Inquire, Interact, and Integrate: A Proactive Agent Collaborative Framework for Zero-Shot Multimodal Medical Reasoning

Anonymous ACL submission

Abstract

The adoption of large language models (LLMs) in healthcare has attracted significant research interest. However, their performance in healthcare remains under-investigated and potentially limited, due to i) they lack rich domain-specific knowledge and medical reasoning skills; and ii) most state-of-the-art LLMs are unimodal, text-only models that cannot directly process multimodal inputs. To this end, we propose a multimodal medical collaborative reasoning framework **MultiMedRes**, which incorporates a learner agent to proactively gain essential information from domain-specific expert models, to solve medical multimodal reasoning problems. Our method includes three steps: i) **Inquire**: The learner agent first decomposes given complex medical reasoning problems into multiple domain-specific sub-problems; ii) **Interact**: The agent then interacts with domain-specific expert models by repeating the "ask-answer" process to progressively obtain different domain-specific knowledge; iii) **Integrate**: The agent finally integrates all the acquired domain-specific knowledge to accurately address the medical reasoning problem. We validate the effectiveness of our method on the task of difference visual question answering for X-ray images. The experiments demonstrate that our zero-shot prediction achieves state-of-the-art performance, and even outperforms the fully supervised methods. Besides, our approach can be incorporated into various LLMs and multimodal LLMs to significantly boost their performance.[1]

1 Introduction

In the domain of medical informatics, the integration of deep learning techniques with extensive hospital database resources has been rapidly evolving these years, particularly in the multimodal analysis of chest X-ray images (Wang et al., 2017; Johnson et al., 2019; Irvin et al., 2019). This effort has attracted the attention of researchers working on various medical multimodal reasoning tasks, including the automatic generation of radiological reports (Biswal et al., 2020; Yin et al., 2019; Liu et al., 2021a,b), and answering pre-defined medical inquiries (Zhang et al., 2023; Chen et al., 2022). In particular, the task of difference visual question answering (DVQA) in radiology (Hu et al., 2023) addresses the challenge of analyzing sequential images of the same patient taken at different times. This requires the model to answer comparative medical questions that involve assessing changes between images, for example, 'What has changed compared to the previous image?'. Hence, to accurately address these questions, models must not only understand each image effectively and identify abnormalities to assist in diagnosis but also precisely describe the differences between the input images, specifically, the progression of disease. Thus, DVQA in radiology is emphasized as a critical task that mirrors the real-life diagnostic processes closely: clinicians routinely compare sequential X-ray images of the same patient to monitor disease progression.

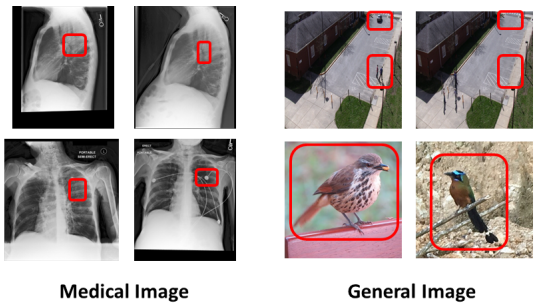


Figure 1: Medical image comparison vs. general image comparison.

Answering differences in medical images inherently poses more complexity, especially compared to general images. As illustrated in Figure 1, typ-

[1]The codes and models are available at https://anonymous.4open.science/r/multiMedAgent-7E57

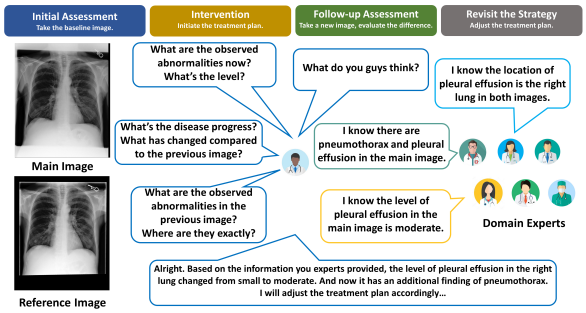


Figure 2: Illustration of the expert consultation.

ical DVQA tasks for general images exhibit two characteristics: different images i) may display pronounced differences in their main objects, such as in the case of two distinct birds (Forbes et al., 2019), or ii) may share the same, usually fixed, viewpoints as those from a video camera, featuring significant changes in salient content (Jhamtani and Berg-Kirkpatrick, 2018). Consequently, clear differences in visual features between images are present, enabling the model to effectively capture their differences. However, in the context of medical imaging, on the one hand, the main objects (e.g., the abnormalities) remain identical, with sought-after differences residing in subtle details (e.g., the severity and the size of the abnormalities). Moreover, it is the identical regions that predominantly feature across images, rather than the differing ones. As shown in Figure 1, the notable differences, highlighted by red bounding boxes, constitute only a minor portion of the overall images. This leads to the visual differences between medical images being relatively understated, thereby hindering the model's ability to discern differences through direct feature comparison. Instead, a model must employ a higher, more abstract level of understanding and comparison to accurately identify changes (Hu et al., 2023). On the other hand, X-ray images of the same patient, even when taken in the same body position, can vary in viewpoints and scale, further making discrimination more difficult. Therefore, to answer differences between X-ray images, a model necessitates a comprehensive understanding of varied domain-specific knowledge to accurately identify a wide range of clinical findings, which may present more significant differences than those discernible through visual features alone. Furthermore, the model must also be capable of recognizing the subtle yet critical differences within the main objects (e.g., disease progress), which are pivotal for effective disease monitoring.

To effectively capture domain-specific knowledge and the subtle yet critical differences between chest X-ray images, we propose a multimodal medical collaborative reasoning framework (MultiMedRes), which imitates the clinicians' working patterns. As shown in Figure 2, the common practice of a patient treatment process (Delrue et al., 2011) typically commences with the acquisition of a baseline image to serve as diagnostic evidence and the foundation for initiating a treatment plan. Subsequently, a follow-up image is obtained to monitor and assess the intervention's efficacy throughout the treatment period. This process often involves consultations with multiple experts across various domains, which may lead to modifications of the initial treatment plan. To model the above working patterns, MultiMedRes introduces three steps led by a learner agent: i) **Inquire**: The learner agent aims to understand the broad challenging difference questions and ask multiple types of domain-specific sub-questions that target specific aspects of the images, e.g., detecting abnormalities, identifying the level of severity and the location of the disease; ii) **Interact**: The learner agent feeds the sub-questions into expert models (i.e., specialists) to obtain their answers. These expert models are pre-trained on the domain-specific task and data to store rich specific knowledge for each type of domain-specific question. Our method will then raise new questions based on the given answers. Through repeating the "ask-answer" interaction process, our agent can progressively obtain sufficient knowledge for different domain-specific questions from the expert models; and iii) **Integrate**: The learner agent finally integrates all knowledge from the domain-specific specialists to address the input difference question accurately. These three steps can also provide readily interpretable for human radiologists.

We conduct the experiments on the benchmark MIMIC-Diff-VQA (Hu et al., 2023) dataset. Experiments demonstrate that our approach can achieve state-of-the-art performances. Moreover, MultiMedRes can be incorporated into diverse LLMs and multimodal LLMs to significantly boost their performance. The contributions of our study are outlined as follows:

- We introduce a collaborative reasoning framework, MultiMedRes, which enables an LLM learner agent to acquire essential domain-specific knowledge from specialized expert models to perform zero-shot medical multimodal reasoning.
- The zero-shot prediction provided by our frame-

work achieves state-of-the-art performance, and even outperforms the fully supervised methods.

- Experiments demonstrate that our framework is compatible with various LLMs, including both text-based LLMs and multimodal LLMs.

2 Related Work

Currently, multi-modal LLMs have also demonstrated promising results in tasks related to medical imaging (Lee et al., 2023; Xu et al., 2023; Thawakar et al., 2023), including multimodal reasoning. There have been explorations leveraging the extensive knowledge base and advanced reasoning capabilities of large language models (LLM) (Lu et al., 2022; Gui et al., 2022; Yang et al., 2021). While these large-scale models indeed shed light on tackling multiple multimodal tasks in the medical domain all at once, they have not surpassed traditional state-of-the-art models for multimodal reasoning on widely recognized datasets (Chen et al., 2024). Besides, such large-scale models all necessitate extensive training or fine-tuning of enormous amounts of parameters and are not tailored for low-resource settings (e.g., rare diseases and novel diseases), where the labeled data for training is limited. In this work, we propose a collaborative agent framework to enable the LLMs to accurately perform zero-shot multimodal reasoning.

Recently, with the growing research interest in LLMs, the concept of LLM-driven collaborative multi-agent systems has garnered significant attention (Wang et al., 2023; Park et al., 2023). LLM agents have also shown considerable promise in managing complex tasks in medical imaging (Allioui et al., 2021; Bennai et al., 2020; Allioui et al., 2022). Nevertheless, to the best of our knowledge, research on leveraging this paradigm for medical multimodal reasoning remains sparse.

3 Methods

In this section, we introduce MultiMedRes, a novel collaborative framework for medical reasoning tasks. Initially, we train a cohort of domain expert models, each tailored to address a distinct category of questions (e.g., abnormality detection and localization), to serve as domain-specific experts. Subsequently, an LLM learner agent is designed to generate inquiries and interact with these experts to gather the essential information for the medical reasoning questions. Finally, once the agent obtains enough information, it will stop asking questions and integrate the conversation for answering. Figure 3 displays the overview of the proposed MultiMedRes.

3.1 Problem Statement

For common medical questions involving a single image (VQA), we adhere to the standard practice of VQA tasks by implementing classification models. Our trained domain expert models, denoted as $\mathcal{F}_{sp}$, are designed to receive an image $i \in I$ and a question $q \in Q$ as inputs, and to subsequently predict the label $a \in A$, which corresponds to answers observed in the dataset. Further, in this study, we mainly focus on the challenging yet essential medical reasoning task: difference question answering (i.e., DVQA). This type of questions inquires about the comparison between current and previous images of the same patients and studies the treatment or disease progress, which is more consistent with real-life radiologists' practice. As illustrated in Figure 3, our framework would take two images $i_1, i_2 \in I$ and a difference question $q_d \in Q$ as input and output the difference captioning regarding the question. More specifically, the agent $\mathcal{F}_{learner}$ will generate questions concerning single images iteratively and interact with the corresponding specialist models to derive an answer in each round of the conversation, terminating the process upon acquiring sufficient information to formulate a response to the initial difference question.

In the subsequent section, we aim to design our prompting system to assist the LLM agent $\mathcal{F}_{learner}$ in comprehending the task, formulating the appropriate questions, and concluding the conversation appropriately.

3.2 Learner Agent

In MultiMedRes, we introduce an automatic question generation mechanism that leverages LLM's in-context learning and reasoning abilities. We utilize the LLM as a learner agent to generate informative questions as well as the corresponding question types about chest X-ray images, and direct these questions to well-trained domain-expert models for answer retrieval.

Mimicking the reasoning process of a human being when answering a difference question, a learner agent must first gather information of the main image and the reference image individually, and then proceed to summarize the difference. Specifically, we incorporate ChatGPT (Brown et al., 2020) or LLaMa (Touvron et al., 2023) as the learner agent

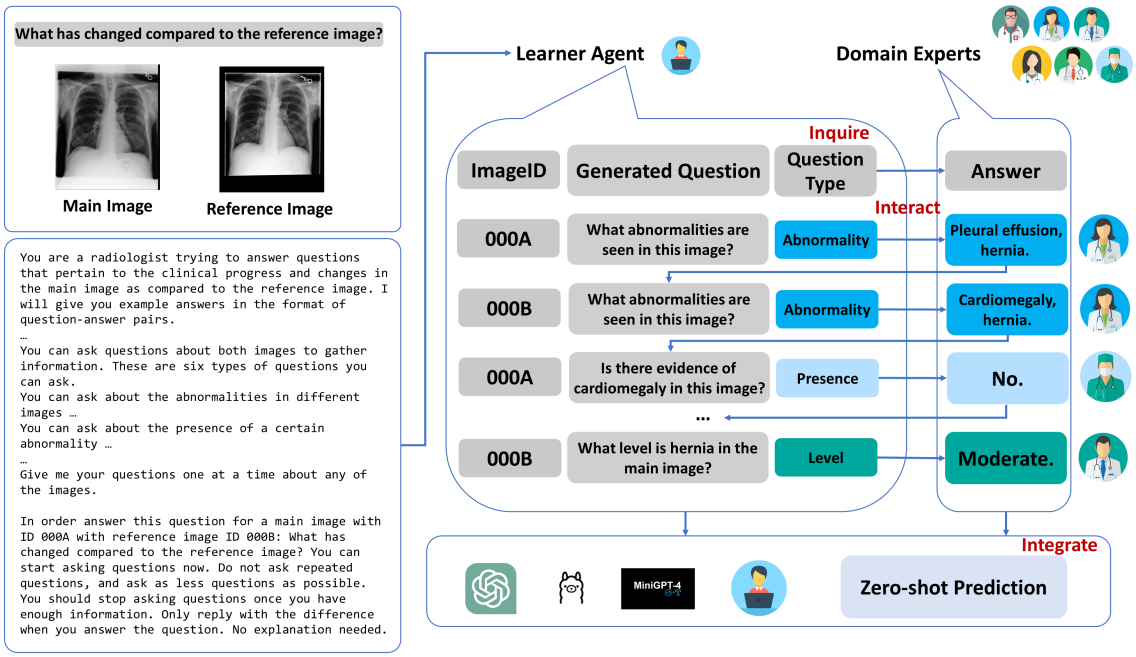


Figure 3: The proposed MultiMedRes framework. Upon receiving questions comparing two images, the learner agent employs an iterative approach, generating questions related to either the main image or the reference image, before consulting the appropriate domain expert (i.e., specialists). Upon collecting sufficient information, the learner agent is prompted to cease question generation and integrate the information to provide a zero-shot prediction.

responsible for generating questions. To optimize the in-context learning and reasoning abilities of the LLMs, we design the prompting system with three parts: a systematic instruction ρ_{task} to describe the difference question-answering task, a context-learning instruction ρ_Q to guide question generation, and an appended instruction ρ_i to signal the end of the questioning process and initiate summarization. Note that during the conversations, the generated question-answer pairs in earlier rounds (denoted as ρ_{log}) are also visible to the agent. Consequently, every question is generated using the combined prompt of $\rho_{task} + \rho_Q + \rho_i + \rho_{log}$.

3.2.1 Task Instruction ρ_{task}

The task instruction ρ_{task} outlines the task that the learner is required to perform and provides the material for in-context learning of the difference question answering. It guides the LLM to utilize the related knowledge base and generate questions to gather necessary information on images to compare the differences. ρ_{task} is designed as follows:

You are a radiologist trying to answer questions that pertain to the clinical progress and changes in the main image as compared to the reference image. I will give you example answers in the format of question-answer pairs. [context] Please answer this question for a main image A with reference image B: [difference questions]

The initial sentence directs the learner agent to access and apply radiology-related knowledge bases, while the context will demonstrate the desired answer formats (as shown in Appendix D).

3.2.2 Question Generation Instruction ρ_Q

To assist the learner agent in question generation, we supply both the categories and contents of the questions, which facilitates the identification of the most appropriate domain expert model. ρ_Q is structured as follows:

You can ask questions about both images to gather information, but do not ask redundant questions. You can ask about the abnormalities in different images, like [...]. You can ask about the presence of a certain abnormality in an image, like [...]. You can ask about the level of a certain abnormality in an image with [...]. Give me your questions one at a time about any of the images. Only return the generated question, the question type, and the corresponding image ID.

[...] above represents the corresponding question format of the given types, which will guide the LLM to generate the desired questions.

3.2.3 Appended Instruction ρ_i

From an efficiency perspective, the agent should stop asking questions once it gathers enough information, and answer with simplicity. Thus, ρ_i as the

last part of the prompt is structured as follows:

You should answer the question like the previous examples once you have enough information. Do not make any assumptions by yourself. Only reply with the difference when you answer the question. No explanation is needed.

Please refer to the Appendix D for a full version of the finalized prompt.

3.3 Domain Expert Models

As discussed in section 2, despite the significant efforts towards developing medical vision LLMs recently, the performance of these models on specialized medical VQA datasets has failed to surpass that of the traditional state-of-the-art classification models. Consequently, in this work, we continue to rely on classification models as the domain experts. However, take MIMIC-DIFF-VQA (Hu et al., 2023) dataset shown in Table 3 in subsection A.1 as an example, there are totally over 9000 answer candidates for all types of questions, not including those for difference questions. This vast number of labels could potentially overwhelm a classification model, particularly when the candidates are only marginally distinct from one another. Moreover, the amount of potential answers for abnormality-related questions is enormous as over 8000. This is particularly evident in responses to the question, "What abnormalities are seen in this image?", where answers comprise various combinations and permutations of 33 abnormalities, further complicating the differentiation process among these answers. For example, the answer "atelectasis, penumothorax" and the answer "penumothorax, atelectasis" are the same thing, and they are only slightly different than "penumothorax, atelectasis and pleural effusion".

To mitigate this challenge, our framework employs a divide-and-conquer strategy. Specifically, we develop a distinct expert model for each question type. As in MIMIC-DIFF-VQA, we will train an expert model for the "Abnormality", "Presence", "View", "Location", "Type" and "Level" type of questions separately as shown in the $2^{th} - 7^{th}$ rows in Table 3. This approach significantly narrows the scope of potential answers, enabling these specialists to concentrate exclusively on their respective domains. Moreover, regarding the 'abnormality' question, which essentially involves detecting abnormalities throughout the entire image, we shift our strategy. Rather than trying to match answers from an extensive training set containing a huge amount of responses, we focus on creating an abnormality detection model tailored specifically to this task, akin to multi-label classification. This model processes the image input and generates a list of identified abnormalities, as exemplified in the chest X-ray images. The response to this question is then formed by simply concatenating the predicted labels. Consequently, in MIMIC-DIFF-VQA, the number of possible answers for subsequent 'abnormality' questions is reduced to 25, a volume well-suited for a classification model.

3.4 Agent Behavior

As illustrated in Figure 3, for questions involving the comparison of two images, or difference questions, our proposed MultiMedRes follows an iterative pattern. During the initial iteration, the learner agent formulates a query for the expert models based on the question only. For the following iterations, the agent not only considers the difference questions but also integrates information already acquired from earlier conversations. This strategy enables the LLM agent to generate the next question accordingly and determine the appropriate time to cease inquiries, thereby avoiding the generation of redundant or irrelevant questions.

4 Experiment

4.1 Experimental Setting

Dataset and Setup In this study, we conduct comprehensive experiments on the newly-released DVQA dataset of chest x-ray images, MIMIC-Diff-VQA(Hu et al., 2023), which was derived from the open benchmark dataset MIMIC-CXR (Johnson et al., 2019). It includes a total of seven types of questions, including 164, 324 questions that require the comparison of two images, referred to as difference questions. We provide a more detailed description of the dataset in subsection A.1.

We employ both GPT (Brown et al., 2020) accessed through the OPENAI API [2] and LLaMa 2 (Touvron et al., 2023) via the Replicate API [3] as the learner agents. For the domain expert models, we integrate MMQ (Do et al., 2021) for the VQA tasks and DenseNet (Huang et al., 2016) for abnormality detection. Due to the space restriction, implementation details of each integrated model are given in subsection A.2. As for the evaluation,

[2] https://openai.com/
[3] https://replicate.com/

Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
EKAID (Hu et al., 2023)	0.569	0.498	0.438	0.382	0.304	0.547	0.823
EKAID_diff (Hu et al., 2023)	0.606	0.529	0.468	0.410	0.350	0.572	0.827
UIO (Lu et al., 2022)	0.360	0.309	0.267	0.223	0.220	0.426	0.388
MiniGPT-v2 (Chen et al., 2023)	0.291	0.237	0.190	0.146	0.333	0.391	0.110
LLaVa (Liu et al., 2023)	0.411	0.333	0.257	0.185	0.320	0.452	0.162
MultiMedRes (GPT-3.5)	0.497	0.423	0.361	0.303	0.317	0.526	0.451
MultiMedRes (LLaMa2)	0.537	0.465	0.391	0.345	0.373	0.554	0.483
MultiMedRes (GPT-4.5)	**0.610**	**0.535**	**0.473**	**0.418**	**0.357**	**0.586**	**0.843**

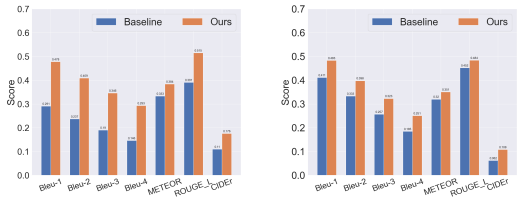
Table 1: Comparative performance of various models on the difference question answering task. EKAID and EKAID_diff represent the state-of-the-art task-specific models, which adopt the labeled data to perform fully-supervised training. The last three rows represent the zero-shot performance of MultiMedRes with the model being GPT 3.5, LLaMa2, and GPT 4.5 respectively.

following previous work, we use the popular natural language processing metrics BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE_L (Lin, 2004), CIDEr (Vedantam et al., 2015) for the performance on difference question answering. For the zero-shot prediction provided by the LLMs, we report the performance of a single run due to budget constraints. For other question types, we adhered to the common practice of using accuracy as the evaluation metric.

Baselines In this work, we consider a state-of-the-are medical VQA model, MMQ (Do et al., 2021), a state-of-the-art medical DVQA model, EKAID (Hu et al., 2023) and three multi-modal LLMs, UNIFIED-IO(Lu et al., 2022), MiniGPT-v2 (Chen et al., 2023) and LLaVA (Liu et al., 2023) as baselines. Please refer to subsection A.3 for details.

4.2 Main Results

We report the models' performance with difference questions in Table 1, and the performance of expert models on questions regarding single images in Table 4 in Appendix C. It can be observed that, training the EKAID model solely with difference questions results in a moderate improvement in performance compared to the baseline established by the original work. As for the large-scale multimodal models, specifically Unified-IO(Lu et al., 2022), MiniGPT-v2(Chen et al., 2023), and LLaVa(Liu et al., 2023), they all struggle to attain competitive performance on both DQA task and VQA task, likely attributable to a deficiency in domain knowledge. Conversely, leveraging the significantly enhanced capabilities of domain expert models, facilitated by our divide-and-conquer strategy, the zero-shot predictions generated by GPT-3.5, LLaMa2-70B, and GPT-4-Turbo through our proposed MultiMedRes all demonstrate competitive performance. Notably, GPT-4-Turbo significantly surpasses the previous state-of-the-art without requiring specific training on DVQA data. Furthermore, as demonstrated in Figure 4, incorporating the dialogue between the learner agent and the specialists into the prompts for the vision LLMs, e.g. MiniGPT-v2 and LLaVa, markedly enhances their performance. This improvement, combined with the outstanding results from the zero-shot prediction capability of MultiMedRes, suggests that the dialogue between the learner agent and the specialists indeed encapsulates essential domain knowledge absent in the LLMs, shedding light on the adaptive usage of LLMs in specific domains.



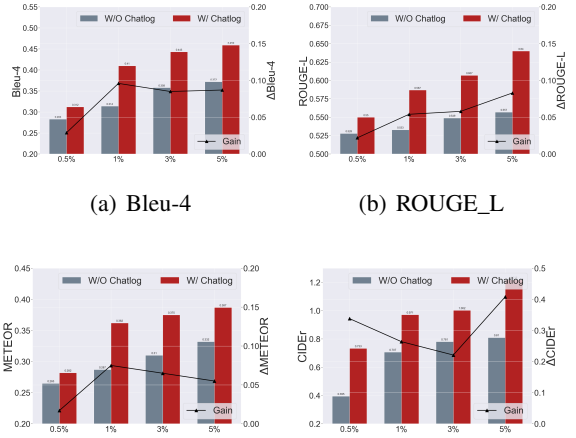
(a) MiniGPT (b) LLaVa

Figure 4: Performance comparison for vision LLMs with or without our MultiMedRes. As we can see, our method significantly boosts the performance of vision LLMs across all metrics.

5 Analysis

5.1 Augmented Training Data

To demonstrate the generalizability of our proposed framework, we further incorporate our generated conversation between the learner agent and the specialists as part of the training data for the supervised model EKAID. Specifically, instead of only considering the difference question as the text input, we also input the conversation chatlog for the training



(a) Bleu-4 (b) ROUGE_L

(c) METEOR (d) CIDEr

Figure 5: Bleu-4, METEOR, ROUGE_L and CIDEr score of the few-shot prediction generated by EKAID with respect to various ratios of labeled data (i.e., question-answer pairs) for training. The comparison is between the model trained with or without the augmented training data. The differences at different ratios are depicted using a polyline and the scales are indicated on the right y-axis.

and test samples. We present the performance of EKAID with varying proportions of the training set in Figure 4. It is observed that our generated chatlog enables the model to consistently outperform one trained without it, and it further improves its performance with an increase in training data. Notably, with only 1% of training data, the EKAID model with augmented training data achieves comparable performance with the one trained with all of the training data without chatlog. With 5% of the training data, it also outperforms the zero-shot prediction generated by the MultiMedRes. This is because the enhancement of zero-shot learning benefits from, and at the same time, is limited by the answers provided by the domain expert models. In other words, its performance is largely determined by the answer accuracy of questions regarding single images. On the other hand, while LLMs tend to accept whatever the specialists return and include everything in their answers, the supervised few-shot prediction model with augmented training data has the opportunity to identify and correct the incorrect, retaining only the essential information in the final answer. We will further discuss this difference in the case study in subsection 5.2.

5.2 Conversation Study

We present two case studies in Figure 6 and Figure 7 (see Appendix B) to more effectively illustrate our proposed framwork. In addressing the

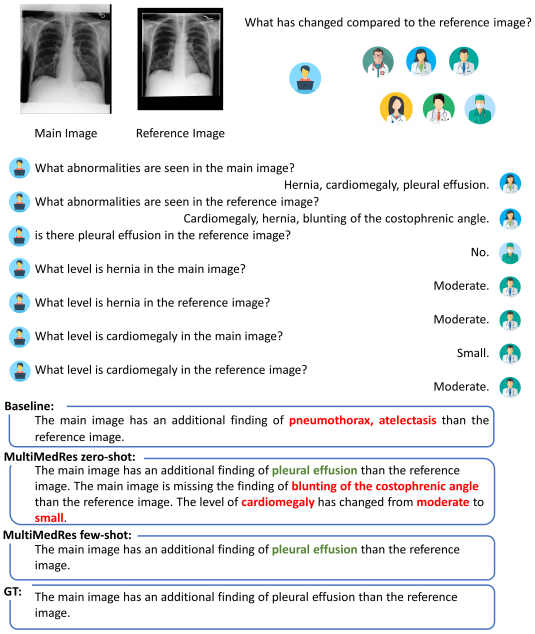


Figure 6: Case study for the commonly-seen question: "What has changed compared to the reference image?". We highlight the same information mentioned in the ground truth answer with the color green, and the redundant or incorrect one with the color red.

common query, 'What has changed compared to the reference image?', as depicted in Figure 6, the learner agent initially asks about image abnormalities from a global perspective, responses to which are provided by the abnormality detection specialist. Upon identifying the abnormalities in both images, the learner agent then consults the abnormality level specialist to examine the severity of recurring abnormalities, such as *hernia* and *cardiomegaly* in this instance, discovering a change in the level of *cardiomegaly*. Finally, having gathered sufficient information, the agent concludes its inquiries and formulates a comprehensive response incorporating all discussed information. However, this approach may not always yield the most accurate results. For instance, the specialists' responses might occasionally be inaccurate or highlight irrelevant details, as shown in Figure 6. The change in *cardiomegaly* severity from small to moderate may be less critical than the new finding of *pleural effusion*, and the noted *blunting of the costophrenic angle* might be a false positive. Thus, including all these details, as the learner agent does, could lead to a lower evaluation score. Conversely, the supervised few-shot prediction model can potentially rectify these inaccuracies with direct access to the images. This improvement is contingent on access to the chatlog, which enables the model to efficiently identify the correct information through

Gender	Female				Male			
	Bleu-4	METEOR	ROUGE_L	CIDEr	Bleu-4	METEOR	ROUGE_L	CIDEr
EKAID	0.398	0.342	0.556	0.805	0.417	0.357	0.583	0.860
MultiMedRes (Zero-shot)	0.414	**0.358**	0.579	0.830	0.416	0.355	0.595	0.858
MultiMedRes (Few-shot)	**0.454**	0.356	**0.596**	**1.261**	**0.457**	**0.364**	**0.608**	**1.141**

Age	Age<55				55<=Age<70				70<=Age			
	Bleu-4	METEOR	ROUGE_L	CIDEr	Bleu-4	METEOR	ROUGE_L	CIDEr	Bleu-4	METEOR	ROUGE_L	CIDEr
EKAID	0.364	0.350	0.537	0.887	0.426	0.352	0.585	0.849	0.445	0.348	0.597	0.743
MultiMedRes (Zero-shot)	0.406	0.356	**0.597**	0.968	0.418	0.360	0.581	0.847	0.424	0.356	0.575	0.676
MultiMedRes (Few-shot)	**0.432**	**0.358**	0.586	**1.472**	**0.467**	**0.362**	**0.609**	**1.096**	**0.471**	**0.362**	**0.619**	**0.915**

Table 2: Comparison of baseline against MultiMedRes for two different kinds of split.

the conversation, as opposed to extracting it by itself. With the chatlog, the search scope for the supervised model is significantly narrowed, yielding enhanced performance with a reduced reliance on training data.

It is worth noticing that the LLM learner agent can also rectify errors made by the specialists. In our observations, the learner agent occasionally poses what appears to be a redundant question, such as verifying the presence of a disease previously indicated by the abnormality specialist, or mentioning an abnormality not identified earlier in the discussion. Interestingly, these additional queries may lead the specialists to reconsider their previous statement, potentially leading to a more accurate final answer. We hypothesize that such questions derive from the extensive knowledge base of the LLMs. For instance, if the LLM rarely encounters instances of *blunting of the costophrenic angle* co-occurring with *cardiomegaly* in prior radiology reports, it might seek verification from the specialists, thereby potentially rectifying any inaccuracies. Indeed, addressing a simple yes-or-no question about a specific abnormality is significantly more straightforward than identifying all abnormalities, and it is evident that our specialists respond more accurately to these direct inquiries, as evidenced in Table 4. Additionally, for less common inquiries such as 'What has changed in the right lung?' shown in Figure 7, the LLM learner agent skillfully tailors its questions to acquire more targeted information. In these scenarios, the zero-shot answer from the learner agent is often more relevant, since the supervised few-shot answer may default to more general response patterns due to limited exposure to such uncommon cases. Given this analysis, the LLM learner agent is undoubtedly an integral component of our proposed framework. And, more importantly, our MultiMedRes is equipped to effectively harness its reasoning abilities and extensive knowledge base to tackle specific domain tasks.

5.3 Bias Evaluation

In this section, we assess the DVQA performance of our approach compared to the previous state-of-the-art, utilizing more fine-grained datasets. Specifically, we further segregate our test set into two subsets based on gender and age. To ensure a relatively even distribution of data and adhere to medical intuition, we categorize ages into three groups: $Age < 55(29.1\%)$, $55 \leq Age < 70(33.9\%)$, and $Age \geq 70(37.0\%)$. The results are presented in Table 2. As observed, the prior method, EKAID, exhibits substantial bias when encountering different age and gender groups, achieving significantly better performance towards elder patients and male patients. Conversely, both the zero-shot predictions made by LLMs and the few-shot predictions made by EKAID with augmented training data effectively reduce such bias. This demonstrates the generalization capability and effectiveness of our method across diverse datasets and patient groups.

6 Conclusion

In this study, we explore the demanding yet intricate setting of the medical image difference VQA problem. To navigate the complexities of identifying differences in medical images in response to specific questions, we introduce the multimodal medical collaborative reasoning framework, MultiMedRes, leveraging the superior reasoning capabilities of text-based LLMs. We first train multiple domain-specific expert models to address questions related to single images. Then, we incorporate the LLM as a learner agent to iteratively generate questions for the expert models, aiming to gather the necessary information in order to address the difference questions. Our extensive empirical analysis confirms the effectiveness of our proposed approach, which consistently improves upon the performance of baseline models under various metrics.

Limitations

Despite the exceptional performance, our proposed framework still encounters some limitations. Firstly, the performance of MultiMedAgent is constrained by the accuracy of the specialist models. If these domain experts fail to provide the correct knowledge for the learner agent, the conversation could lead more to confusion than clarification. Although we have significantly improved the performance of VQA models through a divide-and-conquer approach, a more robust VQA model would invariably enhance the reliability of our framework. Therefore, developing more advanced VQA models remains a promising direction for future research. Secondly, we observe that sometimes the learner agent delivers all the essential information correctly in its response, yet it receives a lower score than the supervised model simply because it does not adhere to the expected answer format. This issue arises because we are currently employing only natural language processing metrics for evaluation. Incorporating domain-specific evaluation metrics could offer insights from a more practical standpoint. We defer the exploration of such metrics to future studies.

Ethics Statement

In this work, we propose an AI agent collaborative framework to provide analysis of medical images in natural text. Our framework enables LLMs, which traditionally lack domain-specific knowledge, to acquire such critical insights regarding medical images from well-trained specialist models and to convey the analysis in the form of conversational dialogue to radiologists. Consequently, our system is not intended to replace radiologists in real-world settings but to act as a supportive assistant, aiding them in clinical decision-making and providing interpretable evidence for the analyses. This approach ensures radiologists can make fully informed decisions independently. We performed our experiments using the public dataset MIMIC-Diff-VQA, which is based on MIMIC-CXR. To comply with the Health Insurance Portability and Accountability Act (HIPAA) standards, all protected health information has been anonymized. Furthermore, all necessary patient/participant consents were obtained, and the appropriate institutional forms have been archived. In adherence to strict privacy protocols, we prohibit the sharing of raw data with any third-party APIs, such as OpenAI and Replicate. For image reference, we use "000A" and "000B" to denote the main image and the reference image, respectively, ensuring that only the VQA answers generated by our local model are shared.

References

Hanane Allioui, Mazin Abed Mohammed, Narjes Benameur, Belal Al-Khateeb, Karrar Hameed Abdulkareem, Begonya Garcia-Zapirain, Robertas Damaševičius, and Rytis Maskeliūnas. 2022. A multi-agent deep reinforcement learning approach for enhancement of covid-19 ct image segmentation. *Journal of Personalized Medicine*.

Hanane Allioui, Mohamed Sadgal, and Aziz Elfazziki. 2021. Optimized control for medical image segmentation: improved multi-agent systems agreements using particle swarm optimization. *Journal of Ambient Intelligence and Humanized Computing*.

Mohamed T. Bennai, Zahia Guessoum, Smaine Mazouzi, Stéphane Cormier, and Mohamed Mezghiche. 2020. A stochastic multi-agent approach for medical-image segmentation: Application to tumor segmentation in brain mr images. *Artificial Intelligence in Medicine*.

Siddharth Biswal, Cao Xiao, Lucas Glass, Michael Brandon Westover, and Jimeng Sun. 2020. Clara: Clinical report auto-completion. *Proceedings of The Web Conference 2020*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NIPS*.

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

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B.

Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv: 2401.12208*.

Louke Delrue, Robert Gosselin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck. 2011. Difficulties in the interpretation of chest radiography. *Comparative interpretation of CT and standard radiography of the chest*.

Tuong Do, Binh X. Nguyen, Erman Tjiputra, Minh Tran, Quang D. Tran, and Anh Nguyen. 2021. Multiple meta-model quantifying for medical visual question answering. In *MICCAI*.

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural naturalist: Generating fine-grained image comparisons. In *EMNLP*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. KDD.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely connected convolutional networks. *CVPR*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David Mong, Safwan Halabi, Jesse Sandberg, Ricky Jones, David Larson, Curtis Langlotz, Matthew Lungren, and Andrew Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *EMNLP*.

Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. 2023. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv: 2305.11490*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv: 2304.03442*.

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

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

and fine-tuned chat models. *arXiv preprint arXiv: 2307.09288*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Kuan Wang, Yadong Lu, Michael Santacroce, Yeyun Gong, Chao Zhang, and Yelong Shen. 2023. Adapting llm agents through communication.

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 *CVPR*.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S. Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. 2023. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv: 2308.01317*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021. An empirical study of GPT-3 for few-shot knowledge-based VQA. *AAAI*.

Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. *ICDM*.

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

A Implementation Details

A.1 Statistics of the DQA Dataset

We examined our proposed technique on the public DVQA dataset, MIMIC-Diff-VQA(Hu et al., 2023), the statistics of which is presented in Table 3. We adhered to the official splits, with the train/validation/test split being 8/2/2. To ensure the integrity of the evaluation process, the images in different sets (train, validation, and test) do not overlap, effectively preventing any leakage of test images into the training dataset of the domain experts. Additionally, for each question type, we provide two examples. These examples also serve as context-learning material for the LLMs.

A.2 Implementation Details

We incorporate gpt-3.5-turbo-0125, gpt-4-1106-preview, and LLaMa2 70B as the learner agents. To enhance the reproducibility and stability of their performance, we set the temperature parameter at 0.2 for both LLMs. In terms of specialist agents, we integrated the MMQ (Do et al., 2021) as the domain expert VQA model and a 121-layer DenseNet (Huang et al., 2016) as the multi-label prediction model for abnormality detection. Regarding the training hyperparameters, we adhered to the configurations specified in the original studies and retained the default settings. We train expert models and perform evaluation on an NVIDIA Tesla P100 GPU with 16 GB memory.

A.3 Baselines

- **MMQ** MMQ(Do et al., 2021) is a recently proposed VQA model designed for medical images, which adopts Model Agnostic Meta-Learning (MAML) through pretraining multiple meta-models on natural images and fine-tuning on the medical images.

- **EKAID** EKAID(Hu et al., 2023) aligns the high dimensional feature of different x-ray images through an expert knowledge graph. The model is designed to adaptively choose either to focus on the subtractive difference feature or the main image feature by utilizing the attention mechanism.

- **UIO** UNIFIED-IO(Lu et al., 2022) (UIO) is a unified vision model designed to handle a broad range of vision-language tasks, including VQA, by standardizing diverse inputs and outputs into a sequence of tokens, which enables the model to be trained on over 90 different datasets using a unified transformer-based architecture.

- **MiniGPT-V2** MiniGPT-v2 (Chen et al., 2023) is a unified large language model designed to efficiently handle a variety of vision-language tasks based on Llama2. By using unique identifiers for different vision-language tasks during training, the model achieves strong performance on multiple benchmarks.

- **LLaVa** LLaVA (Liu et al., 2023), representing a popular practice for training multi-modal

Question Type	Example	# QA pairs	# answer candidates
Abnormality	what abnormalities are seen in this image?	59,435	8316
Abnormality*	what abnormalities are seen in the upper lungs?	85,986	25
Presence	is there evidence of atelectasis in this image? Is there edema?	155,726	2
View	which view is this image taken? is this AP view?	56,265	4
Location	where in the image is the pleural effusion located? is the atelectasis located on the left side or right side?	84,193	735
Type	what type is the opacity? what type is the atelectasis?	27,478	87
Level	what level is the cardiomegaly? what level is the pneumothorax?	67,296	112
All	-	536,379	9,256
Difference	what has changed compared to the reference image? what has changed in the right lung area?	164,324	-

Table 3: Statistics of the MIMIC-CXR-VQA dataset. *Abnormality** stands for the abnormality questions excluding the question "what abnormalities are seen in this image?". *All* stands for the questions excluding the difference type of questions. Note that the examples shown here are the learning context we provide to the learner LLM models.

large language models, combines a vision encoder with a language model, which utilizes GPT-4 to generate language-image instruction data for improving the zero-shot capabilities.

Models	Open	Close	All
MMQ	11.5	10.8	11.5
EKAID	26.4	79.9	52.5
UIO	0.04	53.6	35.4
MiniGPT-v2	15.5	57.3	45.2
LLaVa	13.1	55.8	42.6
Ours	**54.3**	**84.8**	**70.6**

Table 4: Comparative performance of various models on questions concerning single images.

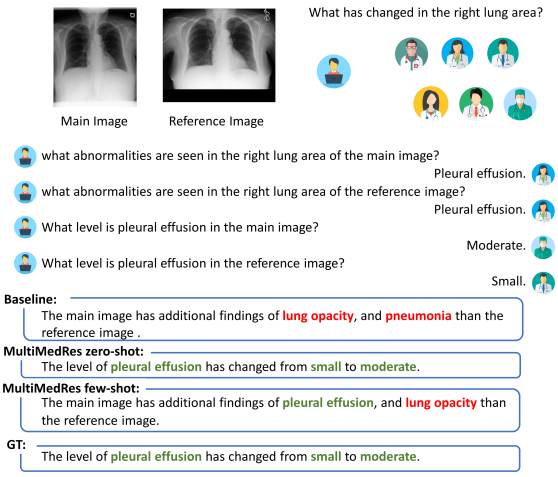


Figure 7: Case study for the rare question: "What has changed in the right lung?". We highlight the same information mentioned in the ground truth answer with the color green, and the redundant or incorrect one with the color red.

B Rare Case Study

We present a case study focusing on a relatively rare question: "What has changed in the right lung?" in Figure 7. This example conversation demonstrates that the LLM learner agent adapts its generated questions to gather more specific information effectively. We visually differentiate the correspondence between the model-generated responses and the ground truth. Information aligning with the ground truth answer is highlighted in green, whereas information that is either redundant or incorrect is marked in red.

C VQA Performance

We present the performance of various VQA models on single-image-related questions in Table 4. It can be observed that employing a divide-and-conquer strategy significantly improves performance, thereby enabling our MultiMedRes to acquire accurate information throughout the communication process.

D Prompts

We exhibit two example prompts for GPT and LLaMa 2 in Figure 8 and Figure 9 respectively. We change the difference questions (highlighted in blue) accordingly at runtime.

Example Prompt for GPT

You are a radiologist trying to answer questions that pertain to the clinical progress and changes in the main image as compared to the reference image. I will give you example answers in the format of question-answer pairs.

'what has changed compared to the reference image? the main image has an additional finding of pneumothorax than the reference image. the main image is missing the findings of fracture, lung opacity, and pleural effusion than the reference image.'
'what has changed compared to the reference image? the main image is missing the findings of lung opacity, consolidation, pleural effusion, and pleural thickening than the reference image.'
'what has changed in the right lung area? the level of pleural effusion has changed from small to moderate.'
'what has changed compared to the reference image? nothing has changed.'

You can ask questions about both images to gather information.
These are six types of questions you can ask.

You can ask about the abnormalities in an image with the key being 'abnormality', like 'what abnormalities are seen in this image?', 'what abnormalities are seen in the upper lungs?'.
You can ask about the presence of a certain abnormality in an image with the key being 'presence', like 'is there evidence of atelectasis in this image?', 'is there edema?'.
You can ask about the level of a certain abnormality in an image with the key being 'level', like 'what level is the cardiomegaly?', 'what level is the pneumothorax?'.
You can ask about the location of a certain abnormality in an image with the key being 'location', like 'where in the image is the pleural effusion located?', 'is the atelectasis located on the left side or right side?'.
You can ask about the type of a certain abnormality in an image with the key being 'type', like 'what type is the opacity?', 'what type is the atelectasis?'.
You can ask about the view of an image with the key being 'view', like 'which view is this image taken?', 'is this AP view?'.

Give me your questions one at a time about any of the images in the format of a Python dictionary with keys recording the imageID, question_type and question_content. Only return the Python dictionary.
In order answer this question for a main image with ID 000A with reference image ID 000B: what has changed compare to the reference image? You can start asking questions now.

Do not ask repeated questions, and ask as less questions as possible. You should stop asking questions once you have enough information.
Only reply with the difference when you answer the question. No explanation is needed.

Figure 8: An example of the prompt for GPT.

Example Prompt for LLaMa 2

System Prompt:

You are a helpful radiologist assistant trying to answer questions that pertain to the clinical progress and changes in the main image as compared to the reference image. I will give you example answers in the format of question-answer pairs.

'what has changed compared to the reference image? the main image has an additional finding of cardiomegaly than the reference image.
'what has changed compared to the reference image? the main image is missing the finding of lung opacity than the reference image.'
'what has changed compared to the reference image? the main image has additional findings of cardiomegaly, and vascular congestion than the reference image. the main image is missing the findings of pleural thickening than the reference image. '
'what has changed in the right lung area? the level of pleural effusion has changed from small to moderate.'
'what has changed compared to the reference image? nothing has changed.'

In order to answer the question like these examples, you can ask questions about both images to gather information.

You can ask about the abnormalities in different images with the question_type being 'abnormality', like 'what abnormalities are seen in this image?', 'what abnormalities are seen in the upper lungs?'.
You can ask about the presence of a certain abnormality in an image with the question_type being 'presence', like 'is there evidence of atelectasis in this image?', 'is there edema?'.
You can ask about the level of a certain abnormality in an image with the question_type being 'level', like 'what level is the cardiomegaly?', 'what level is the pneumothorax?'.
You can ask about the location of a certain abnormality in a image with the key being 'location', like 'where in the image is the pleural effusion located?', 'is the atelectasis located on the left side or right side?'.
You can ask about the type of a certain abnormality in a image with the key being 'type', like 'what type is the opacity?', 'what type is the atelectasis?'.
You can ask about the view of an image with the key being view', like 'which view is this image taken?', 'is this AP view?'.

Give me your questions one at a time about any of the images in the format of a Python dictionary with keys recording the imageID, question_type and question_content. Only return the Python dictionary. You can only ask about the 'abnormality', the 'presence', the 'level', the 'location', the 'type' or the 'view'.

Initial Prompt:

[INST] In order to answer this question for a main image with ID 000A with reference image ID 000B: what has changed compare to the reference image? You can start asking questions about both images now. You should stop asking questions once you have enough information and, starting with the keyword 'Answer:', only reply with the difference when you answer the question. Please follow the format of the provided question-answer pairs. No explanation is needed. [/INST].

Figure 9: An example of the prompt for LLaMa.