiomedGPT [53], have also contributed to the adaptation of general-domain VLMs for medical applications.

However, most of these work underexplored the reasoning behind prediction and concentrate on short answer generation or multiple choice, reducing their reliability for clinical decision-making. In addition, medical VLMs remain limited to image-text interactions, which may not be convenient in scenarios where text input is unavailable or impractical. For example, in surgical environments, speech-based interactions could be more effective, as verbal communication is often preferred over manual text input.

2.2. Medical Datasets and Benchmarks

In addition to model development, researchers have made efforts to create medical VQA datasets to support the ongoing advancements in the field. Several fundamental and widely used medical VQA datasets have been developed, including SLAKE [29], VQA-RAD [23], PathVQA [17], VQA-Med (2018–2021) [7], and PubMedQA [21], EHRXQA [5]. However, these datasets are often limited in size or lack diversity in medical imaging modalities. To address these limitations, recent studies have attempted to scale dataset size using GPT-assisted models and prompting techniques. For example, models such as LLaVA-Med [25] and MedTrinity [48] have leveraged large-scale dataset generation through synthetic data augmentation. Furthermore, OmniMedVQA [18] combines both published and restricted datasets to provide a diverse and large-scale medical VQA benchmark, primarily focusing on multiple-choice questions. In addition, PMC-VQA [55] was generated using self-instruction on PMC-OA [28], offering a comprehensive dataset for biomedical VQA tasks.

Despite these efforts, current medical VLMs still struggle with reasoning-based predictions, resulting in medical VQA models excelling at image captioning tasks but lacking structured reasoning mechanisms to justify their outputs. Moreover, existing evaluation methods primarily focus on text similarity and alignment metrics (n-grams) such as accuracy, BLEU, and ROUGE, without adequately assessing the depth of reasoning in model predictions. These metrics may also fail to capture the semantic quality and logical coherence of the model's reasoning process.

To address these challenges, in this work, we propose SilVar-Med, an end-to-end speech-instructed medical VLM that enhances multimodal interactions and supports structured reasoning for abnormality detection. In addition, we focus on reasoning-based abnormality detection that improves model transparency and decision-making reliability. To this end, we introduce a demonstration dataset for reasoning-based abnormality detection. In term of evaluation, we propose using LLMs as a judge framework with a focus on reasoning responses.

3. Data Processing

3.1. Reasoning Abnormality Dataset

To achieve our study's objective - developing a medical assistant that understands medical images and enables users to interact with it through voice queries - we created a demo dataset addressing two key challenges: (1) understanding the reasoning behind each abnormality detection and (2) enabling voice-based instructions or queries. Particularly, we focus on abdominal and thoracic abnormalities detection across three imaging modalities: MRI, CT, and X-ray. Our dataset includes abnormalities in six organs, including heart, liver, kidney, lung, spleen, because they align well with the expertise of the physicians on our team.

Recognizing the ability of large language models (LLMs) to effectively learn from visual features in images and their corresponding reasoning descriptions, we intentionally created a small, specialized dataset tailored for our downstream task. We manually selected abnormal samples from the SLAKE dataset [29] and then constructed our explainable abnormality detection dataset. Initially, three medical image analysis specialists from our team manually annotated the dataset. However, after annotating a substantial number of samples, we found that we could leverage GPT-4o with Chain-of-Thought (CoT) prompting, combined with our medical expertise, to enhance quality and efficiency. Eventually, the annotation process is described into four steps as follows:

- **Data Selection**: Identifying and verifying abnormal samples extracted from the SLAKE dataset. Following the initial annotation, we select samples that exhibit abnormalities and confirmed disease diagnoses.
- **Annotation & Labeling**: After selecting the desired samples, we identify bounding boxes and label abnormalities directly on the images. Then, we leverage GPT-4o and CoT prompting to generate preliminary annotations as mentioned above. The purpose of this step is to leverage the knowledge of GPT-4o to assist in labeling.
- **Specialists Validation**: Using the original labels and preliminary annotations, our team of three medical image analysis specialists manually reviewed and relabeled the data to generate high-quality annotations, ensuring the correctness of all labels (866 samples).
- **Synthetic Voice Generation**: We normalize the text in the questions to ensure that it generates smooth, natural speech outputs, enhancing the clarity and coherence of the spoken content before using Google Cloud APIs to generate synthetic voice. Regarding the naturalness of the sound, we believe it is more related to sentiment and emotional recognition, which is beyond the scope of this study. Therefore, we assume that the nuance and tone of the synthetic voice do not significantly affect our task, abnormality prediction.

Finally, our dataset has the following imaging modality distribution: MRI (22.4%, 194 samples), CT (16.5%, 143 samples), and X-ray (61.1%, 529 samples). The dataset is split into training and testing sets: train (162 MRI, 122 CT, 432 X-ray) and test (32 MRI, 21 CT, 97 X-ray).

The dataset is designed to support reasoning-based abnormality detection, with questions structured to extract meaningful insights about abnormalities in an organ.
Example questions include:
- What diseases are present in the image?
- Does the specific organ appear normal or abnormal?
- Is the specific organ healthy?
- Are there any abnormalities in this image?
- Which organ is abnormal: {specific organ} or {specific organ}?

For the answer structure, we ensure that:
- The first sentence directly answers the question.
- The subsequent sentences provide reasoning, explaining the signs of abnormality.

Sample 1: No abnormality detection.

Question: Is the lung healthy?

Answer: The lung appears healthy based on the provided imaging. 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.

Sample 2: Abnormality detection.

Question: What diseases are included in the picture?

Answer: The image indicates the presence of lung cancer, characterized by abnormal growths in the lung tissue. The CT scan shows potential lesions or nodules that may suggest malignancy.

Table 1. Examples of reasoning abnormal detection.

Dataset	Train	Validation	Test
VQA-RAD [23]	1.7k	None	451
SLAKE (English)[29]	4.9k	1k	1k
VQA-Med 2019 [7]	12.7k	2k	500
SilVar-Med (ours)	716	-	150

Table 2. Summary of Medical VQA Datasets.

3.2. Data Collection

Recognize the size limitation of our reasoning dataset, we leverage some published datasets including SLAKE [29], VQA-RAD [23], VQA-Med 2019 [7] to enhance its learning knowledge. For these dataset, we follow the synthetic voice generation pipeline as mentioned in Sec. 3.1. For the SLAKE dataset, we use only english samples.

4. Reasoning Abnormality Detection

4.1. Speech-Driven Medical VLM

Our SilVar-Med is a multimodal model built upon to SilVar [37], for medical abnormality detection by integrating speech and image inputs. Unlike traditional medical visual language models that rely on text-based instructions, SilVar-Med introduces an end-to-end speech-driven approach, making it more suitable for real-world clinical environments where text interaction is impractical, such as in surgical settings. Inherit from the flexibility of the modules in the SilVar model, we designed the model with three key components: an audio encoder that extracts speech features, a visual encoder that processes medical images, and a large language model that fuses multimodal inputs to generate reasoned text responses for abnormality detection. In addition, we modified the vision encoder with PubMedCLIP [16], and the language model with Deepseek R1 (Distill-8B-Llama) [14], a rising star for reasoning response. By combining speech and vision-based reasoning, SilVar-Med enhances interpretability in medical image analysis, providing a more interactive and transparent diagnostic support system.

4.2. Training Pipeline

The training of SilVar-Med follows a two-stage process, as shown in Fig. 1. In the first stage, general-to-medical adaptation, we train the Whisper [38] model with a speech-to-text task in the medical domain to ensure it effectively extracts meaningful features from spoken instructions. Once trained, the Whisper encoder is integrated into SilVar-Med, where it works alongside the medical visual encoder and language model to process multimodal inputs. After that, we train the SilVar-Med with 19.5k English medical VQA samples as we mentioned in Tab. 2. In the second stage, we continue trainning the model with our dataset, specializing in medical abnormality detection and reasoning-based medical image interpretation.

In terms of training configuration, we conducted experiments with the Tiny and Small Whisper models for 20 epochs using a batch size of 8. For SilVar-Med, we employ a weight decay of 0.05 and train the model for 20 epochs, with each epoch consisting of 177 iterations. The learning rate is set to 1e-5 and remains constant throughout training, with both the minimum and warmup learning rates also set to 1e-5. Each training batch consists of four samples, and the training utilizes two workers to optimize computational efficiency. This structured training approach ensures that SilVar-Med effectively learns from diverse medical datasets and refines its reasoning capabilities through targeted fine-tuning.

5. Evaluation Metrics and Reasoning Criteria

To evaluate the performance of SilVar-Med, we used both traditional text-based evaluation metrics and a novel LLM-as-Judge assessment. Traditional metrics include BLEU, ROUGE, and BERTScore, which measure the textual similarity between the model's generated responses and ground truth references. However, these methods may not fully capture the accuracy and reasoning quality of medical abnormality detection.

To address this limitation, we propose an LLM-as-Judge evaluation framework to evaluate the reasoning of SilVar-Med's performance in medical domain. To make the justification clear and consistent, we define two key criteria: (1) the accuracy of abnormality predictions and (2) the reasoning behind each prediction. Here, we measure two factors which are the structure of the answer and the accuracy of the answer. In terms of accuracy, the framework categorizes model responses into four levels:

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 both a clear answer and a well-reasoned explanation.

By adopting this approach, we move beyond a strict right/wrong classification and enable medical professionals to interpret model outputs, particularly in cases where the model exhibits uncertainty. To implement this evaluation, we use several commercial large language models including GPT-4o and Gemini Flash 1.5, to assess the responses. We then compute Pearson Correlation and Spearman Correlation to analyze the consistency between the LLM-based assessments and traditional metrics.

Beyond automated evaluations, three medical imaging specialists from our team independently assess SilVar-Med's predictions. We then compare their evaluations with the results obtained from GPT-4o and Gemini, ensuring a comprehensive assessment that combines both expert judgment and automated analysis.

6. Experimental Result

6.1. Speech-To-Text Quality

Before integrating Whisper to SilVar-Med, we fine-tuned it using a combination of the VQA-RAD, English SLAKE, and VQA-Med 2019 datasets, as outlined in Sec. 4.2. It is important to note that

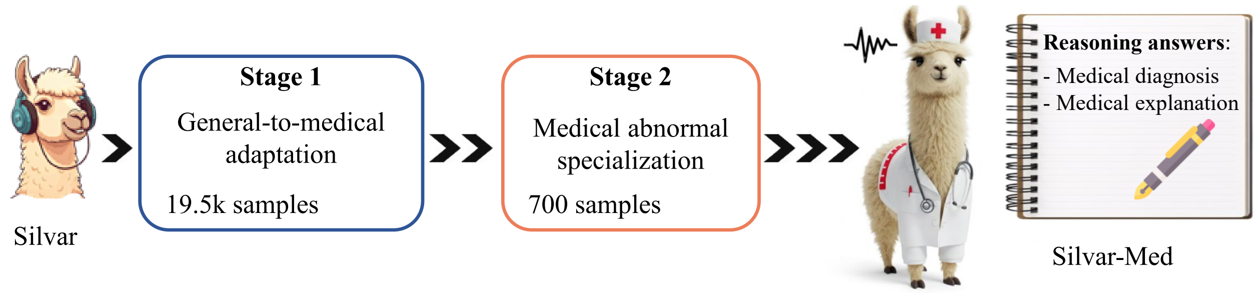


Figure 1. SilVar-Med training pipline.

we convert the questions of these datasets to speech to train the model because we want to maximize the performance of audio encoder in the medical domain. Here, we evaluate the performance of two Whisper models (Tiny and Small) using Word Error Rate (WER) and Character Error Rate (CER), which are standard benchmarks for speech-to-text accuracy [12, 34].

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 3. Evaluation of audio encoder baselines (Whisper Tiny and Whisper Small) using WER and CER on the combined dataset (VQA-RAD, English SLAKE, and VQA-Med 2019).

The results, presented in Tab. 3, indicate that Whisper Tiny and Whisper Small achieve comparable performance, with variations across WER and CER metrics. Specifically, Whisper Tiny achieves a WER of 2.01% (train) and 2.67% (test), along with a CER of 2.01% (train) and 2.99% (test). Whisper Small, on the other hand, reports a WER of 2.02% (train) and 4.57% (test), with a CER of 1.59% (train) and 3.50% (test).

Interestingly, while Whisper Small attains a lower CER during training (1.59% vs. 2.01%), it exhibits a significantly higher WER on the test set (4.57% vs. 2.67%), suggesting that it may be more prone to overfitting compared to Whisper Tiny. This discrepancy indicates that while the Small model has better character-level accuracy in training, its generalization to unseen test data is weaker. Given this observation, the Whisper Tiny model appears to be the more stable choice, balancing both WER and CER more consistently across training and testing phases. Moreover, since Whisper Tiny has a smaller number of parameters compared to Whisper Small, it is computationally more efficient. This makes it a more practical choice for our end-to-end fine-tuning process, as it reduces training time and resource consumption while still maintaining strong performance. Furthermore, these results reinforce the feasibility of using Google Cloud APIs to generate synthetic voice data without considering emotional expressiveness, as the overall error rates remain relatively low.

6.2. Speech-Driven Medical VLMs

To evaluate SilVar-Med's performance, we evaluated it on the test set using BERTScore, BLEU, and ROUGE as standard text generation metrics. Since there are no established benchmarks for speech-driven VLMs in the medical domain and only a few existing speech-driven VLM models, we compared SilVar-Med's performance against SilVar and commercial speech-driven vision-language models (VLMs), including GPT-4o Mini and Gemini Flash 1.5. Unlike SilVar-Med, which is an end-to-end speech-driven VLM, GPT-4o Mini and Gemini Flash 1.5 follow a cascaded approach, requiring an intermediate step to convert audio into text before processing.

Models	**BertScore**	**BLEU**	**ROUGE**
SilVar-Med (Llama 3.1)	0.82	20.87 %	55.18 %
GPT-4o mini	0.76	7.25 %	46.33 %
Gemini Flash 1.5	0.75	3.32 %	34.07 %
Ablation study with different language models for SilVar-Med			
SilVar-Med (Deepseek)	0.81	20.43 %	54.45 %

Table 4. Comparison between the SilVar-Med and speech-driven VLMs on the test set.

The results, summarized in Tab. 4, indicate that SilVar-Med consistently outperforms GPT-4o mini and Gemini Flash 1.5 across all evaluated metrics. With a BERTScore of 0.82, SilVar-Med demonstrates a stronger semantic alignment with ground truth responses compared to GPT-4o mini (0.76) and Gemini Flash 1.5 (0.75), reflecting its ability to generate contextually accurate medical explanations. In terms of BLEU, SilVar-Med achieves 20.87%, significantly surpassing GPT-4o mini (7.25%) and Gemini Flash 1.5 (3.32%), indicating superior syntactic and lexical accuracy in structured medical reasoning. Additionally, SilVar-Med attains the highest ROUGE score of 55.18%, outperforming GPT-4o mini (46.33%) and Gemini Flash 1.5 (34.07%). This suggests that SilVar-Med more effectively captures key phrases and maintains coherence with reference texts.

Overall, these findings indicate that SilVar-Med's domain-specific fine-tuning enables it to generate clinically relevant and semantically precise explanations, making it highly suitable for medical VQA tasks with end-to-end speech queries.

6.3. Reasoning Ability and Human Evaluation

GPT4o	Gemini	Exp. 1	Exp. 2	Exp. 3
143/148	145/148	146/148	145/148	145/148

Table 5. Evaluation of the prediction structure of SilVar-Med models using GPT-4o mini, Gemini 1.5 Flash, and human experts (denoted as Exp. in the table).

To evaluate SilVar-Med's reasoning capabilities, we evaluate its response structure and reasoning accuracy of predictions using the LLM-as-Judge framework (GPT-4o and Gemini Flash 1.5) together with expert evaluations. For **structural responses**, we first analyze whether SilVar-Med's responses follow a coherent and structured format, as this is essential for medical interpretability. The results is shown in Tab. 5, in which, the scores of GPT-4o mini and Gemini Flash 1.5 are 143/148 and 145/148, respectively. In addition, our expert evaluations further reinforce these findings, with scores reaching 146/148, 145/148, and 145/148, indicating that the model generally maintains a structured response format that aligns with human expectations.

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
Ablation studies of SilVar-Med with the langue module of Deepseek R1 Distill 8B					
Completely Incorrect	12	10	10	40.00	20.67
Significantly Incorrect	39	41	47	8.67	23.00
Partially Correct	13	11	21	41.00	52.67
Fully Correct	84	86	70	58.33	51.67

Table 6. Assessment of SilVar-Med's reasoning accuracy behind abnormality prediction. The table compares expert evaluations (Exp. 1–3) with LLM-as-Judge assessments (GPT-4o and Gemini Flash 1.5). It is important to note that, Fully Correct denotes predictions that are both accurate and well-explained.

To assess the **reasoning accuracy** of SilVar-Med, we evaluate how well the model provides observations and justifications for its predictions. The reasoning accuracy is categorized into four levels: Completely Incorrect, Significantly Incorrect, Partially Correct, and Fully Correct, as shown in Tab. 6. A model is considered capable of reasoning-based abnormality detection if it can accurately respond to speech-driven medical queries while providing a coherent and justifiable explanation. Given the inherent variability in text generation by GPT-4o mini and Gemini Flash 1.5, we conducted three independent evaluation rounds per model and averaged the results to ensure consistency. In addition, three experts independently assessed the model outputs to provide a human benchmark for comparison.

Table 6 indicates notable discrepancies between expert evaluations and LLM-based assessments. Experts rate more responses as Fully Correct (74–97 for Llama 3.1 and 70–86 for Deepseek R1 Distill) compared to GPT-4o (59.67–58.33) and Gemini (48.33–51.67). Gemini is more conservative, labeling a higher number of responses as Partially Correct, while GPT-4o assigns more Completely Incorrect ratings. Overall, Table 6 shows that SilVar-Med demonstrates strong reasoning accuracy, effectively answering speech-driven medical queries with high prediction accuracy and well-structured explanations.

Despite the self-corrected and distilled learning in the general domain of Deepseek R1 8B Distill, we found that it achieves modest performance when integrated into SilVar-Med for medical abnormality detection. Additionally, the inconsistencies between GPT-4o, Gemini, and expert evaluations highlight the limitations of the LLM-as-Judge framework. While automated assessments provide useful insights, expert evaluation remains essential to ensure a balanced and clinically relevant assessment.

6.4. Evaluation on image-text VLMs Benchmarks

While SilVar-Med is a speech-driven medical VLM, no medical speech-driven VLMs currently exist for direct comparison. Our objective in this evaluation is not to achieve SOTA performance but rather to demonstrate the potential of voice-based medical communication with VLMs. To provide context for SilVar-Med's performance, we compare it with existing text-based medical VLMs across multiple datasets, including SLAKE, VQA-RAD, and Medical VQA 2019. Although our primary focus is on developing a speech-driven instruction-based medical VLM, we also include comparisons with its text-based counterparts. The evaluation results are presented in Tab. 7 and Tab. 8.

Performance on SLAKE: SilVar-Med (Llama 3.1-8B, speech-based) achieves an accuracy of 74.08% on SLAKE (Open QA) and 79.44% on SLAKE (Closed QA). Compared to LLaVA-Med++ (Medtrinity), which achieves 86.2% (Open) and 89.2% (Closed), SilVar-Med still has room for improvement, particularly in open-ended responses. However, the gap is smaller when comparing against LLaVA-Med, where SilVar-Med's performance remains competitive. It is important to note that most of the models in Tab. 7 are text based, and not able to generate reasoning behind prediction. Additionally, we also use text as direct input for language models. As a result, there is a small performance gap when using text-based input versus audio-based input for SilVar-Med.

Performance on VQA-RAD: Similarly, SilVar-Med achieves an accuracy of 55.34% on VQA-RAD (Open QA) and 62.56% on VQA-RAD (Closed QA). Compared to LLaVA-Med++ (Medtrinity), which achieves 77.1% (Open) and 86.0% (Closed), SilVar-Med exhibits lower performance, particularly in open-ended responses. However, when compared to earlier LLaVA-Med models, such as LLaVA-Med (BioMed CLIP) with 64.75% (Open)

Models	Instruction	SLAKE			VQA-RAD		
		Ref	Open	Closed	Ref	Open	Closed
Representatives of existing studies in the literature							
LLaVA [30]	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 7. Comparison of SilVar-Med with various text-based medical VLMs on the SLAKE and VQA-RAD datasets. Results are reported for both open-ended and closed-ended questions, with reference-based scores where applicable. LLaVA-based and other state-of-the-art (SoTA) models rely on text input, while SilVar-Med processes speech-driven queries.

and 83.09% (Closed), the performance gap is narrower. Notably, SilVar-Med's performance surpasses several traditional VLMs, such as MEVF+SAN and is competitive with models like CR.

Performance on Medical VQA 2019: For the Medical VQA 2019 dataset (Tab. 8), SilVar-Med achieves an accuracy of 64.99%, outperforming models like ImageCLEF (62.4%) and VGG16+BERT (62.4%), while being competitive with MMBERT (67.2%). In terms of BLEU score (62.24), SilVar-Med performs comparably to other models, indicating strong textual coherence. The BERT similarity score (0.80) is higher than MedVINT (0.63) and Med-Flamingo (0.65), suggesting that SilVar-Med's responses are more semantically aligned with the ground truth. These results demonstrate that SilVar-Med's performance is strong among speech-based models and is comparable to leading text-based models. The inclusion of a speech interface provides additional usability advantages in medical applications where hands-free interactions are crucial.

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

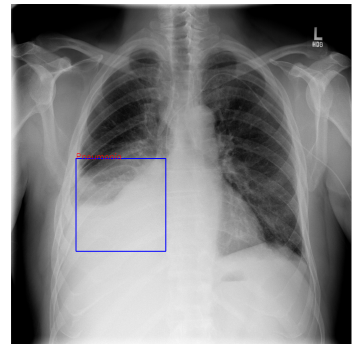
Table 8. Evaluation results of SilVar-Med with different text-based medical VLMs on the Medical VQA 2019 dataset.

7. Ablation Study

We conducted experiments with the language models in SilVar-Med by using different models, including Llama 2 and DeepSeek R1 (Distill-Llama-8B), to analyze their impact on SilVar-Med's performance. By testing SilVar-Med with multiple LLMs, we aim to identify the optimal configuration for medical abnormality detection and reasoning tasks. We evaluate the model's performance on both our dataset and published datasets, as shown in Tab. 6, Tab. 7, and Tab. 8.

Example 1: Abnormality detection.

Does the lung look abnormal?



Grounth 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.

Table 9. Comparison of prediction between our models and the other speech-driven model on the reasoning abnormal detection. Unlike GPT-4o and Gemini 1.5 Flash, our SilVar-Med is an end-to-end speech-driven VLM. For more demonstration, please visit SilVar-Med.

In Tab. 6, there is a small discrepancy between the performance of SilVar-Meds when using Llama 3.1 and DeepSeek R1 (Distill-8B), despite DeepSeek's distilled knowledge and reasoning ability in the general domain. In addition, the performance of SilVar-Med with Llama 2 is competitive compared to DeepSeek and Llama 3.1 on the Slake, VQA-RAD, and VQA 2019 datasets, as shown in Tab. 7 and Tab. 8. Compared to direct text input, the audio-driven approach achieves comparable performance or performs on par, despite challenges related to speech conversion errors, variations in spoken queries, and audio embeddings. These results highlight the robustness and adaptability of SilVar-Med, demonstrating its effectiveness across different language models and datasets.

Prompts	Reasoning Accuracy		BERTScore	BLEU
	GPT4o	Gemini		
Zero-shot	58.33	51.67	0.80	21.43%
COT	61	50	0.81	22.16%
TOT	59	47	0.80	21.44%

Table 10. Comparison of SilVar-Med using different prompts.

Furthermore, to investigate the reasoning ability of SilVar-Med in medical reasoning tasks, we conduct an ablation study by employing Chain-of-Thought (CoT) and Tree-of-Thought (ToT) prompting techniques. As shown in Tab. 10, we use GPT-4o and Gemini Flash 1.5 to evaluate the model's performance. The results indicate that structured reasoning techniques such as CoT and ToT might improve the model's performance compared to zero-shot prompting, although not significantly in our study.

8. Conclusion

In this study, we demonstrate a proof-of-concept study for speech-driven medical VLMs, focusing on reasoning for abnormality detection and interpretable AI assessments. We address two key challenges: (1) enabling voice communication in medical VLMs and (2) providing reasoning for each abnormality prediction. To this end, we also introduce a reasoning dataset for training and testing. The result is evaluated by three physicians along with a proposed LLM-as-Judge evaluation framework to assess both the accuracy and reasoning quality of its predictions.

Our experiments with reasoning interpretation, demonstrate the effectiveness of SilVar-Med in generating structured, accurate, and interpretable medical responses. Despite the challenge of speech-driven input, the model performs on par with other models. In terms of reasoning, although our work is limited by the dataset and the MRI, CT, and X-ray modalities, it provides reliable reasoning and demonstrates its potential in the medical domain, addressing the weaknesses of SOTA models. We also demonstrate that by minimizing speech-to-text errors, the model yields high-quality audio embeddings, leading to performance comparable to text-based models. Additionally, we found a lack of available speech-driven datasets benchmark for medical VLMs, highlighting a critical gap in the advancing field.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2

[3] 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. *arXiv preprint arXiv:2407.04106*, 2024. 2

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1

[5] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880, 2023. 2

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[7] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019*, Lugano, Switzerland, 2019. CEUR-WS.org. 1, 2, 4, 7

[8] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pages rs–3, 2024. 2

[9] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[10] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun 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. 2

[11] Qiuhui Chen and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian Conference on Computer Vision*, pages 2404–2420, 2024. 1, 2

[12] Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. Evaluation metrics for language models. 1998. 5

[13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[14] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 4

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[16] 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. 2, 4, 7

[17] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1, 2

[18] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024. 1, 2

[19] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023. 1

[20] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2

[21] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019. 2

[22] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021. 7

[23] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 1, 2, 4

[24] Khai Le-Duc, Ryan Zhang, Ngoc Son Nguyen, Tan-Hanh Pham, Anh Dao, Ba Hung Ngo, Anh Totti Nguyen, and

Truong-Son Hy. Litegpt: Large vision-language model for joint chest x-ray localization and classification task. *arXiv preprint arXiv:2407.12064*, 2024. 1, 2

[25] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 7

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[27] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 7

[28] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023. 2

[29] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 1, 2, 3, 4

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1, 2, 7

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[32] Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Conference on Information Processing in Medical Imaging*, pages 445–456. Springer, 2023. 7

[33] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1, 2, 7

[34] Andrew Cameron Morris, Viktoria Maier, and Phil D Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768, 2004. 5

[35] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019. 7

[36] OpenAI. Gpt-4. Available at `https://openai.com/gpt-4`, 2024. Model used for dataset generation. 1

[37] Tan-Hanh Pham, Hoang-Nam Le, Phu-Vinh Nguyen, Chris Ngo, and Truong-Son Hy. Silvar: Speech driven multimodal model for reasoning visual question answering and object localization. *arXiv preprint arXiv:2412.16771*, 2024. 1, 4

[38] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4

[39] Kanchana Ranasinghe and Michael S Ryoo. Language-based action concept spaces improve video self-supervised learning. *Advances in Neural Information Processing Systems*, 36:74980–74994, 2023. 1

[40] Fuji Ren and Yangyang Zhou. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636, 2020. 7

[41] Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*, 2024. 1

[42] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022. 1

[43] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[46] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Vision-language modelling for radiological imaging and reports in the low data regime. In *Medical Imaging with Deep Learning*, 2023. 2

[47] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2

[48] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. 1, 2, 7

[49] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. Zhejiang university at imageclef 2019 visual question answering in the medical domain. *CLEF (working notes)*, 85, 2019. 7

[50] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2

[51] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 2

[52] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020. 7

[53] Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints*, pages arXiv–2305, 2023. 2

[54] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 2, 7

[55] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 2, 7

[56] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2

[57] 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. 2