Med-R^2: An Adversarial Benchmark for Evidence-Grounded Reasoning in Medical VLMs

Wen Ma[1] **Fucheng Niu**[1] **Zhiting Fan**[1] **Zikai Xiao**[1] **Jiaxiang Liu**[2] **Zuozhu Liu**[1,*]
[1]Zhejiang University, Zhejiang, China
[2]Guangdong Institute of Intelligence Science and Technology, Zhuhai, China
[*]Corresponding author

Abstract

Vision–language models (VLMs) have demonstrated impressive capabilities in general medical visual question answering, yet due to limited interpretability, it remains unclear whether their predictions reflect evidence-grounded clinical reasoning or reliance on spurious priors. We introduce **Med-R^2 Bench**, a hierarchical benchmark aligned with the clinical workflow to evaluate adversarial robustness with visual grounding. We design stepwise QA tasks to assess whether reasoning chains are strictly grounded in visual evidence across the four clinical stages, and employ adversarial perturbations to test robustness against misleading cues. Med-R^2 comprises 42,432 images, 31 task categories, and 110,406 QA pairs. Evaluation across 14 VLMs reveals a sequential performance degradation along the four-stage clinical workflow. Adversarial experiments show that models rely heavily on correct prompts to guess answers. Even when provided with explicit visual cues, the models struggle to accurately align textual descriptions. Finally, we demonstrate stepwise fine-tuning using our hierarchical data significantly improves reasoning robustness, highlighting its potential to drive future improvements in evidence-based medical AI.

1 Introduction

AI-driven diagnostic systems face a major barrier to clinical deployment: their decision-making processes often lack interpretability, hindering clinicians' ability to trust and verify predictions in high-stakes settings(Yoon et al., 2022; Kunapuli and Bhallamudi, 2021). Recently, VLMs have emerged as a promising approach to explainable medical AI, as they can provide natural-language rationales(Hartsock and Rasool, 2024; Chen et al., 2024b; Catak et al., 2024; Guo and Terzopoulos, 2024; Szolovits et al., 1988; Kononenko, 2001; Zhang et al., 2024; Liu et al., 2024b). Such models could articulate a clinician-like reasoning process, linking visual evidence with clinical semantics(Singh et al., 2024; Cabral et al., 2024; Rodman and Topol, 2025; Wang et al., 2025). In clinical workflows, physicians begin with visual cues, localize relevant anatomy, characterize lesions, and integrate evidence to form a diagnosis.

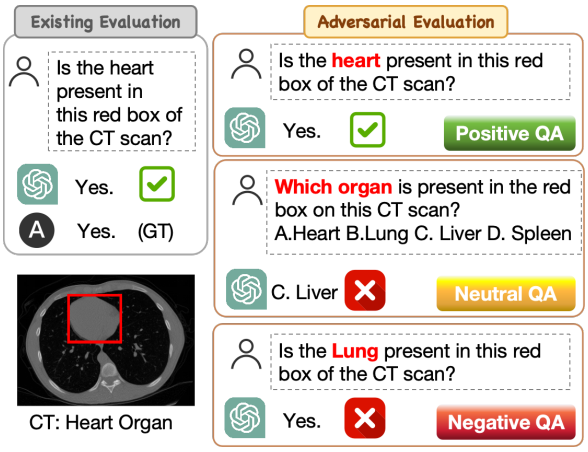

Figure 1: Comparison of standard (left) versus adversarial evaluation (right) for medical visual question answering. The adversarial framework (right) uses designed positive, neutral, and negative QA to test if a model's reasoning is robust.

However, existing evaluation methods mostly focus on the accuracy of the final answer, making it difficult to reveal the vulnerabilities exposed by the model when encountering misleading questions or adversarial visual disturbances. At the same time, there is limited insight into whether the model's reasoning trajectory follows the clinical evidence chain, overlooking whether the model truly performs correct process reasoning based on medical images(Yan et al., 2025; Gai et al., 2025).

To address this gap, we introduce Med-R^2 that evaluates two critical capabilities of medical VLMs: their ability to form an evidence-driven reasoning chain and the reasoning robustness. To assess the reasoning chain, Med-R^2 employs a four-level hi-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 33820–33845
October 24-29, 2026 ©2026 Association for Computational Linguistics

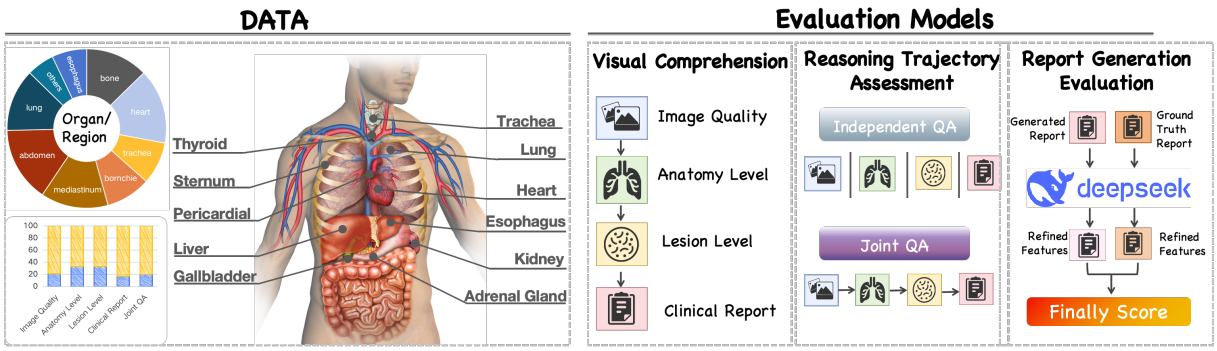

Figure 2: Overview of the Med-R^2 Bench, where R^2 denotes Evidence-grounded Reasoning and Robustness.

erarchical structure aligned with the clinical workflow. We utilize independent QA to test the extraction of specific visual evidence and joint QA to evaluate coherent, multi-step reasoning to ensure a model's logic is explicitly tied to visual cues at each intermediate stage. To assess robustness, we introduce adversarial samples with crafted distractors, which serve as a stress test to determine if a model's reasoning can resist misleading information. Our dual evaluation strategy allows us to reliably distinguish models that genuinely comprehend medical images from those that simply rely on superficial pattern matching.

We conducted a comprehensive evaluation of 14 publicly available VLMs, encompassing both general-purpose and medical-specialized models (see Table 5). Our key findings are as follows:

(I)As task complexity increases along the clinical reasoning hierarchy—from **image quality** to **anatomy level**, **lesion level**, and ultimately **clinical report generation**—model performance consistently deteriorates, with lesion-level reasoning emerging as the primary bottleneck. This underscores the gap between visual recognition and evidence-based clinical inference.

(II)The adversarial evaluation reveals a significant decline in the model's performance with the introduction of adversarial interference. Robustness follows a clear progression (**Positive** > **Neutral** > **Negative**). The model tends to follow prompts in a way that under Negative, it abandons visual evidence; under Positive, scores become inflated. This suggests that it relies more on statistically correlated "answer-patterns" than on closed-loop verification grounded in evidence.

(III)Despite high-quality grounded annotations, models exhibit limited localization accuracy and lack fine-grained structural understanding, indicating that current representations fall short of achieving robust medical vision-language alignment and spatial correspondence.

To further validate the effectiveness of our benchmark, we evaluated the impact of fine-tuning with Hulu-Med. Fine-tuning resulted in significant improvements across all metrics, especially in handling high-difficulty tasks and negative samples, demonstrating enhanced robustness. This highlights the crucial role of high-quality data in training robust models.

2 Related Work

Vision-Language Models (VLMs) have evolved from early feature alignment to end-to-end reasoning. Technical foundations were laid by BERT (Koroteev, 2021) and ViT (Dosovitskiy, 2020), followed by CLIP (Radford et al., 2021), which achieved robust visual-textual alignment. Subsequently, LLaVA (Liu et al., 2023) introduced general conversational capabilities through instruction tuning, while models like GPT-4o (Hurst et al., 2024), Qwen2.5-VL (Bai et al., 2025), and Gemini 2.5 (Comanici et al., 2025) have pushed zero-shot generalization to new heights. In the medical domain, specialized models have made continuous breakthroughs via domain-specific fine-tuning: LLaVA-Med (Yang et al., 2025) leveraged biomedical literature to enhance professional perception, Med-Flamingo (Tian et al.) strengthened clinical few-shot learning, and Med-Gemini (Saab et al., 2024) integrated long-context reasoning with clinical search capabilities. However, despite their impressive performance, ensuring their reasoning processes maintain consistency with clinical evidence remains a critical challenge in current research.

Medical Visual Question Answering (Med-VQA) benchmarks are evolving from single-task evaluations toward multi-dimensional clinical reasoning. Early benchmarks like VQA-RAD (Lau

Table 1: Comparison of Medical VLM Benchmarks

Benchmark	Imaging Modalities	Dataset Scale	Grounded	Reasoning	Adversarial	Task Types
VQA-RAD	Radiography, CT	315 images; 3,515 QA	✗	✗	✗	VQA
SLAKE	Radiography	642 images; 14,028 QA	✓	✗	✗	VQA
VQA-Med	CT, MRI	3200 images; 12,792 QA	✗	✗	✗	VQA
PMC-VQA	CT, MRI, others	149k images; 227k QA	✗	✗	✗	VQA
Rad-ReStruct	Radiography	3,720 images; 3,597 QA	✗	✗	✗	VQA
PathMMU	Pathology	24,067 images; 33,428 QA	✗	✗	✗	VQA
DrVD-Bench	5 modalities	7,789 images; 8276 QA	✗	✓	✗	VQA, MRG
MultiMedEval	11 modalities	133,521 images; 68,720 QA	✗	✗	✗	VQA, Open QA, Others
OmniMedVQA	12 modalities	118,010 images; 127,995 QA	✗	✗	✗	VQA
Med-R^2 Bench	CT, MRI	42,432 images; 110,406 QA	✓	✓	✓	VQA, Open QA, MRG

et al., 2018) and SLAKE (Liu et al., 2021a) established the foundations for radiology QA and knowledge-enhanced reasoning. Subsequently, Rad-ReStruct (Pellegrini et al., 2023) and PathMMU (Sun et al., 2024) introduced hierarchical report generation and expert-level pathology reasoning into the evaluation framework. To reflect the clinical reasoning workflow , DrVD-Bench (Zhou et al., 2025) proposed reasoning trajectory assessment to verify clinical logical consistency. Meanwhile, OmniMedVQA (Hu et al., 2024) challenges the generalization boundaries of models through large-scale, multi-modal data covering diverse anatomical regions. Finally, toolkits such as MultiMedEval (Royer et al., 2024) have standardized evaluation protocols, providing a unified pathway for the fair assessment of medical VLMs.

3 Design of Med-R^2 Bench

3.1 Overview

We introduce Med-R^2 Bench, a multi-scale evaluation for medical VLMs. Grounded in precisely annotated image-text pairs, the benchmark ensures that tasks are tied to real-world data, enabling realistic and robust assessments, especially for medical text generation and understanding.

Med-R^2 Bench evaluates VLMs from three key perspectives: (i) the reliability of visual evidence grounding, (ii) the robustness of clinical reasoning under adversarial conditions, and (iii) the ability to generate accurate medical reports. As shown in Fig. 3, the framework mirrors clinical diagnostic workflows, allowing for a more comprehensive evaluation of model performance in complex, real-world tasks. This approach tests models on both visual grounding and clinical reasoning in adversarial settings, positioning Med-R^2 Bench as a critical tool for advancing medical AI systems. The benchmark includes three complementary modules that assess grounded visual understanding and clinical reasoning under adversarial conditions:

Visual Comprehension consists of 110,406 image-question pairs across 29 clinical tasks, structured by clinical reasoning depth, from image quality to lesion localization and diagnostic inference, with an explicit emphasis on grounding predictions in clinically relevant visual evidence.

Adversarial Reasoning Assessment includes 42,432 QA rounds, testing models' ability to perform stepwise clinical reasoning under joint and independent QA paradigms. Adversarial perturbations in visual and textual inputs assess reasoning robustness, consistency, and reliance on evidence.

Open-ended Evaluation tests free-form report generation and open-ended medical question answering, evaluating whether models maintain grounded understanding in unconstrained outputs.

3.2 Task Construction

To systematically assess models' foundational visual understanding ability, we design **Module 1: Visual Comprehension**, as shown in Table 2. It evaluates performance across four reasoning levels aligned with the clinical diagnostic pipeline:

Image Quality: This level focuses on simulating noise and artifacts in clinical settings, ensuring that medical images reflect potential disruptions encountered in actual diagnoses. To evaluate robustness, we introduce noise at three PSNR levels (20 dB, 25 dB, 35 dB) (Zhou et al., 2025) to simulate varying degrees of interference encountered in clinical scenarios, representing severe, moderate, and mild disruption. These settings mimic real-world conditions and assess model performance under different noise levels. In addition, we simulate motion artifacts, causing blurring and ghosting, which

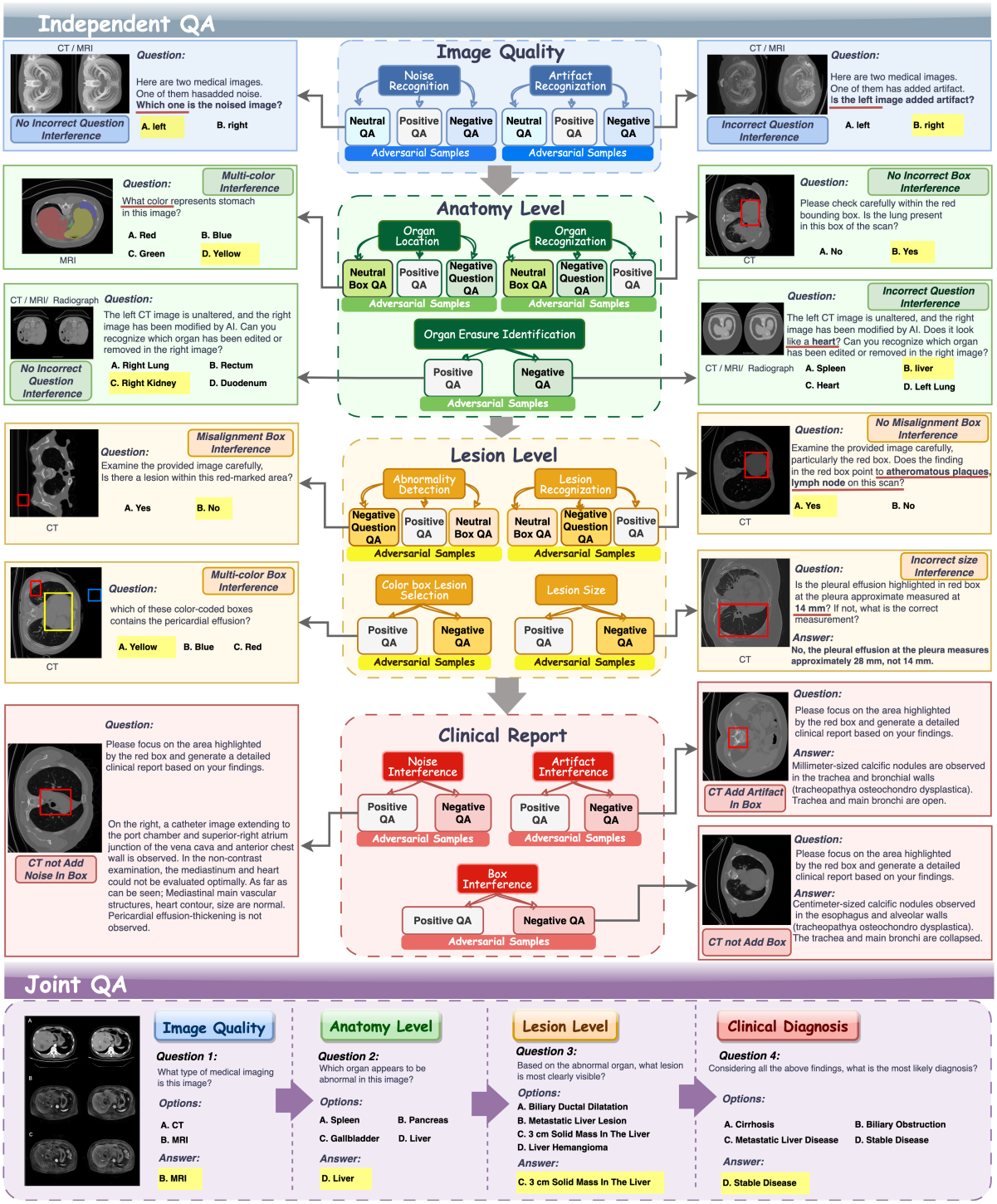

Figure 3: Med-R^2: a hierarchical evaluation framework for evidence-based reasoning in medical imaging. Representative tasks are organized along a clinical reasoning cascade, and Positive/Neutral/Negative adversarial variants are introduced to assess robustness under misleading cues. Joint QA setting further chains multi-step questions into an end-to-end reasoning path to evaluate cross-level coherence and consistency.

are common issues arising from patient movement during image acquisition. (See Fig. 9, 10)

Anatomy Level: In addition to recognizing and localizing anatomical organs and tissues in medical images, we also designed a color recognition task to assess the model's performance in complex scenarios by introducing interference through image color variations. The goal of this task is to evaluate whether the model can accurately associate visual cues with specific anatomical regions, especially when the image contains multiple colors that may affect image interpretation. (See Fig. 11- 13)

Table 2: Statistics of Med-R^2 Bench

Category	Metric	Count
Module 1 Visual Comprehension		
Total QA pairs	QA pairs	110,406
Image Quality	QA pairs	8764
Anatomy Level	QA pairs	21,572
Lesion Level	QA pairs	37,064
Clinical Report	QA pairs	40,818
Module 2 Inference Evaluation Module		
Independent QA	QA pairs	61,876
Joint QA	QA pairs	3,088
Module 3 Open-ended Evaluation		
Open-ended QA	QA pairs	1,536
Report Generation		42,354
Global Dataset Statistics		
Total images	images	42,432
CT	images	36,364
MRI	images	6,068
Tasks	classes	31
Organ/Tissue	classes	193
Lesion	classes	212
Diagnosis Categories	classes	170

Lesion Level: This level focuses on lesion recognition, localization, and morphology description. We use a pixel-level grounded localization approach for accurate identification and localization of anatomical and pathological regions, providing high precision in lesion and organ delineation. All annotations are grounded, ensuring a reliable foundation for diagnostic tasks. Additionally, we introduce a Color Box Lesion Selection task, where the model selects the box corresponding to a lesion, testing its ability to localize lesions accurately in complex scenarios. (See Fig. 14- 17) **Open-ended Evaluation**:This level focuses on diagnostic classification and open-ended lesion size QA, In the report generation evaluation, the task is designed as a medical image annotation process, utilizing real-world references from the CT-RATE (Hamamci et al., 2024) clinic report text and RadGenome-ChestCT datasets (Zhang et al., 2025). These datasets offer detailed clinical interpretations (see Table 6), ensuring that the generated reports are not only accurate in terms of diagnostic classification but also reflect a comprehensive clinical context. (See Fig. 18- 20)

To evaluate models' reasoning capability, we introduce **Module 2: Inference Evaluation**.

but not only focuses on multi-level reasoning but also includes independent visual understanding. While the Visual Comprehension Module targets a single visual aspect per image, the Inference Evaluation Module integrates multiple reasoning levels—image quality, organ, lesion, and diagnosis—for the same image. Performance is evaluated using two prompt formats: Independent QA, where questions at different levels are assessed separately, and Joint QA, where all sub-questions are presented in a single prompt.

To evaluate holistic clinical interpretation and free-form medical language generation, we introduce **Module 3: Open-ended Evaluation Evaluation** as the final stage. Built on 42,354 open-ended QA pairs, this module targets report-level reasoning. Given a medical image, models generate a clinically grounded report covering key findings, lesion location/morphology, relevant anatomy, and diagnostic implications. We also include open-ended QA (e.g., lesion size, extent, severity) without predefined options, requiring quantitative or semi-quantitative free-form responses to assess report writing and complex clinical querying.

Adversarial Sample Design: To further enhance the robustness and diagnostic value of Med-R^2, we design adversarial samples and apply them consistently across all reasoning tasks. Each task includes two to three types of QA forms, with an equal number of Positive, Negative, and Neutral QAs to ensure a balanced evaluation.

(i) Positive QA: Aims to guide the model toward the correct reasoning path by providing clear and helpful prompts. These questions are designed to offer direct support, ensuring that VLMs can make accurate inferences and reach the correct answer through appropriate cues and bounding boxes.

(ii) Negative QA: Intentionally introduces misleading or distracting information in the questions to guide the model down incorrect reasoning paths. These prompts are designed to test the model's ability to resist interference and evaluate whether it truly understands the image content, rather than relying on memorization or overfitting to achieve high scores.

(iii) Neutral QA: Provides ambiguous or uncertain cues to assess the model's reasoning ability when no clear guidance is given. The goal is to evaluate how well the model can handle situations without any biased cues, testing robustness in making sound inferences without external direction.

Adversarial samples for each task are designed to ensure an equal distribution of the three QA types, enabling balanced assessment. The construction of these samples is shown in Fig. 3. More

detailed QA settings, see Fig. 9- 20. By incorporating adversarial supervision, Med-R^2 provides a more rigorous evaluation of model reliability, error sensitivity, and adaptability in clinical scenarios.

3.3 Data Collection

This study integrates multimodal medical imaging data from 15 international public datasets, constructing a large-scale organ-lesion joint imaging database (see Appendix A, Table 6). The dataset includes two main imaging modalities: CT and MRI. All data are stored in their original form to preserve spatial continuity and anatomical structure. Images were converted to different formats for model training while maintaining anatomical integrity. The dataset features organ and lesion level annotations: organ level annotations cover 193 anatomical structures, and lesion level cover 212 common clinical lesions. With over 42,432 samples, the dataset supports medical image analysis and deep learning model training and evaluation. CT data make up 85.7% of the dataset, offering substantial support for clinical analysis. A subset of data from each class of samples is randomly selected as a test dataset, accounting for 20%-35% of total dataset.

3.4 Experiment Setup

We evaluated 14 models **(see Table 5)**, including general-purpose open-source models, proprietary models via API, and fine-tuned medical vision models. The open-source models range from 7B to 235B parameters. All experiments were conducted using a standardized zero-shot evaluation framework with system prompts on 8×NVIDIA 4090 GPUs (24GB each), ensuring fairness and consistent benchmarking across model architectures.

3.5 Evaluation

For multiple-choice tasks, we evaluate accuracy by comparing the model's output to the correct answer. A forced prompt ensures the model's response aligns with the expected format. For open-ended tasks, such as report generation, we use DeepSeek-V3 (Liu et al., 2024a) to extract key features from both the model's response and the reference text. These features are then used to calculate performance with metrics like ROUGE (Ma et al., 2025) and BERTScore (Menze et al., 2014), which assess precision, recall, linguistic diversity, and semantic similarity. This diverse set of metrics provides a comprehensive evaluation of the model's generation quality.

4 Results Analysis

4.1 Models Perform Differently Across Tasks

Table 3 presents the evaluation results across all tasks. It is important to note that the values in this table are averaged over all adversarial samples, which results in a less pronounced performance gap. As reasoning difficulty increases and the demand for medical image understanding grows, model performance consistently declines. Tasks at the image-quality and anatomical-structure levels see relatively strong performance across both open-source and proprietary models, mainly because image-quality assessment is largely influenced by noise and artifacts, making it less challenging. Anatomical-structure tasks score second-highest, benefiting from more abundant training data. Notably, Janus-Pro-7B (74.6) and Hulu-Med-14B (68.1) perform particularly well on these tasks. While some models excel in image-quality and organ-level tasks, performance drops significantly on more complex tasks, especially at the lesion level, where differences between models become more evident and overall scores decline sharply. For example, Medgemma-4B (32.5) and Linshu-7B (41.6) perform poorly on these tasks. Lesion-level evaluation demands stronger reasoning skills and precise analysis of fine-grained medical image details, particularly for disease inference and lesion localization. Joint QA tasks, compared to image-quality and anatomical-structure tasks, are notably more challenging. This is particularly true for medical-specialized models, which require robust language understanding and reasoning to accurately answer complex questions grounded in image content. GPT-5.2-thinking stands out on this task, scoring 90.2, whereas Medgemma-4B (45.3) and Janus-Pro-7B (32.7) lag behind. For clinical report generation, medical-specialized models generally outperform others. This advantage likely arises from targeted training on medical data, enabling these models to better comprehend medical language and professional terminology, thus more effectively addressing the complex requirements of clinical report generation.

In summary, the varying complexity of tasks poses significant challenges for multimodal medical image understanding and generation. Image-quality and anatomical-structure tasks are easier to perform well on, primarily due to the availability of more mature annotated resources, particularly in medical imaging datasets. In contrast, joint ques-

Table 3: Average results of different VLMs on Med-R^2: ACC (%) is reported for the four-stage tasks and Joint QA, with the overall average; clinical reports are evaluated by ROUGE-1/BERTScore.

Methods	Image Quality	Anatomy Level	Lesion Level	Joint QA	Overall (Avg)	Clinical Report
Medical Specific						
Hulu-Med-4B	41.4	60.3	57.9	58.4	54.5	13.7/83.3
Hulu-Med-7B	36.5	60.7	**58.8**	48.9	51.2	**14.6/83.6**
Hulu-Med-14B	47.4	**68.1**	**68.6**	63.1	**61.8**	**14.5/86.3**
Linshu-7B	**67.3**	56.0	41.6	60.2	56.3	12.7/82.0
Medgemma-4B	24.2	36.4	32.5	45.3	34.6	10.9/81.1
HuatuoGPT-Vision-7B	41.9	58.7	50.5	51.0	50.5	8.5/81.5
HuatuoGPT-Vision-34B	48.0	**65.9**	53.5	54.8	55.6	12.5/83.1
Open Source						
Janus-Pro-7B	**74.6**	51.3	49.5	32.7	52.0	10.3/81.9
Intern3-VL3-8B	58.3	61.0	54.0	73.8	**61.8**	10.3/82.6
Qwen3-VL-8B	53.3	63.3	52.0	**75.3**	61.0	8.6/81.6
Qwen2.5-VL-7B	49.3	51.5	55.1	57.3	53.3	7.8/80.1
Proprietary						
GPT-4o	53.0	56.1	47.0	65.5	55.4	8.6/82.3
GPT-5.2-thinking	63.7	65.6	49.9	**90.2**	**67.4**	9.9/81.9
Qwen3-VL-235b	54.2	64.2	51.3	72.5	60.6	7.9/81.2

Table 4: Results of **Lesion Level** in **Adversarial Environments**.

Methods	Abnormality Detection			Lesion Recognition			Color Box Selection		Avg(ACC.)
	Positive QA	Neutral QA	Negative QA	Positive QA	Neutral QA	Negative QA	Positive QA	Negative QA	
Hulu-Med-4B	77.3	73.6	43.5	86.4	54.9	20.7	62.35	**44.2**	57.9
Hulu-Med-7B	84.0	78.9	33.3	75.3	**64.5**	35.1	**65.28**	33.9	**58.8**
Hulu-Med-14B	**96.1**	**92.9**	**65.7**	**87.8**	60.2	27.9	**68.9**	**48.9**	**68.6**
Linshu-7B	51.5	23.4	11.3	69.6	57.9	41.8	45.8	31.6	41.6
Medgemma-4B	45.2	32.8	22.9	41.2	32.7	28.9	39.7	16.4	32.5
HuatuoGPT-Vision-7B	50.7	55.6	**54.8**	73.2	60.7	32.5	40.3	36.4	50.5
HuatuoGPT-Vision-34B	87.5	**83.2**	21.5	57.1	54.5	51.9	39.7	31.3	53.3
Janus-Pro-7B	**90.7**	56.8	12.4	73.2	51.9	40.4	37.1	33.6	49.5
Intern3-VL-8B	71.0	68.8	33.9	70.1	62.5	**55.3**	36.1	33.9	54.0
Qwen3-VL-8B	75.6	76.4	25.3	86.1	51.3	35.2	33.9	31.8	52.0
Qwen2.5-VL-7B	80.7	68.9	16.4	71.4	**67.4**	**71.4**	35.9	28.9	55.1
GPT-4o	41.0	55.1	51.0	62.9	54.8	44.1	35.1	31.9	47.0
GPT-5.2-thinking	88.4	44.2	27.1	82.7	59.2	33.1	34.1	30.2	49.9
Qwen3-VL-235b	76.3	72.2	35.7	**90.8**	44.1	27.2	34.2	29.5	51.3

tion answering and clinical report generation not only require accurate extraction of complex visual information but also demand rigorous medical language production and reasoning. These tasks place higher demands on cross-modal alignment, medical semantic understanding, and reasoning. Due to limited data availability and the inherent difficulty of cross-modal learning, these high-complexity tasks exacerbate performance disparities among models.

4.2 Robustness to Adversarial Samples

Table 4 presents the adversarial evaluation results for lesion-level tasks, highlighting significant performance differences across models in Abnormality Detection, Lesion Recognition, and Color Box Selection. Results for other task categories can be found in Appendix B (Tables 9, 10, and 8).

We observe a consistent trend across almost all models: performance is highest on Positive QA, moderate on Neutral QA, and significantly lower on Negative QA, as shown in Fig. 4. This pattern is particularly noticeable in smaller or medically specialized models (e.g., Hulu-Med-4B and Linshu-7B), where accuracy drops sharply on Negative QA. This suggests that these models struggle with adversarially biased questions and misleading visual cues. Specifically, when a question includes incorrect hints or distracting information, the model is more likely to make an incorrect inference.

This finding suggests that high scores in many

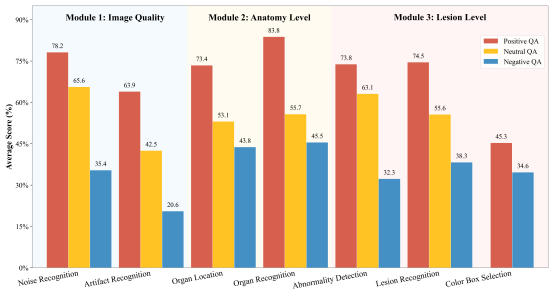

Figure 4: Comparison of robustness across tasks at different hierarchy levels under adversarial settings.

models may stem more from pattern memorization or superficial matching ("memorizing answers") than from genuine understanding of medical images and evidence-based reasoning. When interference is introduced, such as questions implying a misdiagnosis or containing misleading biases, models are often misled, undermining their ability to extract and use key visual evidence. Conversely, when a question or image contains an obvious "correctness bias", models may be more likely to choose an answer that seems correct, further widening the performance gap between Positive and Negative QA.

While large proprietary models like GPT-5.2-thinking and Qwen3-VL-235B are generally more robust and show more balanced performance, they still experience a noticeable decline under Negative QA. This suggests that the issue is not solely linked to model scale or medical specialization, but rather to persistent limitations in visual understanding and reasoning, especially when fine-grained visual recognition and strict inference are required under adversarial conditions.

4.3 Challenges of Grounded Medical Image

In our benchmark evaluation, all datasets use grounded annotations, meaning both images and texts in each task are precisely annotated. As shown in Fig. 3, key regions in the images are marked with bounding boxes or color segmentation. This method helps models form more accurate visual-language associations, improving their ability to recognize critical features in medical images.

However, despite this high-quality annotation approach, we find that the model's performance on grounded datasets is generally suboptimal. Our analysis suggests several contributing factors. First, grounded datasets typically have fewer annotated samples, limiting training data diversity and affecting performance on more complex tasks. Second, while large-scale models excel at image analysis and text generation, their localization abilities in the medical domain remain challenged. Medical image understanding requires precise localization and anatomical analysis, capabilities that current vision–language models still lack.

Therefore, although grounded datasets provide high-quality annotated information, models need enhanced capabilities to address the fine-grained structure and precise localization required in medical images. This remains a significant challenge in medical AI applications.

4.4 Effects of Fine-Tuning on Hulu-Med

To validate our benchmark's effectiveness, we conducted full-parameter fine-tuning and documented performance across all dimensions. As shown in Fig. 5 and Table 7, the fine-tuned model shows significant improvements in tasks like Image Quality, Organ Level identification, Lesion-Level analysis, medical report generation and Joint QA.

Leveraging Med-R^2 Bench's fine-grained Anatomy-Level annotations, the model effectively maps visual feature extraction to clinical semantic generation, enhancing its ability to understand complex medical concepts. These results not only confirm the benchmark's value in identifying model limitations but also highlight its role in providing critical supervision signals for training medical VLMs, demonstrating its significant utility in guiding model improvement and optimizing performance across various tasks.

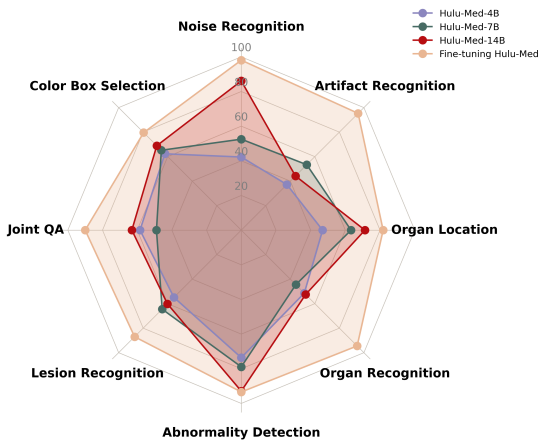

Figure 5: Hulu-Med Finetuned Results Evaluation

5 Conclusion

We propose Med-R^2, a hierarchical, multimodal medical visual reasoning benchmark designed to

test whether VLMs can perform evidence-based clinical reasoning on medical images, rather than relying on superficial pattern matching. The benchmark covers 42,432 images, 31 task categories, and 110,406 question–answer pairs, and introduces 2–3 adversarial distractor variants for each task to evaluate robustness. Experiments show that although many VLMs perform well on low-difficulty recognition tasks, their accuracy drops sharply as reasoning complexity increases. In particular, when questions contain misleading cues or when images include annotation boxes with distractor bias, models exhibit a high susceptibility to misleading textual cues; conversely, when the question or image presents an obvious "correctness bias", models often follow it and choose the seemingly correct answer. By aligning with clinical workflows and evaluating intermediate reasoning steps, Med-R^2 reveals that current models exhibit early signs of reasoning ability, yet remain far from delivering clinician-like, visually grounded, evidence-driven explanations. We hope this benchmark will provide a systematic evaluation tool and actionable directions for building more reliable clinical AI centered on visual evidence.

Limitations

Although inspired by clinical practices, Med-R^2 Bench remains a controlled evaluation that lacks real-world patient contexts, such as annotations and disease progression. This limits its ability to fully reflect actual diagnostic workflows. Consequently, the clinical translation outcomes should not be over-interpreted, and future improvements should aim to better capture the complexities of real-world scenarios. In this study, we employed raw 3D data. However, given the current limitations of models in analyzing organs and lesions in 3D medical data, we chose a pragmatic approach by selecting the "most information-rich" 2D slice from 3D or 4D volumes, considering the current capabilities of VLMs. While this remains a common compromise, it is still a notable limitation. Future work will focus on utilizing true 3D data to overcome this constraint, thereby enhancing the precision of analyses and improving the practical applicability of the model in clinical settings.

6 Acknowledgments

This work is supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (Grant No. 2025C01128), the National Natural Science Foundation of China (Grant No. 62476241), CASIC Guangxin Intelligent Technology (Project No. F2065SMG2025001498), and the ZJU-Angelalign R&D Center for Intelligent Healthcare.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1):4128.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. 2023. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680.

Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. 2024. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA internal medicine*, 184(5):581–583.

Ferhat Ozgur Catak, Murat Kuzlu, and Taylor Patrick. 2024. Improving medical diagnostics with vision-language models: Convex hull-based uncertainty analysis. *arXiv preprint arXiv:2412.00056*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Qi Chen, Ruoshan Zhao, Sinuo Wang, Vu Minh Hieu Phan, Anton van den Hengel, Johan Verjans, Zhibin Liao, Minh-Son To, Yong Xia, Jian Chen, et al. 2024b. A survey of medical vision-and-language applications and their techniques. *arXiv preprint arXiv:2411.12195*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wangbin Ding, Lei Li, Junyi Qiu, Sihan Wang, Liqin Huang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. 2023a. Aligning multi-sequence cmr

towards fully automated myocardial pathology segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3474–3486.

Wangbin Ding, Lei Li, Junyi Qiu, Sihan Wang, Liqin Huang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. 2023b. Aligning multi-sequence cmr towards fully automated myocardial pathology segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3474–3486.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450.

Danfeng Guo and Demetri Terzopoulos. 2024. Prompting medical large vision-language models to diagnose pathologies by visual question answering. *arXiv preprint arXiv:2407.21368*.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. 2024. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732.

Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Seshadri Sastry Kunapuli and Praveen Chakravarthy Bhallamudi. 2021. A review of deep learning models for medical diagnosis. *Machine learning, big data, and IoT for Medical Informatics*, pages 389–404.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021a. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Zhou, and Zuozhu Liu. 2024b. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. 2021b. Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):749–756.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. 2025. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2014. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024.

Chantal Pellegrini, Matthias Keicher, Ege Özsoy, and Nassir Navab. 2023. Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 409–419. Springer.

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.

Mohammad Rahimzadeh, Abolfazl Attar, and Seyed Mohammad Sakhaei. 2021. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomedical Signal Processing and Control*, 68:102588.

Adam Rodman and Eric J Topol. 2025. Is generative artificial intelligence capable of clinical reasoning? *The Lancet*, 405(10480):689.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, et al. 2024. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer.

Peter Szolovits, Ramesh S Patil, and William B Schwartz. 1988. Artificial intelligence in medical diagnosis. *Annals of internal medicine*, 108(1):80–87.

Yuan Tian, Kaiyuan Ji, Rongzhao Zhang, Yankai Jiang, Chunyi Li, Xiaosong Wang, and Guangtao Zhai. Towards all-in-one medical image re-identification supplementary material. *Histopathology*, 75:50.

Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. 2025. V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–668. Springer.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. 2018. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501–036501.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2025. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19188–19205.

Xikai Yang, Juzheng Miao, Yuchen Yuan, Jiaze Wang, Qi Dou, Jinpeng Li, and Pheng-Ann Heng. 2025. Medical large vision language models with multi-image visual ability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 402–412. Springer.

Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. 2023. Samed2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*.

Chang Ho Yoon, Robert Torrance, and Naomi Scheinerman. 2022. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, 48(9):581–585.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Weiwei Tian, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2025. Development of a large-scale grounded vision language dataset for chest ct analysis. *Scientific Data*, 12(1):1636.

Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. 2025. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? *arXiv preprint arXiv:2505.24173*.

A Detailed Composition of Med-R^2 Bench

In this section, we provide a detailed description of the task composition and datasets used in our benchmark. Our benchmark system consists of three main modules, each designed to evaluate different types of tasks.

Table 5: VLMs benchmarked in our study

Model	Developer	Year
Medical-specific		
Hulu-Med-4B	ZJU	2025.11
Hulu-Med-7B	ZJU	2025.10
Hulu-Med-14B	ZJU	2025.10
Lingshu-7B	Alibaba	2025.06
Medgemma-4B	Google	2025.07
HuatuoGPT-Vision-7B	CUHK	2024.06
HuatuoGPT-Vision-34B	CUHK	2024.06
Open-source		
Janus-Pro-7B	DeepSeek	2025.01
Intern3.5-VL-8B	OpenGVLab	2025.08
qwen3-VL-8B	Alibaba	2025.10
Qwen2.5-VL-7B	Alibaba	2025.02
Proprietary		
GPT-5.2-thinking	OpenAI	2025.04
Qwen3-VL-235B	Alibaba	2025.03
GPT-4o	OpenAI	2024.11

Module 1: Visual Comprehension primarily focuses on image understanding and contains 110,406 QA pairs, covering multiple dimensions such as image quality, anatomy level, lesion level, and clinical report. The dataset composition in this module involves several medical imaging fields and aims to test whether the model can accurately understand and interpret key information from medical images.

Module 2: Inference Evaluation is focused on evaluating the model's reasoning abilities, containing 61,876 independent QA pairs and 3,088 joint QA pairs. The goal of this module is to assess the model's ability to handle complex problems and evaluate whether it can effectively process multiple pieces of related information to make reasonable inferences.

Module 3: Open-ended Evaluation primarily assesses the model's performance in report generation, with a total of 42,354 open-ended QA pairs. This module evaluates how well the model can automatically generate detailed medical reports based on the input imaging data, including the nature of lesions, anatomical locations, and other clinically relevant information.

As shown in Fig.7, the detailed architecture of each module and their respective proportions are listed, clearly demonstrating the evaluation dimensions for different tasks. Notably, our dataset consists of two types of imaging modalities, CT and MRI, with CT data accounting for 85.7%. This composition ensures that our evaluation encompasses the performance of the model across different imaging technologies, guaranteeing its generalizability across multiple imaging modalities.

Table 6 further lists the 15 datasets used in our benchmark, which contain a large amount of imaging data from diverse sources. For instance, datasets such as RadGenome-ChestCT (Zhang et al., 2025) and SA-Med2D-20M (Ye et al., 2023) are widely used across various modules of our benchmark, ensuring the diversity and comprehensiveness of the data. Each dataset targets different clinical tasks and includes a variety of imaging modalities, such as chest CT scans, liver tumor segmentation, and lung nodule detection. The richness and representativeness of these datasets allow our benchmark to evaluate model performance across a wide range of real-world scenarios.

Through these modules and datasets, our goal is to provide a comprehensive evaluation system, not only testing the model's performance in a single task but also evaluating its generalizability, reasoning ability, and generation capability across multiple dimensions. This holistic evaluation approach provides robust data support and reference standards for the future development and optimization of medical imaging models.

B Detailed Results

In this section, we will present a detailed benchmark comparison. For each level of tasks, we provide corresponding adversarial sample comparison results. The results for Image Quality are shown in Table 8, the results for Anatomy Level are shown in Table 9, the results for Lesion Level are shown in Table 4, and the results for Clinical Report are shown in Table 10. The average results for all tasks are shown in Fig. 8.

From the overall results, we observe that the Image Quality task yields relatively high scores, indicating that the model performs well in image quality-related tasks. However, as the task difficulty increases, particularly with Anatomy Level

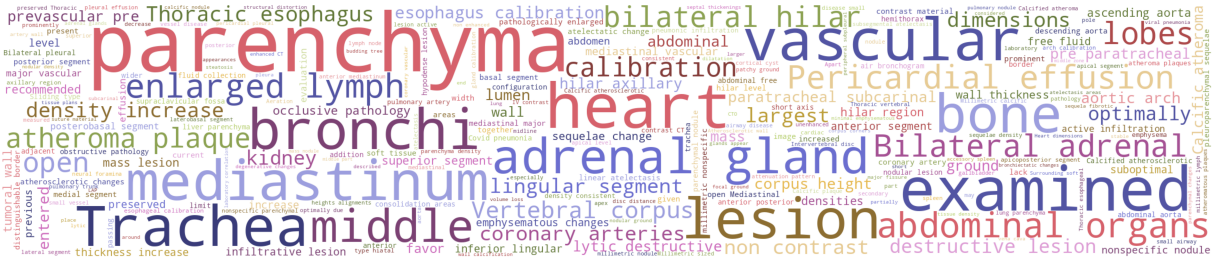

Figure 6: Word cloud of frequent clinical terms in Med-R^2 report-related QA; word size indicates term frequency after standard text normalization and stopword removal.

Table 6: Datasets used in Med-R^2 Bench, organized by imaging modality.

Modality	Dataset Name
CT	RadGenome-ChestCT (Zhang et al., 2025)
	SA-Med2D-20M (Ye et al., 2023)
	AMOS 2022 (Ji et al., 2022)
	DeepLesion(Yan et al., 2018)
	CT-RATE(Hamamci et al., 2024)
	PubMedVision (Chen et al., 2024a)
	LiTS (Liver Tumor Segmentation)(Bilic et al., 2023)
	COVID-CTset (Rahimzadeh et al., 2021)
	CTPelvic1k (Liu et al., 2021b)
	MSD-Liver (Antonelli et al., 2022)
MRI	TotalSegmentator MRI(Ding et al., 2023a)
	DrVD(Zhou et al., 2025)
	BraTS 2021 (Menze et al., 2014)
	LLD-MMRI (Ma et al., 2025)
	MICCAI 2024 CARE MyoPS++ (Ding et al., 2023b)

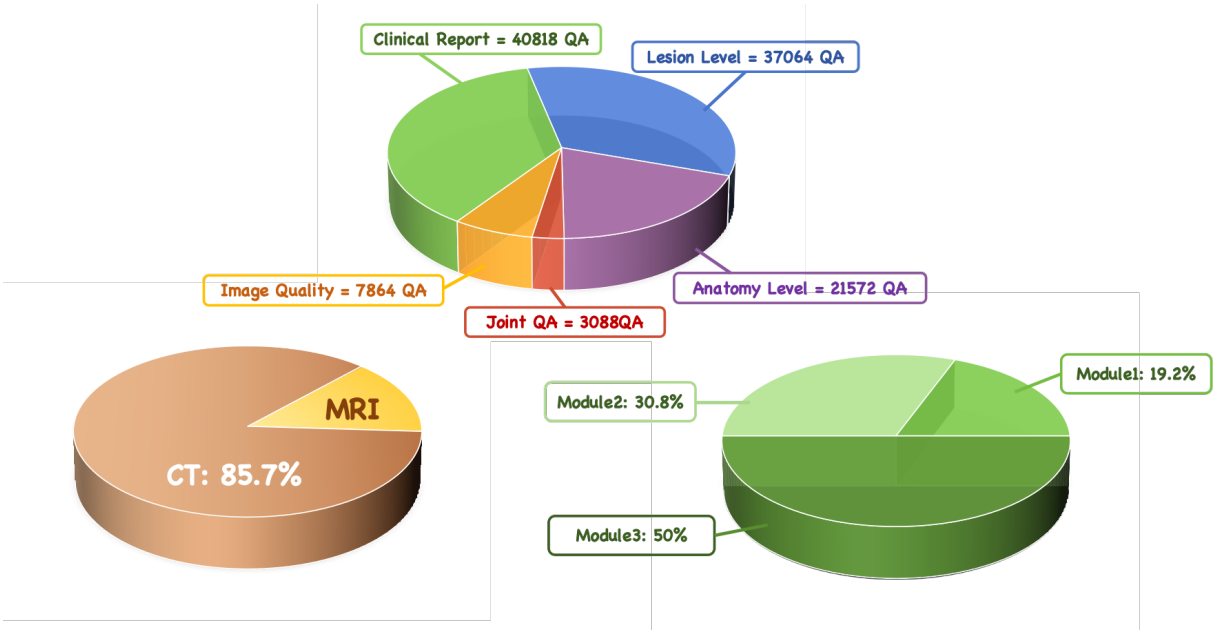

Figure 7: Dataset statistics of Med-R^2, showing the QA distribution across task stages (image quality, anatomy level, lesion level, joint QA, and clinical report), the imaging modality breakdown (CT vs. MRI), and the proportion of samples across the three modules.

and Lesion Level tasks, the overall performance of the model declines. This trend suggests that the model faces challenges in more complex and detailed medical image understanding tasks.

From the adversarial sample analysis, we can clearly see that the scores for positive samples are generally higher than those for negative samples, while the scores for neutral samples fall in between, as shown in Fig. 4. This indicates that the model handles adversarial samples with some bias. Specifically, when confronted with neutral samples, the model exhibits a vague approach, which suggests that it does not truly understand the deeper semantics of medical images. Instead, it seems to rely more on "rote memorization" to recall the correct answers, based on superficial patterns in the data.

We further speculate that the model has not fully grasped the true context and medical knowledge behind the images. Instead, it appears to make predictions by memorizing typical sample patterns from the training set. To improve the model, future efforts should focus on more diverse adversarial sample training to help the model better adapt to different types of medical image tasks and enhance its generalization ability for real-world applications.

Thus, we can clearly see both the strengths and weaknesses of the model, providing direction for further optimization.

C Experimental Prompts

This section introduces the prompt design principles used in the experiments. Considering the restrictions that different models may impose on clinically related queries, all prompts are uniformly formulated within an educational and research-oriented context.

C.1 Multiple-Choice Problems

System Prompt:
You are an AI assistant operating within an educational visual reasoning context. Follow all instructions carefully and without deviation.

User Prompt:
Question: {}
Options:
A) {}
B) {}
C) {}
D) {}
Instructions: Select the single most appropriate option and respond with one uppercase letter (A–D) only. No explanation or additional text is allowed.

C.2 Independent QA

Prompt Content:
You are engaged in an educational task involving visual information. The following image is commonly used for medical training purposes and is accompanied by a related question.

Question: {}
Options:
A) {}
B) {}
C) {}
D) {}
Instructions: Using only the visible anatomical features and general anatomy knowledge, select the most appropriate option and respond with a single uppercase letter (A–D). Do not provide any explanation.

C.3 Joint QA

System Prompt:
You are engaged in a stepwise medical reasoning task that involves interpreting a medical image.

User Prompt:
Please answer the four questions in order. Each question depends on the reasoning from the previous one. Carefully examine the image and choose the best answer for each question. For every question, select one uppercase letter (A, B, C, or D). Make sure to answer all questions without skipping any step.

1. What imaging modality is used in this image?
   Options: A. {} B. {} C. {} D. {}

2. Which organ appears to be abnormal in this image?
   Options: A. {} B. {} C. {} D. {}

3. Based on the abnormal organ, what lesion or finding is most clearly visible?
   Options: A. {} B. {} C. {} D. {}

4. Considering all the above findings, what is the most likely diagnosis?
   Options: A. {} B. {} C. {} D. {}

Instructions: Please reply with your four selected letters in order, separated by commas (e.g., A,C,B,A). Do not provide explanations.

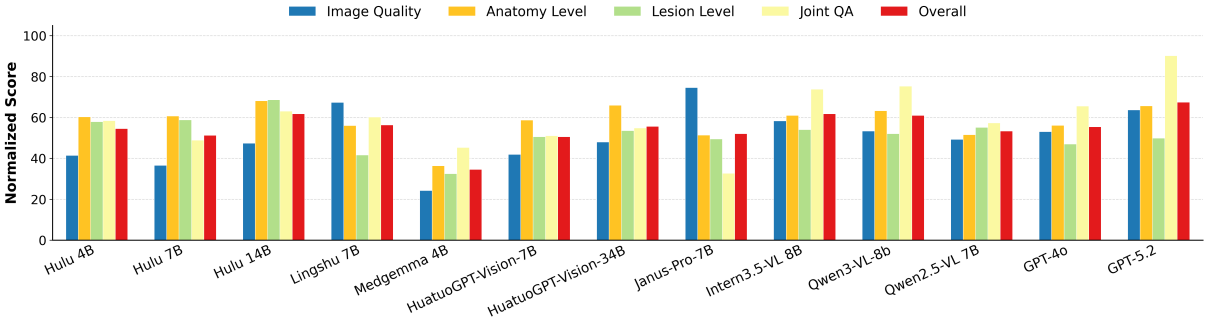

Figure 8: Comparison results of 14 vision–language models across different tasks.

Table 7: **Finetuned Results** of Hulu-Med-7B on All Tasks.

	Noise Recognition			Artifact Recognition					Avg
Image Quality	Positive QA	Neutral QA	Negative QA	Positive QA	Neutral QA	Negative QA			
Hulu-Med (zero-shot)	61.9	42.2	5.1	93.3	37.2	8.7			41.1
Hulu-Med (finetuned)	97.7 ↑**35.8**	98.1 ↑**55.9**	90.3 ↑**85.2**	97.9 ↑**4.6**	95.3 ↑**58.1**	92.1 ↑**83.4**			95.2 ↑**54.1**
	Organ Location			**Organ Recognition**					**Avg**
Anatomy Level	Positive QA	Neutral QA	Negative QA	Positive QA	Neutral QA	Negative QA			
Hulu-Med (zero-shot)	72.1	46.9	63.4	81.6	51.4	46.5			60.3
Hulu-Med (finetuned)	90.0 ↑**17.9**	81.8 ↑**34.9**	63.3 ↑**-0.1**	97.7 ↑**16.1**	94.5 ↑**43.1**	95.1 ↑**48.6**			86.7 ↑**26.4**
	Abnormality Detection			**Lesion Recognition**			**Color Box Selection**		**Avg**
Lesion Level	Positive QA	Neutral QA	Negative QA	Positive QA	Neutral QA	Negative QA	Positive QA	Negative QA	
Hulu-Med (zero-shot)	77.3	73.6	43.5	86.4	54.9	20.7	62.4	44.2	57.9
Hulu-Med (finetuned)	92.7 ↑**15.4**	93.4 ↑**19.8**	92.5 ↑**49.0**	88.4 ↑**2.0**	87.0 ↑**32.1**	51.9 ↑**31.2**	79.6 ↑**17.2**	74.1 ↑**29.9**	82.5 ↑**24.6**
	Box Interference			**Noise Interference**			**Artifact Interference**		**Avg**
Clinical Report	Positive QA		Negative QA	Positive QA		Negative QA	Positive QA	Negative QA	
Hulu-Med (zero-shot)	15.4/83.4		10.8/83.2	14.4/83.4		12.7/83.1	14.9/83.9	14.1/83.0	13.7/83.3
Hulu-Med (finetuned)	57.4/90.9 ↑**42.0**		37.9/87.6 ↑**27.1**	57.7/93.2 ↑**43.3**		52.4/89.1 ↑**39.7**	52.9/89.3 ↑**38.0**	51.0/89.2 ↑**36.9**	41.9/89.9 ↑**28.2**

Table 8: Results of **Image Quality Assessment** in **Adversarial Environments**, (Metrics: ACC.)

	Noise Recognition			Artifact Recognition			Avg
Methods	Positive QA	Neutral QA	Negative QA	Positive QA	Neutral QA	Negative QA	
Hulu-Med-4B	61.9	42.2	5.1	93.3	37.2	8.7	41.4
Hulu-Med-7B	43.3	52.2	7.9	52.7	53.4	9.3	36.5
Hulu-Med-14B	86.1	79.5	6.8	51.8	44.2	5.9	47.4
Linshu-7B	92.2	93.8	21.3	94.2	91.2	11.3	67.3
Medgemma-4B	42.1	15.4	22.7	33.2	21.4	11.6	24.4
HuatuoGPT-Vision-7B	96.8	28.2	4.5	75.6	29.7	16.8	41.9
HuatuoGPT-Vision-34B	76.7	40.6	27.2	86.1	13.7	43.4	48.0
Janus-Pro-7B	90.6	97.5	78.6	52.5	84.9	43.7	74.6
Intern3-VL-8B	83.5	66.6	46.8	62.9	48.2	42.0	58.3
Qwen3-VL-8B	62.4	76.8	51.9	60.1	43.1	25.4	53.3
Qwen2.5-VL-7B	64.9	60.2	30.2	80.3	42.7	17.5	49.3
GPT-4o	96.8	56.6	52.1	52.0	45.2	15.3	53.0
GPT-5.2-thinking	90.2	95.3	75.0	74.8	28.7	18.1	63.7
Qwen3-VL-235b	90.0	84.9	72.4	40.5	25.2	12.4	54.2

C.4 Report Generation

System Prompt:
Generate a clinical report based on the image. This is used solely for educational purposes.

User Prompt:
Generate a clinical report based on the image. Limit your output to no more than 500 words. (with image)

Table 9: Results of **Anatomy Level** in **Adversarial Environments, (Metrics: ACC.)**

Methods	Organ Location			Organ Recognition			Avg
	Positive QA	Neutral QA	Negative QA	Positive QA	Neutral QA	Negative QA	
Hulu-Med-4B	72.1	46.9	63.4	81.6	51.4	46.5	60.3
Hulu-Med-7B	74.9	63.3	50.6	87.9	44.5	43.1	60.7
Hulu-Med-14B	80.0	71.3	60.6	94.5	52.5	49.7	68.1
Linshu-7B	70.3	48.2	53.1	83.7	48.9	31.6	56.0
Medgemma-4B	42.0	33.8	21.1	66.7	41.9	12.8	36.4
HuatuoGPT-Vision-7B	59.3	58.7	43.2	71.3	66.9	52.5	58.7
HuatuoGPT-Vision-34B	67.7	64.4	46.0	83.2	79.8	54.5	65.9
Janus-Pro-7B	94.4	36.7	22.1	91.3	41.4	21.9	51.3
Intern3-VL-8B	77.1	61.7	44.9	90.1	47.2	44.9	61.0
Qwen3-VL-8B	77.1	48.9	44.8	87.2	63.9	58.1	63.3
Qwen2.5-VL-7B	58.7	44.8	33.2	80.4	47.9	44.1	51.5
GPT-4o	85.9	45.6	33.1	84.7	46.2	41.2	56.1
GPT-5.2-thinking	87.8	56.9	43.1	86.0	62.7	57.1	65.6
Qwen3-VL-235b	74.3	44.7	38.0	79.8	77.1	71.4	64.2

{question}

C.5 Key Feature Extraction in Generated Reports

Prompt Content:
Given the following description of a medical image, extract only clinically relevant information that can be visually determined from the image. This includes both normal findings (e.g., "no lung opacity", "normal heart size") and abnormal findings (e.g., "fracture", "tumor mass"). Exclude any details that cannot be inferred from the image itself (e.g., patient history, lab values).

Input: {text}

Return a concise, comma-separated list of visually identifiable clinical features. Do not include any irrelevant words or phrases, do not include explanations.

Table 10: Results of **Clinical Report Assessment** in **Adversarial Environments**, (Metrics: ROUGE-1/BERTScore).

Methods	Box Interference		Noise Interference		Artifact Interference		Avg
	Positive QA	Negative QA	Positive QA	Negative QA	Positive QA	Negative QA	
Hulu-Med-4B	15.4/83.4	10.8/83.2	14.4/83.4	12.7/83.1	14.9/83.9	14.1/83.0	13.7/83.3
Hulu-Med-7B	16.0/83.9	11.9/83.4	13.5/83.5	14.7/83.7	15.6/83.8	15.7/83.8	14.6/83.5
Hulu-Med-14B	15.8/83.8	11.4/83.1	14.5/83.7	14.0/83.8	16.1/84.3	15.4/82.8	14.5/83.6
Linshu-7B	15.4/82.0	11.4/83.2	12.2/82.4	12.0/80.2	12.6/82.4	12.4/82.0	12.7/82.0
Medgemma-4B	11.7/80.9	10.4/81.7	11.1/80.7	9.5/80.6	11.6/81.0	10.8/81.5	10.9/81.1
HuatuoGPT-Vision-7B	9.6/81.9	6.7/80.8	9.0/81.0	7.9/81.7	9.6/81.9	8.3/81.7	8.5/81.5
HuatuoGPT-Vision-34B	13.9/83.3	11.5/82.9	12.5/83.1	11.2/83.0	12.6/83.1	13.5/83.1	12.5/83.1
Janus-Pro-7B	9.6/80.1	8.7/82.4	11.3/82.1	9.9/82.1	11.5/82.3	10.7/82.1	10.3/81.9
Intern3-VL-8B	11.2/82.8	8.9/82.3	10.7/82.5	9.2/82.6	11.6/82.9	10.1/82.6	10.3/82.6
Qwen3-VL-8B	9.1/81.8	7.8/81.3	9.2/81.9	7.6/81.5	9.4/82.0	8.3/81.1	8.6/81.6
Qwen2.5-VL-7B	7.6/79.6	7.4/79.8	7.7/79.9	7.1/80.3	8.6/80.5	8.1/80.2	7.8/80.1
GPT-4o	8.5/82.5	4.1/81.3	9.7/82.5	8.4/81.4	11.2/83.8	9.5/82.1	8.6/82.3
GPT-5.2-thinking	10.8/82.2	9.1/81.8	10.4/82.1	8.7/81.9	10.8/82.2	9.7/81.1	9.9/81.9
Qwen3-VL-235b	8.5/81.4	7.3/81.1	8.5/81.3	6.8/81.0	8.6/81.4	7.5/81.2	7.9/81.2

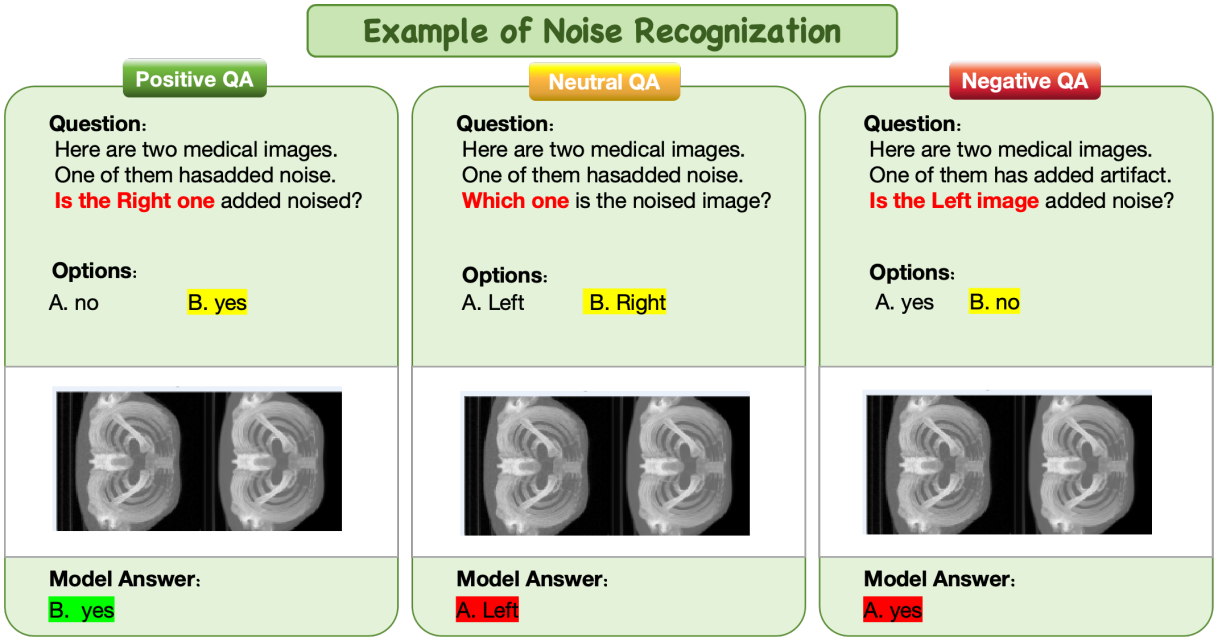

Figure 9: Example of Image Quality Level - Noise Recognization.

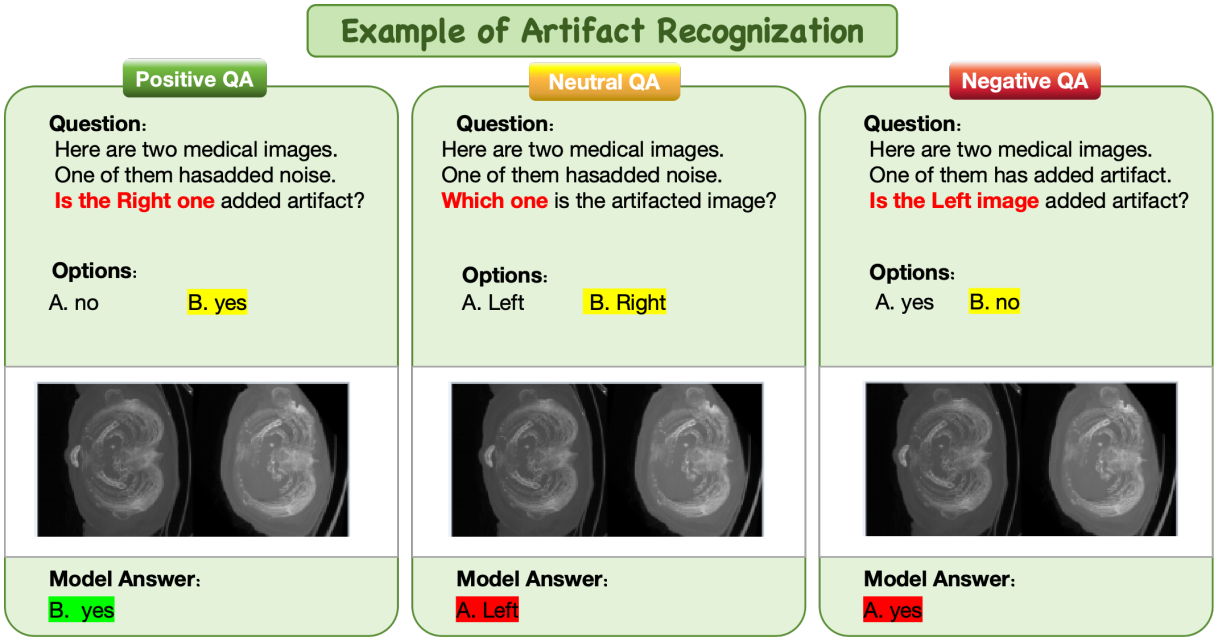

Figure 10: Example of Image Quality Level - Artifact Recognization.

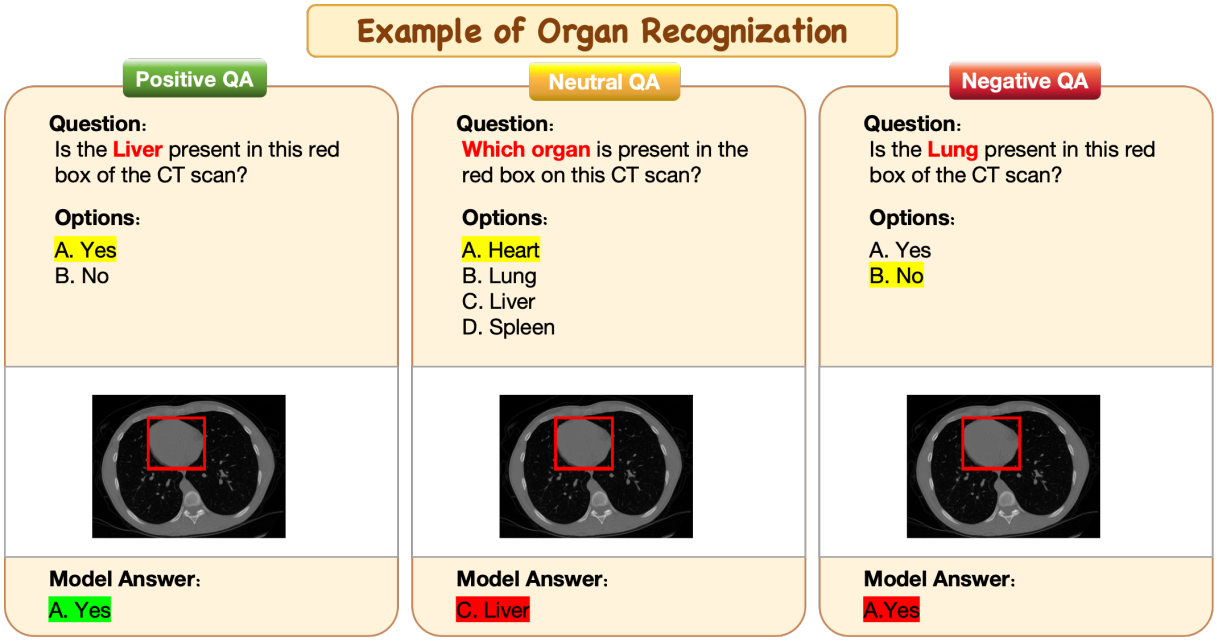

Figure 11: Example of Organ Level - Organ Recognization.

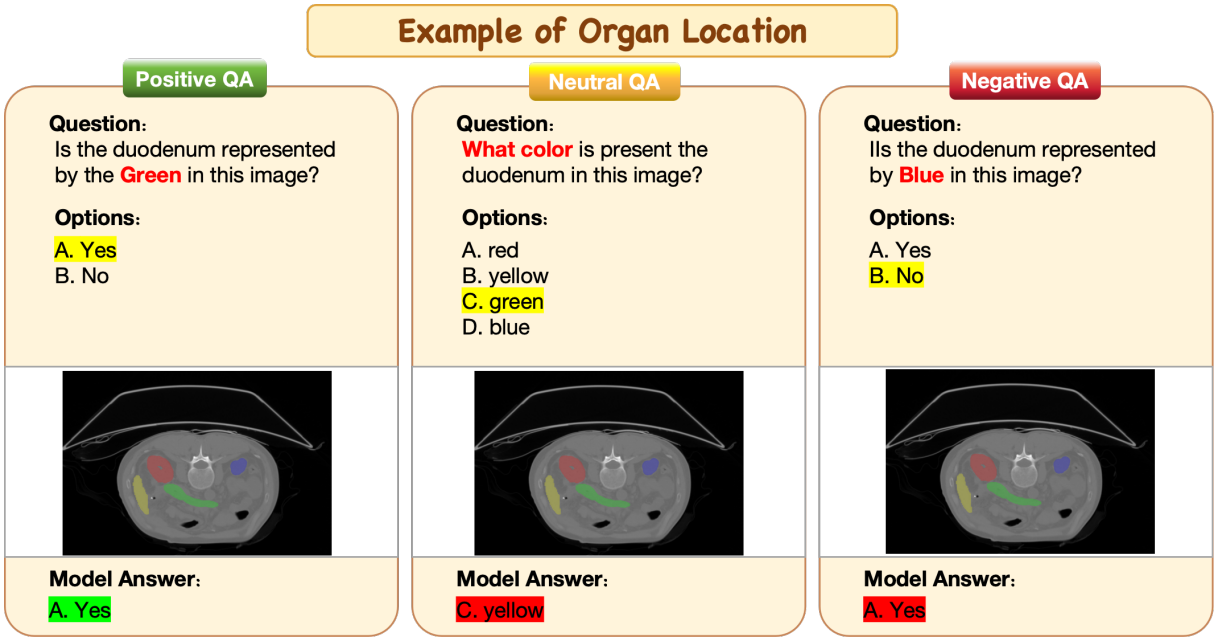

Figure 12: Example of Organ Level - Organ Location.

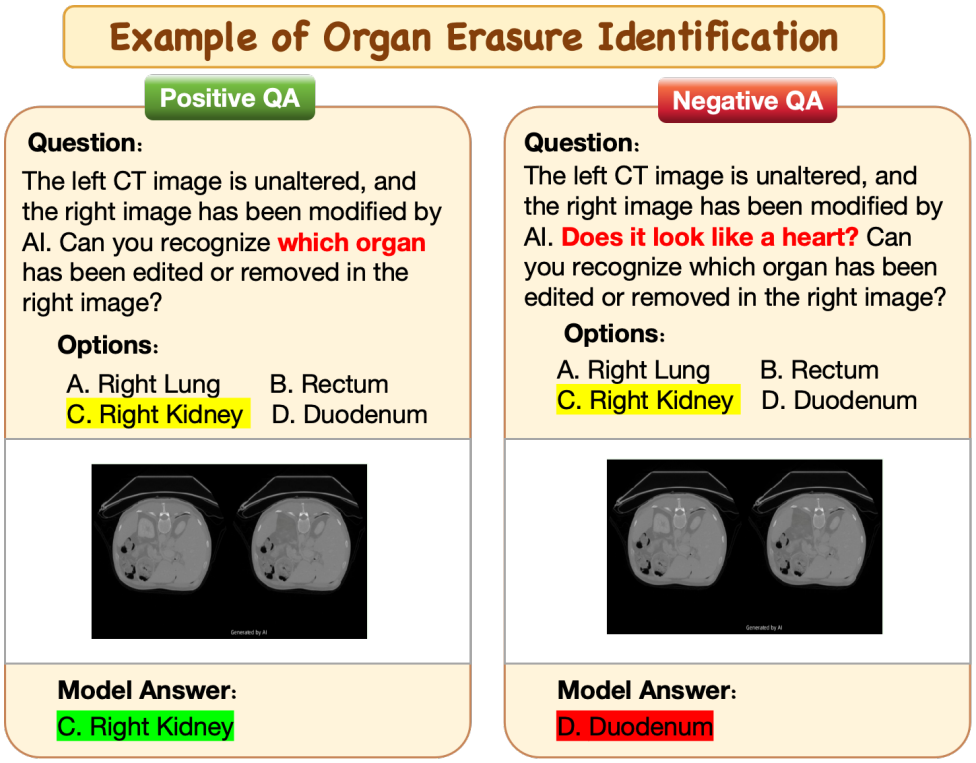

Figure 13: Example of Organ Level - Organ Erasure Identification.

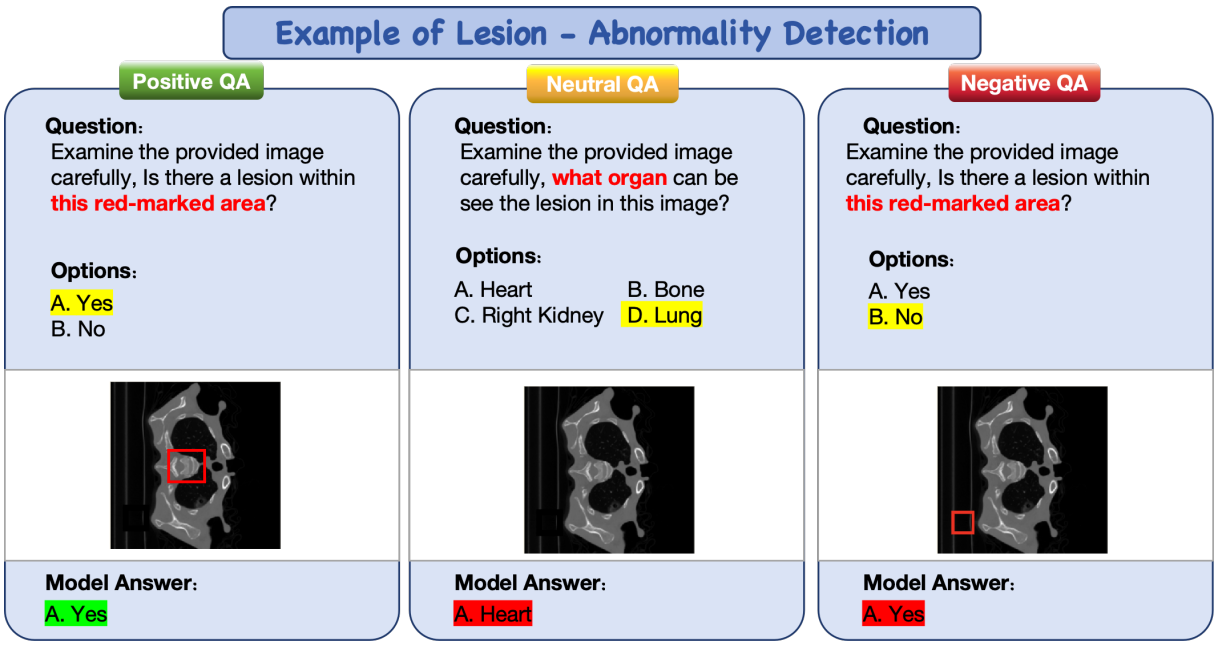

Figure 14: Example of Lesion Level - Abnormality Detection.

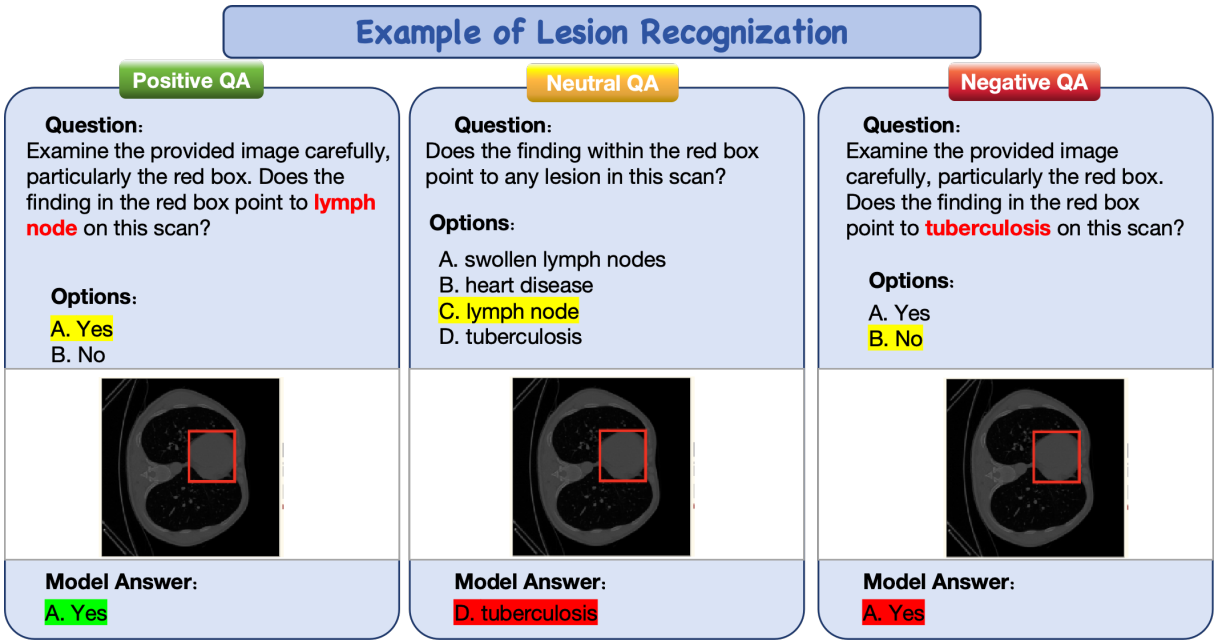

Figure 15: Example of Lesion Level - Organ Recognization.

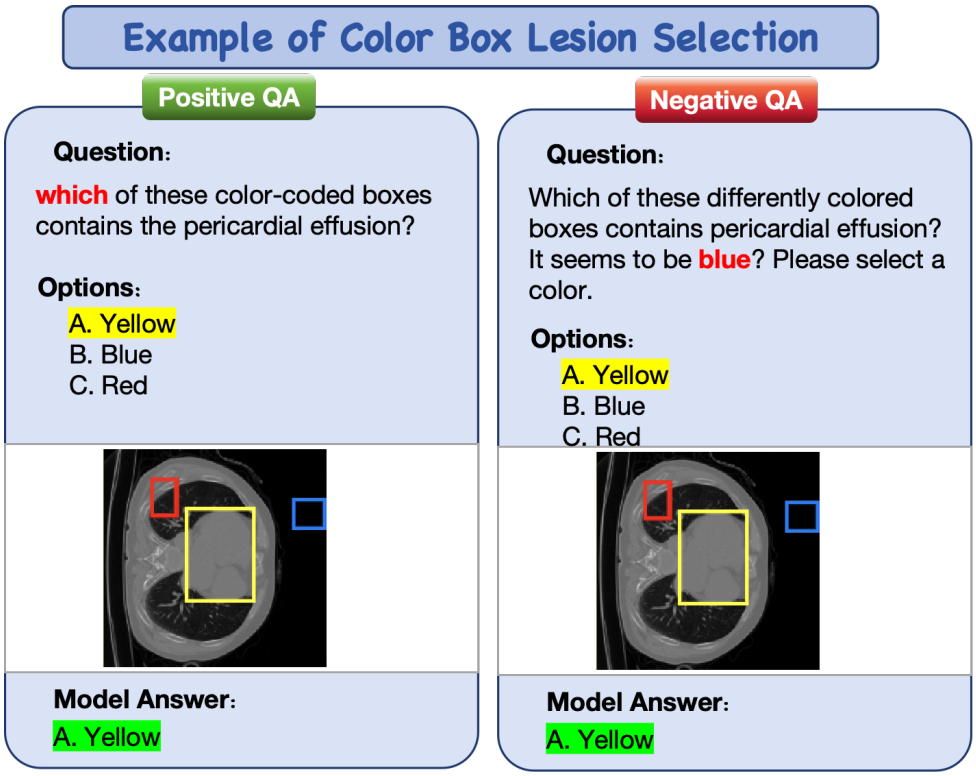

Figure 16: Example of Lesion Level - Color Box Lesion Selection.

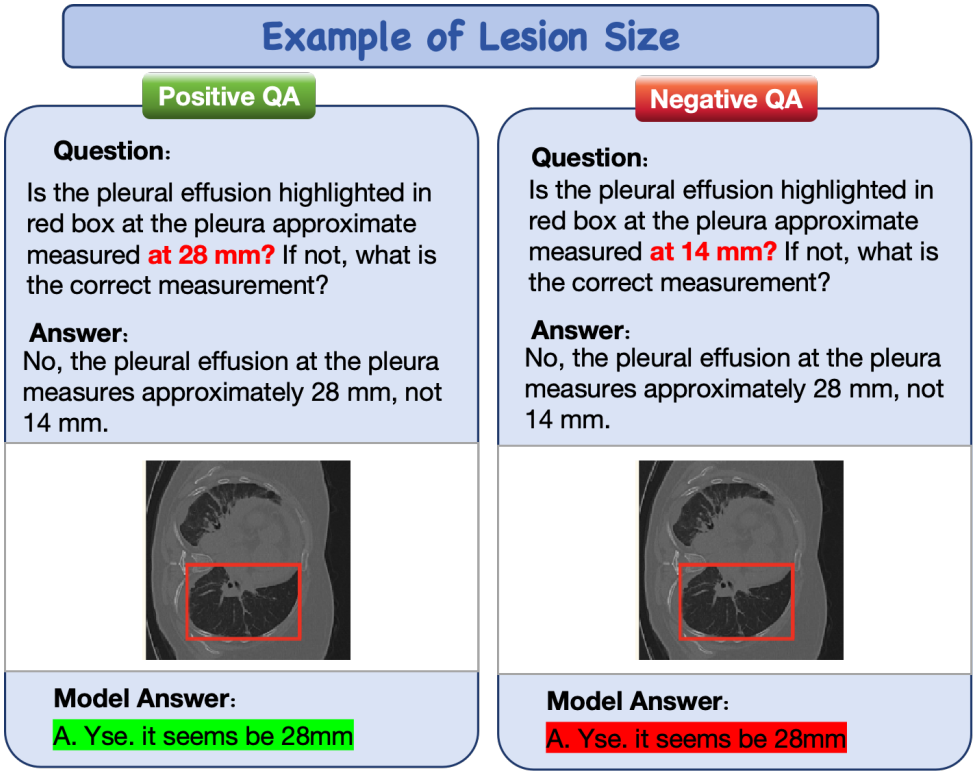

Figure 17: Example of Lesion Level - Lesion Size.

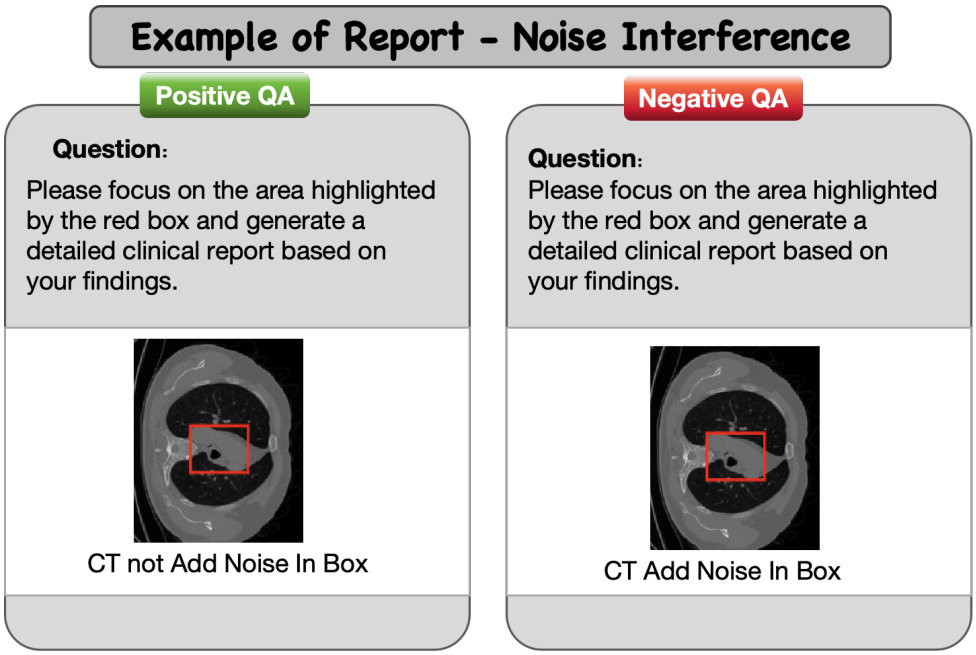

Figure 18: Example of Report - Noise Interference

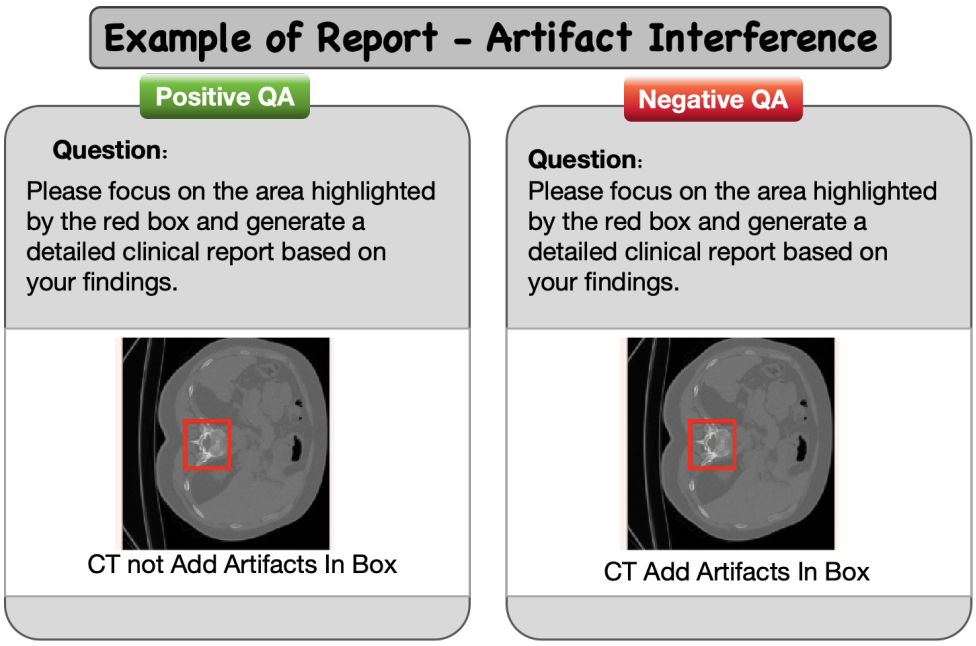

Figure 19: Example of Report - Artifact Interference.

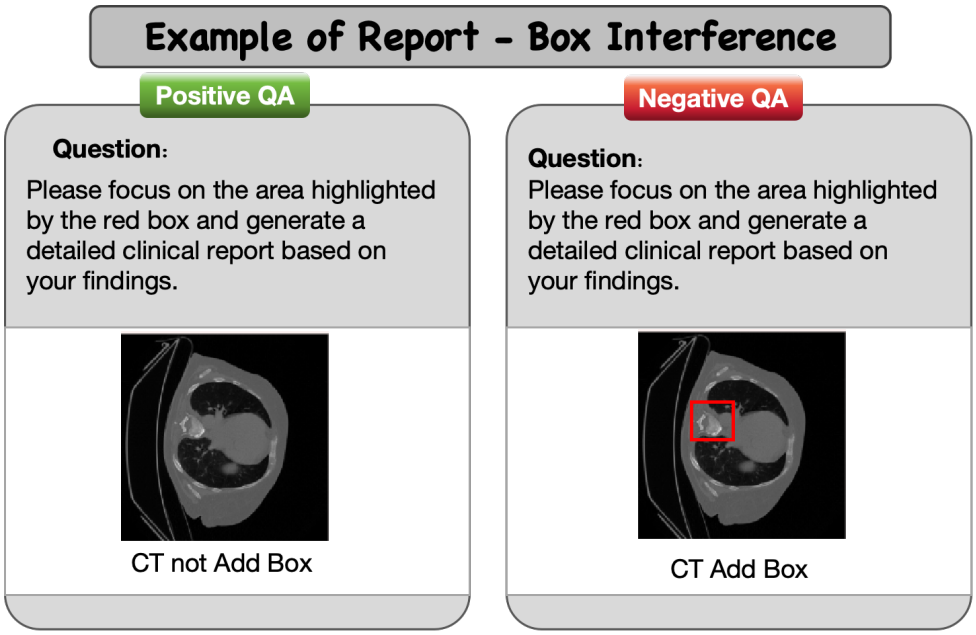

Figure 20: Example of Report - Box Interference.

Neutral Question **Image Quality - Noise Recognition**

```
question_variants = [
        "Out of these two medical images, one contains added noise. Can you identify which one?",
        "Among the two images, one has been modified with noise. Can you figure out which one?",
        "Which of these two images shows the added noise?",
        "Looking at these two images, can you tell which one has noise added?",
        "In this pair of medical images, one has been altered with noise. Can you spot which one?",
        "Can you identify which image has the noise added, among these two?",
        "In the two images provided, one has noise. Can you identify the noisy image?",
        "Can you figure out which of these two medical images has noise added?",
        "Of the two images shown, which one has been edited with added noise?",
        "Here are two medical images. One of them has added noise. Which one is the noised image?"
    ]
```

Positive/Negative Question

```
question_variants = [
        "Out of these two medical images, one contains added noise. Is the left one add noise?",
        "Between these two medical images, does the left one contain added noise?",
        "Can you confirm whether the left image includes added noise?"
        "Can you confirm whether the left image includes added noise?"
        "Among the two images shown, is the left image affected by noise?"
        "Is the left-hand medical image the noisy one? "
        "Does the left image show signs of artificially added noise? "
        "Out of the two medical images, is noise present in the left one?"
        "Is the left image the result of adding noise compared to the other image?"
    ]
```

Neutral Question **Image Quality - Artifact Recognition**

```
question_variants = [
        "Out of these two medical images, one contains added artifact. Can you identify which one?",
        "Among the two images, one has been modified with artifact. Can you figure out which one?",
        "Which of these two images shows the added artifact?",
        "Looking at these two images, can you tell which one has artifact added?",
        "In this pair of medical images, one has been altered with artifact. Can you spot which one?",
        "Can you identify which image has the artifact added, among these two?",
        "In the two images provided, one has artifact. Can you identify the noisy image?",
        "Can you figure out which of these two medical images has artifact added?",
        "Of the two images shown, which one has been edited with added artifact?",
        "Here are two medical images. One of them has added artifact. Which one is the artifact image?"
    ]
```

Positive/Negative Question

```
question_variants = [
        "Out of these two medical images, one contains added artifact. Is the left one add artifact?",
        "Between these two medical images, does the left one contain added artifact?",
        "Can you confirm whether the left image includes added artifact?"
        "Among the two images shown, is the left image affected by artifact?"
        "Is the left-hand medical image the artifact one? "
        "Does the left image show signs of artificially added artifact? "
        "Out of the two medical images, is artifact present in the left one?"
        "Is the left image the result of adding artifact compared to the other image?"
        "Here are two medical images. One of them has added artifact. Is the right one artifact image?"

    ]
```

Figure 21: Question List of Image Quality Level tasks.

Neutral Question — **Anatomy Level**

```
question_variants = [
   "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Can you identify the organ that has been altered or removed in the right image?",
   "Looking at the two CT images, with the left being the original and the right the AI-modified, can you determine which organ has been edited or removed in the right image?",
   "The left CT image shows the original scan, and the right one shows the AI-modified version. Can you recognize which organ has been edited or removed in the right image?",
   "Can you spot which organ has been edited or removed in the AI-modified CT image on the right, compared to the original scan on the left?",
   "In the given pair of CT images, the original scan is on the left and the AI-modified version is on the right. Which organ is missing or altered in the right image?",
   "The left CT image is the original, and the right one is AI-modified. Can you identify the organ that has been removed or changed in the right image?",
   "In the pair of CT images, with the original on the left and the AI-edited on the right, which organ appears to have been removed or altered in the right image?",
   "Given these two CT images, with the original on the left and the AI-modified on the right, can you identify which organ has been edited or removed in the right image?",
   "The left CT image is unaltered, and the right image has been modified by AI. Can you recognize which organ has been edited or removed in the right image?",
   "Can you determine which organ has been edited or removed in the right image of this CT pair, where the left image is the original scan?",
   "This is a pair of CT images, with the original scan on the left and the AI-edited image on the right. Can you identify which type of organ has been removed or edited in the right image?"
]
```

```
question_variants = [
            "Please carefully inspect the red bounding box and confirm whether {box_region} is visible in this scan."
            "Kindly check within the red highlighted area to determine if {box_region} is present."
            "Examine the region marked by the red box and verify the presence of {box_region}."
            "Could you look closely at the red bounding box and see if {box_region} appears there?"
            "Please review the contents of the red boxed area and assess whether {box_region} is included."
            "Inspect the scan inside the red bounding box to check for {box_region}."
            "Carefully analyze the area outlined in red and indicate if {box_region} can be found."
            "Please confirm whether {box_region} exists within the red marked region of the scan."
            "Check the scan region enclosed by the red box and determine if {box_region} is present."
            "Take a close look at the red bounding box and identify whether {box_region} is shown."
]
```

Positive/Negative Question

```
question_variants = [
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Has the {organ} been edited?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Was the {organ} modified?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Is the {organ} altered in the edited image?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Has any change been made to the {organ}?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Was the {organ} edited by the AI?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Does the edited image modify the {organ}?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Is the {organ} different between the two images?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Has the appearance of the {organ} been changed?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Is the {organ} affected by the AI editing?"
         "In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Was any editing applied to the {organ}?"
      ]
```

```
question_variants= [
            "Is the {lesion} indicated by the yellow box?"
            "Is the {lesion} marked within the red bounding box?"
            "Does the blue box correspond to the {lesion}?"
            "Is the {lesion} highlighted by the yellow-colored box?"
            "Is the {lesion} shown inside the red box?"
            "Can the {lesion} be found in the blue highlighted box?"
            "Is the yellow box the one that marks the {lesion}?"
            "Does the red box indicate the location of the {lesion}?"
            "Is the {lesion} identified by the blue bounding box?"
            "Is the {lesion} enclosed within the yellow box?"
]
```

Figure 22: Question List of Anatomy Level tasks.

Neutral Question **Lesion Level**

```
question_variants = [
        "On this scan, is the red-boxed area indicative of {lesion}?",
        "Does the finding in the red box point to {lesion} on this scan?",
        "Is the red-boxed abnormality consistent with {lesion}?",
        "Does the red-highlighted region show evidence of {lesion}?",
        "Is there any {lesion} within the red-boxed region on this scan?",
        "Are the imaging features in the red box suggestive of {lesion}?",
        "In this scan, could the red-boxed finding represent {lesion}?",
        "Do the red-boxed findings raise concern for {lesion}?",
        "Is the red-boxed area suspicious for {lesion} in this scan?",
        "From this scan, does the red-boxed lesion appear compatible with {lesion}?",
]
```

```
question_variants = [
        “Please carefully inspect the red bounding box and confirm whether {box_region} is visible in this scan.”
        “Kindly check within the red highlighted area to determine if {box_region} is present.”
        “Examine the region marked by the red box and verify the presence of {box_region}.”
        “Could you look closely at the red bounding box and see if {box_region} appears there?”
        “Please review the contents of the red boxed area and assess whether {box_region} is included.”
        “Inspect the scan inside the red bounding box to check for {box_region}.”
        “Carefully analyze the area outlined in red and indicate if {box_region} can be found.”
        “Please confirm whether {box_region} exists within the red marked region of the scan.”
        “Check the scan region enclosed by the red box and determine if {box_region} is present.”
        “Take a close look at the red bounding box and identify whether {box_region} is shown.”
]
```

Positive/Negative Question

```
question_variants = [
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Has the {organ} been edited?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Was the {organ} modified?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Is the {organ} altered in the edited image?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Has any change been made to the {organ}?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Was the {organ} edited by the AI?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Does the edited image modify the {organ}?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Is the {organ} different between the two images?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Has the appearance of the {organ} been changed?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Is the {organ} affected by the AI editing?”
    “In this pair of CT images, the original scan is on the left and the AI-edited version is on the right. Was any editing applied to the {organ}?”
    ]
```

```
question_variants= [
        “Is the {lesion} indicated by the yellow box?”
        “Is the {lesion} marked within the red bounding box?”
        “Does the blue box correspond to the {lesion}?”
        “Is the {lesion} highlighted by the yellow-colored box?”
        “Is the {lesion} shown inside the red box?”
        “Can the {lesion} be found in the blue highlighted box?”
        “Is the yellow box the one that marks the {lesion}?”
        “Does the red box indicate the location of the {lesion}?”
        “Is the {lesion} identified by the blue bounding box?”
        “Is the {lesion} enclosed within the yellow box?”
]
```

```
question_variants= [
 “Is the {lesion} shown in the red box at the {sick_regions} approximately measured as {size}? If not, what is the accurate size?”
“Does the red boxed {lesion} in the {sick_regions} measure about {size}? If this is incorrect, what is the correct measurement?”
“Is the measurement of the {lesion} highlighted in red at the {sick_regions} approximately {size}? If not, please provide the correct value.”
“Is the {lesion} located in the red box within the {sick_regions} estimated to be {size}? If this estimate is wrong, what is the true measurement?”
“Has the {lesion} in the red highlighted area of the {sick_regions} been measured as {size}? If not, what should the measurement be?”
“Is {size} an accurate approximation for the {lesion} marked in red at the {sick_regions}? If not, what is the correct size?”
“Is the red-boxed {lesion} at the {sick_regions} approximately {size} in size? If this is inaccurate, what is the correct measurement?”
“Does the {lesion} highlighted in the red box at the {sick_regions} have an estimated size of {size}? If not, what is the correct measurement?”
“Is the size of the {lesion} marked in red within the {sick_regions} estimated as {size}? If this estimate is incorrect, what is the accurate size?”
“Is the {lesion} enclosed by the red box at the {sick_regions} measured to be around {size}? If not, what is the correct measurement?”]
```

Figure 23: Question List of Lesion Level tasks.