MedAtlas: Evaluating LLMs for Multi-Round, Multi-Task Medical Reasoning Across Diverse Imaging Modalities and Clinical Text

Ronghao Xu[*1,2], **Zhen Huang**[*3,4], **Yangbo Wei**[4,5], **Xiaoqian Zhou**[1,2], **Zikang Xu**[6], **Ting Liu**[1,2], **Zihang Jiang**[1,2†], **S. Kevin Zhou**[1,2‡]

[1]School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230026, P.R. China
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215123, P.R. China
[3]School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, P.R. China
[4]School of Information Science and Technology, Eastern Institute of Technology, Ningbo 315200, P.R. China
[5]Shanghai Jiao Tong University, Shanghai 200030, P.R. China
[6]Anhui Province Key Laboratory of Biomedical Imaging and Intelligent Processing, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230026, P.R. China

Abstract

Artificial intelligence has demonstrated significant potential in clinical decision-making; however, developing models capable of adapting to diverse real-world scenarios and performing complex diagnostic reasoning remains a major challenge. Existing medical multi-modal benchmarks are typically limited to single-image, single-turn tasks, lacking multi-modal medical image integration and failing to capture the longitudinal and multi-modal interactive nature inherent to clinical practice. To address this gap, we introduce MedAtlas, a novel benchmark framework designed to evaluate large language models on realistic medical reasoning tasks. MedAtlas is characterized by four key features: multi-round visual question answering (VQA), Joint reasoning of multiple modalities of medical images, multi-task integration, and high clinical fidelity. It supports four core tasks: open-ended multi-round VQA, closed-ended multi-round VQA, multi-image joint reasoning, and comprehensive disease diagnosis. Each case is derived from real diagnostic workflows and incorporates temporal interactions between textual medical histories and multiple imaging modalities, including CT, MRI, PET, ultrasound, X-ray, etc., requiring models to perform deep integrative reasoning across images and clinical texts. MedAtlas provides expert-annotated gold standards for all tasks. Furthermore, we propose two novel evaluation metrics: Stage Chain Accuracy (SCA) and Error Propagation Suppression Coefficient (EPSC). Benchmark results with existing multi-modal models reveal substantial performance gaps in multi-stage clinical reasoning. MedAtlas establishes a challenging evaluation platform to advance the development of robust and trustworthy medical AI.

[*]These authors contributed equally.
[†]Corresponding author. Email: jzh0103@ustc.edu.cn
[‡]Corresponding author. Email: skevinzhou@ustc.edu.cn

Introduction

Clinical multimodal reasoning is a core challenge in intelligent healthcare, requiring AI models to simultaneously process imaging examinations, clinical history, and progressive inquiry information to support diagnostic decision-making. In real-world clinical scenarios, physicians often reach a final diagnosis through multiple rounds of inquiry and various imaging examinations: alternately reviewing multi-modal medical images (e.g., CT, MRI, ultrasound), probing key clinical manifestations, and integrating longitudinal examination results to form a comprehensive judgment.

Existing medical AI benchmarks are mostly limited to single-round, single-image visual question answering (VQA) settings. This limitation prevents models from addressing complex diagnostic needs in real-world scenarios, such as differential diagnosis based on multi-image comparison ("Compared to three months ago, the enlargement of this lesion suggests what possibility?") or tasks requiring stepwise reasoning ("Based on the current MRI findings, should the next step be a biopsy or a PET-CT?"). The absence of such benchmarks constrains the development of practical medical assistants.

The limitations of existing benchmarks are reflected in four aspects: (1) **Task simplicity**: VQA datasets (e.g., VQA-RAD (Lau et al. 2022)) only support single-round question answering, while classification datasets (e.g., CheXpert (Irvin et al. 2019)) lack multi-round, free-form image-text interaction. (2) **Modality isolation**: Most benchmarks do not require cross-modal association (e.g., simultaneously parsing multi-modal images and textual information). (3) **Reasoning discontinuity**: Even if multi-task datasets exist (e.g., PMC-VQA (Zhang and et al. 2023b) includes question answering and report generation tasks), they fail to construct coherent chains of clinical reasoning. Such frag-

mented evaluation deviates significantly from real diagnostic workflows—radiologists need to compare patients' historical images (multi-image), incorporate pathology reports (multi-text), and perform multi-round VQA for stepwise diagnosis, capabilities that remain untested in current benchmarks. (4) **Multi-image integration**: There is a lack of evaluation for the integrated understanding and judgment of medical images from various modalities (e.g.,X-ray, MRI, CT, PET).

This work introduces **MedAtlas**, a benchmark dataset specifically designed to address these shortcomings. **MedAtlas** captures the full complexity of real-world diagnostic workflows by supporting multi-round, cross-modal, and multi-image reasoning, which ensures coherent clinical reasoning across different diagnostic stages, integrates diverse modalities (such as images and text), and evaluates the combined interpretation of medical images from various modalities.

Each case is organized as a sequence of question-answer (VQA) pairs spanning multiple diagnostic stages. It begins with the clinical history (a textual description of the patient's background), followed by several imaging stages. The first stage presents an initial set of images (e.g., radiographs) along with corresponding questions; the second stage introduces advanced imaging (e.g., contrast-enhanced MRI/CT) and follow-up questions that depend on information from the previous stage; if necessary, the sequence extends to a third stage (e.g., specific-position MRI accompanied by new queries), thereby simulating the progressive process of clinical examination and inquiry.

All VQA pairs collectively form a reasoning chain: subsequent questions build upon earlier findings, and the narrative of each case unfolds step by step, culminating in a final diagnostic or therapeutic question. The dataset encompasses multiple imaging modalities, including X-ray, CT, MRI, ultrasound, and PET (with some cases containing mixed modalities). Many stages require the simultaneous interpretation of multiple images (e.g., multi-plane MRI or multi-view CT), thereby necessitating cross-image integration capabilities.

Additionally, each case is annotated with structured information: imaging findings (textual descriptions of key observations), diagnostic outcomes, and detailed disease information (including pathophysiology, epidemiology, clinical manifestations, imaging characteristics, and treatment plans). Such annotations enable extended evaluations of models on tasks such as generating comprehensive clinical summaries and recalling domain-specific knowledge.

By systematically addressing the four major limitations of existing datasets, **MedAtlas** provides a benchmark that is more closely aligned with real-world clinical scenarios than current alternatives.

Main Contributions:

- **A novel benchmark dataset:** The first multi-task medical benchmark featuring multi-round VQA, multi-modal multi-image inputs, and explicitly defined clinical reasoning chains. **MedAtlas** consists of a diverse collection of cases covering a wide range of conditions. Each case is richly annotated with patient history, multiple imaging studies (Rounds), sequential QA pairs, and additional context like radiological findings and disease background. This dataset aims to bridge the gap between simplified VQA tasks and the complexities of actual clinical diagnosis.
- **Innovative evaluation metrics:** In addition to conventional QA accuracy, two new metrics are introduced: *Stage-Chain Accuracy*, which measures the longest sequence of consecutive correct diagnostic stages, and the *Error Propagation Suppression Coefficient*, which evaluates a model's ability to maintain downstream performance despite upstream errors.
- **Comprehensive evaluation of state-of-the-art vision-language models:** A systematic assessment of the latest 2025 versions of prominent multi-round, multi-image capable large models, including GPT, Deepseek, Qwen, Claude, LLaVA, InternVL and Kimi.

By providing the **MedAtlas** dataset, an accompanying evaluation framework, and baseline results, this work aims to drive medical VQA models toward more complex, interpretable, and robust clinical reasoning. The benchmark highlights the limitations of current models in handling complex real-world scenarios and provides clear directions for future research.

Related Work

Our work builds on prior research in Med-VQA, vision-language models (VLMs), multi-round VQA, and evaluation. Early Med-VQA datasets such as VQA-RAD (Lau et al. 2022), SLAKE (Liu and et al. 2021), and PathVQA (He, Wang, and et al. 2020) introduced radiology and pathology QA but were limited in scale, modality, and lacked multi-round or multi-image reasoning. Larger datasets like PMC-VQA (Zhang and et al. 2023b), OmniMedVQA (Liu and et al. 2024b), and MIMIC-Diff-VQA (Hu and et al. 2023) improved coverage and complexity, yet still mostly featured single-turn. Some works like RadVisDial explored synthetic multi-round VQA but remained narrow in scope. Parallel progress in general VLMs (e.g., GPT-4 (OpenAI 2023), Flamingo, BLIP-2 (Li and et al. 2023b), InstructBLIP (Dai and et al. 2023), LLaVA (Li and et al. 2023a)) and medical-specific variants (e.g., Med-Flamingo (Moor and et al. 2023), LLaVA-Med, RadFM (Zhang and et al. 2023a)) has enabled stronger multimodal reasoning. For evaluation, benchmarks like ConvBench (Liu and et al. 2024a), SparklesEval (Huang and et al. 2023), Mantis-Eval (Jiang and et al. 2024), and MMDU (Chen and et al. 2024) extend to multi-round and multi-image contexts but remain largely non-medical. Our work addresses these gaps by integrating multi-round VQA and multi-image reasoning within a medical setting.

MedAtlas Benchmark

MedAtlas is designed to rigorously evaluate the ability of VLMs to engage in multi-round, multi-modal, and multi-image reasoning within the context of realistic clinical diagnostic workflows. It moves beyond static, single-instance

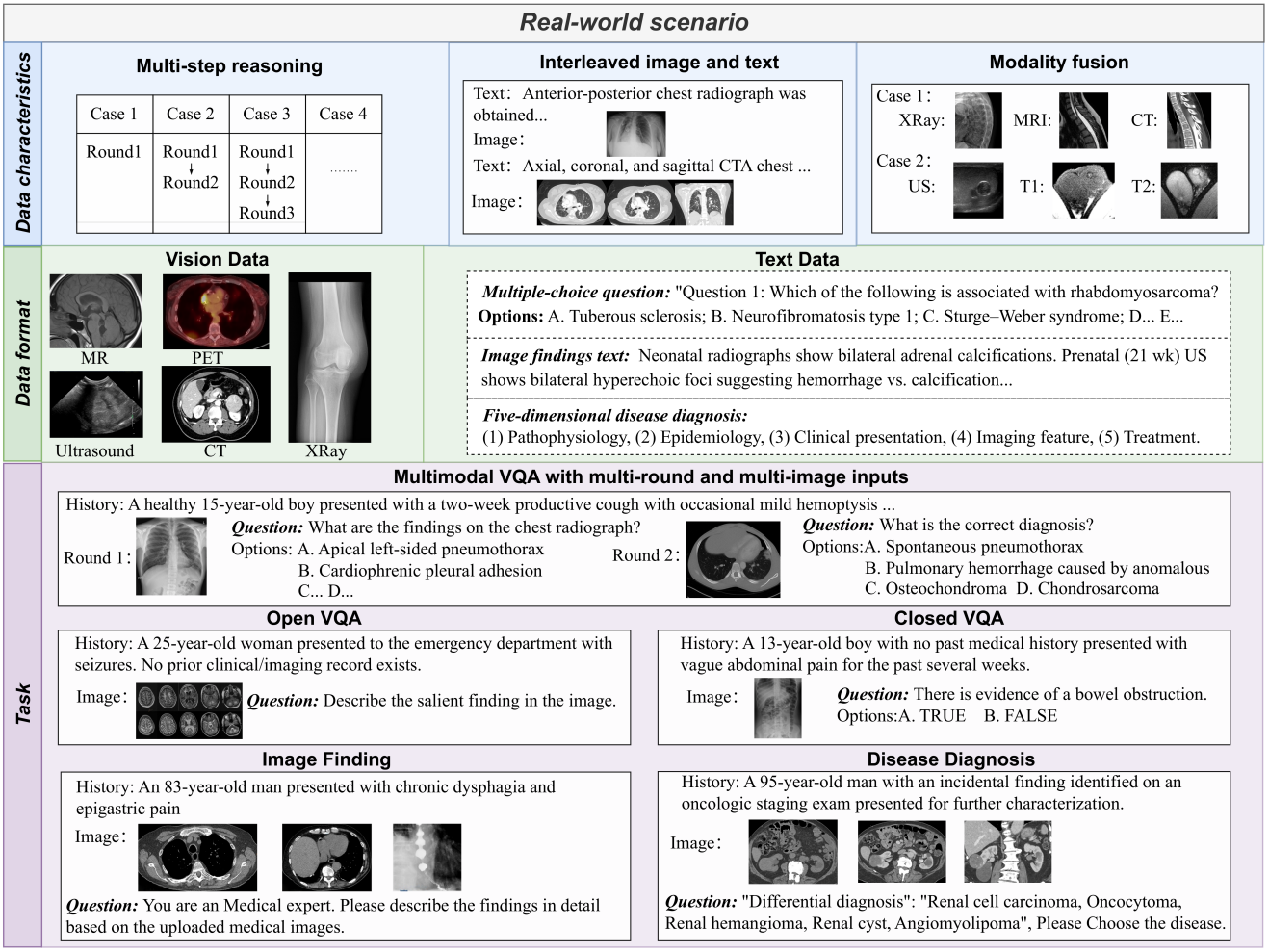

Figure 1: MedAtlas Benchmark Overview. MedAtlas simulates real-world diagnostic workflows through multi-round, multi-modal, multi-image reasoning, where each case encodes a patient's longitudinal trajectory (e.g., CT→MRI). Tasks span open/-closed VQA, multi-image reasoning, and diagnosis. Evaluation uses Stage Chain Accuracy (SCA) and Error Propagation Suppression Coefficient (EPSC) to assess reasoning consistency and error propagation.

VQA towards simulating the dynamic progression of patient cases.

Dataset Statistic

In this section, we provide a comprehensive statistical overview of the multi-round, multi-modal VQA reasoning task within our proposed benchmark, **MedAtlas** as shown in Figure 1. This task is specifically designed for evaluating the capabilities of Visual Language Models (VLMs) in realistic clinical diagnostic scenarios involving sequential reasoning across multiple VQA rounds, imaging modalities, and images.

Overall Dataset Statistics

As shown in Figure 3,the multi-round, multi-modal task in **MedAtlas** comprises 804 clinically realistic patient cases, encompassing a total of 5632 medical images. Each case is structured into multi-round VQAs with clinical reasoning tasks, containing a total of 1516 VQA rounds and 4015 question-answer pairs. On average, each case includes approximately 1.94 VQA rounds, with a median of 3 rounds per case; complex scenarios can extend up to about 10 rounds.

Question and Answer Characteristics

The questions in this task are predominantly structured as multiple-choice, with single-choice questions constituting the majority (1641 questions, accounting for 72.29%), complemented by True/False questions (629 questions, 27.71%). The average number of questions per case is approximately 4.96, mirroring realistic clinical diagnostic workflows.

The dataset features multiple-choice questions with varied numbers of options (ranging from 2 to 17), reflecting the diverse complexity of clinical scenarios. A detailed distribution is illustrated separately in our appendix, highlighting the predominance of questions with 4 or fewer options, typical

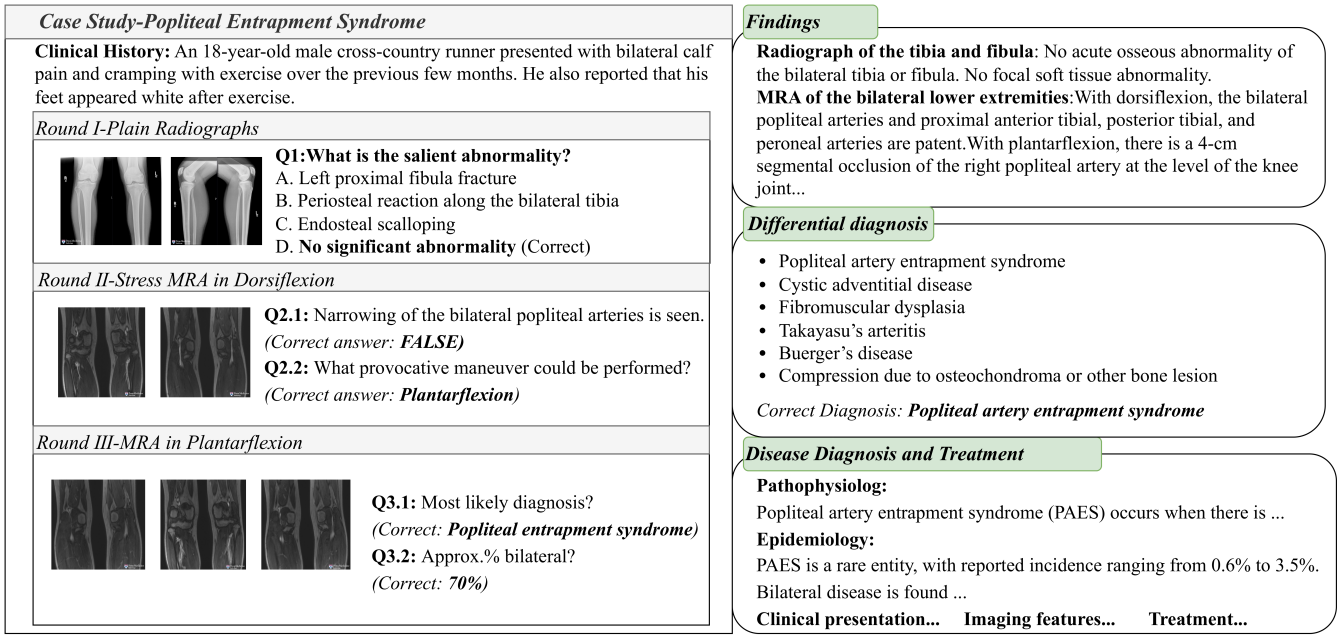

Figure 2: An example of a dataset includes multiple rounds of conversations and multi-image understanding.

of clinical diagnostic and educational scenarios. Additionally, cases with higher numbers of options represent more complex scenarios that require extensive differential diagnosis and advanced clinical reasoning.

Imaging Modalities

As shown in Figure 3, **MedAtlas** spans diverse imaging modalities, led by **CT (36.6%)** and **MRI (35.2%)**, followed by ultrasound (14.0%) and X-ray (11.2%). Low-frequency types (CTA, nuclear medicine, PET/CT, endoscopy; each ∼0.6–0.9%) add further diversity. This broad procedure-level coverage imposes strong modality-robust reasoning demands on VLMs.

Question and Image Count Distribution

The distribution of question counts per case in this task shows most cases (approximately 76.6%) contain between 4 and 6 questions, with 42.2% of cases containing exactly 5 questions. Only a minor proportion of cases involve fewer than 3 or more than 8 questions, ensuring both clinical relevance and complexity variability.

Regarding image count per case, most cases (50.1%) have between 6 and 10 images, followed by 36.7% with 1 to 5 images. Approximately 10.5% of cases are more complex with 11–15 images, and only a minimal fraction (0.4%) exceed 20 images. This structured diversity ensures comprehensive coverage of both routine and complex clinical scenarios.

Clinical Reasoning and Complexity

Questions in the multi-round, multi-modal VQA task as shown in Figure 2 within **MedAtlas** follow a clear clinical reasoning progression, starting from initial observations, advancing through analysis and differential diagnosis, and culminating in final diagnostic conclusions and treatment recommendations. The dataset inherently facilitates multi-hop reasoning, as later questions frequently depend on accurate interpretation of previous images and findings. This design closely mirrors authentic clinical reasoning processes.

Overall, this task within **MedAtlas** represents a rigorous and clinically realistic benchmark. Its unique multi-round VQA structure, diverse multi-modal inputs, and requirement for integrating multi-image information provide a comprehensive evaluation platform for testing VLMs' capabilities in clinical reasoning, medical knowledge retrieval, and diagnostic decision-making.

Evaluation Metrics

As discussed earlier, evaluating a model on **MedAtlas** requires both classical metrics and new, specialized metrics to capture reasoning chains. We detail our metrics and how they are computed.

Standard VQA Metrics

For each question in **MedAtlas**, if a standard short answer or multiple-choice answer is available, binary correctness is computed. The overall accuracy is defined as the proportion of fully correct answers; specifically, multiple-choice questions require selecting all correct options, and single-word or short-phrase responses must match the expected standardized strings. This approach aligns with existing VQA evaluation methodologies (Lau et al. 2022; Liu and et al. 2021). For open-ended questions requiring sentence-level answers, GPT-4o is employed to evaluate the semantic consistency between the generated answers and the reference answers, mitigating erroneous judgments caused by differences in phrasing or synonymous expressions.

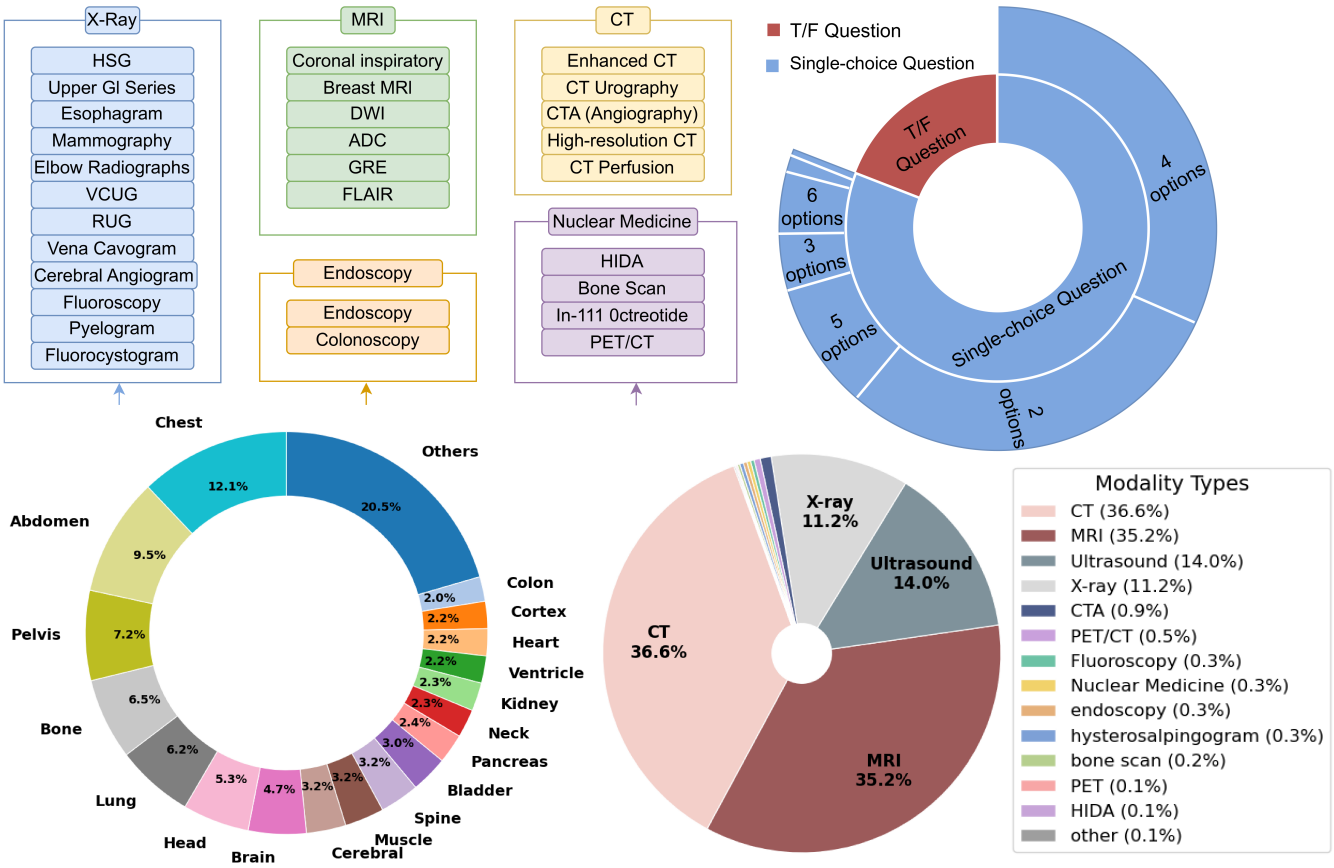

Figure 3: Medical Imaging Data Distribution and Analysis. The figure summarizes the imaging modalities covered in MedAtlas and their subcategories (top-left), the distribution of question formats and option counts (top-right), the frequency of examined body regions (bottom-left), and the overall proportions of modality types (bottom-right). CT and MRI account for the largest shares of studies, and chest and abdomen are among the most frequently examined regions.

Stage Chain Accuracy (SCA). SCA measures how far a model can progress through a multi-round task without error, emphasizing sequential consistency. For each case c, we identify the longest prefix of Rounds answered completely correctly. Let $L(c)$ denote this chain length, and let $\{w_1, w_2, \ldots, w_K\}$ be Round weights satisfying $w_1 < w_2 < \cdots < w_K$ to emphasize later Rounds. The per-case SCA score is then:

$$\mathrm{SCA}(c) = w_{L(c)}. \tag{1}$$

The overall SCA is computed as the average across all cases:

$$\mathrm{SCA} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{SCA}(c), \tag{2}$$

where $\mathcal{C}$ is the set of all cases. We also report the distribution of chain lengths (e.g., proportion of cases reaching Round I, Round II, etc.) to characterize where models tend to fail.

Interpretation. A high SCA indicates that the model can sustain correct answers over multiple sequential Rounds, whereas a steep drop in chain length distribution highlights systematic failures in later Rounds. Unlike per-question accuracy, SCA reflects both accuracy and error ordering, thus directly measuring sequential reasoning consistency.

Error Propagation Suppression Coefficient (EPSC)

To quantify the degree to which early-stage errors affect downstream performance, we compute the *Error Propagation Suppression Coefficient (EPSC)* across all multi-Round cases. We focus on error propagation from Round I to Round II, which is the most critical and widely available Round pair in our dataset.

Definition. Let $A(\text{correct\_prev})$ denote the Round II accuracy across all cases where the model answered all Round I questions correctly. Similarly, let $A(\text{wrong\_prev})$ denote the Round II accuracy across all cases where the model made at least one error in Round I. The EPSC is then defined as:

$$\mathrm{EPSC} = \frac{A(\text{wrong\_prev})}{A(\text{correct\_prev})}. \tag{3}$$

Here,

$$A(\text{correct\_prev}) = \frac{1}{|\mathcal{C}_{\text{correct}}|} \sum_{c \in \mathcal{C}_{\text{correct}}} \frac{\sum_{q \in \mathcal{P}_2(c)} \mathbf{1}\left[\hat{y}_q = y_q\right]}{|\mathcal{P}_2(c)|}, \tag{4}$$

$$A(\text{wrong\_prev}) = \frac{1}{|\mathcal{C}_{\text{wrong}}|} \sum_{c \in \mathcal{C}_{\text{wrong}}} \frac{\sum_{q \in \mathcal{P}_2(c)} \mathbf{1}\left[\hat{y}_q = y_q\right]}{|\mathcal{P}_2(c)|}, \tag{5}$$

where $\mathbf{1}[\cdot]$ is an indicator function that equals 1 if the condition is true and 0 otherwise, $\mathcal{C}_{\text{correct}}$ is the set of cases where Round I is fully correct, $\mathcal{C}_{\text{wrong}}$ is the set of cases where Round I contains any incorrect answer, $\mathcal{P}_2(c)$ denotes all questions in Round II of case c, $\hat{y}_q$ is the model's prediction, and y_q is the ground truth.

Interpretation. An EPSC close to 1.0 indicates that Round II accuracy is largely unaffected by errors in Round I, implying strong error suppression and robustness to upstream mistakes. Conversely, a low EPSC (e.g., < 0.7) indicates substantial error propagation: when the model errs in Round I, its subsequent reasoning degrades significantly.

Implementation Detail. Unlike prior definitions restricted to per-case conditional analysis, we compute EPSC *globally* across all cases in the dataset. This approach aggregates evidence from all available samples without requiring both "correct" and "wrong" Round I outcomes to appear within the same case, thereby yielding a more statistically stable estimate:

$$\mathcal{C}_{\text{correct}} \cap \mathcal{C}_{\text{wrong}} = \emptyset. \tag{6}$$

This global formulation better reflects overall model robustness to early-stage errors in multi-Round reasoning tasks.

Findings Evaluation

For the **findings generation** task, we evaluate model outputs using four widely adopted language metrics: BLEU (Papineni and et al. 2002) (measuring n-gram overlap), ROUGE-L (Lin 2004) (capturing longest common subsequence similarity), METEOR (Banerjee and Lavie 2005) (accounting for stemming and synonymy), and BERTScore computed using **BioClinicalBERT** (Lee et al. 2020), a biomedical domain–specific encoder that better reflects semantic similarity in clinical contexts. These metrics collectively assess linguistic fidelity and semantic alignment between generated findings and expert-written references as shown in Table 2.

To obtain a single unified score for evaluation and comparison, we compute a weighted average of these four metrics, which we refer to as the overall **Findings Score(Findings)**. The weighted aggregation balances lexical overlap (BLEU, ROUGE-L), semantic similarity (METEOR), and embedding-level clinical coherence (BERTScore). The final Findings Score reported in Table 1 corresponds to this aggregated metric.

Diagnostic Evaluation

Unlike findings generation, diagnostic evaluation focuses on determining whether a model's predicted disease or condition matches the ground-truth diagnosis. Because model responses may contain clinically reasonable synonyms, paraphrases, or descriptive variants (e.g., "ACL tear" vs. "torn ACL"), direct string matching is insufficient for reliable assessment.

To ensure robust evaluation, we employ **GPT-4o** as an automated expert judge. Given a model-generated diagnosis and the reference answer, GPT-4o determines whether the two are clinically equivalent based on medical semantics, terminology normalization, and contextual consistency.

The resulting **Diagnosis** metric reflects the model's diagnostic reasoning capability. It is computed as the proportion of cases for which the model's prediction is judged correct out of the total number of diagnostic cases. This score serves as the primary indicator of Diagnostic Evaluation ability and is reported in Table 1.

Knowledge Accuracy (Knowledge Acc.) Calculation

To better capture the composite capability of visual-language models (VLMs) across multiple dimensions, we introduce **Knowledge Accuracy (Knowledge Acc.)**, which aggregates four core metrics: Open QA Accuracy, Closed QA Accuracy, Multi-Image Findings, and Diagnosis Accuracy.

Step 1: Metric Aggregation. We first compute the mean performance μ_i of model i across these four metrics:

$$\begin{aligned}\mu_i = \frac{1}{4}\Big(&\text{OpenQA}_i + \text{ClosedQA}_i \\ &+ \text{MultiImage}_i + \text{Diagnosis}_i\Big).\end{aligned} \tag{7}$$

Step 2: Z-score Standardization. To normalize and compare models on a common scale, we apply Z-score standardization:

$$Z_i = \frac{\mu_i - \overline{\mu}}{\sigma}, \tag{8}$$

where $\overline{\mu}$ and σ represent the mean and standard deviation of μ_i across all models.

Step 3: Logistic Mapping. Finally, to emphasize performance differences between high-performing and low-performing models, we map Z_i into the $[0, 1]$ range using a logistic transformation:

$$\text{KnowledgeAcc}_i = \frac{1}{1 + e^{-k \cdot Z_i}}, \tag{9}$$

where k is a scaling hyperparameter (we set $k = 3$ in our experiments) controlling the steepness of the logistic curve.

Experimental Analysis

We evaluate eleven vision-language models (VLMs), spanning both general-purpose and medical-specialized systems, on the proposed **MedAtlas benchmark**. The results are summarized in Table 1 (core benchmark) and Table 2 (findings similarity metrics). This section presents a detailed analysis across four dimensions: overall knowledge reasoning (*Knowledge Acc.*), task-specific performance, cross-metric correlations, and linguistic fidelity in medical report alignment.

Model	Open VQA	Closed VQA	Findings	Diagnosis	Knowledge Acc.	SCA	EPSC
InternVL3-78B	10.80	25.21	32.03	6.50	0.01	0.25	0.16
Qwen2.5-VL-7B	25.21	14.80	32.03	6.50	0.01	0.30	0.20
Qwen2.5-VL-32B	20.00	57.70	27.85	7.40	0.09	0.60	0.35
Kimi-latest	15.30	63.61	33.30	8.10	0.15	1.94	0.72
Qwen2.5-VL-72B	22.40	68.18	35.43	30.60	0.72	1.50	0.65
Claude-3.7	28.40	71.06	34.77	32.50	0.85	1.16	0.70
Qwen-VL-Max	30.71	68.35	35.25	33.30	0.86	1.96	0.74
Deepseek V3	31.15	71.10	34.95	34.00	0.89	1.94	0.68
GPT-4o	34.21	73.52	35.27	30.70	0.90	1.93	1.00
llama-4-maverick	32.50	71.69	36.12	34.70	0.91	1.70	0.82
claude-sonnet-4	36.73	75.76	35.30	36.60	**0.95**	1.30	0.78

Table 1: Normalized performance of general-purpose and medical-specific VLMs on MedAtlas, sorted by Knowledge Acc.

Model	BS	R-L	B-1	MET
Qwen2.5-VL-32B	70.0	12.6	7.9	20.9
Qwen2.5-VL-7B	72.7	18.6	17.4	19.4
InternVL3-78B	72.7	18.6	17.4	19.4
Kimi-latest	73.6	21.0	16.7	21.9
GPT-4O	75.1	22.2	19.7	24.1
Qwen2.5-VL-72B	75.0	22.9	19.7	24.1
Claude-3.7	74.8	22.0	18.9	23.4
Qwen-VL-Max	75.0	22.6	19.3	24.1
Deepseek V3	74.5	22.2	20.4	22.7
claude-sonnet-4	75.6	22.4	17.2	26.0
llama-4-maverick	75.3	24.6	21.0	23.6

Table 2: Performance of findings similarity metrics: BERTScore(BS), ROUGE-L(R-L), BLEU-1(B-1), METEOR(MET) [in %]

Overall Knowledge Reasoning Performance

The *Knowledge Acc.*, derived from Z-score normalization followed by logistic mapping across *Open QA*, *Closed QA*, *Multi-Image Findings*, and *Diagnosis*, offers a unified view of knowledge-grounded multimodal reasoning. As shown in Table 1, **Claude-sonnet-4** achieves the highest score (0.95), followed closely by **LLaMA-4-Maverick** (0.91), **GPT-4o** (0.90), and **Deepseek V3** (0.89). These models consistently demonstrate strong general reasoning while retaining sufficient medical grounding. In contrast, smaller models such as **InternVL3-78B** and **Qwen2.5-VL-7B** record near-zero scores, underscoring the substantial gap between state-of-the-art and early-generation or undertrained systems. Statistical tests (one-way ANOVA, $p < 0.01$) confirm significant separation between these performance tiers.

Task-Specific Analysis

Open vs. Closed QA. Although most models achieve reasonable scores in Closed QA (e.g., Claude-sonnet-4: 75.76), Open QA remains challenging, with a notable drop (Claude-sonnet-4: 36.73). This disparity indicates that while factual recall anchored in fixed knowledge is well addressed, open-ended clinical reasoning involving implicit context remains underexplored.

Multi-Image Findings and Temporal Fusion. Performance on Multi-Image Findings—computed as the weighted average of BERTScore, ROUGE-L, BLEU-1, and METEOR—strongly correlates with diagnosis accuracy (Pearson $r = 0.82$). Models such as **LLaMA-4-Maverick** (36.12) excel, suggesting improved temporal fusion and spatial reasoning, whereas **Qwen2.5-VL-32B** (27.85) lags, indicating insufficient cross-slice integration typical of CT or MRI interpretation workflows.

Diagnosis Accuracy. Diagnosis performance mirrors the stratification in Knowledge Acc, with **Claude-sonnet-4** (36.6), **Deepseek V3** (34.0), and **LLaMA-4-Maverick** (34.7) outperforming others. This alignment underscores that diagnostic reasoning benefits from both language fidelity and multimodal fusion, rather than superficial report similarity alone.

Findings Similarity and Language Fidelity

Table 1 (findings similarity) evaluates models on report alignment. **Claude-sonnet-4** again leads in BERTScore (75.6) and METEOR (26.0), reflecting superior semantic alignment with radiology lexicons. Conversely, while **LLaMA-4-Maverick** achieves the highest BLEU-1 (21.0), lexical overlap alone proves less predictive of diagnostic accuracy, reinforcing the importance of embedding-based semantic metrics over token-level comparisons in clinical text evaluation.

Sequential Consistency and Error Propagation Analysis.

As shown in Table 1, models with higher **Knowledge Acc.** generally achieve stronger **Stage Chain Accuracy (SCA)**, indicating better consistency in sustaining correct reasoning across multi-Round tasks. For example, GPT-4o and Qwen-VL-Max exhibit near-maximal SCA ($\approx$2.0), demonstrating their ability to maintain correctness across consecutive Rounds. In contrast, weaker models (e.g., InternVL3-78B) fail early in the reasoning chain, leading to much lower SCA (< 0.3).

Similarly, **Error Propagation Suppression Coefficient (EPSC)** correlates with model robustness: high-performing models such as GPT-4o (EPSC=1.0) show minimal degradation in later Rounds even after early errors, whereas low-performing models (EPSC < 0.3) exhibit strong error propagation, compounding mistakes across Rounds. These results highlight that advanced VLMs not only improve single-Round accuracy but also exhibit enhanced sequential reasoning resilience, a critical property for multi-step medical decision-making.

Cross-Metric Correlations and Error Patterns

Cross-metric analysis highlights several systematic weaknesses. First, models with poor Multi-Image Findings (e.g., Qwen2.5-VL-7B) display parallel deficits in Diagnosis, revealing that limited cross-view integration constrains downstream reasoning. Second, models with high Closed QA but low Open QA (e.g., Qwen2.5-VL-72B) exhibit retrieval-oriented behavior, relying on memorized facts rather than generative inference. These patterns suggest that advancing temporal attention and incorporating retrieval-augmented strategies may be critical for further gains.

Key Insights

We draw three primary insights: (1) **Scaling alone is insufficient.** While larger general models like GPT-4o achieve competitive reasoning, specialized pretraining (e.g., Deepseek V3) is crucial for clinical fidelity. (2) **Multi-image reasoning is a bottleneck.** Strong correlation with diagnosis highlights this as an indispensable benchmark axis for future VLM development. (3) **Semantic metrics better capture clinical alignment.** BERTScore and METEOR more closely track diagnostic performance than lexical overlap metrics, suggesting their suitability for medical language evaluation.

Conclusion

Overall, while frontier models approach robust medical reasoning, substantial gaps persist in open-ended VQA, multi-image contextualization, and semantically grounded language generation. The proposed **MedAtlas benchmark** thus not only ranks existing systems but also exposes modality-specific limitations, offering a rigorous basis for guiding next-generation medically grounded vision-language models.

Acknowledgments

This work was supported in part by the Jiangsu Province Science Foundation for Youths under Grant BK20240464, in part by the Natural Science Foundation of China under Grant 62271465, and in part by the Suzhou Basic Research Program under Grant SYG202338.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop*.

Chen, S.; and et al. 2024. MMDU: A Benchmark for Multi-turn Multi-image Dialog Understanding. *arXiv preprint arXiv:2402.00102*.

Dai, X.; and et al. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*.

He, T.; Wang, R.; and et al. 2020. PathVQA: Visual question answering for pathology images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Hu, W.; and et al. 2023. MIMIC-Diff-VQA: A Longitudinal Visual Question Answering Benchmark for Radiology. *arXiv preprint arXiv:2303.16208*.

Huang, Y.; and et al. 2023. SparklesEval: An LLM-Based Multimodal Dialogue Evaluation Benchmark. *arXiv preprint arXiv:2305.06772*.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Jiang, J.; and et al. 2024. MantisEval: A Multi-Image Reasoning Evaluation Benchmark. *arXiv preprint arXiv:2402.00771*.

Lau, J.; Gayen, S.; Abedin, A. R.; and et al. 2022. VQA-RAD: Visual Question Answering Dataset for Radiology. *arXiv preprint arXiv:1909.11315*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, H.; and et al. 2023a. LLaVA: Large Language and Vision Assistant. *arXiv preprint arXiv:2304.08485*.

Li, J.; and et al. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*.

Liu, J.; and et al. 2024a. ConvBench: A Benchmark for Evaluating Multi-turn Dialogue of Large Vision-Language Models. *arXiv preprint arXiv:2401.11538*.

Liu, J.; and et al. 2024b. OmniMedVQA: A Comprehensive Benchmark for Medical Vision Question Answering. *arXiv preprint arXiv:2401.13946*.

Liu, Z.; and et al. 2021. SLAKE: A benchmark dataset for medical visual question answering. *Medical Image Analysis*, 70: 102024.

Moor, M.; and et al. 2023. Med-Flamingo: A multimodal medical few-shot learner. *arXiv preprint arXiv:2306.05425*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Papineni, K.; and et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Zhang, H.; and et al. 2023a. RadFM: A Foundation Model for Radiology. *arXiv preprint arXiv:2312.05406*.

Zhang, T.; and et al. 2023b. PMC-VQA: A large-scale biomedical visual question answering benchmark. *arXiv preprint arXiv:2306.16746*.