MIRA: Evaluating Multimodal AI on Complex Clinical Reasoning in Interventional Radiology

Jingxiong Li[1], Chenglu Zhu[2], Sunyi Zheng[3], Yuxuan Sun[2], Yifei Wang[4], He Liu[5], Yunlong Zhang[2], Yixuan Si[2], Lin Yang[2], Liang Xiao[1,6*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]School of Engineering, Westlake University, Hangzhou 310024, China
[3]Tianjin Medical University Cancer Institute and Hospital, Department of Radiology, Tianjin 300060, China
[4]Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China
[5]Department of Cardiology, Xuzhou Central Hospital, Xuzhou 221000, China
[6]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China
lijingxiong@njust.edu.cn, zhuchenglu@westlake.edu.cn, zhengsunyi@tmu.edu.cn, sunyuxuan@westlake.edu.cn, happywangyf2018@163.com, liuhe520666@126.com, {zhangyunlong,siyixuan,yanglin}@westlake.edu.cn, xiaoliang@mail.njust.edu.cn

Abstract

We present MIRA (**M**ultimodal **I**nterventional **RA**diology evaluation), a comprehensive benchmark for evaluating large multimodal models in expert-level interventional radiology tasks requiring specialized domain knowledge and advanced visual reasoning capabilities. Unlike existing medical benchmarks that primarily provide binary labels without contextual depth, MIRA offers diverse question formats, including open-ended, closed-ended, single-choice, and multiple-choice categories, each accompanied by detailed expert-validated explanations. The benchmark incorporates approximately 184K high-quality medical images spanning multiple imaging modalities with 1.2M meticulously generated question-answer pairs across various anatomical regions. These pairs were created through a sophisticated cascade methodology involving expert interventional radiologists at both the data collection and validation stages. Our comprehensive evaluation, encompassing zero-shot testing and fine-tuning experiments of large multimodal models, revealing significant performance gaps between AI systems and human specialists. Fine-tuning experiments demonstrate substantial improvements, with models achieving up to 0.80 accuracy on single-choice questions. MIRA establishes a challenging benchmark that suggests promising directions for developing specialized clinical AI systems for interventional radiology.

Code — https://github.com/pompom6/MIRA

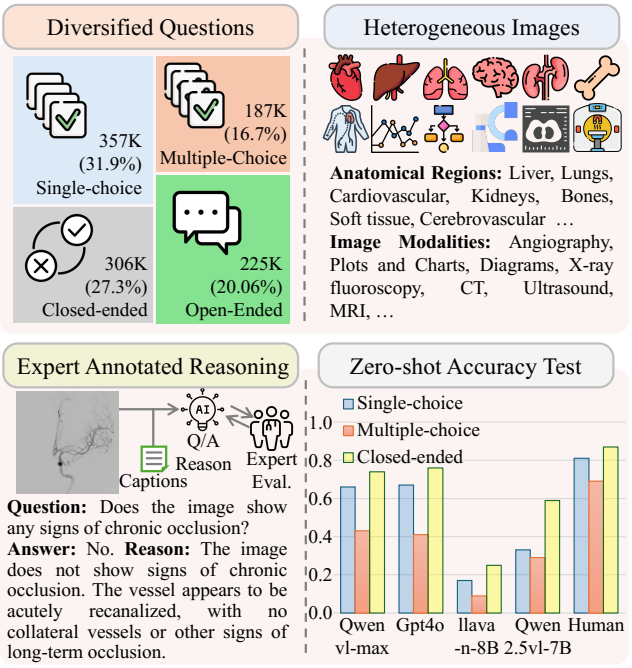

Figure 1: An overview of MIRA. The dataset encompasses reasoning text with QA pairs, spanning multiple anatomical regions and imaging modalities. Modern LMMs exhibit a substantial performance gap compared to human experts.

Introduction

Interventional Radiology (IR) is an advanced medical discipline that utilizes image-guided, minimally invasive techniques for both diagnostic and therapeutic interventions (Kaufman and Lee 2013; Bundy et al. 2019). The adoption of IR procedures has expanded rapidly across modern healthcare systems (Guan et al. 2025; Cleary and Peters 2010), driven by their efficacy, and reduced recovery times. Practitioners in this domain are required to analyze a wide array of imaging modalities, including ultrasonography, fluoroscopy, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI), in conjunction with rich textual clinical documentation to formulate accurate diagnoses and optimize treatment strategies (Ji et al. 2021; El-Fakdi et al. 2014). This inherently multimodal nature of IR introduces

*Corresponding author

significant complexity to clinical decision-making (El-Fakdi et al. 2014), necessitating systems capable of understanding across modalities.

Recent advances in Artificial Intelligence (AI), particularly in deep learning, have shown significant promise in enhancing IR workflows (Glielmo et al. 2024). Successful applications include automated vessel segmentation (Yan, Yang, and Cheng 2018), real-time navigation support (Maybody, Stevenson, and Solomon 2013), and computer-assisted procedure planning (Sardar et al. 2019). However, the majority of existing approaches rely on narrowly defined tasks and structured datasets (Aliferis and Simon 2024; Alowais et al. 2023), limiting their scalability and generalizability in complex, real-world IR environments where multimodal integration and domain expertise are crucial.

The increasing clinical impact of IR underscores the urgent need for rigorous evaluation of Large Language Models (LLMs) and Large Multimodal Models (LMMs) before their deployment in healthcare applications (Hager et al. 2024; Li et al. 2024; Workum et al. 2025). While recent studies have demonstrated the potential of LMMs in various medical imaging tasks (Tian et al. 2023; Zhou et al. 2023), challenges persist due to the procedural complexity, context-dependent reasoning requirements, and the scarcity of high-quality, expertly annotated datasets (Nazi and Peng 2024; Luo et al. 2024). Although Visual Question Answering (VQA) in medical field has emerged as a promising paradigm for evaluating the reasoning and interpretability of AI systems in clinical contexts (Sun et al. 2024; Liu et al. 2024a; Yue et al. 2024), existing benchmarks suffer from several limitations: they often lack procedural depth, fail to incorporate multimodal reasoning across imaging and text, and are rarely curated or validated by clinical experts—limitations that are particularly critical in the context of IR, where decision-making is highly specialized and risk-sensitive.

To address these gaps, we introduce **MIRA**, the first comprehensive benchmark explicitly tailored to the domain of interventional radiology. Key contributions are:

- We present a high-quality dataset with expert-annotated clinical images paired with diverse questions and detailed explanations, reflecting the multimodal reasoning demands of real IR procedures.
- We introduce close collaboration with board-certified interventional radiologists in both data curation and quality assurance phase, ensuring clinical validity and interpretability throughout dataset.
- Extensive evaluations across multiple state-of-the-art vision-language models, revealing significant performance gaps between current AI systems and human experts in specialized IR tasks.

Related Works

Large Fundation Models

Large Foundation Models, particularly Large Language Models (LLMs) and Large Multimodal Models (LMMs), have significantly advanced the field of artificial intelligence by demonstrating strong generalization capabilities across a wide range of tasks (Devlin et al. 2019; Brown et al. 2020; Achiam et al. 2023; Touvron et al. 2023). Transformer-based architectures pre-trained on massive web-scale corpora have enabled emergent abilities in language understanding, reasoning, and few-shot adaptation (Wei et al. 2022). Recent efforts have extended these capabilities to the medical domain, where domain-specific LLMs such as BioGPT (Luo et al. 2022) and Med-PaLM (Singhal et al. 2025) have been trained on biomedical literature to improve factual accuracy and clinical relevance.

On the vision-language front, LMMs such as Flamingo (Alayrac et al. 2022), Gemini Pro Vision (Team et al. 2023), LLaVA (Liu et al. 2023b), InstructBLIP (Dai et al. 2023) and Qwen2.5-VL (Bai et al. 2025) have shown promise in visual reasoning tasks by jointly encoding image and textual inputs. In the medical domain, Med-Flamingo (Moor et al. 2023) and LLaVA-Med (Li et al. 2023) have adapted these architectures to handle medical imaging modalities and clinical language. However, their performance in procedure-centric specialties such as Interventional Radiology remains largely untested, partly due to the lack of domain-specific, multimodal datasets with fine-grained clinical supervision.

VQA Benchmarks

Visual Question Answering (VQA) has become a standard framework for evaluating the multimodal reasoning abilities of AI models (Antol et al. 2015; Yin et al. 2023; Liu et al. 2024c). In the biomedical domain, several benchmarks have been proposed to assess the performance on specialized image modality, such as VQA-RAD (Lau et al. 2018) for radiographs; PathVQA (He et al. 2020), PathMMU (Sun et al. 2024) for pathology slides; and ScienceQA (Saikh et al. 2022) for biomedical figures. More recently, general-purpose medical benchmarks such as MedVQA (Karaca and Aydin 2025) and MMMU (Yue et al. 2024) have emerged to provide large-scale, open-ended QA datasets across multiple specialties.

Despite these advances, existing benchmarks often suffer from one or more of the following limitations: (i) relying on static image-text pairs without reasoning or procedural context; (ii) limited coverage of high-stakes, decision-critical tasks; and (iii) lack of expert involvement in question formulation or validation. These limitations make them suboptimal for evaluating AI models in interventional radiology, where multimodal reasoning is tightly coupled with domain-specific anatomical knowledge and procedural workflow.

Our MIRA benchmark addresses these gaps by incorporating expert-generated, task-oriented questions grounded in real clinical procedures, offering a rigorous platform for evaluating AI models in specialized IR settings.

Construction of MIRA

MIRA represents the first comprehensive dataset for evaluating vision-language models across clinical procedures in IR. The framework builds on three principles: (1) **Expert-centered Data Retrieval** through board-certified radiologists curating diverse imaging scenarios; (2) **Reasoning-focused Q&A Pair Generation** using vision-language

Features of MIRA	
Total Images	184479
Total Questions	1161366
IR-specific Keywords	100
Matching Threshold	2
Question Types	4
Open-ended Questions	323483(27.9%)
Closed-ended Questions	330496(28.5%)
Single-choice Questions	325867(28.06%)
Multiple-choice Questions	181520(15.63%)
Images(Questions) for Training	178003(1120031)
for Validation	5039(33056)
for Test	1437(8279)

Table 1: Key features of MIRA dataset.

Term	Category	Freq.
artery	Vascular Anatomy	2749
angiography	Imaging Modality	1013
stent	Pathological Descriptor	520
balloon	Device	240
occlusion	Pathological Descriptor	301
injection	Procedural Terms	151

Table 2: Example IR-specific terms with semantic categories and frequency.

models to create clinically meaningful questions with explanations; and (3) **Hybrid Quality Assurance Protocol** combining algorithmic verification with expert clinical review.

Expert-centered Data Retrieval

We construct our dataset by sourcing image-text pairs from the PubMed Central Open Access Subset (Bethesda 2003), ensuring compliance with CC-BY or CC0 licensing terms. To enable high-recall identification of IR-related samples, we design a multi-stage expert-in-the-loop term curation pipeline. Two board-certified interventional radiologists independently annotated 9,551 image-text pairs to identify IR-relevant content, forming the foundation for downstream keyword mining.

We rank terms by frequency and manually select 100 high-precision domain-specific keywords as filtering anchors. These terms span five semantically meaningful categories — *vascular anatomy, imaging modality, pathological descriptor, device*, and *procedural terms* — which align with core clinical concepts in interventional radiology. Table 2 provides representative examples of curated terms across categories, and additional examples are included in the appendix.

To ensure both domain relevance and semantic richness, we classify a sample as IR-related if it contains at least two distinct terms from the curated list. Figure 2 illustrates the distribution of keyword diversity and usage across categories. Notably, vascular anatomy dominates in both term count and frequency, underscoring its central role in IR-related documentation. This structured categorization not

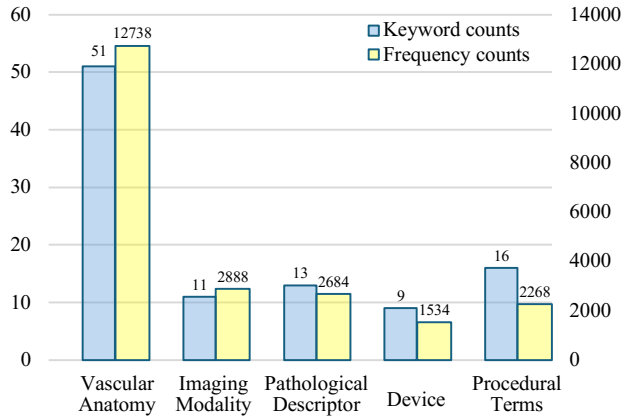

Figure 2: Keyword and frequency distribution across five semantic categories in expert-selected IR-related captions. "Keyword counts" (left y-axis) represent the number of distinct terms identified per category, while "Frequency counts" (right y-axis) reflect the total occurrences of those terms in the dataset.

only enhances coverage breadth but also provides a principled basis for semantic filtering in multimodal data collection.

Reasoning-focused Q&A Pair Generation

To generate high-quality question-answer pairs that incorporate explicit clinical reasoning, we develop a sophisticated question-answer generation pipeline. The pipeline addresses a critical challenge in medical VQA datasets: the need for questions that not only test factual knowledge but also require structured reasoning about visual features, anatomical relationships, and clinical implications. By leveraging state-of-the-art vision-language models, specifically GPT4o(Achiam et al. 2023) and Qwen-VL-Max(Bai et al. 2023), our approach generates questions that demand comprehensive analysis of both visual and textual inputs, ensuring the reasoning process is explicitly captured in the corresponding answers. For each image-text sample, we instruct the VLMs to generate a diverse set of questions, including open-ended, closed-ended, single-choice and multiple-choice QAs. Testing anatomical understanding, diagnostic reasoning, procedural knowledge, and complex decision-making. The prompting strategy, as demonstrated in Figure 3, is designed to generate responses that demonstrate not just factual knowledge but also clinical reasoning processes, incorporating visual features, anatomical relationships, and procedural considerations in the generated content.

Hybrid Quality Assurance Protocol

Given the critical nature of clinical decision-making in interventional radiology and the need for high-quality training data, we implement a hybrid validation approach combining expert judgment with advanced language models. Expert interventional radiologists first curates 40 questions (10 per

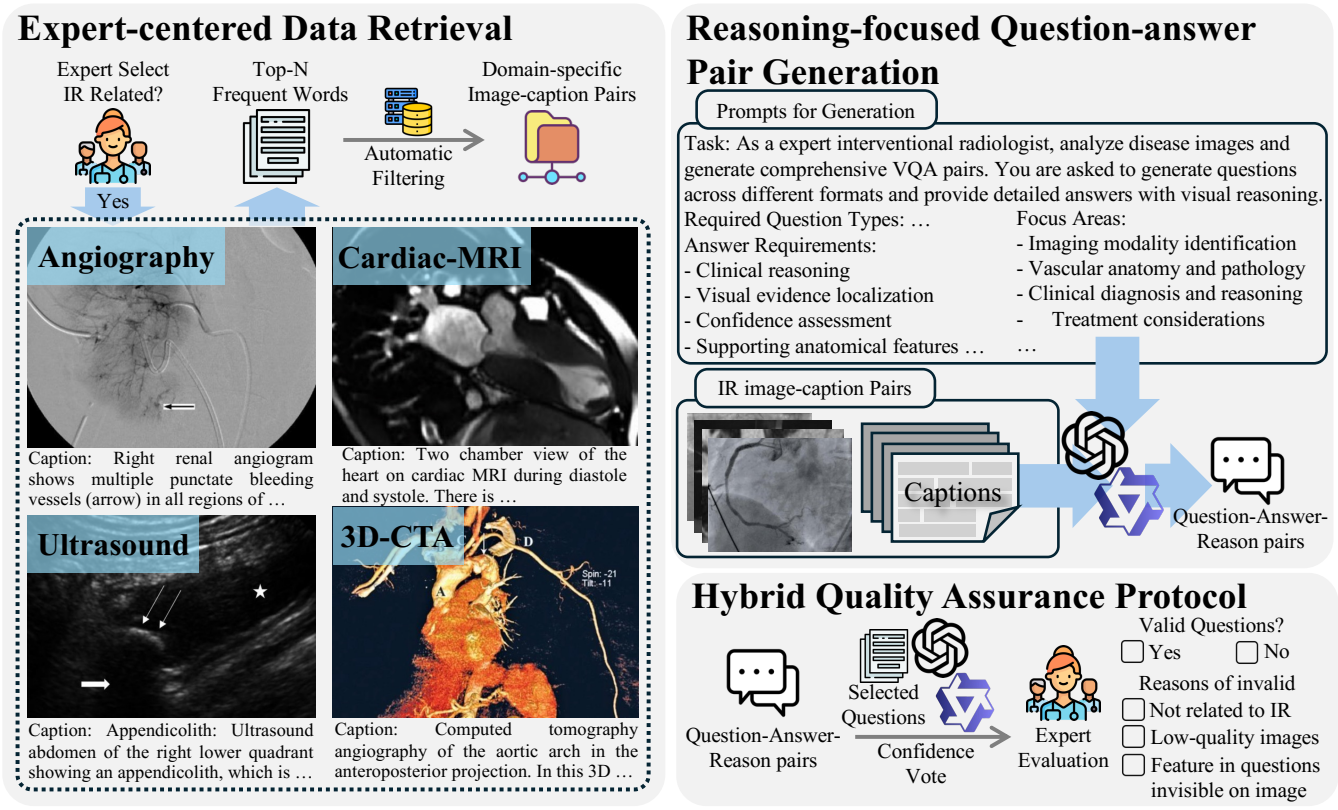

Figure 3: An illustration of 3 primary processes in MIRA dataset construction: (1) Expert-centered Data Retrieval for high-quality image-text pair selection; (2) Reasoning-focused Question-answer Pair Generation encompassing multiple question types; and (3) Hybrid Quality Assurance Protocol which combines expert radiologist assessment and large multimodal model evaluation.

category) to establish gold standards that serves as demonstration prompts for vision-language models. We then employ GPT-4o (Achiam et al. 2023) and Qwen-VL-max (Bai et al. 2023) to evaluate generated questions based on clinical relevance, reasoning complexity, and linguistic clarity. Questions receiving high scores from both models are submitted for expert verification. The evaluation questions are demonstrated in Figure 3, human experts have to evaluate the questions and images based on a combination of visual clarity, diagnostic relevance, and caption informativeness, the GUI developed for human expert evaluation is included in appendix.

Comparisons with Existing Benchmarks

As illustrated in Figure 3 and detailed in Table 1, our MIRA dataset establishes a substantial advancement over existing medical VQA benchmarks across scale, domain specificity, reasoning complexity, and expert involvement.

Early medical VQA datasets such as VQA-RAD (Lau et al. 2018) and VQA-Med (Ben Abacha et al. 2019) offered domain relevance and expert validation but remained limited in both size and diversity of reasoning types. PMC-VQA (Zhang et al. 2023), while significantly larger in scale, lacks expert validation and includes minimal interventional radiology (IR) content, making it less suitable for evaluating procedure-intensive reasoning. More recent efforts, such as GEMeX (Liu et al. 2024a) and RG-CCT (Zhang et al. 2024), have introduced structured rationales and chain-of-thought annotations, yet still lack comprehensive coverage of IR-specific tasks.

In contrast, MIRA is the first large-scale benchmark explicitly tailored for interventional radiology. It comprises 1.16 million question-answer pairs across 184,479 clinical images, covering a wide spectrum of procedural and diagnostic reasoning. Specifically, the training set includes 1,120,031 Q&A pairs over 178,003 images, with an additional 33,056 Q&As (5,039 images) for validation and 8,279 Q&As (1,437 images) for testing. The dataset features four distinct question types—including open-ended, closed-ended, single-choice, and multiple-choice formats—accompanied by expert-validated rationales that reflect real-world decision-making in IR practice.

Unlike prior datasets, MIRA uniquely integrates (i) large-scale multimodal coverage, (ii) reasoning annotations for multi-type questions, and (iii) expert-in-the-loop validation, offering a robust foundation for evaluating advanced models in specialized interventional radiology domain.

Dataset	Q&As /Images	Question Type	IR-Support?	Reasons?	Expert Validated?
VQA-RAD	3.5K/315	O.,M.	✗	✗	✓
VQA-Med	5K/5K	O.,C.	✗	✗	✓
PMC-VQA	227K/149K	O.,C.,S.	✓	✗	✗
GEMeX	1.6M/151K	O.,C.,S.,M.	✗	✓	✗
RG-CCT	1.3M/50K	O.,C.	✗	✓	✓
MIRA	1.2M/184K	O.,C.,S.,M	✓	✓	✓

*O.: Open-ended, C.: Closed-ended, S.: Single-choice, M.: Multiple-choice

Table 3: Comparison between MIRA and existing medical VQA benchmarks. MIRA distinguishes itself as the only benchmark specifically designed for interventional radiology, evaluated by human experts with reasoning cotext while maintaining significant scale.

Evaluation of MIRA

In this section, we conduct comprehensive experiments to evaluate model performance on the MIRA benchmark. Initially, we assess LMMs under both zero-shot and fine-tune settings. To validate the effectiveness of our multi-type question approach, we perform ablation studies to validate the multi-type question strategy, and analyze the impact of top-n word selection in data retrieval. All fine-tuning experiments are implemented on 4 NVIDIA A100 GPUs for 2 epochs with a batch size of 1. The fine-tuning protocol employs a LoRA scheme, with warm-up phase during the initial 0.05 epochs with a linear learning rate increase to 1e-5, followed by cosine schedule decay, optimizing convergence while preventing overfitting on specialized medical content.

Zero-Shot Baseline Assessment

We evaluate the zero-shot capabilities of 9 contemporary large multimodal models (LMMs) on the MIRA test set: 6 open-source models (Qwen2-VL-2B/7B (Wang et al. 2024), Qwen2.5-VL-7B (Bai et al. 2025), LLaVA-1.5-7B (Liu et al. 2023a), LLaVA-1.5-13B (Liu et al. 2023a), LLaVA-next-8B (Liu et al. 2024b)) and 3 proprietary models (Qwen-VL-max (Bai et al. 2023), Qwen-VL-plus (Bai et al. 2023), GPT4o (Achiam et al. 2023)). Performance is measured using BLEU-4 and ROUGE-L metrics across all question types, calculated between ground truth and model-generated answers including reasoning explanations. We further evaluate accuracy on single-choice, multiple-choice, and closed-ended questions to assess clinical decision-making capabilities. These experiments aim to identify capability gaps in current models when addressing specialized interventional radiology tasks. Key findings are as follows:

Advanced LMMs struggle with MIRA. Despite strong general VQA performance, advanced LMMs underperform on the MIRA benchmark. GPT4o scores only 10.31/25.09 (BLEU-4/ROUGE-L) on open-ended questions, while the best open-source model achieves just 6.46/22.80. For multiple-choice questions, GPT4o attains 0.67 accuracy, and models like llava1.5-13B reach only 0.26 on single-choice tasks. This performance gap compared to human experts underscores the difficulty of IR-specific reasoning, which requires integrating procedural knowledge, anatomical understanding, and clinical reasoning.

Performance varies across question types. The models demonstrate heterogeneous capabilities when processing different question formats. Notably, as illustrated in Table 4, closed-ended questions yield superior BLEU-4 scores in 7 out of 9 evaluated models. In contrast, single-choice and multiple-choice questions present significant analytical challenges, with markedly reduced accuracy documented in Table 4, where 6/9 models achieve accuracy scores below 0.5 for single-choice questions, and same number of models fall below 0.4 accuracy for multiple-choice questions. This pattern suggests pronounced limitations in discriminative reasoning tasks that require precise differentiation among visually similar interventional radiology scenarios.

Human experts maintain a significant advantage. To facilitate comparative evaluation against human experts, we establish a specialized MIRA-tiny subset containing 450 carefully selected Q&As equally distributed into single-choice, multiple-choice and closed ended type. According to Table 4, A substantial performance differential persists between LMMs and clinical professionals. Board-certified interventional radiologists achieves 0.81, 0.69, and 0.87 accuracy on single-choice, multiple-choice, and closed-ended questions respectively, compared to maximal zero-shot model performance of 0.67, 0.43, and 0.76 across the same categories. These performance disparity underscores the indispensability of specialized human expertise and the significant opportunities for further advancement in medical AI systems.

Fine-Tuned Model Performance

As demonstrated in Table 4, LoRA fine-tuning significantly enhances the performance of all open-source models across the dataset, often doubling or tripling their zero-shot metrics. Qwen2.5-VL-7B excels in single-choice (33.25 BLEU-4) and multiple-choice questions (36.03 BLEU-4), while LLaVA-next-8B achieves the highest ROUGE-L scores across most categories (36.19 for closed-ended, 53.71 for single-choice, and 54.38 for multiple-choice questions). Table 4 demonstrates particularly dramatic accuracy improve-

	Open-ended		Closed-ended			Single-choice			Multiple-choice		
	Bleu4	RougeL	Bleu4	RougeL	ACC	Bleu4	RougeL	ACC	Bleu4	RougeL	ACC
Image and Text as Input											
Zero-shot											
Qwen2-VL-2B	6.12	22.41	9.86	22.37	0.33	3.45	14.12	0.23	5.84	17.13	0.25
Qwen2-VL-7B	5.73	22.80	14.62	24.39	0.59	6.13	16.39	0.33	7.12	18.68	0.29
Qwen2.5-VL-7B	6.43	20.36	13.56	22.79	0.52	7.29	21.32	0.45	6.82	18.04	0.33
LLaVA1.5-7B	5.50	21.03	7.20	23.03	0.44	3.32	11.75	0.22	8.08	15.37	0.31
LLaVA1.5-13B	6.46	22.39	7.08	24.39	0.50	4.91	15.28	0.26	7.25	19.61	0.24
LLaVA-next-8B	1.25	5.98	3.97	7.69	0.25	2.65	2.87	0.17	1.34	3.27	0.09
Qwen-VL-max	9.64	24.91	**15.27**	27.65	0.74	11.69	36.03	0.66	**11.65**	37.43	**0.43**
Qwen-VL-plus	7.33	23.80	14.81	26.20	0.68	11.37	34.01	0.53	11.31	36.92	0.38
GPT4o	**10.31**	**25.09**	15.04	**28.07**	**0.76**	**12.04**	**36.95**	**0.67**	10.95	**38.08**	0.41
Fine-tune											
Qwen2-VL-2B	16.95	32.98	27.14	32.92	0.75	29.25	44.93	0.71	31.40	44.38	0.62
Qwen2-VL-7B	**18.76**	34.87	**27.84**	43.29	0.82	32.68	48.57	0.78	35.23	48.30	0.69
Qwen2.5-VL-7B	18.47	34.95	25.79	**44.33**	0.81	**33.25**	48.12	**0.80**	**36.03**	49.38	**0.70**
LLaVA1.5-7B	15.94	32.23	23.46	37.50	0.76	29.27	45.40	0.69	30.35	44.21	0.61
LLaVA1.5-13B	16.31	32.35	25.34	38.06	**0.84**	31.63	44.93	0.74	32.07	46.86	0.63
LLaVA-next-8B	17.71	**36.19**	24.05	39.40	0.65	29.86	**53.71**	0.63	32.41	**54.38**	0.32
Text Only as Input											
Zero-shot											
Qwen3-8B	2.91	19.71	6.85	22.87	0.31	6.01	15.32	0.25	6.29	21.03	0.22
Qwen3-32B	3.39	**22.04**	6.91	22.48	0.32	6.95	19.57	0.29	8.41	22.31	0.30
GPT4.1	2.33	10.97	6.25	17.29	0.14	7.08	22.41	0.35	9.25	24.91	0.33
GPT4.1-mini	3.02	16.53	6.34	22.20	**0.38**	**7.76**	24.09	**0.38**	**9.96**	**25.42**	**0.35**
Qwen-max	2.06	12.10	5.58	18.63	0.15	6.73	23.03	0.34	8.99	24.82	0.32
Qwen-plus	**4.36**	21.90	**7.22**	**24.25**	0.37	7.37	**25.01**	0.35	9.13	24.95	0.33
Expert Evaluation (Image and Text)											
Human Expert	-	-	-	-	0.87	-	-	0.81	-	-	0.69

Table 4: Validation results of LMMs on the MIRA test set in different types of questions. The best performing LMM in each subset is in-bold.

ments: Qwen2-VL-2B achieves a 0.48 increase, reaches 0.71 accuracy on single-choice questions. LLaVA1.5-13B achieves the highest accuracy on closed-ended questions at 0.84 with 0.34 increase. These results demonstrate that while fine-tuning significantly narrows the gap between closed-source and proprietary models, specialized medical reasoning in interventional radiology remains challenging.

Analysis under Incomplete Input Modality

To evaluate the reasoning capabilities of large language models in the absence of visual context, we conduct a zero-shot assessment where only the textual question is provided as input. As shown in Table 4, this setting reveals consistent limitations across all evaluated models comparing with the proprietary LMMs, indicating a clear reliance on visual grounding for accurate and specific answers. Notably, on open-ended questions, all models exhibit low BLEU-4 and ROUGE-L scores, suggesting that generated answers tend to be vague or generic. Accuracy metrics further corroborate this trend: even high-capacity models such as GPT-4.1 and Qwen-max remain below 0.35 in most settings, underscoring their difficulty in grounding domain-specific predictions without image evidence.

Interestingly, GPT-4.1-mini and Qwen-plus attain most of the highest scores across multiple metrics, surpassing larger models such as GPT4.1 and Qwen-max. To better understand model behavior under incomplete input, we perform an error analysis comparing Qwen-plus and Qwen-max in the text-only setting, as shown in Figure 4. Qwen-plus exhibits a higher tendency to "guess" answers (37%) and produces fewer refusals, suggesting aggressive decoding or reliance on language priors. In contrast, Qwen-max is more conservative: 52% of its errors stem from failure due to missing visual input, it shows a higher refusal rate (11%) and less frequently to hallucinate visual features, indicating stronger uncertainty calibration. Interestingly. These findings highlight differing modality coping strategies: smaller models may overcommit when uncertain, while larger models demonstrate caution but are still vulnerable to speculative reasoning without visual grounding.

Ablation Studies

Data Curation Settings: To filter domain-relevant data, we establish a minimum occurrence threshold of two distinct IR terms per caption. We conduct a post-hoc analysis on a validation set of 4000 samples, with 1:1 IR and non-IR.

Qwen2-VL -7B	Closed-ended		Single-choice		llava1.5 -7B	Closed-ended		Single-choice	
	BLEU4	ROUGEL	BLEU4	ROUGEL		BLEU4	ROUGEL	BLEU4	ROUGEL
zero-shot	14.62	24.39	6.13	16.39	zero-shot	7.08	24.39	4.91	15.28
OE	20.37	31.28	7.86	22.39	OE	15.27	31.07	10.52	21.47
OE+MC	21.40	32.25	23.95	37.61	OE+MC	18.43	32.72	22.51	35.83

Table 5: Performance comparison of LMMs on MIRA test set. Performance is evaluated on closed-ended and single-choice questions under various training configurations: zero-shot (no fine-tuning), OE (fine-tuned on open-ended questions), and OE+MC (fine-tuned on combined open-ended and multiple-choice questions).

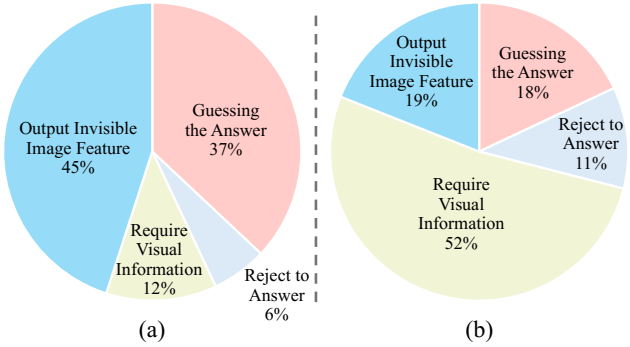

Figure 4: Error distribution over 100 annotated (a) Qwen-plus and (b) Qwen-max errors using text-only input on closed-ended questions.

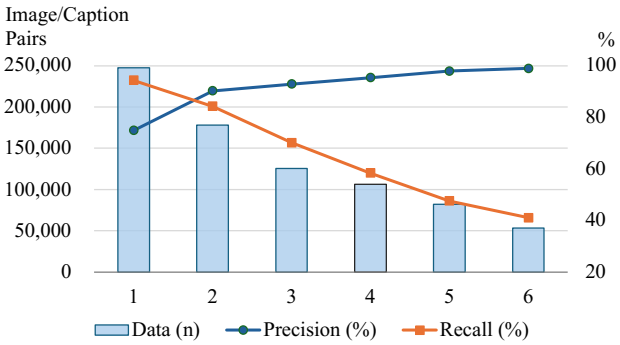

Figure 5: Filtering performance across different keyword thresholds. The x-axis represents the minimum number of IR-specific terms required in each image-text pair for inclusion.

Accoring to Figure 5, As the threshold increases from 1 to 6, the total number of selected pairs decreases, while precision improves steadily from 74.91% to 98.92%. However, recall drops significantly from 94.35% to 41.1%, reflecting a trade-off between dataset purity and coverage. The chosen threshold of two terms balances high precision (90.21%) with adequate recall (84.25%), providing a reliable filtering criterion for constructing the dataset.

Question Type Efficacy: We evaluate the effect of question type diversity by training Qwen2-VL-7B and LLaVA1.5-7B on subsets of the MIRA dataset and testing on closed- and single-choice questions using BLEU-4 and ROUGE-L (Table 5). Results show that training on open-ended questions notably boosts closed-ended performance (e.g., BLEU-4 of LLaVA1.5-7B : 7.08→15.27). Combining open-ended and multiple-choice questions yields the best gains — Qwen2-VL-7B achieves a BLEU-4 of 23.95 and ROUGE-L of 37.61 on single-choice tasks. These findings suggest that diverse training formats enhance cross-question generalization in medical VQA.

Limitations and Future Work: While MIRA offers broad coverage, it underrepresents rare interventional procedures, limiting model generalizability. It also lacks integration with patient-specific data and does not fully address uncertainty estimation or demographic bias. Future iterations of MIRA may include integration with temporal imaging and patient metadata to further emulate real-world clinical workflows.

Conclusion

MIRA represents a significant advancement in developing and benchmarking multimodal AI systems for interventional radiology, establishing a challenging framework that effectively evaluates both visual perception and clinical reasoning capabilities through expert-centered data retrieval, reasoning-focused question generation, and rigorous quality assurance protocols. Our evaluations reveal substantial performance gaps between current models and human specialists. Ablation studies demonstrate the value of our proposed curation method in retrieving IR-related data, and diverse question formulations in enhancing model reasoning capabilities across metrics. Fine-tuning experiments validate the potential for significant enhancement through specialized medical training.

To our knowledge, MIRA is the first large-scale multimodal benchmark curated for interventional radiology with expert-in-the-loop validation and reasoning-grounded QA pairs. It paves the way for the development of trustworthy AI systems capable of supporting real-world diagnostic and therapeutic decision-making in interventional medicine.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant 62471235). We also thank our collaborators for their valuable feedback and computational resources throughout the development of this work.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Aliferis, C.; and Simon, G. 2024. Lessons Learned from Historical Failures, Limitations and Successes of AI/ML in Healthcare and the Health Sciences. Enduring Problems, and the Role of Best Practices. *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, 543–606.

Alowais, S. A.; Alghamdi, S. S.; Alsuhebany, N.; Alqahtani, T.; Alshaya, A. I.; Almohareb, S. N.; Aldairem, A.; Alrashed, M.; Bin Saleh, K.; Badreldin, H. A.; et al. 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1): 689.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Ben Abacha, A.; Hasan, S. A.; Datla, V. V.; Demner-Fushman, D.; and Müller, H. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

Bethesda. 2003. PMC Open Access Subset [Internet].Bethesda (MD): National Library of Medicine. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/ [Accessed: 2025-07-18].

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bundy, J. J.; Hage, A. N.; Chick, J. F. B.; Gemmete, J. J.; Srinivasa, R. N.; Lee, E.; Johnson, E.; Hussain, J.; Cline, M.; Patel, N.; et al. 2019. Trends in interventional radiology through the eye of the journal of vascular and interventional radiology: a 27-year history. *Current Problems in Diagnostic Radiology*, 48(4): 353–358.

Cleary, K.; and Peters, T. M. 2010. Image-guided interventions: technology review and clinical applications. *Annual review of biomedical engineering*, 12(1): 119–142.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

El-Fakdi, A.; Gamero, F.; Melendez, J.; Auffret, V.; and Haigron, P. 2014. eXiTCDSS: A framework for a workflow-based CBR for interventional Clinical Decision Support Systems and its application to TAVI. *Expert Systems with Applications*, 41(2): 284–294.

Glielmo, P.; Fusco, S.; Gitto, S.; Zantonelli, G.; Albano, D.; Messina, C.; Sconfienza, L. M.; and Mauri, G. 2024. Artificial intelligence in interventional radiology: state of the art. *European Radiology Experimental*, 8(1): 62.

Guan, J. J.; Elhakim, T.; Matsumoto, M. M.; McKeon, T.; Laage-Gaupp, F.; Iqbal, S.; Patel, P. J.; Pereira, P.; Tam, A. L.; Binkert, C.; et al. 2025. Results of a Global Survey on the State of Interventional Radiology 2024. *Journal of Vascular and Interventional Radiology*.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9): 2613–2622.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Ji, J.; Fang, S.; Chen, W.; Zhao, Z.; Cheng, Y.; et al. 2021. Precision interventional radiology.

Karaca, Z.; and Aydin, I. 2025. Med-VQA: Performance Analysis of Question-Answering Systems on Medical Images. In *2025 29th International Conference on Information Technology (IT)*, 1–4. IEEE.

Kaufman, J. A.; and Lee, M. J. 2013. *Vascular and interventional radiology: the requisites*. Elsevier Health Sciences.

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

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, J.; Deng, Y.; Sun, Q.; Zhu, J.; Tian, Y.; Li, J.; and Zhu, T. 2024. Benchmarking large language models in evidence-based medicine. *IEEE Journal of Biomedical and Health Informatics*.

Liu, B.; Zou, K.; Zhan, L.; Lu, Z.; Dong, X.; Chen, Y.; Xie, C.; Cao, J.; Wu, X.-M.; and Fu, H. 2024a. GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA

Benchmark for Chest X-ray Diagnosis. *arXiv preprint arXiv:2411.16778*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6): bbac409.

Luo, X.; Deng, Z.; Yang, B.; and Luo, M. Y. 2024. Pre-trained language models in medicine: A survey. *Artificial Intelligence in Medicine*, 102904.

Maybody, M.; Stevenson, C.; and Solomon, S. B. 2013. Overview of navigation systems in image-guided interventions. *Techniques in vascular and interventional radiology*, 16(3): 136–143.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Nazi, Z. A.; and Peng, W. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, 57. MDPI.

Saikh, T.; Ghosal, T.; Mittal, A.; Ekbal, A.; and Bhattacharyya, P. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3): 289–301.

Sardar, P.; Abbott, J. D.; Kundu, A.; Aronow, H. D.; Granada, J. F.; and Giri, J. 2019. Impact of artificial intelligence on interventional cardiology: from decision-making aid to advanced interventional procedure assistance. *Cardiovascular interventions*, 12(14): 1293–1303.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 1–8.

Sun, Y.; Wu, H.; Zhu, C.; Zheng, S.; Chen, Q.; Zhang, K.; Zhang, Y.; Wan, D.; Lan, X.; Zheng, M.; et al. 2024. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, 56–73. Springer.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tian, D.; Jiang, S.; Zhang, L.; Lu, X.; and Xu, Y. 2023. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 14(1): 1108.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Workum, J. D.; Volkers, B. W.; van de Sande, D.; Arora, S.; Goeijenbier, M.; Gommers, D.; and van Genderen, M. E. 2025. Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. *Critical Care*, 29(1): 72.

Yan, Z.; Yang, X.; and Cheng, K.-T. 2018. Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 65(9): 1912–1923.

Yin, Z.; Wang, J.; Cao, J.; Shi, Z.; Liu, D.; Li, M.; Huang, X.; Wang, Z.; Sheng, L.; Bai, L.; et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36: 26650–26685.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zhang, X.; Wu, C.; Zhao, Z.; Lei, J.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhou, H.; Liu, F.; Gu, B.; Zou, X.; Huang, J.; Wu, J.; Li, Y.; Chen, S. S.; Zhou, P.; Liu, J.; et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.