MedThink: Explaining Medical Visual Question Answering via Multimodal Decision-Making Rationale

Anonymous ACL submission

Abstract

Medical Visual Question Answering (MedVQA) provides language responses to image-based medical inquiries, facilitating more accurate diagnoses. However, existing MedVQA methods lack interpretability and transparency. To address this, we introduce a semi-automated annotation process and create new benchmark datasets, R-RAD and R-SLAKE, incorporating multimodal language models and human annotations. Additionally, we develop a framework, MedThink, to fine-tune lightweight generative models with medical decision-making rationales. This framework employs three distinct strategies to generate decision outcomes and corresponding rationales, effectively showcasing the medical decision-making process during reasoning. MedThink achieves 83.5% accuracy on R-RAD and 86.3% on R-SLAKE, outperforming current baselines. Datasets and code will be released.

1 Introduction

The Medical Visual Question Answering (MedVQA) task uses images to answer medical queries, aiding diagnosis, and reducing misdiagnosis risk (Hasan et al., 2018; Liu et al., 2023b; Zhan et al., 2020). However, existing MedVQA faces two challenges. First, datasets lack the decision-making process between questions and answers, hindering model interpretability (Lau et al., 2018; Liu et al., 2021b; Lu et al., 2022; Liu et al., 2023c; Lai et al., 2024a). However, manual rationale annotation for decision-making process is time-consuming and requires in-depth understanding of medical knowledge (Litjens et al., 2017; Liu et al., 2023a). Second, models need to resolve MedVQA tasks quickly, accurately, and interpretably. Current methods use retrieval, contrastive, or classification objectives (Nguyen et al., 2019; Zhang et al., 2022; Liu et al., 2021a; Eslami et al., 2023). Multimodal large language models (MLLMs) handle text and image inputs but are impractical due to high costs and latency (Nori et al., 2023; Lai et al., 2024b; OpenAI, 2023; Team et al., 2023).

In this paper, we introduce new benchmark datasets and novel solutions for MedVQA. We design a semi-automated annotation method leveraging the powerful inference capabilities of MLLMs, creating the R-RAD and R-SLAKE datasets with Medical Decision-Making Rationales. Besides, we develop our framework, MedThink, to fine-tune T5-base generative models (Raffel et al., 2020), to output decision outcomes and rationales, proposing three generative modes: "Explanation", "Reasoning", and "Two-Stage Reasoning". MedThink show 83.5% accuracy on R-RAD and 86.3% on R-SLAKE, improving over PubMedCLIP (Eslami et al., 2023) by 4.0% and 3.8%. Ablations on different large language models (LLMs), such as GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023), further validate MedThink. Our contributions are as follows:

- We develop a semi-automated process for annotating MedVQA data with decision-making rationale. To the best of our knowledge, the R-RAD and R-SLAKE datasets represent the first multimodal MedVQA benchmark datasets that encompass rationales for answers.
- We propose a lightweight framework with three answering strategies, enabling faster and more accurate MedVQA with enhanced interpretability.
- We conduct extensive experiments and ablations that demonstrate the usefulness of the R-RAD and R-SLAKE datasets and superiority of our method.

2 Methodology

2.1 Dataset Collection

We establish two datasets, R-RAD and R-SLAKE, based on VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021b). VQA-RAD, from MedPix®, contains 315 images and 3,515 questions, split into closed-end" and open-end" categories. We follow the official dataset split for evaluation. The

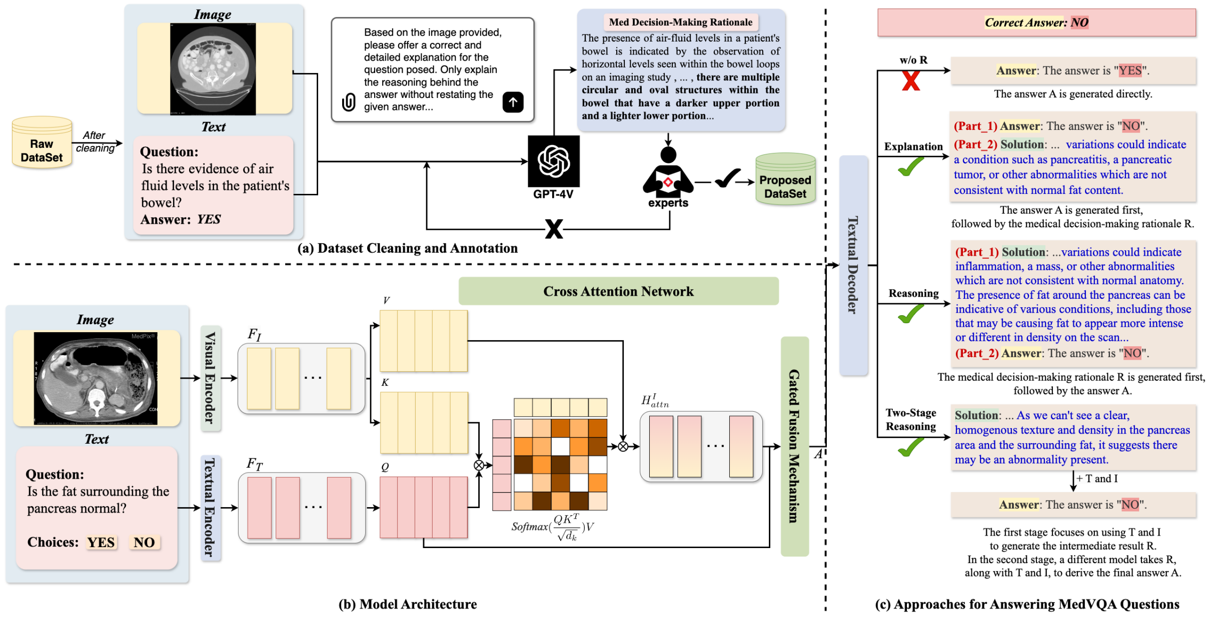


Figure 1: Overview of MedThink. **(a)** Data annotation. **(b)** Model architecture. **(c)** Reasoning strategies.

SLAKE dataset, from ChestX-ray8 (Wang et al., 2017), CHAOS Challenge (Kavur et al., 2021), and Medical Segmentation Decathlon (MSD) (Simpson et al., 2019), contains 642 medical images and around 14,000 questions. We use only the "English" component and follow the original split. The datasets we used are rigorously desensitized.

After cleaning and annotation, R-RAD has 3,515 questions and 314 images, and R-SLAKE has 5,980 questions and 546 images. Both datasets include open-ended and closed-ended questions, with statistics in Table 3 in the Appendix.

2.2 Dataset Cleaning and Annotation

We integrate GPT-4V (OpenAI, 2023) into SLAKE and VQA-RAD data cleaning and annotation to streamline workflows. GPT-4V identifies errors for expert review. After cleaning, GPT-4V generates medical decision-making rationales (Figure 1 (a)), enhancing model reasoning without revealing answers. Fixed prompts guide GPT-4V, with domain experts validating and regenerating rationales as needed. If unsuccessful after three attempts, an expert creates it manually. We select physicians with clinical experience as domain experts to ensure the professional and accurate annotation of data. Recognizing the diversity of opinions among physicians, we establish review criteria to guide our annotation process: a) The rationale generated by GPT-4V must enable experts to deduce the correct answer to the question. b) The rationale should be free of common sense and medical errors, and directly related to the question.

2.3 Dataset Analysis

We segment the rationales into words, excluding common stop words. The word cloud (Figure 3) highlights terms like "brain", "chest", "lung", "located", "transverse", and "density", reflecting medical knowledge and enhancing AI performance in MedVQA. The rationale length distribution (Figure 4) ranges from 60-110 words for organ-related questions, indicating balanced annotations.

2.4 Dataset Annotation Reliability

We validate R-RAD/R-SLAKE annotations through expert verification of GPT-4 rationales (Figure 1(a)) and test their reliability. Using an answer stage (Two-Stage Reasoning strategy in Figure 1(c)) with GPT4 rationale, we achieve over 99% test set accuracy, which confirms the reliability and validity of annotations.

2.5 Problem Formulation

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 samples. The goal of the MedVQA is to develop a function $f(\cdot)$ that generates textual answers to medical questions:

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

Here, I denotes the medical image (X-ray, CT, or MRI). T is the natural language question about I. The model output $f(\cdot)$, $\{A, R\}$, includes: A (the predicted answer to T), and R (the medical decision-making rationale), explaining A by detailing the model's processing of I and T.

Loss Function. Given the input $X = \{I, T\}$, the model f is trained to maximize the likelihood of

Table 1: Accuracy (%) Comparison on Closed-End Questions in R-RAD and R-SLAKE. Gray and red backgrounds highlight established methods and MedThink.

Methods	**R-RAD**	**R-SLAKE**
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
w/o R	79.0	82.5
Reasoning	73.9	80.8
Two-Stage Reasoning	80.5	79.1
Explanation	**83.5**	**86.3**

predicting the target output $Y = \{A, R\}$. The loss function, primarily the negative log-likelihood of correctly predicting tokens in Y, is:

$$L = -\sum_{n=1}^{N} \log p(Y_n|X, Y^{1:n-1}), \quad (2)$$

where N is the number of tokens in Y, and $p(Y_n|X, Y^{1:n-1})$ is the conditional probability of predicting the n-th token in Y.

Model Architecture. The model architecture comprises five components (Figure 1 (b)): TextualEncoder, VisualEncoder, Cross Attention Network, Gated Fusion Network, and TextualDecoder.

The TextualEncoder converts the input question T into the textual feature $F_T \in \mathbb{R}^{n\times d}$, and the VisualEncoder transforms the input medical image I into vision features $F_I \in \mathbb{R}^{m\times d}$: $F_T = \text{TextualEncoder}(T)$, $F_\text{I} = \text{VisualEncoder}(I)$, where n is the text length, d is the hidden dimension, and m is the number of image patches.

The Cross-Attention Network computes the attention-guided visual feature $H^I_{attn} \in \mathbb{R}^{n\times d}$:

$$H^\text{I}_\text{attn} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \quad (3)$$

where Q, K, and V are derived from F_T and F_I.

The Gated Fusion Mechanism combines F_T and H^I_{attn}, with fusion coefficient λ determined by:

$$\lambda = \text{Sigmoid}(W_l F_T + W_v H^\text{I}_\text{attn}), \quad (4)$$

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

$$F_\text{fuse} = (1-\lambda)\cdot F_\text{T} + \lambda \cdot H^\text{I}_\text{attn}, \quad (5)$$

where W_l and W_v are model parameters. Finally, F_fuse is fed into the TextualDecoder to generate the output A, R:

$$A, R = \text{TextualDecoder}(F_\text{fuse}), \quad (6)$$

Table 2: Impact of Medical Decision-Making Rationales on the Accuracy (%) of Gemini Pro for MedVQA on Closed-End Questions in the R-RAD and R-SLAKE Datasets.

Strategy	R-RAD	R-SLAKE
w/o R	73.2	72.8
w/ Reasoning	76.5(+3.3)	77.6(+4.8)
w/ Two-Stage Reasoning	79.4(+6.2)	77.9(+5.1)
w/ Explanation	**82.0** (+8.8)	**81.3** (+8.5)

Three Generation Strategies. To investigate the impact of the medical decision-making rationale on model performance in MedVQA, we present three generation strategies. These strategies guide the model in generating outputs that reflect different orders of the medical decision-making rationale. The strategies are categorized as "Explanation", "Reasoning" and "Two-Stage Reasoning", as shown in Figure 1 (c).

In the "Explanation", the answer A is generated first, followed by the rationale R. In the "Reasoning", R is generated before A. The "Two-Stage Reasoning" uses a phased approach with two independent models. The first stage uses the medical question T and image I to generate the rationale R. In the second stage, a model with different weights but the same architecture uses R, T, and I to derive the answer A.

3 Experiments

3.1 Setting

In this paper, UnifiedQA (Khashabi et al., 2020) serves as TextualEncoder(·) and TextualDecoder(·), while DETR (Carion et al., 2020) is VisualEncoder(·). We evaluate "Explanation", "Reasoning", and "Two-Stage Reasoning" on R-RAD and R-SLAKE datasets, comparing with PubMedCLIP (Eslami et al., 2023), MMBERT (Tiong et al., 2022), MEVF (Nguyen et al., 2019), BAN (Kim et al., 2018), SAN (Yang et al., 2016), and MFB (Yu et al., 2017), and assess rationale quality.

For MedThink model, the learning rate is set at 5e-4, with 300 epochs for R-SLAKE and 150 epochs for R-RAD. "Two-Stage Reasoning" uses phased fine-tuning: first with these parameters, then with a learning rate of 5e-5 for 20 epochs.

3.2 Main Results

Our performance evaluation is divided into two parts: closed-end and open-end questions. Closed-end questions, structured as multiple-choice with a single definitive answer, are assessed using accuracy, as shown in Table 1. Open-end questions allow for a range of answers, making precise matching difficult. Thus, we use text generation metrics such as Rouge and BLEU to evaluate performance, as exhibited in Table 4 in the Appendix.

For closed-end questions, the performance of “Explanation" outperforms the “Reasoning" and “Two-Stage Reasoning", achieving 83.5% accuracy on R-RAD and 86.3% on R-SLAKE, surpassing the state-of-the-art PubMedCLIP model by 4.0% and 3.8%. Besides, we evaluate MLLMs on the R-RAD and R-SLAKE datasets, presenting zero-shot and fine-tuning results in Table 6 in the Appendix. It can be obseved that MedThink remains competitive with fine-tuned MLLMs.

To observe the role of medical decision-making rationales in MedVQA, Figure 2 shows examples using the "Explanation". For instance, when asked “Are the small bubbles of air seen abnormal?", the model without rationales responds incorrectly. In contrast, MedThink with the “Explanation" method correctly answers and explains, “The small bubbles appear very dark or black on the CT scan" and adds, “there are multiple areas of radiolucency, consistent with small bubbles of air". This highlights the supportive role of rationales in guiding the model to answer MedVQA questions accurately.

3.3 Generalizability Across Diverse Datasets

To evaluate generalizability, additional tests are conducted using other datasets, VQAMed-2019 and PathVQA. Table 5 in the Appendix shows that incorporating rationales improve accuracy by 4.6% and 1.2%.

3.4 Rationale Quality Assessment

We compare “Explanation", “Reasoning", and “Two-Stage Reasoning" strategies for closed-end questions on R-RAD and R-SLAKE datasets, with “w/o R" as a control excluding medical decision-making rationales. Table 1 shows accuracy changes on R-RAD by 4.5%, -5.1%, and 1.5%, and on R-SLAKE by 3.8%, -1.7%, and -3.4%, respectively.

Employing Gemini Pro, we evaluate the assistance of medical decision-making rationales. Initially, Gemini Pro’s input includes only medical queries and image. Subsequently, rationales generated by our strategies are incorporated to assist Gemini Pro. Following the self-consistency protocol (Wang et al., 2022), where Gemini Pro answers each question five times, Table 2 shows initial accuracies of 73.2% on R-RAD and 72.8% on R-SLAKE. The “Explanation" improves accuracy by 8.8% on R-RAD and 8.5% on R-SLAKE.

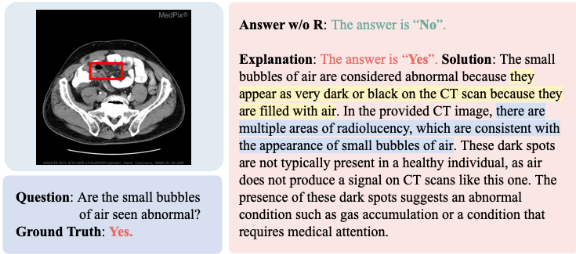


Figure 2: Illustration of enhancing MedVQA with decision-making rationales. The figure shows “Explanation" aiding diagnostics. Yellow highlights medical knowledge, blue image details, and red anatomical features in red box, aligning rationale with visual evidence.

3.5 Zero-shot Capability

We evaluate the model’s zero-shot capability on the VQA-Med-2019 dataset. Pretrained with “Explanation" on R-RAD, MedThink achieves 62.34% accuracy on the closed-end set, close to the highest reported accuracy of 62.40%. Our focus, however, is on integrating medical decision-making rationale into the MedVQA and developing a comprehensive data production and model training system.

3.6 Ablation Study

To validate the impact of MLLMs selection and expert annotation, we design three variations for annotating closed-end questions in the R-RAD dataset: Gemini Pro without expert input, GPT-4V without expert input, and GPT-4V with expert input. Results in Figure 5 in the Appendix show GPT-4V enhances MedVQA performance over Gemini Pro, aligning with prior research (Qi et al., 2023; Fu et al., 2023).

Expert annotations improve data quality. Table 1 shows the baseline (w/o R) at 79.0% accuracy. Models with rationales varied: “Gemini Pro" reached 79.4%, “GPT-4V w/o Expert" 76.5%, and “GPT-4V w/ Expert" 73.9%. Longer rationales may cause hallucinations, shifting focus from the answer to the rationale (Lu et al., 2022).

4 Conclusion

In this paper, we present a generative model-based framework MedThink for MedVQA and construct the R-RAD and R-SLAKE datasets with intermediate reasoning steps to address black-box decision-making. Experimental results show MedThink clarifies the medical decision-making process and significantly enhances performance. Future research will further explore generative models tailored for clinical settings and better evaluate MedVQA models’ performance in open-ended scenarios.

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

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

Aisha Al-Sadi, Bashar Talafha, Mahmoud Al-Ayyoub, Yaser Jararweh, and Fumie Costen. 2019. Just at imageclef 2019 visual question answering in the medical domain. In *CLEF (working notes)*.

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*.

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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer.

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*.

Bumjun Jung, Lin Gu, and Tatsuya Harada. 2020. bumjun_jung at vqa-med 2020: Vqa model based on feature extraction and multi-modal feature fusion. In *CLEF (Working Notes)*.

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.

Zhixin Lai, Jing Wu, Suiyao Chen, Yucheng Zhou, Anna Hovakimyan, and Naira Hovakimyan. 2024a. Language models are free boosters for biomedical imaging tasks. *arXiv preprint arXiv:2403.17343*.

Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024b. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. *arXiv preprint arXiv:2403.13335*.

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.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

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, 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.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. 2023. Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases. *arXiv preprint arXiv:2312.15011*.

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.

Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. 2021. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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.

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.

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. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

A Appendix

A.1 Related Work

1. MedVQA

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

Previous MedVQA solutions (Nguyen et al., 2019; Al-Sadi et al., 2019; Jung et al., 2020; Do et al., 2021; Sharma et al., 2021; Zhang et al., 2022) rely 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) shift 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 MedVQA 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.

This incongruity underscores the necessity for a MedVQA approach that is more adaptive and reflective of the complexities inherent in medical diagnostics and decision-making. In this paper, we redefine MedVQA as the generative task. Within the actual medical environment, when faced with open-ended queries, our proposed MedVQA model can still generate informed responses based on the medical knowledge it learns.

Table 3: Details of Datasets: Distribution of Images and Questions in the R-RAD and R-SLAKE 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

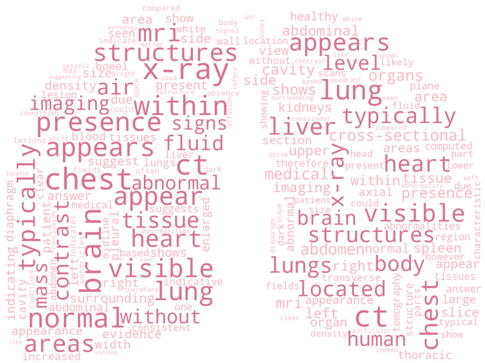


Figure 3: Word Cloud Representation of High-Frequency Terms in Medical Decision-Making Rationales from the R-RAD (left) and R-SLAKE (right) Datasets.

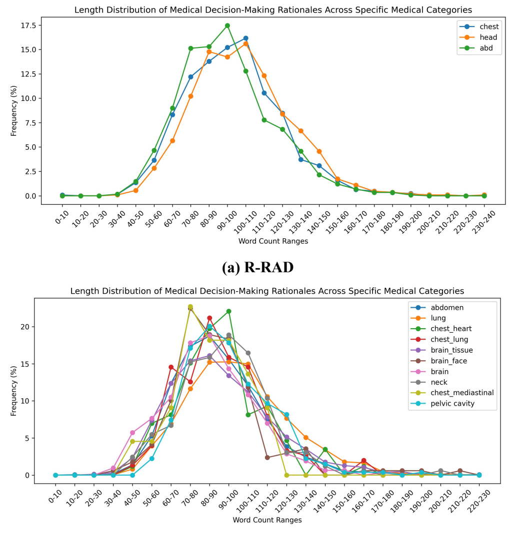
2. The Thought Chain

Recently, natural language processing (NLP) is significantly transformed by language models (Raffel et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2023).

To further enhance the reasoning capabilities of language models, prior works (Cobbe et al., 2021; Wei et al., 2022) incorporate 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., 2023) propose multimodal reasoning methods for VQA. These

Table 4: Performance of Our strategies on Open-End Questions in the R-RAD and R-SLAKE Datasets.

Dataset	Strategy	Rouge-1	Rouge-2	Rouge-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
R-RAD	Explanation	**50.2**	20.2	**29.5**	**38.3**	**22.9**	**14.0**	8.8
	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
R-SLAKE	Explanation	53.1	22.7	31.7	39.2	24.1	15.4	9.9
	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**



(a) R-RAD

(b) R-SLAKE

Figure 4: Length Distribution of Medical Decision-Making Rationales in the R-RAD and R-SLAKE Datasets. The x-axis denotes length ranges, while the y-axis represents the frequency of medical decision-making rationales across various medical categories.

Table 5: Accuracy (%) of strategies on Closed-End Questions in the VQAMed-2019 and PathVQA Datasets.

Strategies	VQAMed-2019	PathVQA
w/o R	68.8	86.0
w/Reasoning	64.1	83.1
w/Explanation	**73.4**	87.0
w/Two-Stage Reasoning	68.8	**87.2**

methods, commonly referred to as “the thought chain", introduces intermediate steps to assist the model in reasoning. In this paper, we present the “Medical Decision-Making Rationale" and apply it to the MedVQA task. We anticipate that MedVQA systems, equipped with the “Medical Decision-Making Rationale", will not only offer support in medical decision-making but also elucidate the underlying rationales behind these decisions.

A.2 More Cases

To observe the assistance of medical decision-making rationales in MedVQA tasks specifically, Figure 6 shows more examples where the model

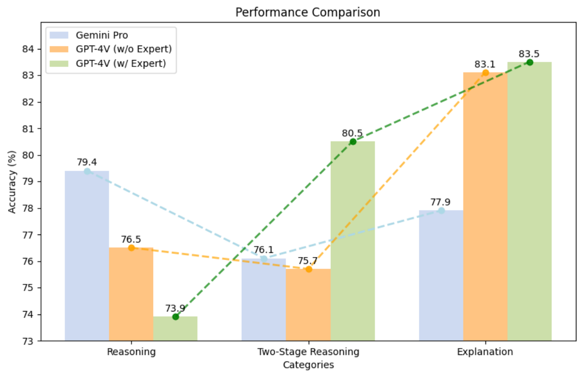


Figure 5: Impact of Selecting MLLMs and Expert Participation during the Data Annotation Process on Med-VQA Task Accuracy (%).

Table 6: Accuracy(%) Comparison of Methods with open MLLMs on Closed-End Questions in the R-RAD and R-SLAKE Datasets.

Type	Method	R-RAD	R-SLAKE
Zero-shot	LLaVA	59.19	50.24
Zero-shot	Qwen-VL-Chat	47.79	57.69
Finetuning	LLaVA	65.07	63.22
Finetuning	LLaVA-Med (From LLaVA)	84.19	85.34
Finetuning	LLaVA-Med (From Vicuna)	81.98	83.17
Finetuning	**Ours**	**83.50**	**86.30**

employs the “Explanation" strategy.

A.3 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.4 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 8. After the GPT-4V modification, for instance, the question

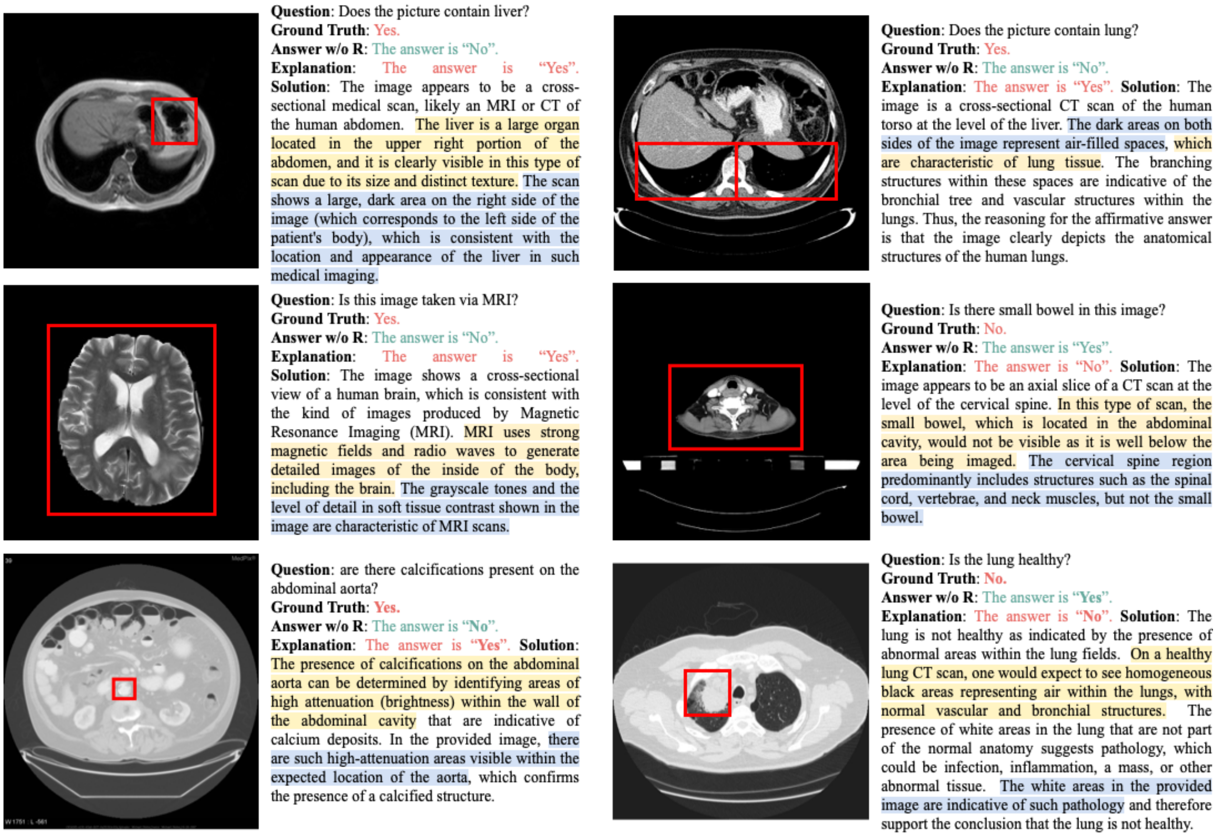


Figure 6: More Cases. The figure showcases four examples where the "Explanation" strategy facilitates the diagnostic process of the model. The yellow highlighted text indicates medically relevant knowledge that aids in answering the question, while the blue highlighted text provides descriptive details 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.

"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 9, 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.5 Details of Dataset Annotation

In this section, we demonstrate what constitutes standard medical decision-making rationales during the annotation process. As shown in Figure 10, 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 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.6 Hallucinations

It is noteworthy that the participation of expert annotations is found to improve data quality. As shown in Table 1 of our manuscript, the baseline (w/o R) on the R-RAD dataset achieves 79.0% ac-

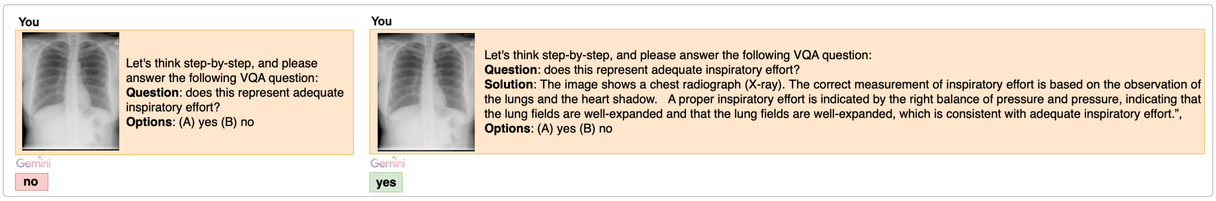


Figure 7: An example of rationale validation using Gemini Pro. The red background text represents the incorrect answer, while the green background text represents the correct answer.

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 8: 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 9: An Example of the Process for Identifying Inconsistent Questions.

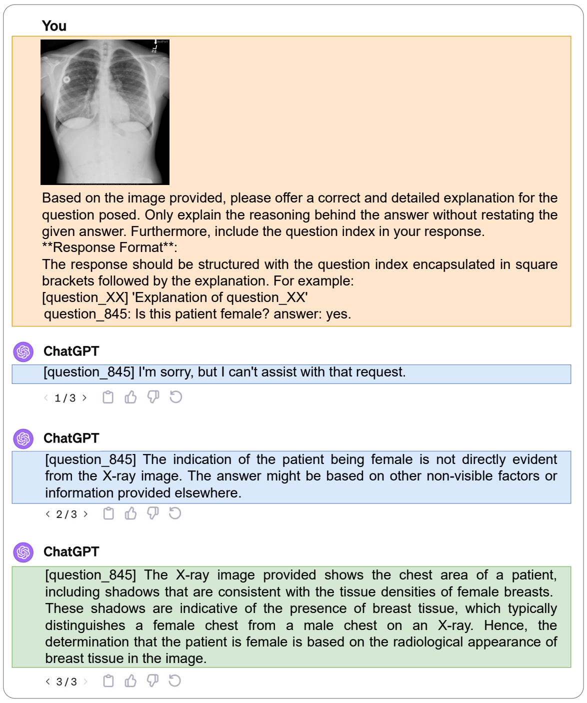


Figure 10: An Example of Annotation Process. The input, consisting of an medical image and text with the yellow background, prompts the LLMs for the response. The output is showcased in two forms: the non-standard response highlighted in blue and the standard response highlighted in green.

curacy. However, the accuracy for models with rationales varied: "Gemini Pro" reached 79.4%, "GPT-4V w/o Expert" achieved 76.5%, and "GPT-4V w/ Expert" recorded 73.9%. This demonstrates that their performance is either close to the baseline or significantly lower. In fact, similar phenomena are observed in general VQA tasks. Researchers speculate that the "Reasoning" strategy may cause severe hallucinations and in the training data, the longer the length of the Rationale, the more it will cause the model to focus on generating the Rationale rather than generating the answer (e.g., (Lu et al., 2022)). Based on our findings from the experiment, we draw the following hypothesis. Shorter rationales from Gemini Pro keep the model's focus on answer generation, reducing hallucinations, so the accuracy of "Gemini Pro" is highest, essentially on par with the baseline. Conversely, longer rationales from GPT-4V(GPT) emphasize rationale generation, leading to increased hallucinations and inferior accuracy. Therefore, the accuracy of "GPT w/ Expert" and "GPT w/o Expert" is significantly lower than that of "Gemini Pro" and the baseline. Among them, the average length of the rationales generated by "w/o Expert" is shorter than that of "w/ Expert", thereby reducing the phenomenon of hallucination. Therefore, the accuracy of "w/o Expert" is higher than that of "w/ Expert". Two-stage Reasoning separates generating rationales and answers, allowing the model to focus on answer generation. Its effectiveness is proven in MMCoT

(Multimodal-CoT).Explanation outputs the Answer first, then the Rationale. This focuses the model on the Answer, with the Rationale explaining it, reducing hallucinations. Two-stage Reasoning and Explanation approaches significantly reduce the impact of hallucinations, hence the results presented in Figure 5.

A.7 Open-end Questions

With open-end questions, different strategies show distinct advantages in Table 4. The Rouge scores, similar to the “Recall", emphasizes the completeness of the generated text, while the BLEU scores, akin to the “Precision", stresses the preciseness of the generated text. The “Explanation" strategy demonstrates higher Rouge and BLEU scores on the R-RAD dataset, with Rouge-1, Rouge-L, BLEU-1, BLEU-2, and BLEU-3 reaching 50.2%, 29.5%, 38.3%, 22.9%, and 14.0%, respectively. The “Two-Stage Reasoning" method showcases higher scores on the R-SLAKE dataset, with Rouge-2, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 at 23.1%, 23.1%, 39.5%, 24.5%, 15.8%, and 10.3%, respectively. The “Reasoning" method maintains robust performance across both the R-RAD and R-SLAKE datasets; on the R-RAD dataset, Rouge-2, BLEU-3, and BLEU-4 reached 20.3%, 14.0%, and 8.9%, respectively, while on the R-SLAKE dataset, Rouge-1, Rouge-L, and BLEU-1 are 53.5%, 32.1%, and 39.5%, respectively. These outcomes highlight the necessity of diverse methods for various types of open-end questions.