MedThink: A Rationale-Guided Framework for Explaining Medical Visual Question Answering

Xiaotang Gai[1,2]*, Chenyi Zhou[1,2]*, Jiaxiang Liu[1,2]*, Yang Feng[3], Jian Wu[2], Zuozhu Liu[1,2]†

[1] ZJU-Angelalign R&D Center for Intelligence Healthcare, ZJU-UIUC Institute, Zhejiang University, China
[2] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Zhejiang University, China
[3] Angelalign Research Institute, Angelalign Technology Inc., China
{xiaotang.23, zuozhuliu}@intl.zju.edu.cn

Abstract

Medical Visual Question Answering (Med-VQA), which offers language responses to image-based medical inquiries, represents a challenging task and significant advancement in healthcare. It assists medical experts to swiftly interpret medical images, thereby enabling faster and more accurate diagnoses. However, the model interpretability and transparency of existing Med-VQA solutions are often limited, posing challenges in understanding their decision-making processes. To address this issue, we devise a semi-automated annotation process to streamline data preparation and build new benchmark Med-VQA datasets R-RAD, R-SLAKE and R-Path. These datasets provide intermediate medical decision-making rationales generated by multimodal large language models and human annotations for question-answering pairs in existing Med-VQA datasets, i.e., VQA-RAD, SLAKE and PathVQA. Moreover, we design a novel framework, MedThink, which finetunes lightweight pretrained generative models by incorporating medical decision-making rationales. MedThink includes three distinct strategies to generate decision outcomes and corresponding rationales, clearly showcasing the medical decision-making process during reasoning. Our comprehensive experiments show that our method achieves an accuracy of 83.5% on R-RAD, 86.3% on R-SLAKE and 87.2% on R-Path. These results significantly exceed those of existing state-of-the-art models with comparable parameters. Datasets and code are available at `https://github.com/Tang-xiaoxiao/Medthink`.

1 Introduction

The Medical Visual Question Answering (Med-VQA) task is designed to take medical images and specialized clinical queries as inputs, and provide accurate answers with texts. Since the inception of the Med-VQA challenge in 2018(Hasan et al., 2018), there has been a significant surge in interest in exploring the capabilities of Med-VQA. Med-VQA not only holds the potential to enhance patient engagement, thereby alleviating patient stress, but also assists physicians in clinical diagnosis, thus conserving valuable medical resources and reducing the risk of misdiagnosis (Zhan et al., 2020; Liu et al., 2023b).

The challenges to resolve the Med-VQA tasks are two-fold. On one hand, though there exist a wealth of datasets composed of medical images and text annotations (Porwal et al., 2018), the decision-making process between the question and answer pairs are usually missing, impeding reliable evaluation of model interpretability. While some recent datasets already incorporated images, specialized medical queries, and answer texts (Lau et al., 2018; Liu et al., 2021b), the corresponding reasoning process to reach certain diagnostic decisions remain unclear, resulting in black-box and clinically inapplicable inference (Lu et al., 2022; Liu et al., 2023c). A straightfoward solution is to integrate expert-level reasoning rationales in these datasets to unravel the underlying reasoning processes. However, manual annotation of such rationales is time-consuming and requires in-depth understanding of medical knowledge (Liu et al., 2025, 2024), while a fast and reliable rationale annotation framework is still missing (Liu et al., 2023a). On the other hand, models which can resolve the Med-VQA tasks in a fast, accurate and interpretable manner is of high necessity in real-world applications. Current Med-VQA methods often model this problem by retrieval and train Med-VQA models with contrastive or classification objectives. For instance, Nguyen et al. (Nguyen et al., 2019) employed a combination of unsupervised convolutional denoising autoencoders and the meta-learning method to learn domain specific

*Equal contribution.
†Correspondence author.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 7453–7465
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

weight initialization of Med-VQA model on external medical datasets. Moreover, Zhang et al. (Zhang et al., 2022) first implemented contrastive learning in the medical domain, presenting ConVIRT, a methodology that utilizes medical text-image contrastive loss for pretraining medical visual representations. Further, Liu et al. (Liu et al., 2021a) proposed CPRD, a two-stage pre-training framework, leveraging representation distillation and contrastive learning to train visual encoder for Med-VQA system on a large corpus of unlabeled radiological images. The recent PubMedCLIP model (Eslami et al., 2023) pioneers the incorporation of the Contrastive Language-Image Pre-Training (Radford et al., 2021) into the Med-VQA tasks by conducting pre-training.

In contrast, the remarkable performance of large language models (LLMs) across various natural language processing (NLP) tasks has been extended to text question-answering in healthcare(Nori et al., 2023) Building upon this, multimodal large language models (MLLMs) (OpenAI, 2023; Team et al., 2023) accept both text and image inputs to generate responses, presenting a novel approach to tackling the Med-VQA tasks. However, applying MLLMs directly to Med-VQA tasks in real medical scenarios is challenging due to their high operational costs and lack of interpretability.

In this paper, we aim to address the aforementioned challenges by providing new benchmark datasets and novel Med-VQA solutions. We design a semi-automated annotation method that leverages the powerful inference capabilities of MLLMs to assist experts during annotation, significantly improving the efficiency. Through our method, we develop the R-RAD, R-SLAKE and R-Path datasets. These datasets provide the intermediate reasoning steps critical for medical decision-making, including necessary medical background knowledge and descriptions of medical images, which we term Medical Decision-Making Rationales (MDMRs). Moreover, we design a novel framework, MedThink, to finetune the pretrained generative models, specifically selecting the T5-base architecture (Raffel et al., 2020) as our base architecture due to its practicality in real-world applications. With only 223M parameters, the architecture adeptly performs generative tasks, balancing cost-effectiveness and practical value. By incorporating MDMRs into the training process, our model outputs not only decision outcomes but also corresponding rationales, thereby clearly showcasing the medical decision-making process during inference. Based on different inputs for MDMRs during training, we further propose three distinct generative modes: “Explanation", “Reasoning", and “Two-Stage Reasoning", as shown in Figure 1.

Extensive experimental results demonstrate that our method achieves an accuracy of 83.5% on R-RAD, 86.3% on R-SLAKE and 87.2% on R-Path. These results show significant enhancements over the existing state-of-the-art models with comparable parameters. Our contributions are as follows:

- We develop a semi-automated process for annotating Med-VQA data with decision-making rationale. To the best of our knowledge, the R-RAD, R-SLAKE and R-Path datasets represent the first Med-VQA benchmark datasets that encompass rationales for answers.
- We propose MedThink, a lightweight framework with three answering strategies, which enables faster and more accurate Med-VQA with interpretability. This has been demonstrated through extensive experiments.

2 Related Work

2.1 Med-VQA

VQA represents a cutting-edge, multimodal task at the intersection of computer vision and natural language processing, drawing significant attention in both domains. Med-VQA applies the principles of VQA to interpret and respond to complex inquiries about medical imagery. A Med-VQA system usually consists of three key components for feature extraction, feature fusion and answer reasoning, which aims to generate answers in text by processing given medical images.

Previous Med-VQA solutions (Nguyen et al., 2019; Zhang et al., 2022) have relied on the CNNs, such as those pretrained on ImageNet like VGGs or ResNets, to extract visual features. Meanwhile, the RNNs are employed to process textual information. With the development of large-scale pretraining, recent works (Liu et al., 2023b; van Sonsbeek et al., 2023; Eslami et al., 2023) have shifted towards the transformer-based models to enhance feature extraction capabilities for both textual and visual modalities. In terms of content, these works still treat the Med-VQA as the classification problem. However, this approach is misaligned with the realities of medical practice, where clinicians rarely face scenarios that can be addressed with predefined answer options.

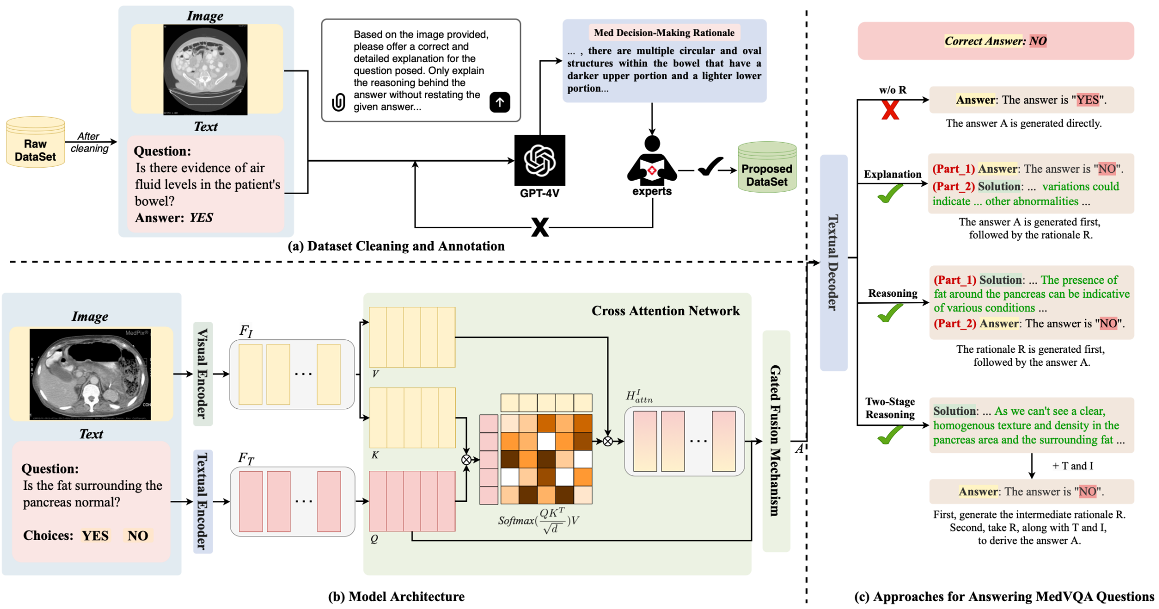


Image
Text
Raw DataSet
After cleaning
Question:
Is there evidence of air fluid levels in the patient's bowel?
Answer: YES
Based on the image provided, please offer a correct and detailed explanation for the question posed. Only explain the reasoning behind the answer without restating the given answer...
GPT-4V
Med Decision-Making Rationale
... , there are multiple circular and oval structures within the bowel that have a darker upper portion and a lighter lower portion...
experts
Proposed DataSet
(a) Dataset Cleaning and Annotation
Image
Text
Question:
Is the fat surrounding the pancreas normal?
Choices: YES NO
Visual Encoder
Textual Encoder
F_I
F_T
Cross Attention Network
V
K
Q
H^I_{attn}
$Softmax(\frac{QK^T}{\sqrt{d}})V$
Gated Fusion Mechanism
A
(b) Model Architecture
Textual Decoder
Correct Answer: NO
w/o R
Answer: The answer is "YES".
The answer A is generated directly.
Explanation
(Part_1) Answer: The answer is "NO".
(Part_2) Solution: ... variations could indicate ... other abnormalities ...
The answer A is generated first, followed by the rationale R.
Reasoning
(Part_1) Solution: ... The presence of fat around the pancreas can be indicative of various conditions ...
(Part_2) Answer: The answer is "NO".
The rationale R is generated first, followed by the answer A.
Two-Stage Reasoning
Solution: ... As we can't see a clear, homogenous texture and density in the pancreas area and the surrounding fat ...
+ T and I
Answer: The answer is "NO".
First, generate the intermediate rationale R. Second, take R, along with T and I, to derive the answer A.
(c) Approaches for Answering MedVQA Questions

Figure 1: Overview of the Data Preparation, Model Architecture and Methods for Answering Med-VQA Questions. **(a)** outlines the dataset cleaning and annotation process, where raw data undergoes refinement and annotation to formulate a new dataset with accurate MDMRs. **(b)** displays the model architecture, which incorporates a textual encoder for processing the medical question, a visual encoder for analyzing medical images, and a cross-attention network with a gated fusion mechanism that synergistically combines textual and visual features to generate informed responses for the Med-VQA task. (Carion et al., 2020; Khashabi et al., 2020; Zhang et al., 2023b) **(c)** is the illustration of various strategies for answering Med-VQA questions. These strategies show how the inclusion and arrangement of MDMRs can influence the model's output. The training process involves three steps. First, the training sets are annotated by the MLLM. Next, we use the training sets to train our models. Third, trained models generate MDMRs for the test sets.

This incongruity underscores the necessity for a Med-VQA approach that is more adaptive and reflective of the complexities inherent in medical diagnostics and decision-making. Our work redefines Med-VQA as the generative task. Within actual medical environment, when faced with open-ended queries, our proposed Med-VQA model can still generate informed responses based on the medical knowledge it has learned.

2.2 The Chain of Thought

Recently, NLP has been significantly transformed by language models (Raffel et al., 2020; Chowdhery et al., 2023). To further enhance the reasoning capabilities of language models, prior works (Cobbe et al., 2021; Wei et al., 2022) have incorporated reasoning rationales during training or inference phases, which guide models to generate the final prediction. On the other hand, in the realm of VQA, it is crucial for VQA systems to understand multimodal information from diverse sources and reason about domain-specific questions. To achieve this goal, several works (Lu et al., 2022; Zhang et al., 2023b) have proposed multimodal reasoning methods for VQA. These methods, commonly referred to as “Chain of Thought", introduces intermediate steps to assist the model in reasoning. In this paper, we present the “Medical Decision-Making Rationale" (MDMR) and apply it to the Med-VQA tasks. We anticipate that Med-VQA systems, equipped with the MDMR, will not only offer support in medical decision-making but also elucidate the underlying rationales behind these decisions.

3 Methodology

3.1 Problem Formulation

In this paper, We denote the medical dataset as $\mathcal{D} = \{(I_m, T_m, A_m, R_m)\}_{m=1}^{M}$ where M is the number of data samples. And the goal of the Med-VQA tasks is to develop a mapping function $f(\cdot)$ that can generate textual answers in response to the medical questions, represented as:

$$\{A, R\} = f(I, T), \tag{1}$$

Here, I denotes the medical image sourced from modalities such as X-ray, CT, or MRI. T represents the natural language question pertaining to the medical image I. The output of the model $f(\cdot)$, represented as $\{A, R\}$, comprises two components. A is the predicted textual answer, directly addressing the query posed in T. R, termed as “medical decision-making rationale", offers a detailed justification for the answer A, elucidating an interpretative insight into how the model processes I and T.

3.2 Model Architecture

The model architecture comprises five components, shown in Figure 1 (b): TextualEncoder, VisualEncoder, Cross Attention Network, Gated Fusion Network, and TextualDecoder. Notably, the TextualEncoder, VisualEncoder and TextualDecoder are all based on the Transformer architecture, renowned for its powerful learning and representational capabilities.

The TextualEncoder vectorizes the input question T into the textual feature space, represented as $F_T \in \mathbb{R}^{n\times d}$, while the VisualEncoder transforms the input medical image I into vision features $F_I \in \mathbb{R}^{m\times d}$. This can be expressed as: $F_{\text{T}} =$ TextualEncoder(T) and $F_{\text{I}} =$ VisualEncoder(I), where n denotes the length of the input text, and d indicates the hidden dimension, m represents the number of image patches.

Upon acquiring the textual representation F_T and visual representation F_I, our model employs the Cross-Attention Network to facilitate interaction between these two modalities. The Cross-Attention Network computes the attention-guided visual feature $H^I_{attn} \in \mathbb{R}^{n\times d}$, which captures the relevant visual features corresponding to the textual query through the following operation:

$$H^{\text{I}}_{\text{attn}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (2)$$

where Q, K, V correspond to the query, key, and value, derived from F_T, F_I, F_I, respectively.

Subsequently, the Gated Fusion Mechanism is utilized to dynamically combine the textual representation F_T and the attention-guided visual feature H^I_{attn}. It determines the fusion coefficient λ through a sigmoid-activated linear combination of the two modalities:

$$\lambda = \text{Sigmoid}(W_l F_T + W_v H^{\text{I}}_{\text{attn}}), \quad (3)$$

The fused output $F_{\text{fuse}} \in \mathbb{R}^{n\times d}$ is computed as a weighted sum of F_T and $H^{\text{I}}_{\text{attn}}$, moderated by λ:

$$F_{\text{fuse}} = (1-\lambda)\cdot F_{\text{T}} + \lambda \cdot H^{\text{I}}_{\text{attn}}, \quad (4)$$

Here, W_l and W_v are the model parameters that are learned during training to optimize the fusion of information between textual and visual streams. Finally, the fused output F_{fuse} is fed into the TextualDecoder to generate the output $\{A, R\}$:

$$\{A, R\} = \text{TextualDecoder}(F_{\text{fuse}}), \quad (5)$$

3.3 Loss Function

Given the input $X = \{I, T\}$, the model f is trained by maximizing the likelihood of accurately predicting the target output $Y = \{A, R\}$. The training involves a loss function, primarily the negative log-likelihood of correctly predicting subsequent tokens in the sequence Y, accumulated over all time steps. This is mathematically formulated as:

$$L = -\sum_{n=1}^{N} \log p(Y_n | X, Y^{1:n-1}), \quad (6)$$

In this context, N represents the total number of tokens in the target answer Y, and $p(Y_n|X, Y^{1:n-1})$ denotes the conditional probability of correctly predicting the n-th token in Y, given the input X and all preceding tokens $Y^{1:n-1}$ in the sequence. This loss function significantly improves the model's capability to accurately forecast each token in the target output, thereby enhancing its overall predictive performance.

3.4 Three Generation Strategies

To investigate the impact of MDMRs on the model performance in the Med-VQA tasks, we present three different generation strategies. These strategies are designed to guide the model in generating various forms of outputs, corresponding to different orders of MDMR in the process of generation. The methods are categorized as “Explanation", “Reasoning" and “Two-Stage Reasoning", as shown in Figure 1 (c).

In the “Explanation" method, the answer A is generated first, followed by the MDMR R. In contrast, the “Reasoning" method reverses this order, generating R before A. The “Two-Stage Reasoning" method follows a phased strategy, where two independent models are trained in distinct stages. The first stage focuses on using the medical question T and the medical image I to generate the intermediate result R. In the second stage, a different model takes R, along with T and I, to derive the final answer A.

Table 1: Details of Datasets: Distribution of Images and Questions in the R-RAD, R-SLAKE and R-Path Datasets.

Dataset	Images	Training set	Test set
R-RAD(closed-end)	300	1823	272
R-RAD(open-end)	267	1241	179
R-SLAKE(closed-end)	545	1943	416
R-SLAKE(open-end)	545	2976	645
R-Path(closed-end)	3361	9806	3391
R-Path(open-end)	3425	9933	3364

4 Dataset Creation

4.1 Dataset Collection

We establish three benchmark datasets R-RAD, R-SLAKE and R-Path based on the VQA-RAD dataset (Lau et al., 2018), the SLAKE dataset (Liu et al., 2021b) and the PathVQA (He et al., 2020), respectively.

Relevant statistics for the R-RAD, R-SLAKE and R-Path datasets are detailed in Table 1. Details about these datasets and their specific splits can be found in the Appendix A.1.

4.2 Dataset Cleaning

We identify noticeable inconsistencies within raw datasets. Specifically, the answers to similar questions about the same medical image are not always consistent. For instance, given a chest X-ray imaging, the response to the question “Is/Are the right hemidiaphragm normal?" is “No", while the answer to “Is this image normal?" is “Yes". This apparent contradiction prompted us to seek further expert medical review for such cases, ensuring the reliability of our dataset.

In light of advancements of MLLMs, we integrate the MLLM into our data cleaning and annotation process, aiming to streamline the workflows. This integration can not only expedite data processing but also unearth subtleties often missed in manual cleaning and annotation practices. To address inconsistencies, we first use the MLLM to systematically review all question-answer pairs for each medical image. After identifying inconsistencies, domain experts revise the answers, ensuring consistency across all questions related to the same medical image.

4.3 Dataset Annotation

After data cleaning, we utilize the MLLM for data annotation, specifically in generating MDMRs for the items within the VQA-RAD, SLAKE and PathVQA datasets, as shown in Figure 1 (a). This involves furnishing the MLLM with the datasets' images, questions, and correct answers. Therefore, we design a fixed prompt to guide the generation process of the MLLM. To ensure the quality of MDMRs, domain experts check MDMRs' validity and applicability. MDMRs not meeting criteria will be regenerated by the MLLM. If a MDMR generated by the MLLM remains below standard even after three attempts, domain experts will personally create an acceptable version, adhering to predefined criteria.

We enlist experienced physicians as domain experts to ensure the professional and accurate annotation of our data. To account for the diversity of medical opinions, we establish rigorous review criteria to guide the annotation process. The criteria are as follows:

(1) Coherence: The MDMR must be logically coherent, with no errors in grammar or spelling.

(2) Relevance: The MDMR must be directly related to the question and pertinent to the clinical context.

(3) Accuracy: The MDMR should be free from common sense and medical knowledge errors.

Only when all three conditions are met will the MDMR be included in our datasets.

5 Experiments

5.1 Training Details

During the datasets construction phase, we select GPT-4V (OpenAI, 2023) from among MLLMs to handle data cleaning and annotation. In our framework, the encoder and decoder from UnifiedQA (Khashabi et al., 2020) are integrated as TextualEncoder(·) and TextualDecoder(·), respectively. Additionally, DETR (Carion et al., 2020) is employed as VisualEncoder(·).

In our experiments, the learning rate is uniformly set at $5e-4$ for the R-SLAKE, R-RAD and R-Path datasets. The number of epochs during fine-tuning varies by dataset: 300 epochs for the R-SLAKE dataset, 150 epochs for the R-RAD dataset and 50 epochs for the R-Path dataset. It is important to note that our “Two-Stage Reasoning" strategy requires a phased fine-tuning process involving two separate models. In the first phase, we follow the parameters mentioned above. In the second phase, we fine-tune with a learning rate of 5e-5 for 20 epochs across all three datasets. The batch size is set to 32.

All experiments reported in this paper are

Table 2: Accuracy (%) Comparison of Methods on Closed-End Questions in the R-RAD, R-SLAKE and R-Path.

Methods	MLLM-Based	**R-RAD**	**R-SLAKE**	**R-Path**	Parameters
Zero-shot results					
Med-MoE(StableLM)(Jiang et al., 2024)	✓	66.9	52.6	69.1	2B
LLaVA-Med(From LLaVA) (Li et al., 2024)	✓	60.2	47.6	59.8	7B
Gemini Pro(Team et al., 2023)	✓	73.5	69.0	64.8	-
Gemini Pro (w/ Reasoning)(Team et al., 2023)	✓	77.2	77.4	70.9	-
Gemini Pro (w/ Two-Stage Reasoning)(Team et al., 2023)	✓	79.4	77.9	72.3	-
Gemini Pro (w/ Explanation)(Team et al., 2023)	✓	79.8	78.1	72.6	-
Representative & SOTA methods (Supervised finetuning results)					
MFB(Yu et al., 2017)		74.3	75.0	-	-
SAN(Yang et al., 2016)		69.5	79.1	-	-
BAN(Kim et al., 2018)		72.1	79.1	-	-
MEVF+SAN(Nguyen et al., 2019)		73.9	78.4	-	-
MEVF+BAN(Nguyen et al., 2019)		77.2	79.8	-	-
MMBERT(Tiong et al., 2022)		-	77.9	-	-
PubMedCLIP(Eslami et al., 2023)		79.5	82.5	-	-
Prefix T. Medical LM(GPT2-XL)(van Sonsbeek et al., 2023)	✓	-	82.1	87.0	1.5B
LLaVA (Li et al., 2024)	✓	65.1	63.2	63.2	7B
Med-Flamingo (Moor et al., 2023)	✓	65.1	63.2	63.2	7B
LLaVA-Med (From LLaVA) (Li et al., 2024)	✓	84.2	85.3	91.2	7B
LLaVA-Med (From Vicuna) (Li et al., 2024)	✓	82.0	83.2	91.7	7B
Med-MoE(StableLM) (Jiang et al., 2024)	✓	80.1	83.4	91.3	2B
Med-Gemini (Yang et al., 2024)	✓	-	84.8	83.3	-
MedThink (w/o R)		**79.0**	**82.5**	**86.0**	**0.2B**
MedThink (w/ Reasoning)		**73.9** (-5.1)	**80.8** (-1.7)	**83.1** (-2.9)	**0.2B**
MedThink (w/ Two-Stage Reasoning)		**80.5** (+1.5)	**79.1** (-3.4)	**87.2** **(+1.2)**	**0.2B**
MedThink (w/ Explanation)		**83.5** **(+4.5)**	**86.3** **(+3.8)**	**87.0** (+1.0)	**0.2B**

Red and blue numbers indicate increases and decreases in accuracy compared to the **MedThink (w/o R) results respectively.*

conducted using PyTorch on an Ubuntu server equipped with four NVIDIA RTX 3090 GPUs. Training on the R-RAD dataset takes about 2.5 hours. In comparison, training on the R-SLAKE dataset requires approximately 5.5 hours, while the R-Path dataset takes around 14 hours. During inference, processing each sample takes about 6 seconds.

5.2 Evaluation Metrics

Our performance evaluation is divided into two parts, focusing on closed-end and open-end questions separately. For closed-end questions, which are formatted as multiple-choice with a single correct answer, we assess performance using accuracy as the metric. For open-end questions, in contrast to previous works (Yang et al., 2016; Kim et al., 2018; Yu et al., 2017; Nguyen et al., 2019; Tiong et al., 2022; Eslami et al., 2023) that often emphasize scoring all possible answers in open-ended Med-VQA datasets to gauge classification accuracy, our work on generative Med-VQA prioritizes clinical utility. Following established studies (Li et al., 2023; Zhang et al., 2023a), we employ BLEU and ROUGE to assess the quality of our method's outputs. The BLEU scores, akin to the "Precision", evaluates the overlap of k-grams between generated and reference sentences, while the ROUGE scores, similar to the "Recall", measures the similarity in word sequences.

5.3 Main Results

In facing closed-end questions, we evaluate the performance of MedThink under various generation strategies, and compare them against several baseline methods on the R-RAD, R-SLAKE, and R-Path datasets. The results are shown in Table 2. MedThink demonstrates varying levels of performance across different generation strategies. To be specific, MedThink with the "Explanation" strategy achieves the highest accuracy on the R-RAD and R-SLAKE datasets, recording 83.5% and 86.3% respectively. Meanwhile, MedThink with the "Two-Stage Reasoning" strategy achieves the best performance on R-Path with an accuracy of 87.2%.

In contrast, the state-of-the-art classification model, PubMedCLIP, achieves accuracies of 79.5% on the R-RAD dataset and 82.5% on the R-SLAKE dataset, which is significantly lower than MedThink's results. This underscores the superior performance of MedThink. Compared to other generative models based on MLLM, MedThink outperforms models such as LLaVA (Li et al., 2024), Med-Flamingo (Moor et al., 2023), and Med-Gemini (Yang et al., 2024), achieving overall accuracies that are on par with the more parameter-

Table 3: Score (%) Comparison of Medthink on Open-End Questions in the R-RAD, R-SLAKE and R-Path Datasets.

Dataset	Strategy	Rouge-1	Rouge-2	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
R-RAD	Reasoning	49.8	**20.3**	29.3	37.8	22.7	**14.0**	**8.9**
	Two-Stage Reasoning	49.1	19.9	28.7	37.7	22.5	13.9	8.8
	Explanation	**50.2**	20.2	**29.5**	**38.3**	**22.9**	**14.0**	8.8
R-SLAKE	Reasoning	**53.5**	22.8	**32.1**	**39.5**	24.3	15.5	10.0
	Two-Stage Reasoning	53.2	**23.1**	32.0	**39.5**	**24.5**	**15.8**	**10.3**
	Explanation	53.1	22.7	31.7	39.2	24.1	15.4	9.9
R-Path	Reasoning	41.5	13.0	24.8	31.8	17.0	9.6	5.7
	Two-Stage Reasoning	41.7	**13.2**	24.9	**32.1**	**17.1**	**9.7**	**5.8**
	Explanation	**41.9**	**13.2**	**25.0**	**32.1**	**17.1**	**9.7**	**5.8**

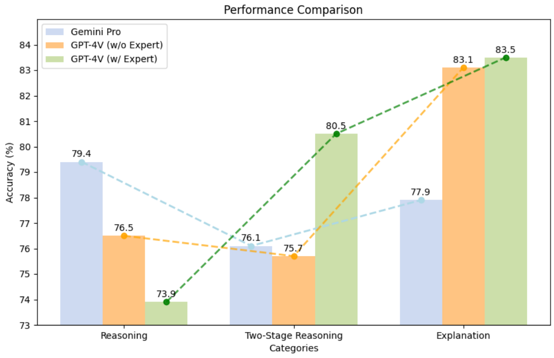
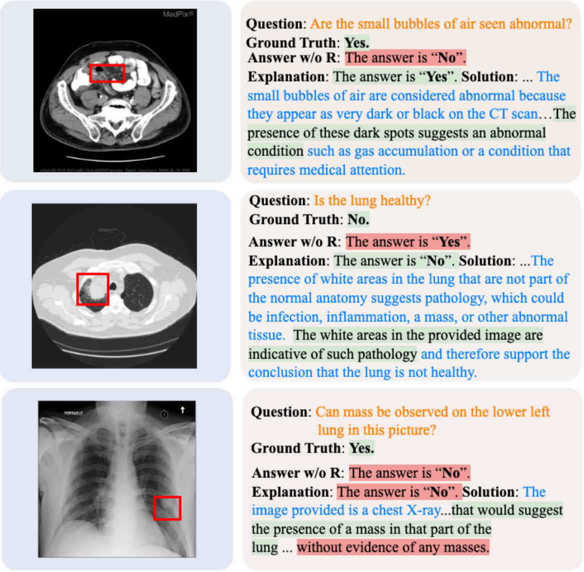


Figure 2: Impact of MLLMs Selection and Expert Participation in Dataset Creation on the Med-VQA Tasks Accuracy (%) on Closed-End Questions in the R-RAD.

heavy LLaVA-Med (Li et al., 2024) and Med-MoE (Jiang et al., 2024) models. Notably, MedThink accomplishes this with a parameter count that is less than one-tenth of these models, demonstrating both its efficiency and effectiveness. Using open-end questions, we conduct a comprehensive evaluation of MedThink's three strategies on the R-RAD, R-SLAKE, and R-Path datasets. The results are summarized in Table 3. For the R-RAD dataset, the "Explanation" strategy outperforms other strategies, achieving the highest scores in five out of seven metrics. It records 50.2% in Rouge-1, 29.5% in Rouge-L, 38.3% in BLEU-1, 22.9% in BLEU-2, and 14.0% in BLEU-3. On the R-SLAKE dataset, the "Two-Stage Reasoning" strategy leads in performance, securing the top scores in five out of seven metrics, with 23.1% in Rouge-2, 39.5% in BLEU-1, 24.5% in BLEU-2, 15.8% in BLEU-3, and 10.3% in BLEU-4. Regarding the R-Path dataset, the "Explanation" strategy once again delivers the highest overall performance, achieving 41.9% in Rouge-1, 13.2% in Rouge-2, 25.0% in Rouge-L, 32.1% in BLEU-1, 17.1% in BLEU-2, 9.7% in BLEU-3, and 5.8% in BLEU-4. These results collectively highlight the importance of selecting appropriate generation strategies tailored to addressing various medical scenarios, ensuring the model generates comprehensive and detailed responses.

5.4 Ablation Study

To explore the influence of various components in MedThink, we conduct a series of ablation experiments. First, we evaluate the effect of different MLLMs used during dataset creation and the contribution of domain experts to data annotation. We implement three variations for annotating the closed-end questions in the R-RAD dataset: using Gemini Pro (Team et al., 2023) without expert involvement, using GPT-4V without expert involvement, and using GPT-4V with expert involvement. As shown in Figure 2, when GPT-4V is used with expert involvement in dataset creation, the “Explanation" and “Two-Stage Reasoning" strategies achieved their highest accuracies of 83.5% and 80.5%, respectively. In contrast, the “Reasoning" strategy performed best with Gemini Pro without expert involvement, reaching an accuracy of 79.4%, which is only slightly above the baseline accuracy of 79.0% when MDMRs are not applied. We attribute this to the instability of the “Reasoning" strategy, which hinders its ability to consistently benefit from MDMRs, aligning with previous research (Lu et al., 2022). Overall, expert involvement enhances the quality of MDMRs, positively impacting MedThink. Additionally, GPT-4V's stronger reasoning ability compared to Gemini Pro (Fu et al., 2023) further suggests that using a more advanced MLLM during data annotation is beneficial. Next, we examine how the introducing of MDMRs impacted results. We introduce a control experiment, MedThink without MDMRs, where the models are trained and inferred without incorporating MDMRs (MedThink w/o R). The “Explanation", “Reasoning" and “Two-Stage Reasoning" strategies are compared with the control experiment. As indicated in Table 2, compared to “MedThink w/o R", the “Explanation", “Two-Stage



Question: Are the small bubbles of air seen abnormal?
Ground Truth: **Yes.**
Answer w/o R: The answer is "**No**".
Explanation: The answer is "**Yes**". **Solution**: ... The small bubbles of air are considered abnormal because they appear as very dark or black on the CT scan…The presence of these dark spots suggests an abnormal condition such as gas accumulation or a condition that requires medical attention.

Question: Is the lung healthy?
Ground Truth: **No.**
Answer w/o R: The answer is "**Yes**".
Explanation: The answer is "**No**". **Solution**: ...The presence of white areas in the lung that are not part of the normal anatomy suggests pathology, which could be infection, inflammation, a mass, or other abnormal tissue. The white areas in the provided image are indicative of such pathology and therefore support the conclusion that the lung is not healthy.

Question: Can mass be observed on the lower left lung in this picture?
Ground Truth: **Yes.**
Answer w/o R: The answer is "**No**".
Explanation: The answer is "**No**". **Solution**: The image provided is a chest X-ray...that would suggest the presence of a mass in that part of the lung ... without evidence of any masses.

Figure 3: Illustration of MDMRs Enhancing Model Responses in the Med-VQA Tasks. The green highlighted text represents medically relevant knowledge that aids in answering the question, while the red highlighted text indicates information that could lead to incorrect conclusions. The red boxes in the images correspond to the described anatomical features, underscoring the alignment between the rationale and the visual evidence.

Reasoning", and "Reasoning" strategies improve accuracy by 4.5%, 1.5%, and -5.1% on the R-RAD dataset, 3.8%, -3.4%, and -1.7% on the R-SLAKE dataset, and 1.0%, 1.2% and -2.9% on the R-Path dataset, respectively.

Finally, we assess the practicality of MDMRs generated by MedThink using different strategies. Initially, Gemini Pro is provided with only medical queries and related imagery. Subsequently, we inincorporate MDMRs generated by MedThink with the "Explanation", "Reasoning" and "Two-Stage Reasoning" strategies to assist Gemini Pro in answering. The results, presented in Table 2, indicate an initial accuracy of 73.5% on the R-RAD dataset, 69.0% on the R-SLAKE dataset and 64.8% on the R-Path dataset for Gemini Pro. The integration of MDMRs has led to significant improvements. Among three strategies, the "Explanation" strategy stands out, enhancing the accuracy by 6.3% on the R-RAD dataset, 9.1% on the R-SLAKE dataset and 7.8% on the R-Path dataset.

5.5 Case Study

To assess the specific impact of MDMRs on the Med-VQA task, Figure 3 shows several examples where MedThink applies the "Explanation" strategy to answer questions from the R-SLAKE datasets.

Table 4: The error rate for each region (Lower values are better)

Anatomical Regions (Number)	w/o Rationale ↓	Explanation ↓
Lung (N=141)	12.06%	9.93%
Abdomen (N=141)	24.11%	19.15%
Head (N=91)	18.68%	13.18%
Neck (N=16)	18.75%	6.25%
Chest (N=5)	20.00%	0.00%
Pelvic Cavity (N=22)	4.55%	13.64%

When the generated MDMR is accurate, MedThink can effectively and precisely answer the related medical question. If the MDMR contains errors, however, it misguides MedThink, leading to a phenomenon known as hallucination, which is a common issue in vision-language models. To investigate the causes of hallucinations in MedThink, we analyze the number of incorrect answers it provided on the R-SLAKE dataset. The R-SLAKE dataset is chosen because it covers medical questions about six anatomical regions, offering a complex and representative challenge.

We perform the analysis through the following steps. First, we categorize the test set questions by the anatomical regions associated with the medical images. Next, we tally the number of incorrectly predicted questions for each anatomical region. Finally, we calculate the proportion of incorrect predictions for each region, as shown in Table 4. The results indicate that MedThink significantly aids in addressing medical issues related to the chest and abdomen. However, these areas still account for the majority of prediction errors. We attribute this to the greater complexity of chest and abdomen images, which contain more organs than other regions, presenting a considerable challenge for the model.

6 Conclusion

In this paper, we present a medical chain of thought method for Med-VQA and construct the R-RAD, R-SLAKE, and R-Path datasets. These datasets include intermediate reasoning steps to address the challenge of black-box decision-making processes in Med-VQA models. Extensive experimental results show that our proposed framework not only elucidates the medical decision-making process of Med-VQA models with clarity but also significantly enhances their performance. Future research will further explore generative models tailored for real clinical settings and how to better evaluate the performance of Med-VQA models in open-ended scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare.

Limitations

Data Security and Privacy: While we use open-source and desensitized datasets like VQA-RAD and SLAKE, there are still concerns about data security with external LLMs. Ensuring encryption, anonymization, and compliance with privacy standards is crucial, especially for private datasets or sensitive medical data.

Trust and Reliability: Our work aims to improve medical decision-making accuracy and model interpretability. However, the extent to which our method can be trusted remains a challenge. The reliability of AI outputs depends on the clinician's expertise, and inexperienced doctors might struggle to identify unreliable outputs. This issue requires further research and collaboration to establish common standards for AI in clinical practice.

References

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia. Association for Computational Linguistics.

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. 2023. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*.

Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew Lungren. 2018. Overview of imageclef 2018 medical domain visual question answering task. *Proceedings of CLEF 2018 Working Notes*.

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

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *Preprint*, arXiv:2404.10237.

A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. 2021. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

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.

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

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer.

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

Jiaxiang Liu, Jin Hao, Hangzheng Lin, Wei Pan, Jianfei Yang, Yang Feng, Gaoang Wang, Jin Li, Zuolin Jin, Zhihe Zhao, et al. 2023a. Deep learning-enabled 3d multimodal fusion of cone-beam ct and intraoral mesh scans for clinically applicable tooth-bone reconstruction. *Patterns*, 4(9).

Jiaxiang Liu, Tianxiang Hu, Jiawei Du, Ruiyuan Zhang, Joey Tianyi Zhou, and Zuozhu Liu. 2025. Kpl: Training-free medical knowledge mining of vision-language models. *arXiv preprint arXiv:2501.11231*.

Jiaxiang Liu, Tianxiang Hu, Huimin Xiong, Jiawei Du, Yang Feng, Jian Wu, Joey Zhou, and Zuozhu Liu. 2024. Vpl: Visual proxy learning framework for zero-shot medical image diagnosis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9978–9992.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, and Zuozhu Liu. 2023b. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–11.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, YANG FENG, and Zuozhu Liu. 2023c. A chatgpt aided explainable framework for zero-shot medical image diagnosis. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

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

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

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. GPT-4V(ision) System Card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`. Accessed: 2023-12-29.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. 2018. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.

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

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. 2022. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. *arXiv preprint arXiv:2303.05977*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

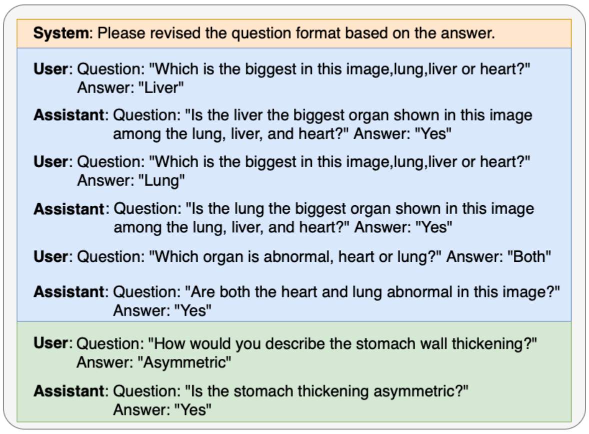
Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

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

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.



System: Please revised the question format based on the answer.

User: Question: "Which is the biggest in this image,lung,liver or heart?" Answer: "Liver"

Assistant: Question: "Is the liver the biggest organ shown in this image among the lung, liver, and heart?" Answer: "Yes"

User: Question: "Which is the biggest in this image,lung,liver or heart?" Answer: "Lung"

Assistant: Question: "Is the lung the biggest organ shown in this image among the lung, liver, and heart?" Answer: "Yes"

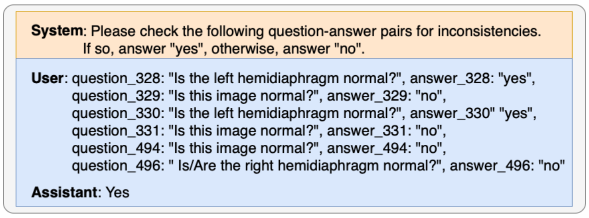
User: Question: "Which organ is abnormal, heart or lung?" Answer: "Both"

Assistant: Question: "Are both the heart and lung abnormal in this image?" Answer: "Yes"

User: Question: "How would you describe the stomach wall thickening?" Answer: "Asymmetric"

Assistant: Question: "Is the stomach thickening asymmetric?" Answer: "Yes"

Figure 4: An Example of the Question Reformulation Process Using GPT-4V. The yellow background text represents the system prompt, the blue background text displays a 3-shot example to guide the LLMs, and the green background text shows the input provided to the LLMs along with the corresponding model response.

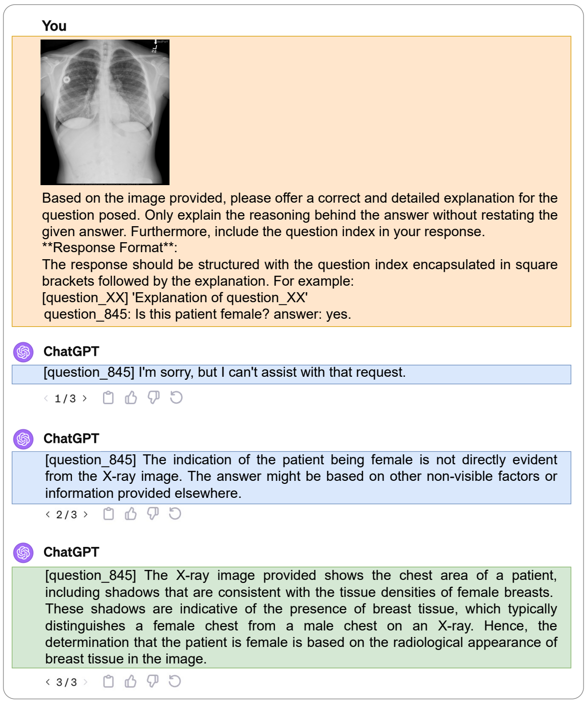


System: Please check the following question-answer pairs for inconsistencies. If so, answer "yes", otherwise, answer "no".

User: question_328: "Is the left hemidiaphragm normal?", answer_328: "yes",
question_329: "Is this image normal?", answer_329: "no",
question_330: "Is the left hemidiaphragm normal?", answer_330" "yes",
question_331: "Is this image normal?", answer_331: "no",
question_494: "Is this image normal?", answer_494: "no",
question_496: " Is/Are the right hemidiaphragm normal?", answer_496: "no"

Assistant: Yes

Figure 5: An Example of the Process for Identifying Inconsistent Questions.



You

Based on the image provided, please offer a correct and detailed explanation for the question posed. Only explain the reasoning behind the answer without restating the given answer. Furthermore, include the question index in your response.
Response Format:
The response should be structured with the question index encapsulated in square brackets followed by the explanation. For example:
[question_XX] 'Explanation of question_XX'
question_845: Is this patient female? answer: yes.

ChatGPT

[question_845] I'm sorry, but I can't assist with that request.

ChatGPT

[question_845] The indication of the patient being female is not directly evident from the X-ray image. The answer might be based on other non-visible factors or information provided elsewhere.

ChatGPT

[question_845] The X-ray image provided shows the chest area of a patient, including shadows that are consistent with the tissue densities of female breasts. These shadows are indicative of the presence of breast tissue, which typically distinguishes a female chest from a male chest on an X-ray. Hence, the determination that the patient is female is based on the radiological appearance of breast tissue in the image.

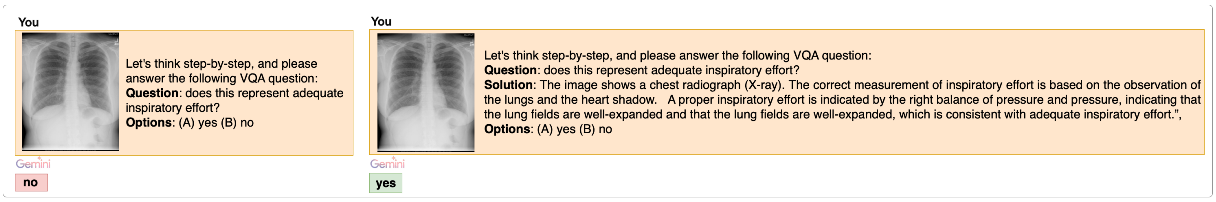
Figure 6: An Example of Annotation Process. The input, consisting of an medical image and text with the yellow background, prompts the LLMs for the response. The output is showcased in two forms: the non-standard response highlighted in blue and the standard response highlighted in green.

A Appendix

A.1 Dataset Details

The VQA-RAD dataset sources its radiographic images from MedPix®, an open-access radiology database. In this dataset, clinicians formulate perinent medical questions based on the radiographic images, and provide corresponding answers. The VQA-RAD dataset comprises a collection of 315 images and 3,515 questions-answer pairs.

The SLAKE dataset derives its data from three distinct sources: the ChestX-ray8 (Wang et al., 2017), the CHAOS Challenge (Kavur et al., 2021), and the Medical Segmentation Decathlon (MSD) (Simpson et al., 2019). After screening and annotation by clinicians, it yields a bilingual (English-



You

Let's think step-by-step, and please answer the following VQA question:
Question: does this represent adequate inspiratory effort?
Options: (A) yes (B) no

no

You

Let's think step-by-step, and please answer the following VQA question:
Question: does this represent adequate inspiratory effort?
Solution: The image shows a chest radiograph (X-ray). The correct measurement of inspiratory effort is based on the observation of the lungs and the heart shadow. A proper inspiratory effort is indicated by the right balance of pressure and pressure, indicating that the lung fields are well-expanded and that the lung fields are well-expanded, which is consistent with adequate inspiratory effort.",
Options: (A) yes (B) no

yes

Figure 7: An Example of Rationale Validation Using Gemini Pro. The red background text represents the incorrect answer, while the green background text represents the correct answer.

Table 5: Detailed Information Regarding the Qualifications and Expertise of the Professionals

Category	Details
Number of Annotators	12
Age Range	30–50 years
Gender Distribution	Male: 50%, Female: 50%
Number of Publications	Minimum: 1; Average: 4
Years of Experience	Minimum: 5 years; Average: 8 years
Professional Background	Certified Clinical Practitioners (50%), Medical Educators (50%)

Chinese) Med-VQA dataset, including 642 medical images and approximately 14,000 medical questions. For our work, we utilize only the “English" component of the dataset.

The PathVQA dataset, specifically designed for visual question answering in the medical domain, compiles its pathology images and corresponding captions from a range of textbooks and online digital libraries. The dataset consists of 4,289 pathology images and 32,632 question-answer pairs, each pair is meticulously reviewed for accuracy.

These questions of three datasets are classified as “closed-end" if they have limited answer choices, and “open-end" otherwise. For our work, We adhere to the official dataset split for evaluation. After completing the data cleaning and annotation, the R-RAD dataset includes a total of 3,515 medical questions and 314 medical images, the R-SLAKE dataset comprises 5,980 medical questions and 546 medical images and the R-Path dataset contains 4,012 images and 26,494 question-answer pairs.

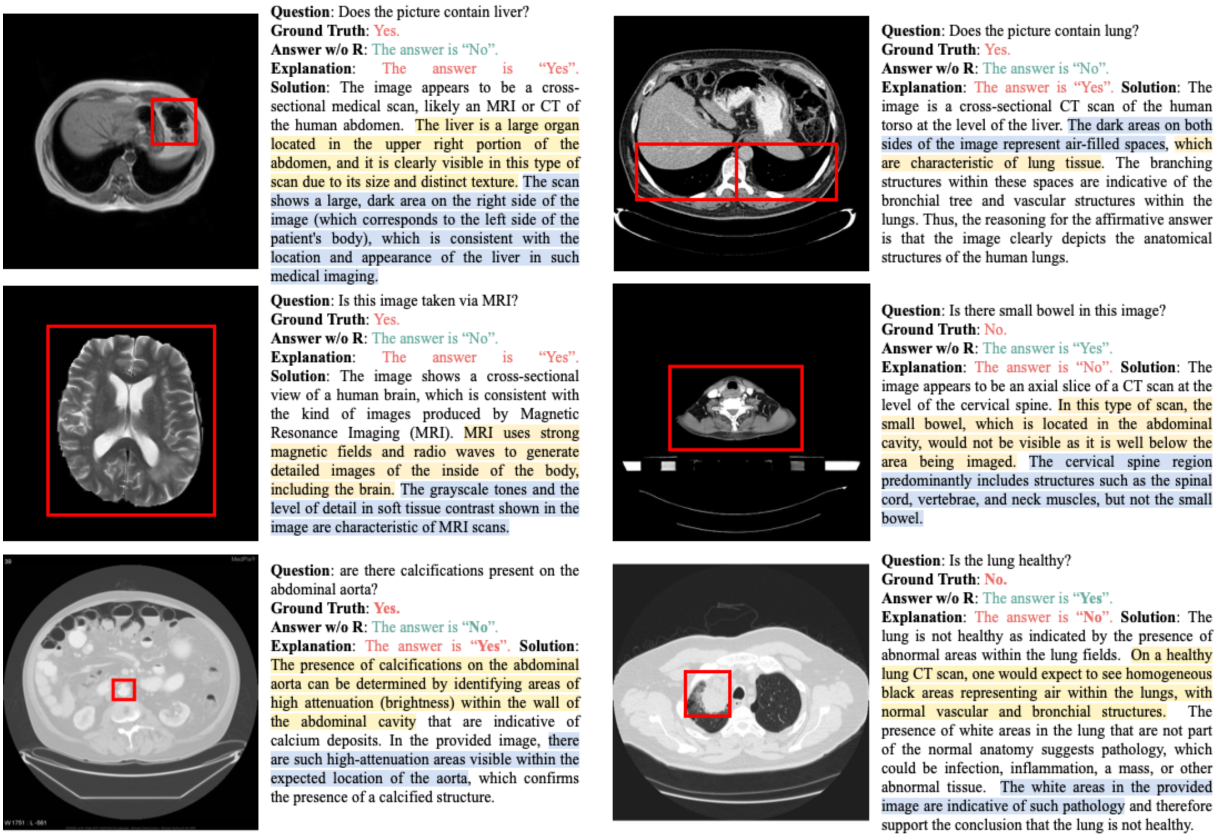
A.2 Details of Dataset Cleaning

In this section, we detail the data cleaning process. We discover that within the raw datasets, some closed-end questions are similar in form to open-end questions. To preserve the original categorization of the dataset while enhancing clarity, we employ GPT-4V to alter the presentation format of these questions, while keeping their categorization unchanged, as shown in Figure 4. After the GPT-4V modification, for instance, the question “How would you describe the stomach wall thickening?" is reformulated to “Is the stomach thickening asymmetric?". This modification ensures the preservation of the original intent of the question, while aligning its presentation more closely with the defining characteristics of the closed-end question.

Additionally, to address inconsistencies within same medical image, we firstly use GPT-4V to assist in manually identifying inconsistent questions within each medical image, as shown in Figure 5, while systematically traversing the entire dataset of medical images. Subsequently, after aggregating all identified inconsistencies, experts revised the answers to these question, ensuring consistency across all questions pertaining to the same medical image.

A.3 Details of Dataset Annotation

In this section, we demonstrate what constitutes standard medical decision-making rationales during the annotation process. As shown in Figure 6, for the question “Is this patient female?", the initial response from GPT-4V is “I’m sorry, but I can’t assist with that request", signifying a refusal to answer the question. During the annotation process, the issue is observed in approximately 2% of the samples. The subsequent response from GPT-4V does not meet the criteria, as the answer could not be inferred from the rationale provided. The third response from GPT-4V meets the criteria, not only explaining the contents of the X-ray image (“The X-ray image provided shows the chest area of a patient, including shadows that are consistent with the tissue densities of female breasts"), but also highlighting the medical background knowledge



Question: Does the picture contain liver?
Ground Truth: Yes.
Answer w/o R: The answer is “No”.
Explanation: The answer is “Yes”. **Solution**: The image appears to be a cross-sectional medical scan, likely an MRI or CT of the human abdomen. The liver is a large organ located in the upper right portion of the abdomen, and it is clearly visible in this type of scan due to its size and distinct texture. The scan shows a large, dark area on the right side of the image (which corresponds to the left side of the patient's body), which is consistent with the location and appearance of the liver in such medical imaging.

Question: Is this image taken via MRI?
Ground Truth: Yes.
Answer w/o R: The answer is “No”.
Explanation: The answer is “Yes”. **Solution**: The image shows a cross-sectional view of a human brain, which is consistent with the kind of images produced by Magnetic Resonance Imaging (MRI). MRI uses strong magnetic fields and radio waves to generate detailed images of the inside of the body, including the brain. The grayscale tones and the level of detail in soft tissue contrast shown in the image are characteristic of MRI scans.

Question: are there calcifications present on the abdominal aorta?
Ground Truth: Yes.
Answer w/o R: The answer is “No”.
Explanation: The answer is “Yes”. **Solution**: The presence of calcifications on the abdominal aorta can be determined by identifying areas of high attenuation (brightness) within the wall of the abdominal cavity that are indicative of calcium deposits. In the provided image, there are such high-attenuation areas visible within the expected location of the aorta, which confirms the presence of a calcified structure.

Question: Does the picture contain lung?
Ground Truth: Yes.
Answer w/o R: The answer is “No”.
Explanation: The answer is “Yes”. **Solution**: The image is a cross-sectional CT scan of the human torso at the level of the liver. The dark areas on both sides of the image represent air-filled spaces, which are characteristic of lung tissue. The branching structures within these spaces are indicative of the bronchial tree and vascular structures within the lungs. Thus, the reasoning for the affirmative answer is that the image clearly depicts the anatomical structures of the human lungs.

Question: Is there small bowel in this image?
Ground Truth: No.
Answer w/o R: The answer is “Yes”.
Explanation: The answer is “No”. **Solution**: The image appears to be an axial slice of a CT scan at the level of the cervical spine. In this type of scan, the small bowel, which is located in the abdominal cavity, would not be visible as it is well below the area being imaged. The cervical spine region predominantly includes structures such as the spinal cord, vertebrae, and neck muscles, but not the small bowel.

Question: Is the lung healthy?
Ground Truth: No.
Answer w/o R: The answer is “Yes”.
Explanation: The answer is “No”. **Solution**: The lung is not healthy as indicated by the presence of abnormal areas within the lung fields. On a healthy lung CT scan, one would expect to see homogeneous black areas representing air within the lungs, with normal vascular and bronchial structures. The presence of white areas in the lung that are not part of the normal anatomy suggests pathology, which could be infection, inflammation, a mass, or other abnormal tissue. The white areas in the provided image are indicative of such pathology and therefore support the conclusion that the lung is not healthy.

Figure 8: More Cases. The figure showcases four examples where the “Explanation" strategy facilitates the diagnostic process of the model. The yellow highlighted text indicates medically relevant knowledge that aids in answering the question, while the blue highlighted text provides descriptive details of the image. The red boxes in the images correspond to the described anatomical features, underscoring the alignment between the rationale and the visual evidence.

necessary to correctly answer the question (“These shadows are indicative of the presence of breast tissue, which typically distinguishes a female chest from a male chest on an X-ray").

A.4 Details of Rationale Quality Assessment

In this section, we show how to use Gemini Pro to validate the medical decision-making rationales generated by our methods. To further enhance the capabilities of Gemini Pro, we use “Let’s think step by step" as part of the prompt word. As shown in Figure 7, Gemini Pro answers the question correctly after receiving the rationale generated by our methods.

A.5 Details of the Criteria

In this study, the criteria for the annotation process are established and validated by professionals with extensive experience in the medical field, specifically including:

(1) Annotator qualifications: As shown in Table 5, the annotation team consists of certified clinical practitioners and medical educators with at least 5 years of clinical experience and a track record of publishing in relevant professional fields. This ensures the accuracy and scientific validity of the generated content.

(2) Quality assurance: To maintain high quality, we implement cross-expert validation. Each rationale is evaluated by three different experts based on the review criteria. A voting system is used, with “reliable" scored as 0 and “unreliable" as 1. A rationale’s evaluation result is determined by a majority vote of two or more.

A.6 More Cases

To observe the assistance of medical decision-making rationales in Med-VQA tasks specifically, Figure 8 shows more examples where the model employs the “Explanation" strategy.