MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration

Yucheng Zhou, Lingran Song, Jianbing Shen*
SKL-IOTSC, CIS, University of Macau
yucheng.zhou@connect.um.edu.mo, jianbingshen@um.edu.mo

Abstract

Recent advancements in medical Large Language Models (LLMs) have showcased their powerful reasoning and diagnostic capabilities. Despite their success, current unified multimodal medical LLMs face limitations in knowledge update costs, comprehensiveness, and flexibility. To address these challenges, we introduce the Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis (MAM). Inspired by our empirical findings highlighting the benefits of role assignment and diagnostic discernment in LLMs, MAM decomposes the medical diagnostic process into specialized roles: a General Practitioner, Specialist Team, Radiologist, Medical Assistant, and Director, each embodied by an LLM-based agent. This modular and collaborative framework enables efficient knowledge updates and leverages existing medical LLMs and knowledge bases. Extensive experimental evaluations conducted on a wide range of publicly accessible multimodal medical datasets, incorporating text, image, audio, and video modalities, demonstrate that MAM consistently surpasses the performance of modality-specific LLMs. Notably, MAM achieves significant performance improvements ranging from 18% to 365% compared to baseline models. Our code is released at https://github.com/yczhou001/MAM.

1 Introduction

Large Language Models (LLMs) have recently demonstrated remarkable reasoning capabilities (Radford, 2018; OpenAI, 2023; Touvron et al., 2023a; Yang et al., 2024; Zhang et al., 2023b; Shao et al., 2024; Zhang et al., 2023a). Beyond demonstrating impressive language reasoning and generation capabilities, LLMs are expanded to process diverse modalities, e.g., images, audio, and video (Liu et al., 2023a; Chu et al., 2023; Zhang et al., 2023a). This formidable reasoning capacity holds significant promise for addressing problems in medical diagnostics.

For medical practice, physicians are confronted with a deluge of heterogeneous medical data, encompassing textual reports, medical images, cardiac sounds, and even surgical video recordings. Accurately extracting critical information from this complex data to arrive at precise diagnoses places a significant cognitive burden and challenge on clinicians. Furthermore, the growth in the volume of medical diagnostic data provides a substantial foundation for the training of LLMs. Consequently, the development of LLMs to enhance medical diagnostic workflows is crucial.

However, many efforts are directed towards constructing unified multimodal medical large models (Li et al., 2023; Thawakar et al., 2024; Deng et al., 2024). While these models have shown some progress in integrating multimodal information, they suffer from two limitations. Firstly, for unified models, each knowledge update is cost, often requiring substantial computational resources to retrain the entire model. Secondly, unified models lack modularity and flexibility, necessitating a single model to exhibit sufficient performance across various medical diagnostic tasks to satisfy demands. To explore the capabilities of existing domain-specific LLMs, we conducted an empirical study. Our findings indicate that role assignment significantly enhances the diagnostic abilities of LLMs, and LLMs possess the potential to discern the correct diagnosis from multiple ones.

To overcome the aforementioned limitations of unified multimodal medical LLMs and to better emulate the collaborative approach of human medical teams, we propose the Multi-Agent Framework for Multi-Modal Medical Diagnosis (MAM). Instead of pursuing an "omnipotent" unified model, MAM framework decomposes the medical diag-

* Corresponding author. This work was supported by the National Natural Science Foundation of China (No. 624B2002) and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

Findings of the Association for Computational Linguistics: ACL 2025, pages 25319–25333
July 27 - August 1, 2025

nostic process into several specialized roles and designs LLM-based agents for each role. These agents include: a General Practitioner agent responsible for initial triage, a Specialist Team agent providing domain-expert opinions, a Radiologist agent specializing in medical image analysis, a Medical Assistant agent aiding information retrieval and knowledge management, and a Director agent responsible for coordinating and synthesizing diagnostic opinions. The core advantages of the MAM framework lie in its modular design and collaborative workflow. The modular agent design enables more granular and efficient knowledge updates and model maintenance, without requiring global retraining. The framework also allows MAM to easily integrate and leverage various existing medical models and specialized knowledge bases.

In the experiments, we evaluate our MAM framework in multimodal medical diagnosis tasks through comprehensive experiments on several publicly available multimodal medical datasets. Experimental results demonstrate that the MAM framework consistently outperforms specific-modal LLMs across various medical datasets and data modalities. In addition, we conduct ablation studies, consistency analysis, and sensitivity analyses regarding the number of discussion rounds and roles to gain deeper insights into the roles of individual components and the operational mechanisms of the framework.

2 Related Work

2.1 LLM-based Multi-Agent

With the rapid advancement of LLMs, their application across various tasks has become increasingly widespread (Radford, 2018; OpenAI, 2023; Touvron et al., 2023a; Yang et al., 2024; Zhang et al., 2023b; Shao et al., 2024; Zhang et al., 2023a; Zhou et al., 2025, 2024a). These LLMs, outstanding in natural language processing, have been widely adapted to different tasks (Wang et al., 2024a; Yue et al., 2024; Yang et al., 2024; Zhou et al., 2024b; Hu et al., 2025). However, for relatively complex tasks, the capabilities of a single LLM may not achieve the desired effect. In order to solve complex tasks beyond the function of a single LLM, LLM-based multi-agent systems are developed (Liu et al., 2023c; Zhao et al., 2024; Talebirad and Nadiri, 2023). In (Wu et al., 2024), the author investigates the effectiveness of using LLM-based Multi-Agent to solve mathematical problems through dialogue, and MathChat is proposed as a conversational problem-solving solution designed for mathematical problems. In software engineering, the MAGIS (Tao et al., 2024) framework enables the collaboration of various agents in the planning and coding process to solve GITHUB problems. In the field of finance, inspired by the organizational structure of effective investment firms in the real world, FinCon (Yu et al., 2024) is developed to accomplish a variety of financial tasks.

The emergence of these LLM-based Multi-Agent systems points to a common conclusion that LLM-based Multi-Agents are well adapted for reasoning (Zhao et al., 2024; Talebirad and Nadiri, 2023; Wu et al., 2024; Tao et al., 2024), decision making (Liu et al., 2023c; Yu et al., 2024), etc. This adaptability makes them essential in the medical domain where interdisciplinary knowledge and multi-step problem solving are required.

2.2 Medical LLM

Due to the remarkable performance of LLMs in different tasks (Zhang et al., 2023b; Shao et al., 2024; Zhang et al., 2023a), various medical LLMs have been developed to solve a wide range of medical problems (Bao et al., 2023; Zhao et al., 2024; Ye et al., 2023; Fleming et al., 2024). As a comprehensive solution, DISC-MedLLM (Bao et al., 2023) utilizes LLM to deliver accurate and realistic medical responses in end-to-end conversational healthcare services. As the first LLaMA based Chinese medical LLM, Zhongjing (Zhao et al., 2024) has implemented the training pipeline from continuous pre-training, SFT, to Reinforcement Learning from Human Feedback (RLHF), where pre-training enhances medical knowledge and RLHF further improves instruction compliance and safety. Through a multi-stage training approach that combines domain-specific continuous pre-training (DCPT), SFT, and Direct Preference Optimization (DPO), Qilin-Med (Ye et al., 2023) shows substantial performance gains as a medical LLM.

Moreover, medical field is characterized by the presence of multimodal information, including diverse data types such as text, images, audios, etc. To make full use of these diverse data types, multimodal medical large models have been created (Li et al., 2023; Thawakar et al., 2024; Deng et al., 2024; Liu et al., 2023b; Chen et al., 2024). In LLaVA-Med (Li et al., 2023), the authors propose a cost-effective way to train a visual language con-

Dataset	Direct	Assigned Roles
MedQA (Jin et al., 2020)	30.8	50.6 (+19.8)
PubMedQA (Jin et al., 2019)	48.5	87.0 (+38.5)
PathVQA (He et al., 2020)	40.1	46.6 (+6.5)
PMC-VQA (Zhang et al., 2023c)	24.0	29.0 (+5.0)
DeepLesion (Yan et al., 2017)	11.1	40.0 (+28.9)
NIH (Wang et al., 2017)	12.6	50.7 (+38.1)
Brain Tumor (Bhuvaji et al., 2020)	80.2	98.2 (+18.0)
Heartbeat (Bentley et al., 2011)	43.9	62.5 (+18.6)
SoundDr (Hoang et al., 2023)	25.0	45.4 (+20.4)
MedVidQA (Gupta et al., 2022)	55.3	69.7 (+14.4)

Table 1: Performance comparison of "Direct" and "Assigned Roles" prompting methods across multi-modal medical tasks. The blue numbers indicate the performance improvement.

versational assistant that can answer open-ended research questions on biomedical images. In the task area of radiology, XrayGPT (Thawakar et al., 2024) was developed, which is a new conversational medical visual language model that can analyze and answer open-ended questions about chest radiographs. In the field of ophthalmology, OphGLM (Deng et al., 2024) has built a large multimodal model of ophthalmology, contributing to the clinical application of ophthalmology.

3 Empirical Study

3.1 The Significance of Assigned Roles in Medical Diagnosis with LLMs

Experimental Setting. We investigate the impact of assigned roles on the performance of Large Language Models (LLMs) in multimodal medical diagnosis. Our experiments leverage a diverse collection of publicly available medical datasets, encompassing text, image, audio, and video modalities. Specifically, we utilize the following datasets: Brain Tumor (Bhuvaji et al., 2020) (394 cases), DeepLesion (Yan et al., 2017) (225 cases), Heartbeat (Bentley et al., 2011) (461 cases), MedQA (Jin et al., 2020) (200 cases), MedVidQA (Gupta et al., 2022) (284 cases), NIH Chest X-rays (Wang et al., 2017) (215 cases), PathVQA (He et al., 2020) (200 cases), PMC-VQA (Zhang et al., 2023c) (200 cases), PubMedQA (Jin et al., 2019) (200 cases), and SoundDr (Hoang et al., 2023) (240 cases). We use these datasets to evaluate the capabilities of LLMs in handling multimodal medical diagnostic tasks. We emploly Qwen-Audio-Chat (Chu et al., 2023) for audio tasks, Medichat-Llama3-8B (sethuiyer, 2024) for text tasks, HuatuoGPT-Vision-7B (Chen et al., 2024) for image tasks, and VideoLLaMA2-7B (Cheng et al., 2024b) for video tasks. LLMs can assign roles in input prompts.

Dataset	Expectation	Reasoning
MedQA (Jin et al., 2020)	46.9	49.3 (+2.4)
PubMedQA (Jin et al., 2019)	54.0	72.6 (+18.6)
PathVQA (He et al., 2020)	50.0	91.7 (+41.7)
PMC-VQA (Zhang et al., 2023c)	41.7	62.5 (+20.8)
DeepLesion (Yan et al., 2017)	52.5	57.5 (+5.0)
NIH (Wang et al., 2017)	44.0	46.0 (+2.0)
Brain Tumor (Bhuvaji et al., 2020)	55.1	73.9 (+18.8)
Heartbeat (Bentley et al., 2011)	50.2	54.0 (+3.8)
SoundDr (Hoang et al., 2023)	48.1	55.7 (+7.6)
MedVidQA (Gupta et al., 2022)	46.0	50.0 (+4.0)

Table 2: LLM Diagnostic Discernment: Comparing Expected vs. Reasoning Accuracy. This table shows the performance of LLMs in selecting the correct diagnosis from a set of plausible alternatives. "Expectation" represents random selection accuracy, while "Reasoning" reflects the LLM's accuracy in identifying the correct diagnosis.

Table 1 compares two prompting strategies: "Direct" and "Assigned Roles". The "Direct" is without role assignment, while the "Assigned Roles" approach creates a physician role using a specific prompt (see Appendix B). The results indicate a consistent and significant performance improvement across all datasets using the "Assigned Roles" prompting strategy, with gains ranging from 5.0% (PMC-VQA) to 38.5% (PubMedQA). This suggests that role context enhances LLMs' ability to interpret and reason about medical data, improving diagnostic accuracy. Even in datasets with high baseline performance (e.g., Brain Tumor), role assignment provides a noticeable benefit. Furthermore, role assignment is particularly effective for tasks requiring deeper medical context and domain-specific knowledge.

3.2 LLMs' Capability to Discern Correct Reasoning Outcomes

To evaluate the reasoning capabilities of LLMs, we designed an experiment to assess their ability to identify the correct diagnosis from a set of plausible alternatives. Using the "Assigned Roles" prompting strategy, we generated three diagnostic outputs for each instance in the datasets. The datasets were pre-filtered to include only instances where at least one of the three diagnoses was correct, establishing a baseline "Expectation" representing random selection accuracy. The "Reasoning" column in Table 2 reflects the LLMs' accuracy in explicitly selecting the correct diagnosis from three generated options. Results show that the "Reasoning" accuracy consistently exceeds the "Expectation" across all datasets, with improvements ranging from 2.0% (NIH) to 41.7% (PathVQA). This demonstrates

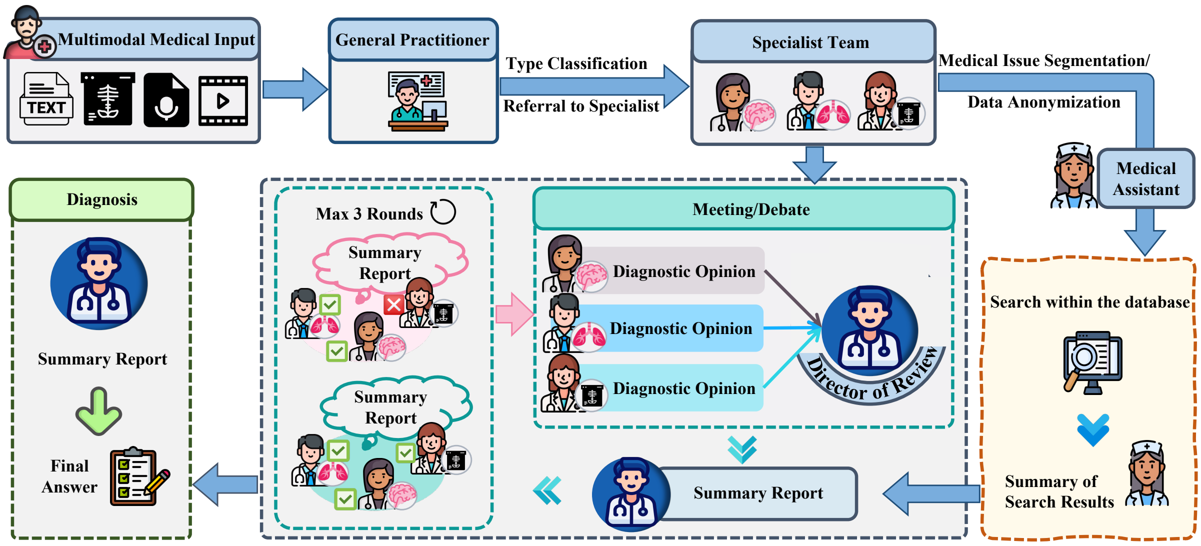

Figure 1: Overview of our MAM Framework for multi-modal medical diagnosis.

that LLMs possess reasoning capabilities beyond random selection, particularly in complex visual question answering tasks like PathVQA and PMC-VQA. The consistent positive delta across datasets indicates LLMs' potential to evaluate and refine their outputs to identify accurate conclusions.

4 Method

Based on our preliminary empirical studies, we observed that augmenting Large Language Models (LLMs) with specific medical roles significantly enhances their diagnostic performance. Furthermore, LLMs demonstrate a notable capacity to reason and synthesize correct diagnoses from diverse diagnostic opinions. Inspired by these findings, we propose the Multi-Agent Medical (MAM) framework, shown in Figure 1. This framework aims to transform multi-modal medical diagnosis into a collaborative endeavor, thereby amplifying the diagnostic capabilities of existing models. It comprises five key roles, each embodied by an LLM-based agent, working synergistically within a defined workflow: ① **General Practitioner**: Responsible for initial *Disease Type Classification* and *Referral to Specialist*. ② **Specialist Team**: Charged with providing *Diagnostic Opinions* on specific medical conditions and actively participating in discussions. ③ **Radiologist**: Tasked with analyzing medical images and contributing to diagnostic discussions. ④ **Medical Assistant**: Responsible for retrieving and summarizing relevant medical information from databases. ⑤ **Director**: Synthesizes discussion reports and reviews the quality of medical diagnoses.

In our framework, multi-modal medical inputs are initially directed to the General Practitioner, who performs disease classification and subsequently refers the case to the relevant Specialist Team. The Specialist Teams will decompose and anonymize the medical problem. The Medical Assistant then retrieves and summarizes pertinent information from medical databases based on the decomposed problem. Subsequently, the Director orchestrates discussions among the Specialist Team, where each specialist presents their diagnostic opinion. The Director then synthesizes these opinions and the database summaries into a comprehensive report. The Specialist Team reviews this report and votes on whether to endorse it. In cases of disagreement, the process iteratively re-enters the Specialist Team discussion phase. Otherwise, upon reaching a consensus, the Director derives the final diagnosis based on the synthesized report.

4.1 Doctor Agent Role Design

General Practitioner The General Practitioner agent is designed to mimic the role of a primary care physician in a clinical setting. Upon receiving multi-modal medical inputs, this agent is responsible for the initial triage, performing *Disease Type Classification* to categorize the medical case. Crucially, it then determines the appropriate *Referral to Specialist*, directing cases to the relevant Specialist Team based on the initial classification.

Specialist Team The Specialist Team is composed of multiple agents, each representing a specialist in a specific medical domain. These agents are tasked with providing *Diagnostic Opinions* relevant to their expertise. They engage in discussions, sharing their perspectives and interpretations of the medical case. Furthermore, Specialist Team members participate in a voting process to reach a consensus on the synthesized diagnostic report.

Radiologist The Radiologist agent specializes in the interpretation of medical images, such as X-rays and CT scans. Its primary responsibility is to analyze these images and provide imaging-based insights to the other agents. The Radiologist communicates with the Specialist Team and the Director, offering expertise in image analysis to aid in diagnosis and treatment planning.

Medical Assistant The Medical Assistant agent plays a crucial role in information management. Its responsibilities include processing medical data to facilitate retrieval of relevant information from medical databases. Furthermore, the Medical Assistant summarizes the retrieved information, providing concise summaries.

Director The Director agent serves as the orchestrator and synthesizer within the MAM framework. This agent is responsible for reviewing the diagnostic opinions provided by the Specialist Team and the Radiologist. It synthesizes the discussion outcomes from the Specialist Team into a comprehensive report. Crucially, the Director derives the final diagnosis based on the synthesized report, specifically when a consensus is reached among the Specialist Team through voting.

4.2 Collaborative Diagnosis Process

The MAM framework orchestrates a collaborative diagnostic process initiated upon receiving multi-modal medical inputs. Let $M = \{m_1, m_2, ..., m_k\}$ represent the multi-modal medical input, where m_i denotes the i-th modality.

Initial Triage and Referral: The General Practitioner agent (G) receives the multi-modal input M. G performs *Disease Type Classification* to categorize the medical case into a disease type d. This can be represented as:

$$d = C^G(M) \tag{1}$$

where C^G denotes the disease type classification function performed by agent G.

Based on the classified disease type d, G determines the appropriate Specialist Team $S = \{s_1, s_2, ..., s_n\}$ for referral. This referral process can be represented by:

$$S = R^G(d) \tag{2}$$

where R^G is the referral function by agent G, and S is the set of specialist agents s_i.

Problem Decomposition and Anonymization: The Specialist Team S receives the medical case and decomposes the problem into a set of sub-problems $P = \{p_1, p_2, ..., p_m\}$. Anonymization is performed concurrently. This decomposition is represented by:

$$P = D^S(M) \tag{3}$$

where D^S is the problem decomposition function by Specialist Team S.

Information Retrieval: The Medical Assistant agent (A) utilizes the decomposed problem P to retrieve relevant medical information. Let I_r represent the retrieved information, obtained through:

$$I_r = \text{Retrieve}^A(P) \tag{4}$$

where Retrieve^A is the information retrieval function by agent A. Since we do not have access to a real hospital database, the retrieval process is conducted using the Google API. The query used for retrieval is based on the decomposed and anonymized problem P, ensuring no privacy leakage. A then summarizes the retrieved information into a concise summary I_s:

$$I_s = \text{Summarize}^A(I_r) \tag{5}$$

where Summarize^A is the summarization function by agent A.

Diagnostic Opinion Generation and Discussion: Each specialist $s_i \in S$ and the Radiologist agent (Rad) independently generate their diagnostic opinions based on the multi-modal input M, and the information summary I_s. Let O_{s_i} be the diagnostic opinion of specialist s_i and O_{Rad} be the opinion of the Radiologist. Opinions are generated through:

$$O_{s_i} = \text{Diag}^{s_i}(M, I_s) \quad \forall s_i \in S \tag{6}$$

$$O_{Rad} = \text{Diag}^{Rad}(M) \tag{7}$$

where Diag^{s_i} and Diag^{Rad} are the diagnostic opinion generation functions for specialist s_i and Radiologist Rad, respectively. The Director agent (Dir) orchestrates a discussion.

Algorithm 1 Consensus and Iteration Process

```
while No Consensus do
    Specialist Team S and Radiologist Rad present and discuss diagnostic opinions {O_{s_i}}_{s_i∈S} and O_{Rad}.
    Director Dir synthesizes a report R_p = Synth_{Dir}({O_{s_i}}_{s_i∈S}, O_{Rad}, I_s).
    Specialist Team S reviews report R_p.
    for each specialist s_i ∈ S do
        Specialist s_i votes v_i ∈ {0, 1} on endorsing R_p.
    end for
    Calculate total endorsement votes V = ∑_{i=1}^{n} v_i.
    if V == n then
        Consensus Reached ← True.
    else
        Consensus Reached ← False.
    end if
end while
return Consensus Reached
```

Report Synthesis and Review: The Director agent (Dir) synthesizes the diagnostic opinions $\{O_{s_1}, O_{s_2}, ..., O_{s_n}, O_{Rad}\}$ and the information summary I_s into a comprehensive diagnostic report R_p. This synthesis is performed by:

$$R_p = \text{Synth}^{Dir}(\{O_{s_i}\}_{s_i \in S}, O_{Rad}, I_s) \tag{8}$$

where Synth^{Dir} is the synthesis function by agent Dir. The Specialist Team S reviews R_p and votes on endorsement. Let $v_i \in \{0, 1\}$ be the vote of specialist s_i.

Consensus and Iteration: The Director agent checks for consensus. Let $V = \sum_{i=1}^{n} v_i$ be the total endorsement votes. If $V = n$ (assuming $n = 3$ in your algorithm description seems to be a typo, it should be consensus of all specialists which is $V = n$), consensus is reached.

Final Diagnosis Derivation: Upon reaching consensus, the Director agent derives the final diagnosis D_{final} based on R_p. This is performed by:

$$D_{final} = \text{Diagnosis}^{Dir}(R_p) \tag{9}$$

where Diagnosis^{Dir} is the diagnosis derivation function by agent Dir. D_{final} is the output of the MAM framework.

Method	MedQA	PubMedQA
LLaMA-7B (Touvron et al., 2023b)	18.6	37.2
DAPT-7B (Gururangan et al., 2020)	25.7	44.1
MedAlpaca-7B (Han et al., 2023)	29.3	51.2
AdaptLLM-7B (Cheng et al., 2024a)	30.5	56.8
LLaMA-3-8B (AI@Meta, 2024)	29.6	43.6
Medichat-Llama3-8B (sethuiyer, 2024)	30.8	48.5
MAM	**40.0**	**84.0**

Table 3: Performance comparison of different LLMs on text-based medical datasets.

5 Experiments

5.1 Setup

Text-based Evaluation. For text-based evaluation, we used MedQA (Jin et al., 2020) and PubMedQA (Jin et al., 2019). We selected 200 English four-option multiple-choice questions from MedQA (Jin et al., 2020) and 200 question-answer pairs from PubMedQA (Jin et al., 2019). We compared MAM against LLMs: LLaMA-7B (Touvron et al., 2023b), DAPT-7B (Gururangan et al., 2020), MedAlpaca-7B (Han et al., 2023), AdaptLLM-7B (Cheng et al., 2024a), LLaMA-3-8B (AI@Meta, 2024), and Medichat-Llama3-8B (sethuiyer, 2024).

Image-based Evaluation. For image evaluation, we used Brain Tumor (Bhuvaji et al., 2020) (test set, 394 cases), DeepLesion (Yan et al., 2017) (225 cases from 9 categories), NIH Chest X-rays (Wang et al., 2017) (215 cases), PathVQA (He et al., 2020) (200 cases), and PMC-VQA (Zhang et al., 2023c) (200 pairs). Compared LVLMs include LLaVA-7B (Liu et al., 2023a), Qwen2-VL-7B (Wang et al., 2024b), LLaVA-Med-7B (Li et al., 2023), Qilin-Med-VL-13B (Liu et al., 2023b), and HuatuoGPT-Vision-7B (Chen et al., 2024).

Audio-based Evaluation. For audio evaluation, we used Heartbeat (Bentley et al., 2011) (clinical trial data, 461 instances) and SoundDr (Hoang et al., 2023) (240 instances). MAM is compared with Qwen-Audio-Chat (Chu et al., 2023).

Video-based Evaluation. For video evaluation, we used MedVidQA (Temporal Segment Prediction test set, 284 data points (Gupta et al., 2022)). We preprocessed it into yes/no questions using original and alternative video segments. MAM is compared with video-LLMs: LLaVA-Next-Video-7B (Zhang et al., 2024), Qwen2-VL-7B (Wang et al., 2024b), and VideoLLaMA2-7B (Cheng et al., 2024b).

Method	Pa	PMC	DL	NIH	BT
LLaVA-7B (Liu et al., 2023a)	7.3	6.2	2.6	4.2	34.6
Qwen2-VL-7B (Wang et al., 2024b)	29.5	10.6	3.6	6.3	52.6
LLaVA-Med-7B (Li et al., 2023)	36.3	19.8	8.5	9.2	73.7
Qilin-Med-VL-13B (Liu et al., 2023b)	39.2	22.5	11.3	11.4	80.6
HuatuoGPT-Vision-7B (Chen et al., 2024)	40.1	24.0	11.1	12.6	80.2
MAM	**47.6**	**32.5**	**35.1**	**58.6**	**97.9**

Table 4: Performance comparison of different LVLMs on various image-based medical datasets."Pa", "PMC", "BT" and "DL" denote "PathVQA", "PMC-VQA", "Brain Tumor" and "DeepLesion".

Method	Heartbeat	SoundDr
Qwen-Audio-Chat (Chu et al., 2023)	34.9	25.0
MAM	**64.0**	**47.9**

Table 5: Performance comparison of audio-LLMs on audio-based medical datasets.

5.2 Main Results

Our comprehensive experiments across text, image, audio, and video medical data demonstrate the superior performance of the MAM framework. As shown in Table 3-6, MAM consistently outperforms strong competitors across all modalities and achieves significant performance improvements ranging from 18% to 365% compared to baseline models. For text-based medical question answering (Table 3), MAM significantly surpasses baseline LLMs on MedQA and PubMedQA datasets, demonstrating enhanced medical text understanding. In image-based diagnosis (Table 4), MAM achieves top accuracy across PathVQA, PMC-VQA, DeepLesion, NIH Chest X-rays, and Brain Tumor datasets, with particularly substantial gains on DeepLesion and NIH Chest X-rays. Audio-based results on Heartbeat and SoundDr datasets (Table 5) show MAM's clear advantage over audio-LLM baselines in medical audio interpretation. For video-based medical question answering on MedVidQA (Table 6), MAM achieves leading accuracy, outperforming all video-LLM competitors. These results collectively demonstrate MAM's efficacy in multi-modal medical diagnosis, highlighting the benefits of its collaborative multi-agent approach.

5.3 Ablation Study

To evaluate the contribution of each component in the MAM framework, we conducted an ablation study, with results shown in Table 7. The study systematically assesses the impact of incrementally adding key functionalities, starting from a baseline "Direct" approach (using the baseline LLM directly for diagnosis) and progressively in-

Method	MedVidQA
LLaVA-Next-Video-7B (Zhang et al., 2024)	51.5
Qwen2-VL-7B (Wang et al., 2024b)	54.8
VideoLLaMA2-7B (Cheng et al., 2024b)	55.3
MAM	**74.3**

Table 6: Performance comparison of different video-LLMs on the Video-based medical dataset.

Dataset	Direct	+Roles	+Discussion	+Retrival
	Incrementally Added Function →			
MedQA	30.8	31.0	32.5	**40.0**
PubMedQA	48.5	69.5	77.0	**84.0**
PathVQA	40.1	46.0	47.0	**47.6**
PMC-VQA	24.0	26.0	32.0	**32.5**
DeepLesion	11.1	33.8	34.7	**35.1**
NIH	12.6	36.0	38.6	**58.6**
Brain Tumor	80.2	92.4	97.0	**97.9**
Heartbeat	34.9	35.1	49.5	**64.0**
SoundDr	25.0	32.9	43.3	**47.9**
MedVidQA	55.3	58.0	60.6	**74.3**

Table 7: Ablation study of our MAM framework. The "Direct" represents the baseline. From left to right, we incrementally add functions. "+Retrieval" is our full MAM framework.

tegrating agent roles (+Roles), inter-agent discussion (+Discussion), and information retrieval (+Retrieval), representing the complete MAM framework. The results reveal consistent performance improvements across all datasets as each component is added. The introduction of agent roles (+Roles) shows significant gains over the baseline, emphasizing the value of role specialization. Enabling discussion (+Discussion) further enhances performance, demonstrating the benefits of collaborative reasoning. Most notably, the full MAM framework (+Retrieval) achieves the highest performance, highlighting the synergistic effects of role specialization, collaborative discussion, and information retrieval. The substantial improvement from "+Discussion" to "+Retrieval" underscores the critical role of the Medical Assistant in enhancing diagnostic accuracy through relevant medical knowledge. These findings confirm the efficacy of each MAM component and their combined impact on multi-modal medical diagnosis.

5.4 Consistency

To evaluate the MAM framework's behavior, we analyzed its prediction consistency compared to the "Direct" approach. Consistency is defined as the percentage of instances where MAM's final prediction aligns with a correct prediction from the "Direct" method. This metric assesses MAM's

Dataset	Consistency	MAM
MedQA (Jin et al., 2020)	34.4	40.0
PubMedQA (Jin et al., 2019)	74.2	84.0
PathVQA (He et al., 2020)	50.0	47.6
PMC-VQA (Zhang et al., 2023c)	14.6	32.5
DeepLesion (Yan et al., 2017)	12.0	35.1
NIH (Wang et al., 2017)	59.3	58.6
Brain Tumor (Bhuvaji et al., 2020)	97.5	97.9
Heartbeat (Bentley et al., 2011)	70.2	64.0
SoundDr (Hoang et al., 2023)	60.0	47.9
MedVidQA (Gupta et al., 2022)	67.5	74.3

Table 8: Consistency of prediction results from baseline (Direct) and MAM. Rows with lighter cyan color indicate datasets where MAM has relatively lower performance.

ability to retain and reinforce correct baseline predictions while correcting errors. Table 8 compares the consistency scores with MAM's overall performance across datasets. Results indicate a positive correlation between MAM's performance and consistency. For instance, datasets, where MAM performs well, show high consistency, suggesting MAM effectively builds on the "Direct" method's correct predictions. In contrast, datasets with lower performance, such as PMC-VQA and DeepLesion lower consistency. This implies that when the "Direct" achieves lower accuracy, MAM may introduce changes that slightly reduce consistency with the original correct predictions. Nevertheless, MAM generally outperforms the "Direct" method overall, as shown in Table 7 and Table 8, indicating that its refinements enhance diagnostic accuracy despite occasional deviations from the baseline's correct predictions. This demonstrates that MAM actively improves predictions through its collaborative, knowledge-augmented framework rather than merely replicating the "Direct" approach.

5.5 Discussion Time and Performance

We investigated the impact of iterative discussions on diagnostic accuracy by evaluating performance across different discussion rounds, as illustrated in Figure 2. For Brain Tumor, performance improved in early rounds, indicating that iterative discussions enhance accuracy for complex cases. However, extending discussions beyond a few rounds did not consistently yield further gains. For MedQA and PathVQA, performance fluctuated, with peak accuracy often achieved within the first two or three rounds. PMC-VQA experienced a performance decline in the final round, suggesting potential overfitting or dataset-specific issues. Results imply that while initial discussions can refine diagnoses, successive rounds introduce noise or dilute accurate initial opinions. Limiting discussions can balance collaborative benefits and avoid over-discussion.

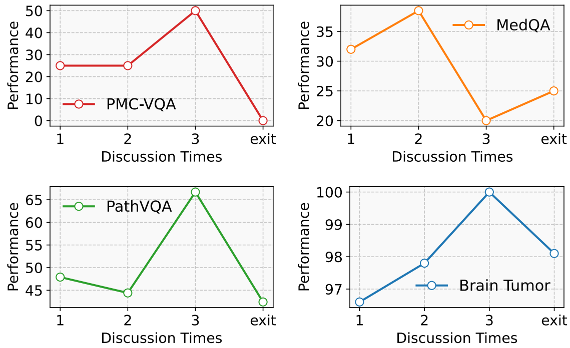

Figure 2: Performance with different times of discussion (≤ 3) in our MAM pipeline across various datasets.

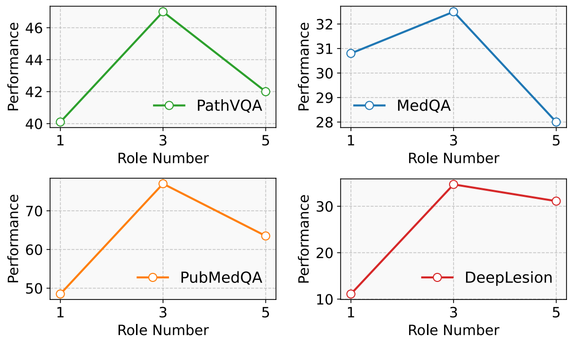

Figure 3: Performance with different number of roles in our MAM pipeline across various datasets.

5.6 Impact of Role Number

We investigate the effect of role granularity by varying the number of agents in MAM (Figure 3). Performance generally followed an inverted U-shape: increasing roles from 1 ("Direct") to 3 significantly improved results, highlighting the benefit of role specialization. However, further increasing to 5 roles led to a performance decrease across datasets. This suggests an optimal level of role granularity exists. While role specialization is beneficial, excessive roles may introduce redundancy or overhead, hindering diagnosis. A moderately specialized framework with 3 roles appears to strike a better balance than either a single-agent approach or an overly complex multi-agent system, indicating that streamlined role specialization is crucial for effective collaborative medical diagnosis.

5.7 Recall of Retrieval

To evaluate the Medical Assistant's information retrieval module, we first measured recall, defined as

Dataset	Recall	Answer Correct
DeepLesion (Yan et al., 2017)	31.7	53.4
Heartbeat (Bentley et al., 2011)	34.0	58.8
NIH (Wang et al., 2017)	12.1	46.2

Table 9: "Recall" indicates the proportion of instances where the retrieved content includes the correct answer. "Answer Correct" represents the accuracy of the final answer under the retrieved content that encompasses the correct answer.

the proportion of retrieved medical documents containing information necessary to correctly answer diagnostic questions. As shown in Table 9, recall varies across datasets, ranging from 12.1% for NIH to 34.0% for Heartbeat. These results indicate that while the module retrieves relevant information in some cases, significant improvement is needed. Imperfect recall may stem from limitations in retrieval algorithms, incomplete medical databases, or challenges in formulating effective search queries for diverse medical questions. Enhancing recall is critical for ensuring the availability of necessary information for downstream diagnostic tasks.

5.8 Impact from Retrieval Content

Complementary to evaluating retrieval recall, we examined the impact of retrieved content on diagnostic accuracy by calculating the conditional probability of obtaining a correct answer when the retrieved documents contained the necessary information. This metric, labeled "Answer Correct" in Table 9, assesses the MAM framework's ability to leverage retrieved information effectively. As shown in Table 9, the "Answer Correct" is consistently higher than the corresponding accuracy of the "+Discussion" (without retrieval) in Table 7. For instance, in the NIH dataset, the "Answer Correct" (46.2%) significantly surpasses the "Discussion" (38.6%). It demonstrates that when relevant information is retrieved, the MAM framework is more likely to arrive at a correct diagnosis. However, the "Answer Correct" is not perfect, demonstrating the importance of improving retrieval and LLM's reasoning capabilities.

5.9 Case Study

Figure 4 shows a case of multimodal input from DeepLesion, comparing the outputs from the baseline and our framework. The baseline model produced incorrect results, whereas our framework delivered correct predictions. Our framework begins by identifying the input modality and determining the data type. Based on this information, it generates three expert roles to engage in up to three rounds of dialogue to discuss potential solutions. Concurrently, the Medical Assistant formulates queries for web retrieval. After processing the retrieved data, the Director reviews the discussion and retrieval records, synthesizes the insights into a summary, and presents it to the expert team for voting. The Director then uses the voting results and summary to make the final diagnosis. As shown in the process discussed in Figure 4, it is evident that although not all expert roles prioritized the correct answer initially, the structured approach of discussions and voting leads to an accurate resolution, which demonstrates effectiveness of our framework, demonstrating its ability of decision-making in complex medical scenarios.

6 Conclusion

This study introduces the Multi-Agent Framework for Multi-Modal Medical Diagnosis (MAM), addressing the limitations of unified multimodal medical LLMs. MAM employs a modular, collaborative approach, assigning specialized roles, i.e., General Practitioner, Specialist Team, Radiologist, Medical Assistant, and Director, to distinct LLM-based agents. This structure enhances knowledge updates, leverages specialized expertise, and adapts to diverse medical tasks and modalities. Extensive evaluations on multimodal medical datasets demonstrate MAM's superiority, outperforming modality-specific LLMs by 18% to 365%. Future work will integrate advanced knowledge retrieval and evaluate MAM in real-world clinical settings.

Limitations

The performance of MAM is fundamentally constrained by the capabilities of the underlying LLMs utilized for each agent role. Inherent limitations such as model biases, knowledge gaps, or reasoning inaccuracies within these LLMs may propagate through the framework, potentially compromising diagnostic outcomes. MAM's architecture allows for flexible switching of base models, which could mitigate some limitations in future applications. The other limitation of the current study is the absence of real-world clinical validation, which presents substantial challenges in terms of resource allocation and human expertise required for comprehensive evaluation. We acknowledge this limitation and propose to address it through clinical validation studies in our future work.

References

AI@Meta. 2024. Llama 3 model card.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *CoRR*, abs/2308.14346.

P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. 2011. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results.

Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. 2020. Brain tumor classification (mri).

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *CoRR*, abs/2406.19280.

Daixuan Cheng, Shaohan Huang, Ziyu Zhu, Xintong Zhang, Wayne Xin Zhao, Zhongzhi Luan, Bo Dai, and Zhenliang Zhang. 2024a. On domain-specific post-training for multimodal large language models. *CoRR*, abs/2411.19930.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919.

Zhuo Deng, Weihao Gao, Chucheng Chen, Zhiyuan Niu, Zheng Gong, Ruiheng Zhang, Zhenjie Cao, Fang Li, Zhaoyi Ma, Wenbin Wei, and Lan Ma. 2024. Ophglm: An ophthalmology large language-and-vision assistant. *Artif. Intell. Medicine*, 157:103001.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Pontes Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher D. Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2024. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22021–22030. AAAI Press.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2022. A dataset for medical instructional video classification and question answering. *CoRR*, abs/2201.12888.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca - an open-source collection of medical conversational AI models and training data. *CoRR*, abs/2304.08247.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *CoRR*, abs/2003.10286.

Truong V. Hoang, Quang H. Nguyen, Cuong Q. Nguyen, Phong X. Nguyen, and Hoang D. Nguyen. 2023. Sound-dr: Reliable sound dataset and baseline artificial intelligence system for respiratory illnesses. *Preprint*, arXiv:2201.04581.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, and Laizhong Cui. 2025. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023,* *NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023b. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *CoRR*, abs/2310.17956.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023c. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *CoRR*, abs/2310.02170.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Alec Radford. 2018. Improving language understanding by generative pre-training.

sethuiyer. 2024. Medichat-llama3-8b. `https://huggingface.co/sethuiyer/Medichat-Llama3-8B`. Built upon LLaMa-3 architecture.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent LLM agents. *CoRR*, abs/2306.03314.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. MAGIS: llm-based multi-agent framework for github issue resolution. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2024, Bangkok, Thailand, August 16, 2024*, pages 440–448. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024a. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3462–3471. IEEE Computer Society.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2024. Mathchat: Converse to tackle challenging math problems with LLM agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. 2017. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *CoRR*, abs/1710.01766.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19368–19376. AAAI Press.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *CoRR*, abs/2310.09089.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *Database Systems for Advanced Applications - 29th International Conference, DASFAA 2024, Gifu, Japan, July 2-5, 2024, Proceedings, Part V*, volume 14854 of *Lecture Notes in Computer Science*, pages 304–321. Springer.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10859–10885. Association for Computational Linguistics.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023c. PMC-VQA: visual instruction tuning for medical visual question answering. *CoRR*, abs/2305.10415.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: LLM agents are experiential learners. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19632–19642. AAAI Press.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024a. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics.

Yucheng Zhou, Jianbing Shen, and Yu Cheng. 2025. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yucheng Zhou, Jihai Zhang, Guanjie Chen, Jianbing Shen, and Yu Cheng. 2024b. Less is more: Vision representation compression for efficient video generation with large language models.

A Case Study

Figure 4 shows the comparison results between the baseline model and our framework under the same sample input from DeepLesion (Yan et al., 2017) dataset.

B Prompt

Image Type Classification Prompt

1.Please answer with a single word: What kind of medical image is this? X-Ray, CT, MRI, Pathology, Biomedical.
2.Please answer with a single word: What part of the human body does this image show? Brain, bone, abdomen, mediastinum, liver, lung, kidney, soft tissue, pelvis.

Audio Type Classification Prompt

Please answer with a single word: What kind of audio is this? Cardiovascular, Respiratory.

Video Type Classification Prompt

Please answer with a single word: What kind of video is this? Sports, Rehabilitation, Emergency.

Text Type Classification Prompt

System prompt: You are given a question, please select a question type according to the given question.
Input: The question is $\{question\_text\}$. Which kind of question is this? Anaesthesia, Anatomy, Biochemistry, Dental, ENT, FM, O&G, Medicine, Microbiology, Ophthalmology, Orthopaedics, Pathology, Pediatrics, Pharmacology, Physiology, Psychiatry, Radiology, Skin, PSM, Surgery, Unknown.
Output example:
The question type is **Anaesthesia**.

Role Generation Prompt

Given a disease type, generate a system prompt that assigns tasks to relevant medical roles, including **Specialist Doctor**, **Radiologic Technologist**, etc, from the perspective of a General Practitioner.
Input: The modality type is $\{modality\_type\}$, the disease type is $\{disease\_type\}$, and the patient question is $\{question\}$.
Output:
A system prompt that:
Identifies the relevant Specialist Doctor(s), Radiologic Technologist(s), and other Specialist(s) for the given disease type.
Assigns tasks to each identified role, specifying the necessary actions, tests, or examinations required for diagnosis and treatment.
Output example:
Specialist Doctor (Pulmonologist):
- Assess Patient's Health: Evaluate patient's function and overall health.
- Use $\{modality\_type\}$ Studies: Utilize expertise in the $\{disease\_type\}$ domains to diagnose diseases.
- Analyze Patient History and Symptoms: Determine the cause and severity of diseases by analyzing patient's medical history and symptoms.

Get Discuss Prompt

You are a $\{role\_name\}$, responsible for the following tasks: $\{role\_responsibilities\}$. Please thoughtfully express your views for the following question.
Input:
Question type: $\{disease\_type\}$.
Question: $\{question\}$.
Example output:
Assessment Steps:
- Initial Assessment: [Provide a detailed overview of the initial assessment process]
- Diagnostic Studies (e.g., imaging, lab tests): [Include relevant details about any studies conducted]
- Additional Considerations: [Mention any other pertinent factors or evaluations]
Possible Answers:
- Answer 1: [Briefly explain answer 1]
Reasoning: [Briefly describe the corresponding reason for answer 1]
- Answer 2: [Briefly explain answer 2]
Reasoning: [Briefly describe the corresponding reason for answer 2]
- Answer 3: [Briefly explain answer 3]
Reasoning: [Briefly describe the corresponding reason for answer 3]
Conclusion: [Summarize the findings and provide a final recommendation or insight]

Get Summarize Prompt

You are a specialized doctor serving as the moderator of this meeting. Please provide a detailed summary of the discussions that have taken place.
Example output:
Possible Answers:
- Answer 1: [Briefly explain answer 1]
- Answer 2: [Briefly explain answer 2]
- Answer 3: [Briefly explain answer 3]
Agreements:
- [Description of any agreements reached]
Disagreements:
- [Description of any disagreements that were noted]
Conclusions:
- [Final thoughts or conclusions drawn from the discussion]
Input: The question is $\{question\}$. The previous discussion of the meeting includes: $\{discussion\}$.

Get Vote Prompt

You are a $\{role\_name\}$, responsible for the following tasks: $\{role\_responsibilities\}$.
Please answer just using "yes" or "no" according to the following questions and the corresponding summery and the contents of the given file(if any).

Input: The question is $\{question\}$, and the summery of the discussion is:
$\{summary\}$
Do you agree with the summery above? Please answer just using "yes" or "no".

Get Review Prompt

Question: Is there any medical reasoning errors, redundant statements, or invalid outputs in the following paragraph?
Please answer just using "yes" or "no".
Please read the rollowing paragraph:
$\{dis\}$

Get Multimodal Description Prompt

Please describe this $\{modality\_type\}$ briefly in 100 words:

Get Search Summarize Prompt

Please summarize the following search results briefly in 200 words: $\{search\_result\}$

Get Diagnosis Prompt

Input: Based on the provided image/video/audio (if applicable) and the meeting record, please provide answer to the following question.
Question: $\{ques\}$.
Meeting record: $\{record\}$.

Get Overall Review Prompt

Input: You're a medical assistant. Please check whether the answer to this question is reasonable, if it is, please answer "yes", if not, please answer "no".
Question: $\{ques\}$.
Answer: $\{record\}$.

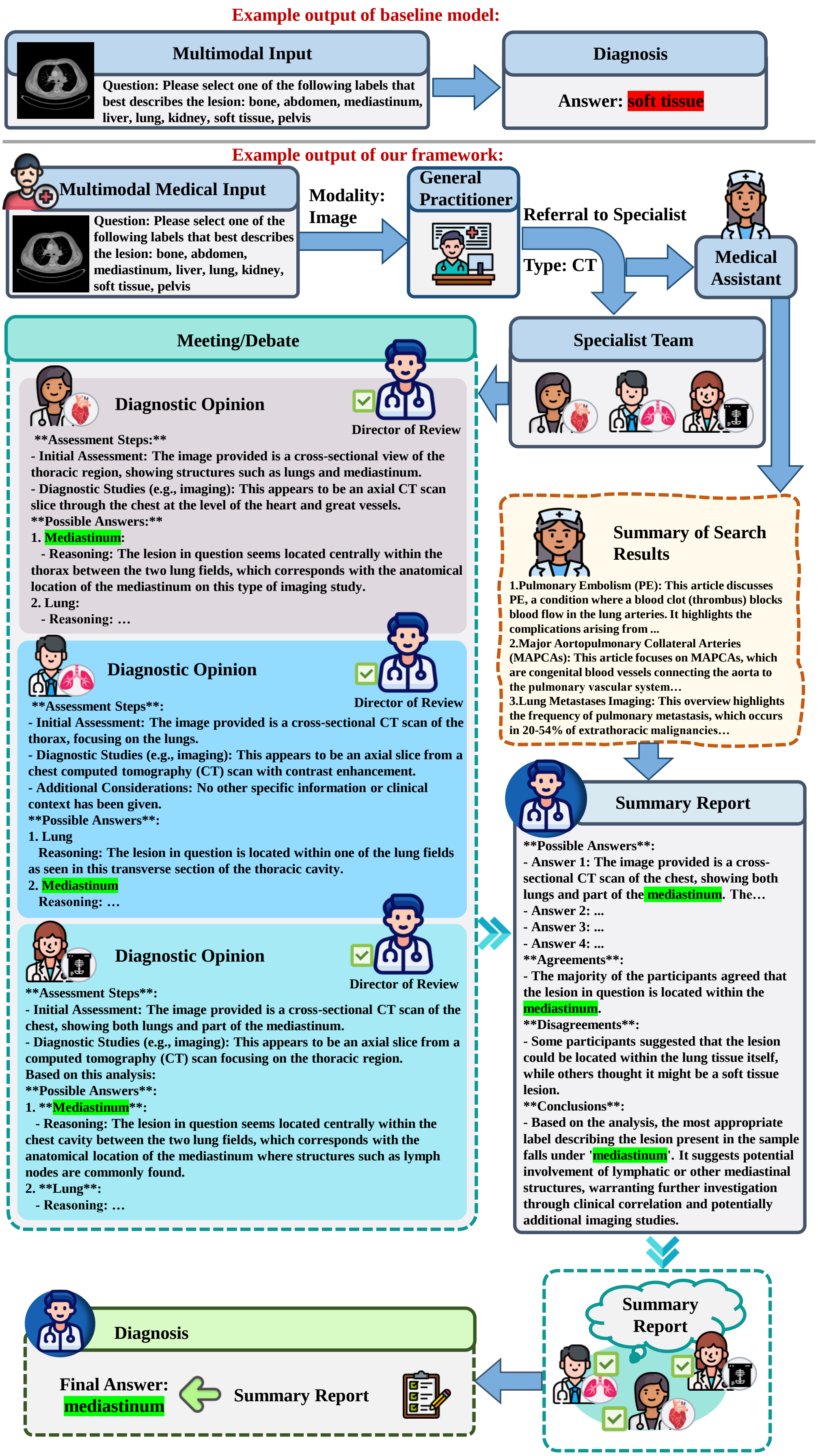

Figure 4: Case Study.