MMedAgent: Learning to Use Medical Tools with Multi-modal Agent

Binxu Li[1], Tiankai Yan[1], Yuanting Pan[1], Jie Luo[2], Ruiyang Ji[3], Jiayuan Ding[4], Zhe Xu[2,5,*], Shilong Liu [6], Haoyu Dong[7*], Zihao Lin[3*], Yixin Wang[1*]

[1]Stanford University, [2]Harvard Medical School, [3]Virginia Tech, [4]MSU, [5]CUHK, [6]Tsinghua University, [7]Duke University

{andy0207,yxinwang}@stanford.edu, haoyu.dong151@duke.edu, zihaol@vt.edu

Abstract

Multi-Modal Large Language Models (MLLMs), despite being successful, exhibit limited generality and often fall short when compared to specialized models. Recently, LLM-based agents have been developed to address these challenges by selecting appropriate specialized models as tools based on user inputs. However, such advancements have not been extensively explored within the medical domain. To bridge this gap, this paper introduces the first agent explicitly designed for the medical field, named **M**ulti-modal **Med**ical **Agent** (MMedAgent). We curate an instruction-tuning dataset comprising six medical tools solving seven tasks across five modalities, enabling the agent to choose the most suitable tools for a given task. Comprehensive experiments demonstrate that MMedAgent achieves superior performance across a variety of medical tasks compared to state-of-the-art open-source methods and even the closed-source model, GPT-4o. Furthermore, MMedAgent exhibits efficiency in updating and integrating new medical tools.

1 Introduction

Multi-modal Large Language Models (MLLMs) have made considerable progress across diverse tasks with inputs from different medical imaging modalities (*e.g.*, Magnetic Resonance Imaging, Computed Tomography, X-ray) in healthcare, including Visual Question Answering (VQA) (Moor et al., 2023a; Zhang et al., 2023a; Li et al., 2023), image segmentation (Ma et al., 2024a), and Medical Report Generation (MRG) (Thawkar et al., 2023; Hamamci et al., 2024), etc. Despite these advancements, MLLMs often exhibit limitations in seamlessly solving multiple tasks across different medical imaging modalities. Although recent large medical models (Zhang et al., 2023b; Tu et al., 2024; Wu et al., 2023; Yang et al., 2024; Zhao et al., 2024a) have attempted to address this challenge, they remain limited to handling a narrow range of tasks across a restricted set of imaging modalities and cannot be efficiently extended to new tasks or more imaging modalities. Furthermore, these generalists typically do not provide expert-level responses comparable to those of specialized MLLMs customized for specific tasks.

One way to address this issue is to build an AI Agent, an AI system driven by Large Language Models (LLMs) that integrates various domain expert models as tools. Such a system can understand user instructions, make decisions, and select the appropriate tools to execute any specific task, thereby generating expert-level responses for any given request (Xie et al., 2024; Chen et al., 2023; Wang et al., 2024; Liu et al., 2023b; Tao et al., 2023). Despite the significant success of AI agents in the general image domain (Tao et al., 2023; Qin et al., 2023; Wang et al., 2023a), there are currently few AI agents developed specifically for the medical domain. Although several works (Tang et al., 2023; Schmidgall et al., 2024; Li et al., 2024; Fan et al., 2024) in the medical field use the term "agent" in their methods, they focus on utilizing LLMs to play various roles and collaborate on complex tasks, in which an "agent" refers to a specific role. PathAsst (Sun et al., 2024) integrates tool utilization into their framework but specifically designed for pathology tasks.

In this work, we aim to build the first AI agent specifically for the medical domain, termed as **M**ulti-modal **Med**ical **Agent** (MMedAgent). We choose LLaVA-Med (Li et al., 2023) as the backbone and aim to extend its capability to handle various language and multi-modal tasks, including grounding, segmentation, classification, MRG, and Retrieval-Augmented Generation (RAG). These tasks encompass multiple medical imaging modalities, such as MRI, CT, and X-ray, allowing

*Corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8745–8760
November 12-16, 2024

MMedAgent to support a wide range of data types typically encountered in clinical practice. The first step to building MMedAgent is to collect the state-of-the-art (SOTA) methods for each task, hereafter referred to as "tools". During this phase, we identify a lack of an effective tool for the grounding task, prompting us to fine-tune Grounding DINO (Liu et al., 2023c) specifically for medical applications. Next, we build an instruction-based dataset that teaches the agent to select the proper tool(s) when encountering a user instruction and aggregate the outputs from tools to reply to users precisely and comprehensively. The core of our approach involves an end-to-end training regimen through visual instruction tuning (Liu et al., 2023b). MMedAgent has demonstrated promising results in various aspects. When evaluated on several complex medical tasks, MMedAgent significantly outperforms sevaral open-source SOTA methods, including LLaVA (Liu et al., 2023a), Flamingo-Med (Moor et al., 2023b), Yi-VL-34B (AI et al., 2024), Qwen-VL-Chat (Bai et al., 2023), LLaVA-Med (Li et al., 2023) and RadFM (Wu et al., 2023), and even surpasses close-source method, GPT-4o (OpenAI, 2024), on average. It also enhances MMedAgent's backbone, *i.e.*, LLaVA-Med, original capability in the VQA task, as well as exhibits efficient capability in learning new tools. Our code and web UI are available at `https://github.com/Wangyixinxin/MMedAgent`, and a demonstration of the user interface is provided in Appendix Figure 6.

Our contributions can be summarized as:

- We propose MMedAgent, the first multi-modal medical AI Agent incorporating a wide spectrum of tools to handle various medical tasks across different modalities seamlessly.

- We build the first open-source instruction tuning dataset for general-purpose multi-modal medical agents.

- Adaptive multi-modal medical tools are incorporated into our Agent. We develop specialized datasets to adapt existing grounding and segmentation tools to the medical domain.

- Extensive experiments demonstrate that MMedAgent surpasses previous SOTA multi-modal medical language models across a range of tasks.

2 Related Work

2.1 Medical MLLMs

LLMs present fertile new ground for research that pushes the frontier of the medical domain. Unlike natural domains, the intrinsic complexity of medical data, which includes multiple sources and modalities, has led most LLMs in the medical field to focus on narrowly defined tasks using language and text alone. Singhal et al. (Singhal et al., 2023) curate MultiMedQA, a benchmark for medical question-answering datasets, and propose Med-PaLM, which utilizes instruction prompt tuning tailored to medical domains based on PaLM (Chowdhery et al., 2023). Med-PaLM performs encouragingly on the axes of the human evaluation framework.

Recent progress on LLMs has been made on multi-modal conversational capability (Moor et al., 2023a; Zhang et al., 2023b; Tu et al., 2024; Zhang et al., 2023c,a; Thawkar et al., 2023; Sun et al., 2024; Wu et al., 2023; Li et al., 2023; Ma et al., 2024a; Yang et al., 2024; Zhao et al., 2024a; Hamamci et al., 2024). Owing to the diversity inherent in medical data and tasks, LLMs have initially been localized to specific imaging domains such as X-ray (Thawkar et al., 2023), CT (Hamamci et al., 2024), and histology (Sun et al., 2024), or tailored for different tasks such as segmentation (Ma et al., 2024a; Lei et al., 2023) and medical report generation (Wu et al., 2023). In contrast, generalist models expand these capabilities by enabling a single LLM to cover a wider range of imaging modalities and tasks by enlarging the pre-training datasets greatly (Zhang et al., 2023b; Li et al., 2023; Zhao et al., 2024a; Liu et al., 2023b; Yang et al., 2024). Although generalist models are capable of handling a wide range of medical modalities and tasks, they face limitations in scalability when incorporating additional skills and lack specialization in specific tasks.

2.2 AI Agent

A multi-modal AI Agent is a system that achieves users' general-purpose goals by perceiving the environment and making decisions based on the perceptions (Xie et al., 2024; Wooldridge and Jennings, 1995). Recent works utilize LLMs as planners to understand multi-modal input from environments and make decisions to call different tools to achieve goals. Based on whether the LLM is open source or not, (Xie et al., 2024) classifies multi-modal

AI Agents into two types: (i) closed-source LLMs as planners, which utilize prompt technique to enable LLMs to make decisions (Chen et al., 2023; Wang et al., 2024); (ii) fine-tuned LLMs as planners, where an LLM is fine-tuned to understand instructions, make decisions, and call tools/APIs (Liu et al., 2023b; Tao et al., 2023; Zhang et al., 2024b). MMedAgent belongs to the second type.

Multi-modal AI Agents have achieved great success in various applications. For example, (Tao et al., 2023; Gur et al., 2023; Zhan and Zhang, 2023) apply agents to control the website or user interface. Some works (Qin et al., 2023; Wang et al., 2023c) focus on robotics or embodied AI which applies multi-modal LLMs to perceive and interact with real environments. Most works concentrate on multi-modal understanding, or generation, especially image, video, or audio (Liu et al., 2023b; Wang et al., 2023a; Zhang et al., 2023e). However, these works are limited to the natural domains. To the best of our knowledge, we are the first to build a more versatile medical AI Agent, which encompasses a broader spectrum of image modalities, including MRI, CT, X-ray, and histology.

3 MMedAgent

Multi-modal Medical Agent (MMedAgent), a system based on an MLLM, is designed to seamlessly manage diverse medical tasks by integrating various open-source medical models. MMedAgent comprises two components: (1) an instruction-tuned multi-modal LLM that functions as an action planner and results aggregator, and (2) a collection of medical tools tailored to the agent, each targeting specific tasks in the medical domain. We first present the fundamental workflow of MMedAgent in Section 3.1, followed by a description of creating an instruction-tuning dataset for training the multi-modal LLM as an action planner in Section 3.2. The details of medical tasks and corresponding tools incorporated in MMedAgent are described in Section 3.3.

3.1 Workflow

Following LLaVA-Plus (Liu et al., 2023b), the objective of MMedAgent is to learn to utilize a wide range of multi-modal medical tools, extending the MLLMs' capabilities to analyze and accomplish various medical tasks. As shown in Figure 1, the workflow consists of four parts: (1) users provide an instruction X_q and a medical image I_q; (2) MLLM works as an action planner, which understands X_q and I_q and then generates a formatted instruction X_{tool} to call a specific tool. (3) The tool is executed given I_q and the output X_{result} of the tool is sent to the MLLM. (4) The MLLM aggregates the output with X_q and I_q and generates the final answer X_{answer} to users. We train the agent end-to-end with an auto-regressive objective on the generated sequence, *i.e.,* X_{tool} and X_{answer}, to enable the model to use correct tools and answer questions based on the tool's results.

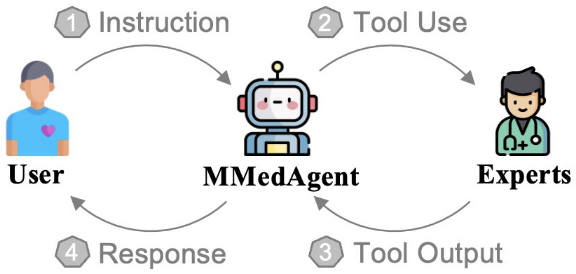

Figure 1: The four-step MMedAgent pipeline.

3.2 Instruction Tuning

In order to ensure MMedAgent simultaneously performs as both action planner and results aggregator, we adopt the unified dialogue format proposed by (Liu et al., 2023b), illustrated in Figure 2. Specifically, upon receiving a user's input, MMedAgent generates three components in its outputs: (1) `Thoughts`, which determine whether MMedAgent can independently solve the user's instructions or if external tools are required, and if so, identifies the appropriate tool; (2) `Actions`, which enumerate a list of API calls necessary to execute the `thought`. This comprises two sub-fields: `API Name` and `API Params`. If the action list is null, no API call is initiated. (3) `Value`, which provides a natural language response from the MLLM. During the first round, it indicates the chosen tool(s); during the second round, it represents the final output that addresses the user's initial request. As depicted in Appendix Figure 5, we construct the instruction data by querying GPT-4o through one-shot learning, presenting an example that demonstrates the input and output of MMedAgent. We set a fixed `System` instruction prompt for each tool and select several examples as conversation templates (`User_1` and `Assistant_1` in Appendix Figure 5). The tool processes the generation of instruction data from the dialogue.

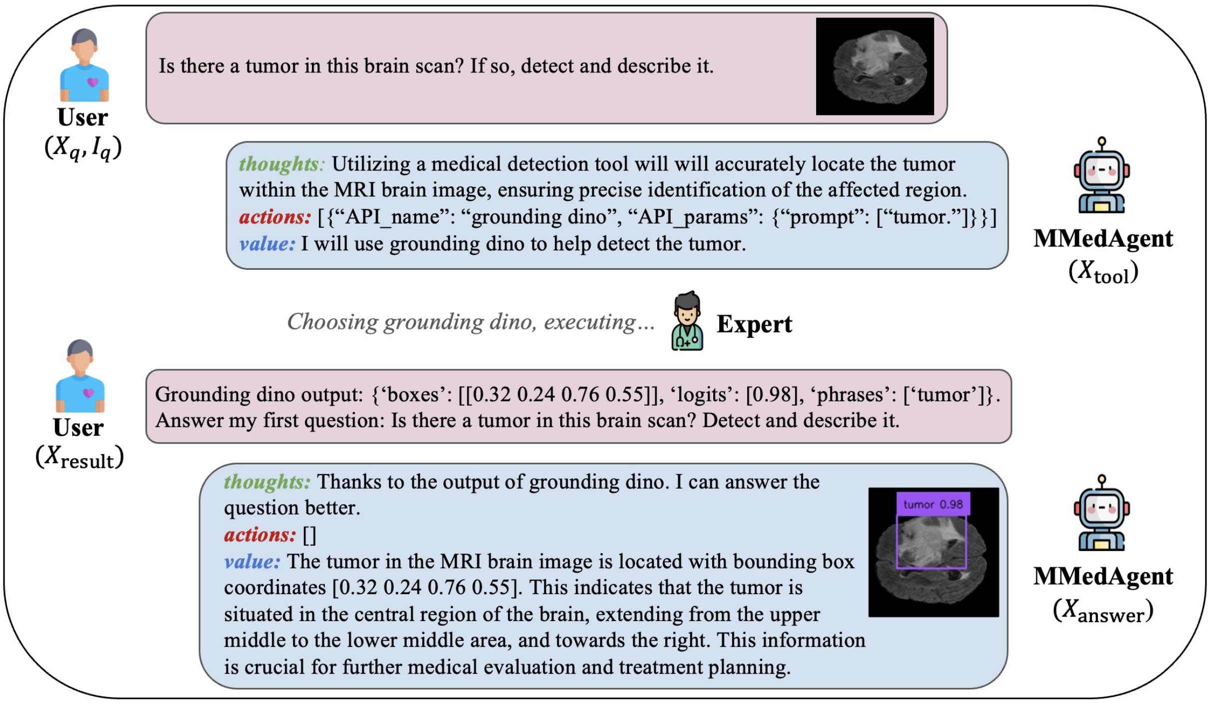

Figure 2: An example of the training data for MMedAgent that learns to use the tool of Grounding DINO for object detection and answer the user's question.

3.3 Medical Tasks and Tools

Our MMedAgent possesses the capability to access a diverse array of tools with the scalability to handle various tasks. As shown in Table 1, we integrate six tools that encompass seven representative tasks in medical domains, *i.e.*, (1) grounding, (2) segmentation with bounding-box prompts (B-Seg), (3) segmentation with text prompts (G-Seg), (4) medical imaging classification, (5) Medical Report Generation (MRG), (6) retrieval augmented generation (RAG), and (7) VQA. Note that no additional tools are required for the VQA task since we utilize LLaVA-Med, which originally supports the task, as the backbone. Each tool functions as a specialist, exhibiting exceptional proficiency in executing a specific task across various medical imaging modalities.

3.3.1 Grounding

Grounding, also known as detection, aims to identify and localize specific objects within an input image by generating the coordinates of bounding boxes containing the objects. To the best of our knowledge, no existing medical models can simultaneously process images from different modalities. Consequently, we propose a generalized grounding tool tailored for the medical domain. Specifically, we choose to fine-tune Grounding DINO (Liu et al., 2023c), an open-set object detector, to the medical imaging field.

Our first step is to collect multiple medical image segmentation datasets, including FLARE2021 (Ma et al., 2022), WORD (Luo et al., 2022), BRATS (Menze et al., 2015), Montgomery County X-ray Set (MC) (Jaeger et al., 2014; Candemir et al., 2014), VinDr-CXR (Nguyen et al., 2022), and multi-modal cell segmentation dataset (Cellseg) (Ma et al., 2024b). As detailed in Appendix Table 5, these datasets target different modalities, organs, or diseases, each including the original imaging along with their corresponding pixel-level segmentation annotations. These segmentation masks are further transformed into bounding boxes by extracting the minimal outer rectangle around each object. The coordinates of the bounding boxes and the corresponding object labels are then recorded as the grounding labels in each dataset.

Based on the released pre-trained weights, we fine-tuned the Grounding DINO with the dataset described above as well as two common datasets in the natural image field, *i.e.*, COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015), to maintain model's ability in detecting common objects.

Task	Tool	Data Source	Imaging Modality
VQA	LLaVA-Med (Li et al., 2023)	PMC article *60K-IM*(Li et al., 2023)	MRI, CT, X-ray, Histology, Gross
Classification	BiomedCLIP (Zhang et al., 2024a)	PMC article *60K-IM*	MRI, CT, X-ray, Histology, Gross
Grounding	Grounding DINO (Liu et al., 2023c)	WORD, etc.*	MRI, CT, X-ray, Histology
Segmentation	MedSAM (Ma et al., 2024a)	WORD, etc.*	MRI, CT, X-ray, Histology, Gross
G-Seg.	Grounding DINO + MedSAM	WORD, etc.*	MRI, CT, X-ray, Histology
MRG	ChatCAD (Wang et al., 2023b)	MIMIC-CXR (Johnson et al., 2019)	X-ray
RAG	ChatCAD+ (Zhao et al., 2024b)	Merck Manual (Porter and Kaplan, 2011)	–

Table 1: The tasks, tools, data source, and corresponding medical imaging modalities incorporated in MMedAgent. "–" means that the RAG task only focuses on natural language without handling images. "WORD, etc.*" indicates various data sources including WORD (Luo et al., 2022), FLARE2021 (Ma et al., 2022), BRATS (Menze et al., 2015), Montgomery County X-ray Set (MC) (Jaeger et al., 2014; Candemir et al., 2014), VinDr-CXR (Nguyen et al., 2022), and Cellseg (Ma et al., 2024b).

3.3.2 Other Tasks

B-Seg involves identifying and delineating the region of interest (ROIs) of an image when a bounding box that covers the ROIs is provided. It is a type of interactive segmentation, which has become popular since the development of Segment Anything (SAM) (Kirillov et al., 2023). We select MedSAM (Ma et al., 2024a), which fine-tunes SAM to the medical field, as our tool. The prompts are limited to bounding boxes because they provide more precise guidance to SAM (Mazurowski et al., 2023). Specifically, in this scenario, we consider the users to provide the position of the bounding box in which MedSAM can be directly applied to obtain the ROI masks.

G-Seg refers to combining grounding with SAM. It aims to address a more common scenario when users specify only a particular object to segment in an image. We first activate the fine-tuned grounding tool to localize the referred object and then provide its location, in box format, to MedSAM. Note that this task also presents MMedAgent's capability to address complex tasks with more than one tools, an ability not presented in previous works.

Classification aims to identify the most appropriate category for a medical image within a closed set. Specifically, we define a closed set of labels L, including organ types, common image modalities, and complex modalities such as ultrasound imaging, hematoxylin, and eosin histopathology. The details of the set L are shown in Appendix A.1. We adopt BiomedCLIP (Zhang et al., 2024a), which exhibits superior performance in zero-shot and fine-grained classification. The image is classified based on the cosine similarity between the image embedding and each text embedding.

MRG involves creating accurate and authentic medical reports from provided medical information or imaging. MMedAgent incorporates ChatCAD (Wang et al., 2023b), an open-source tool designed for generating medical reports for chest X-ray images. The model was trained on the MIMIC-CXR dataset (Johnson et al., 2019) and can provide reports with detailed radiographic analyses, identifying chest-related conditions such as cardiomegaly, edema, consolidation, atelectasis, etc.

RAG refers to enhancing the generated outputs by incorporating the most relevant information acquired from external data sources. We select ChatCAD+ (Zhao et al., 2024b) to implement the medical retrieval process. ChatCAD+ retrieves information from a medical dictionary containing detailed descriptions of 1972 diseases and medical procedures, including their introduction, symptoms, diagnosis, treatment, and causes, sourced from the Merck Manual (Porter and Kaplan, 2011), a professional medical reference. Given the users' input, the model searches for medical entrees that share the highest cosine similarity with the encoded message and retrieves the relevant knowledge from the medical dictionary.

	Grounding			Cls.	MRG	RAG	Overall	Abs.
	Cell	Organ	Disease					
Flamingo-Med (Moor et al., 2023b)	13.11	15.87	15.33	23.56	16.59	-	14.68	1.16
RadFM (Wu et al., 2023)	-	-	-	25.00	68.13	-	45.38	3.59
LLaVA-Med (Li et al., 2023)	51.78	65.48	68.58	53.46	70.10	30.44	60.68	4.80
Yi-VL-34B (AI et al., 2024)	63.23	79.40	68.32	76.02	72.95	14.67	64.08	5.07
LLaVA-Med (Tool in Test)	45.32	52.77	67.91	57.53	74.34	67.55	65.31	5.17
Qwen-VL-Chat (Bai et al., 2023)	61.34	65.90	62.38	88.40	73.41	78.80	76.21	6.03
LLaVA-34B (Liu et al., 2023a)	76.75	84.85	80.75	**96.04**	80.27	**91.64**	86.52	6.84
MMedAgent (ours)	**97.50**	**102.29**	**125.89**	81.11	**121.49**	85.55	**109.48**	**8.66**

Table 2: Performance comparison between MMedAgent and other baselines. Cls. stands for classification. We report the relative scores for all tasks and the absolute (abs.) scores for overall performance in the last column. "-" indicates the tasks that the corresponding model is not applicable to. LLaVA-Med refers to the *60K-IM* version with only the initial query X_q and image I_q as input, while LLaVA-Med (Tool in Test) takes X_q, I_q and also the internal output from tools X_{result} as input.

4 Experimental Settings

MMedAgent is initialized with LLaVA-Med *60K-IM*, instruction-tuned using LoRA (Hu et al., 2021) for 15 epochs, and conducted over approximately 72 hours on two 80G NVIDIA A100 GPUs. The rank of LoRA is set to 128, and the training batch size is set to 48. We employ AdamW (Loshchilov and Hutter, 2019) as the optimizer alongside a cosine learning rate schedule peaking at 2e-4. We generate 48K instruction-tuning data, consisting of 15K augmented VQA instruction following the method from LLaVA-Plus (Liu et al., 2023b) derived from 60K inline mentions (Li et al., 2023), 10K data points for detection, 3K for RAG, 5K each for segmentation, classification, MRG, and G-Seg. Data sources are shown in Table 1.

5 Experimentals

We conduct experiments on MMedAgent to answer three research questions: (1) What is the performance of MMedAgent in addressing diverse medical tasks across various modalities (Section 5.1)? (2) Does the instruction-tuned MMedAgent exhibit superior performance in open-ended biomedical dialogue (Section 5.2)? (3) What is the efficiency of MMedAgent in invoking tools or incorporating new tools (Section 5.3)?

5.1 Various Medical Tasks

5.1.1 Evaluation Criterion

To evaluate the performance of MMedAgent on various complex medical tasks, we create an evaluation dataset consisting of 70 diverse questions. For this dataset, we initially select 10 concepts randomly from the Merck Manual for RAG and 60 unseen images of different tasks from respective data sources. These include 10 images each for organ grounding, disease grounding, and cell grounding, along with 20 X-ray images for MRG and 10 images across various modalities for classification. Notably, the VQA task evaluation is shown in Section 5.2. Due to the inability to describe the segmentation task linguistically, we provide the qualitative results shown in Section 5.1.3. Then we utilize the same prompt as outlined in Section 3.2 to generate the instruction-tuning data for evaluation. Subsequently, we separately feed the data into GPT-4o, MMedAgent and other benchmarks to obtain the outputs. GPT-4o is a newly released multimodal model with strong visual understanding capabilities. According to the testing from OpenAI, it surpasses GPT-4 Turbo and has a faster inference speed. Thus, the output from GPT-4o can be viewed as a strong benchmark. All the outputs will be assessed by GPT-4 and rated on a scale from 1 to 10 based on their helpfulness, relevance, accuracy, and level of details. We provide GPT-4 with figure captions and include inline mentions from *60K-IM* for the VQA task. The detailed prompts are illustrated in Figure 7. For the MRG task, the reports are taken as captions of the input figures. For detection and other tasks without a caption in the original data, we generate the captions by combining the images with the labels, *e.g.,* "A CT scan showing the kidney organ.". Based on the output from GPT-4o, we propose an absolute score, *i.e.,* the score output by GPT-4, and a relative score, defined as $S_*/S_{GPT-4o}(\%)$, which indicates the performance change caused by other MLLMs. Here, S_* refers to the score of outputs generated by $*$, with $* \in$ {MMedAgent, LLaVA-Med,...}. Specifically, we compare with two MLLMs in the medical field, i.e., Med-Flamingo (Moor et al., 2023b) and RadFM (Wu et al., 2023), as well as three generic MLLM, LLaVA-34B (Liu et al., 2023a), Yi-VL-34B (AI et al., 2024) and Qwen-VL-Chat (Bai et al., 2023). A higher score indicates a superior output quality. During the evaluation, MMedAgent dynamically selects, activates, and executes tools in real-time, then aggregates the obtained results from these tools to answer questions.

5.1.2 Experimental Results

As illustrated in Table 2, MMedAgent significantly outperforms all other baselines on various tasks. Note that RadFM cannot handle grounding and RAG tasks, and Flamingo-Med is not applicable for RAG because it cannot process text-only input. It is observed that the overall relative score of MMedAgent (109.48) outperforms all other state-of-the-art MLLMs by a large margin, being 1.8 times higher than that of LLaVA-Med (60.68), which is the backbone of MMedAgent. We also propose LLaVA-Med (Tool in Test), an enhanced version of LLaVA-Med that incorporates the internal output of tools and MMedAgent maintains its superior performance in this case.

Furthermore, the scores for organ grounding, disease grounding, and MRG exceed 100%, indicating that MMedAgent surpasses GPT-4o in these tasks. These results underscore the superior efficiency of MMedAgent in diverse medical tasks across various modalities.

5.1.3 Case Study

A detailed visual comparison between LLaVA-Med and MMedAgent is illustrated in Figure 3. Given the user queries on tasks involving analyzing the images, such as classification, grounding, and segmentation tasks, LLaVA-Med only generates simple conversational responses without solving the given requests (highlighted in Red) and it is unable to generate visualized results. In contrast, MMedAgent effectively addresses these questions by activating the appropriate tools, integrating their outputs, generating accurate responses (highlighted in Green), and visualizing the results. This is guaranteed by the precise selection of tools by MMedAgent and the superiority of the tools themselves. When encountering language generation-based tasks, *i.e.*, MRG and RAG, LLaVA-Med fails to provide an in-depth analysis of the images. However, MMedAgent provides more straightforward and accurate responses by utilizing the tools designed specifically for these tasks.

5.2 Medical VQA

When implementing VQA tasks, MMedAgent can rely on its backbone, i.e., LLaVA-Med (Sections 5.2.1) and could be further enhanced by leveraging other VQA specialists (5.2.2).

5.2.1 Open-ended Medical Dialogue

We follow the setting of open-ended medical dialogue in LLaVA-Med (Li et al., 2023) and use the same test data as LLaVA-Med, consisting of 193 novel questions and 50 unseen images from PMC-15M (Zhang et al., 2024a). This dataset contains 5 modalities and can be divided into two main classes: conversation questions and detailed description questions. We utilize the same relative score in Section 5.1.1 as the evaluation criterion. Since this is a pure language task, we select the output from GPT-4 rather than GPT-4o as the reference score.

As shown in Table 3, performance is evaluated from the perspective of question types (conversation and description) and image modalities (X-ray, MRI, Histology, Gross and CT). After instruction-tuning on the tool learning dataset, MMedAgent performs better on both types of questions. Moreover, MMedAgent outperforms LLaVA-Med in all domains but MRI, demonstrating the efficiency of MMedAgent in open-ended medical dialogue.

5.2.2 VQA Benchmark

MMedAgent is also evaluated on four VQA benchmarks, including VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PMC-VQA (Zhang et al., 2023d) and PATH-VQA (He et al., 2020). We construct the instruction-tuning data for each dataset and fine-tune the model with these newly added VQA tools. For each dataset, We randomly select 4K image-text pairs as the training set and 300 pairs for evaluation. Specifically, VQA-RAD, PATH-VQA, and SLAKE each contain 150 samples from the open set and 150 from the closed set, while PMC-VQA comprises 300 closed multi-choice questions. Figure 4 indicates that MMedAgent significantly outperforms LLaVA-Med across all VQA benchmarks.

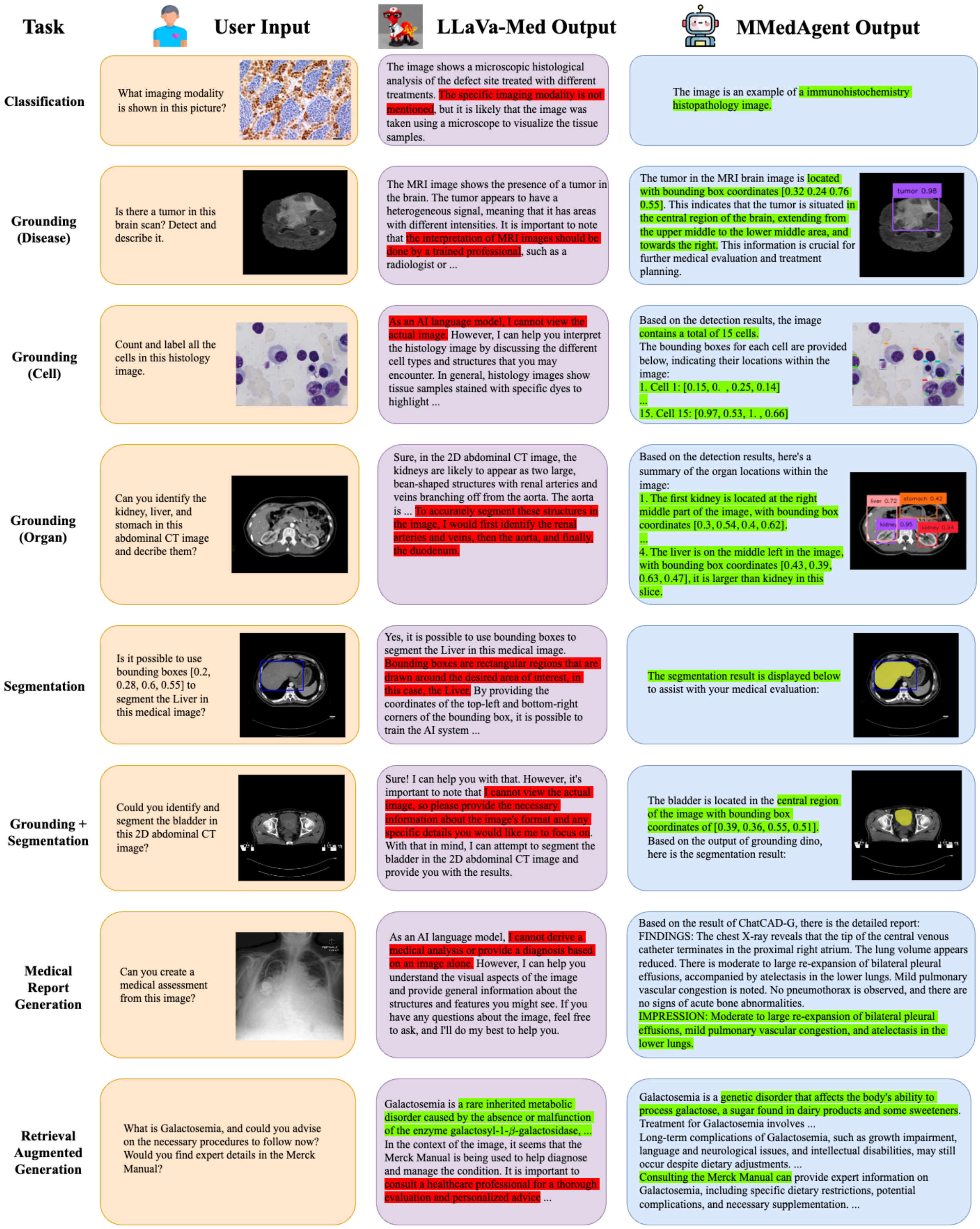

Figure 3: Qualitative comparison between LLaVA-Med and MMedAgent across different tasks. The undesired and desired responses are highlighted in Red and Green respectively.

	Question Types		Imaging Modalities					Overall
	Conversation	Description	X-ray	MRI	Histology	Gross	CT	
(Question Count)	(143)	(50)	(37)	(38)	(44)	(34)	(40)	(193)
LLaVA-Med	53.30	38.90	56.58	**40.84**	54.71	48.47	50.68	50.94
MMedAgent	**54.49**	**39.75**	**58.37**	35.09	**56.88**	**51.88**	**52.79**	**51.42**

Table 3: Comparison of open-ended medical dialogue between MMedAgent and LLaVA-Med.

	RAD-VQA		SLAKE		PATH-VQA		PMC-VQA
	Open	Close	Open	Close	Open	Close	Close
(Question Count)	(150)	(150)	(150)	(150)	(150)	(150)	(300)
LLaVA-Med	28.23	61.40	39.17	52.16	12.30	54.05	27.48
MMedAgent	**58.31**	**86.72**	**79.39**	**86.34**	**39.16**	**90.38**	**39.50**

Table 4: Comparison of VQA tasks between MMedAgent and LLaVA-Med across different VQA benchmarks.

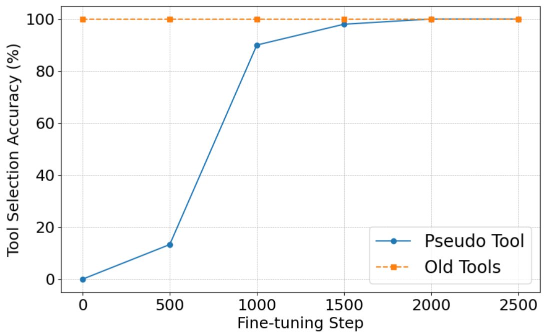

Figure 4: The scalability of MMedAgent.

5.3 Tool Utilization

The superior performance of MMedAgent on the various tasks described above depends on accurately understanding users' inputs and activating the correct tools. After training MMedAgent for 15 epochs, the tool selection accuracy reached 100%, demonstrating MMedAgent's ability to select the appropriate tools without errors.

One significance of MMedAgent is its ability to adapt to new tools. Here, we consider two scenarios. Firstly, when a superior tool for tasks that MMedAgent is already equipped to handle becomes available, the API name of the outdated tool can be seamlessly replaced with that of the new tool, eliminating the need for additional retraining. Secondly, to extend MMedAgent to a new task, it is sufficient to generate a small set of instruction-tuning data for this specific task and fine-tune the agent accordingly, rather than retraining it from the beginning. To verify this capability, we simulate a new tool called "Pseudo Tool", generate an additional 5K instruction-tuning data (following Section 4), and create 30 unseen diverse questions for evaluation following Section 5.1.1. We utilize the same training settings to fine-tune MMedAgent with a smaller learning rate of 1e-6 and a batch size of 10 on one 80G A100 GPU. As shown in Figure 4, the accuracy of selecting a new tool increase to 100% within 2K steps without damaging the performance on selecting old tools.

6 Conclusion

We propose MMedAgent, the first multi-modal medical AI agent that is capable of seamlessly utilizing various medical tools to handle a broad spectrum of medical tasks across different imaging modalities. We create an instruction-tuning dataset that MMedAgent utilize to learn to invoke various medical tools and aggregate results from tools. Comprehensive experiments demonstrate that MMedAgent significantly outperforms open-source baselines and even surpasses GPT-4o across many medical tasks. Furthermore, MMedAgent efficiently integrates with new tools while remaining the capability to activate previously learned tools.

7 Limitation

Our work is currently limited to seven tasks across five modalities. Due to the need for extensive domain knowledge, more specialized tools should be included and MMedAgent's scalability allows for the inclusion of more powerful tools in the future. Additionally, more generalist LLMs in the medical domain could potentially serve as stronger backbone to enhance MMedAgent.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

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

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. 2014. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590.

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer Antani, George Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and Clement J. McDonald. 2014. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

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

Wenhui Lei, Xu Wei, Xiaofan Zhang, Kang Li, and Shaoting Zhang. 2023. Medlsam: Localize and segment anything model for 3d ct images. *Preprint*, arXiv:2306.14752.

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. *Preprint*, arXiv:2306.00890.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. 13.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. 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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N. Metaxas, Guotai Wang, and Shaoting Zhang. 2022. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024a. Segment anything in medical images. *Nature Communications*, 15(1).

Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, Wei Lou, Haofeng Li, Eric Upschulte, Timo Dickscheid, José Guilherme de Almeida, Yixin Wang, Lin Han, Xin Yang, Marco Labagnara, Vojislav Gligorovski, Maxime Scheder, Sahand Jamal Rahi, Carly Kempster, Alice Pollitt, Leon Espinosa, Tâm Mignot, Jan Moritz Middeke, Jan-Niklas Eckardt, Wangkai Li, Zhaoyang Li, Xiaochen Cai, Bizhe Bai, Noah F. Greenwald, David Van Valen, Erin Weisbart, Beth A. Cimini, Trevor Cheung, Oscar Brück, Gary D. Bader, and Bo Wang. 2024b. The multi-modality cell segmentation challenge: Towards universal solutions. *Nature Methods*.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, Shuiping Gou, Franz Thaler, Christian Payer, Darko Štern, Edward G.A. Henderson, Dónal M. McSweeney, Andrew Green, Price Jackson, Lachlan McIntosh, Quoc-Cuong Nguyen, Abdul Qayyum, Pierre-Henri Conze, Ziyan Huang, Ziqi Zhou, Deng-Ping Fan, Huan Xiong, Guoqiang Dong, Qiongjie Zhu, Jian He, and Xiaoping Yang. 2022. Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. *Medical Image Analysis*, 82:102616.

Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. 2023. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. 2015. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

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

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023b. Med-flamingo: A multimodal medical few-shot learner. ArXiv:2307.15189.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. 2022. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-05-26.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. pages 2641–2649.

Robert S. Porter and Justin L. Kaplan. 2011. *The merck manual of diagnosis and therapy, 2011*. Merck Research Laboratories.

Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. 2023. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. *arXiv preprint arXiv:2312.07472*.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. 2024. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5034–5042.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Heyi Tao, Sethuraman TV, Michal Shlapentokh-Rothman, Derek Hoiem, and Heng Ji. 2023. Webwise: Web interface control and sequential exploration with large language models. *arXiv preprint arXiv:2310.16042*.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023a. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023b. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *Preprint*, arXiv:2302.07257.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. 2023c. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*.

Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Preprint*, arXiv:2308.02463.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. 2024. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Zhuosheng Zhan and Aston Zhang. 2023. You only look at screens: Multimodal chain-of-action agents. *arXiv preprint arXiv:2309.11436*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2024a. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *Preprint*, arXiv:2303.00915.

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

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. *arXiv preprint arXiv:2305.10415*.

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

Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. 2023e. Loop copilot: Conducting ai ensembles for music generation and iterative editing. *arXiv preprint arXiv:2310.12404*.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, et al. 2024b. Toolbehonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. *arXiv preprint arXiv:2406.20015*.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, et al. 2024a. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. 2024b. Chatcad+: Towards a universal and reliable interactive cad using llms. *IEEE Transactions on Medical Imaging*, page 1–1.

A Details of Tools

A.1 Classification

We construct a close set of labels L for Biomed-CLIP to search for the most suitable category for the given image.

L ={"adenocarcinoma histopathology", "brain MRI", "covid line chart", "squamous cell carcinoma histopathology", "immunohistochemistry histopathology", "bone X-ray", "chest X-ray", "pie chart", "ultrasound imaging", "hematoxylin and eosin histopathology", "gross"}.

A.2 Retrieval Augmented Generation (RAG)

RAG distinguishes itself from standard report generation by its access to an external knowledge base, such as Merck Manual. We consider the following three common uses of RAG. The instruction-tuning data are generated based on these functionalities.

1. **Chest X-ray image report analysis.** The chest X-ray image report analysis can function to analyze the report on medical images and provide an analysis including the potential diseases and their related retrieved knowledge and source.

2. **General medical report analysis.** The general medical report analysis can take a summarized report on common diseases and generate an analysis with medical advice such as treatments and precautions, together with a link to the retrieved source from the Merck Manual official website.

3. **General medical advice generation.** For general medical advice generation, the user can ask general questions about the diseases, and the model will retrieve and provide related information on them.

For the chest X-ray image report analysis, we generate 1000 chest X-ray reports from the MRG tool described in Section 3.3.2 as the report dataset. For the datasets of general medical report analysis and general medical advice generation, we utilize GPT-4o to generate 1000 medical reports and 1000 patient questions respectively about common diseases sampled from the entrees covered in the Merck Manual.

A.3 Medical Grounding DINO

The datasets used to fine-tune the medical grounding DINO is shown in Appendix Table 5.

B Instruction Tuning Dataset Generation

We represent our prompts for generating an instruction tuning dataset in Appendix Figure 5.

C Agent Serving

MMedAgent operates within the FastChat system , which consists of web UI that interact with users, model workers hosting the language model, and various tools. A controller coordinates the activities between the web UI and model workers. The entire system, including the 7B MMedAgent and all associated tools, can be run on an Nvidia A100 (80GB) GPU.

D Evaluation Prompt

We utilize GPT-4 to assess the answers generated by MMedAgent and other models with prompts shown in Appendix Figure 7.

System

You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture. Design a conversation between you and a person asking about this photo.

You will be given a medical image and one or multiple coordinates, These coordinates are in the form of bounding boxes indicating locations of either organs, disease region, or cells , represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1 as they are the coordinates of upper left corner and lower right corner of the bounding box, normalized by the size of original image.

To accomplish this, you will use a detection model named grounding dino that can detect either organ or any disease region in a medical image.

In the first question, the user ask you to detect some objects in an image. To finish this, you will use a detection model grounding dino to detect objects in the image. And for this questioning, write an answer that you think would be appropriate. It must include the <thoughts>, the <actions>, and <actions>. Your response's format should be:

Question:
In the first question, the user ask you to detect some objects in an image.

Answer:
<thoughts> your thoughts about the question and the answer, and your intended approach including the tool you choose.
<actions> [{'tool name': API_name, 'API_params': some parameters}].
<values> the final results..

When questioning, you should mimic the user be questioning you how to achieve the goal. The API name in this case should be \"grounding dino\". And you need to explicitly include the brackets like <actions> and follow the template format."

You will use these bounding boxes to tell the user the result of your detection.

Ask diverse questions and give corresponding answers. Only include questions that have definite answers. Do not ask any questions that cannot be answered confidently.

Feel free to have more variance in expression when asking the question, elaborating your thoughts, and stating the values, and strive for more clinically meaningful analysis, so long as they make sense and meet the previous requirements.

User_1

an x-ray chest scan showing abnormalities.
cardiomegaly: [0.39, 0.46, 0.79, 0.6]

Assistant_1

Question:
Can you locate the region of cardiomegaly in this image and confirm its presence?

Answer:
<thoughts> Using an open set medical detection model will help in identifying and segmenting the presence and region of cardiomegaly in the chest X-ray scan.
<actions> [{'API_name': 'grounding dino', 'API_params': {'image': '<imagepath>', 'prompts': ['cardiomegaly']}}]
<values>
Based on the grounding dino model output, the presence of cardiomegaly is indicated in the chest X-ray. The region of cardiomegaly is located with bounding box coordinates [0.39, 0.46, 0.79, 0.6]. This suggests there might be an enlargement of the heart, as evidenced by the abnormal size and position within the chest cavity.

User_2

The X-ray image shows a region of consolidation, infiltration.
consolidation: [0.22, 0.21, 0.49, 0.58]
infiltration: [0.22, 0.21, 0.5, 0.59]

Figure 5: Pipeline of generating instruction-tuning dataset for the grounding task.

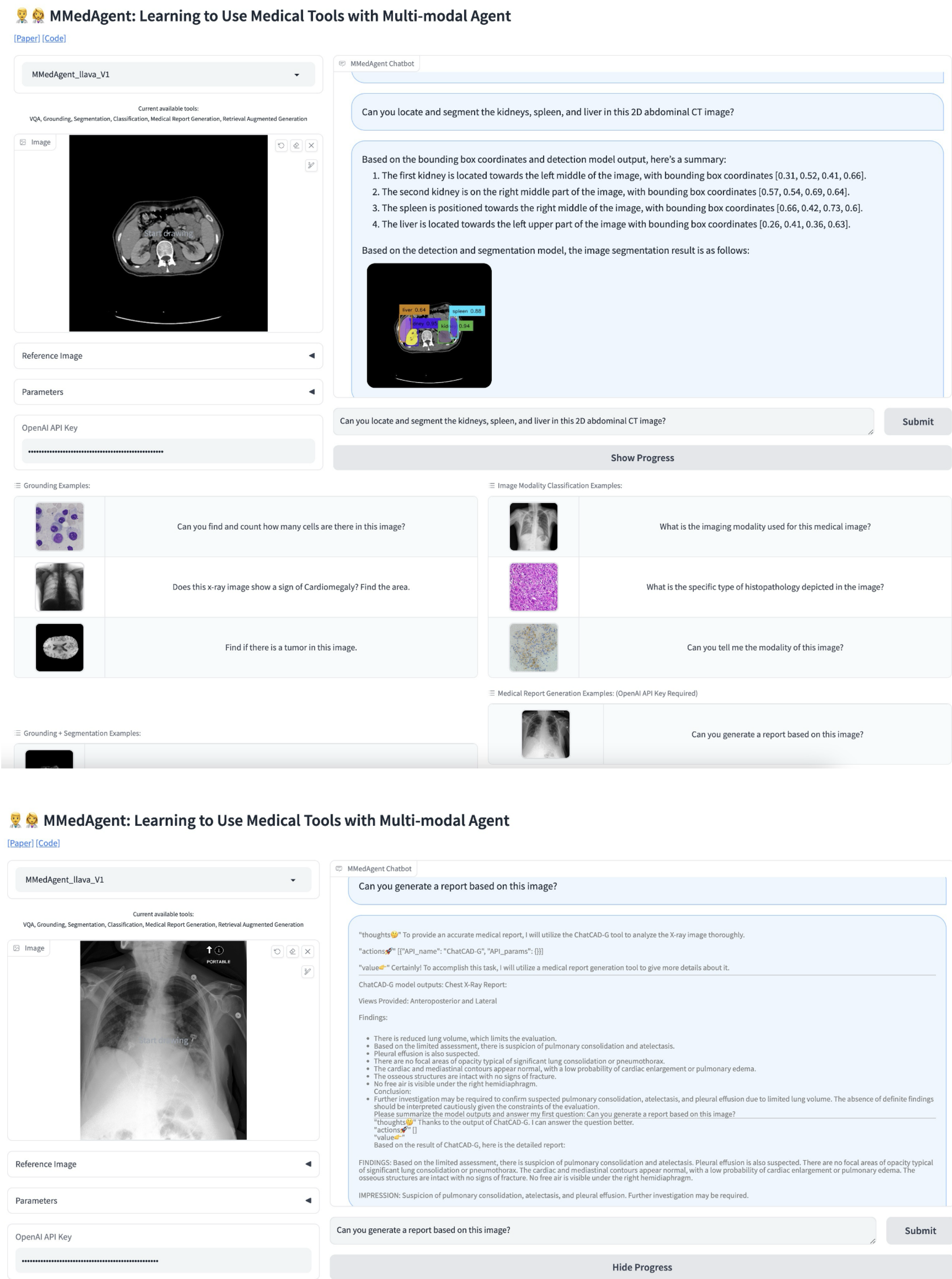

Figure 6: The user interface of the MMedAgent Web UI. Users can upload images and input questions in text, allowing MMedAgent to select the appropriate tool and provide comprehensive answers. The example shown demonstrates a request for segmenting organs in an abdominal CT image and generating diagnostic report for an X-ray image showing the thought progress of MMedAgent.

Dataset	Modality	Anatomy	Image Number	Labels
WORD	CT	Abdomen	9309	Liver, Spleen, Kidney, Stomach, Gallbladder, Esophagus, Pancreas, Duodenum, Colon, Intestine, Adrenal, Rectum, Bladder, Head of femur
FLARE	CT	Abdomen	4797	Liver, Kidney, Spleen, Pancreas, Aorta, IVC, Adrenal Gland, Gallbladder, Esophagus, Stomach, Duodenum
VinDr-CXR	X-ray	Chest	4394	Aortic enlargement, Atelectasis, Calcification, Cardiomegaly, Consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Other lesion, Pleural effusion, Pleural thickening, Pneumothorax, Pulmonary fibrosis
MC	X-ray	Chest	566	Lung
BRATS	MRI	Brain	14720	Tumor
Cellseg	Histology	Cell	229	Cell

Table 5: Dataset overview for fine-tuning Grounding DINO.

System

You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture.
You act as a helpful and precise assistant for checking the quality of the answer.

User

[Context]
Figure Caption:***{fig_caption}***
Figure Context:***{fig_inline_mention}***

[Question]
{question}

[Assistant 1]
{ans1}
[End of Assistant 1]

[Assistant 2]
{ans 2}
[End of Assistant 2]

[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing an image. For your reference, the visual content in the image is represented with caption describing the same image.
Please rate the helpfulness, relevance, accuracy, level of details of their responses.
Please note that both AI assistants are capable of analyzing medical images.
Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Figure 7: Evaluation pipeline. Assistant 1 is the model to be evaluated, which can be MMedAgent or LLaVA-Med and Assistant 2 is GPT-4o in our experiment.