

# Beyond Single View: A Comprehensive Benchmark for Medical Multimodal Large Language Models on Multi-Image Understanding

Dexuan Xu<sup>1</sup>, Jiayin Yuan<sup>2</sup>, Jianing Wang<sup>2</sup>, Yanyuan Chen<sup>3</sup>, Hanpin Wang<sup>1</sup>, Yu Huang<sup>4\*</sup>

<sup>1</sup> School of Computer Science, Peking University

<sup>2</sup> School of Software & Microelectronics, Peking University

<sup>3</sup> School of Data Science, University of Virginia

<sup>4</sup> National Engineering Research Center for Software Engineering, Peking University

## Abstract

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in interpreting single medical images. However, real-world clinical diagnosis is intrinsically a multi-view process, requiring the synthesis of information across volumetric slices, temporal sequences, and comparative modalities. Existing benchmarks fail to capture this complexity, limiting the assessment of models in realistic clinical workflows. To bridge this gap, we introduce MedMultiBench, the first large-scale benchmark specifically designed for medical multi-image understanding. Comprising 11,392 expert-curated samples, MedMultiBench evaluates MLLMs across four distinct dimensions: Joint Reasoning, Comparative Analysis, Comprehensive Perception, and In-Context Learning. We benchmark 13 state-of-the-art MLLMs, revealing that while current models excel in single-view tasks, they struggle significantly with multi-image contexts. Our experiments identify a performance degradation in open-source models when processing increased visual loads, whereas closed-source models demonstrate better scalability. MedMultiBench provides a robust framework to facilitate the development of MLLMs capable of holistic clinical reasoning.

## 1 Introduction

The convergence of computer vision and natural language processing has accelerated the development of medical multimodal large language models (MLLMs). Models such as LLaVA-Med (Li et al., 2023) and Med-PaLM (Tu et al., 2024) have achieved expert-level proficiency on standard visual question answering (VQA) benchmarks. These achievements suggest the imminent potential for assisting radiological interpretation. However, current evaluation benchmarks predominantly treat

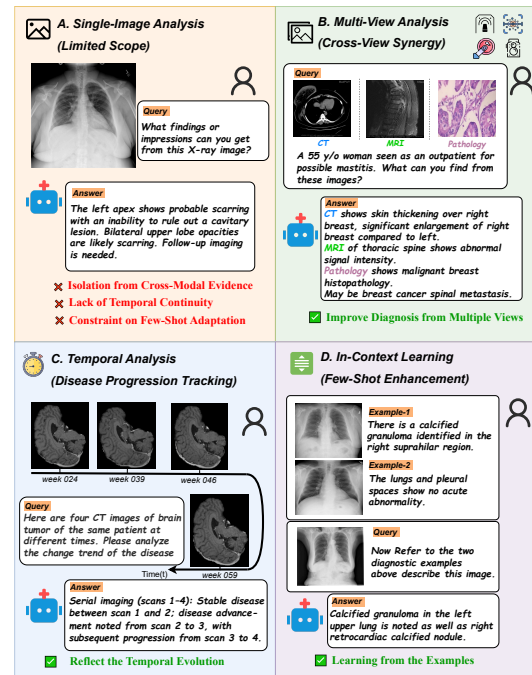


Figure 1: The motivation for multi-image analysis capabilities in medical MLLMs.

medical visual understanding as a static classification task restricted to a single image (Xu et al., 2024), which fundamentally diverges from actual clinical scenarios (Yang et al., 2025b).

As illustrated in Figure 1, a radiologist rarely diagnoses a patient based on a single isolated slice. The diagnostic process requires scrolling through adjacent slices to trace anatomical structures, comparing current scans with priors to track disease progression, or correlating findings across different imaging modalities. A practitioner relying on a single snapshot would be deemed clinically insufficient (Tang et al., 2024). Consequently, relying on single-image evaluation creates a disparity between model performance metrics and real-world utility.

Current medical VQA benchmarks, including VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PMC-VQA (Zhang et al., 2023), remain predominantly restricted to single-image interpre-

\* Corresponding author: Yu Huang, hy@pku.edu.cn.

tation, largely overlooking inter-image relational dynamics. While the general computer vision community has shifted toward multi-image (Zhao et al., 2024; Wang et al., 2025) and video-based understanding (Li et al., 2024), medical adaptations remain limited. Recent efforts like MedFrameQA (Yu et al., 2025) introduce multi-frame inputs but are confined to instructional videos, failing to capture broader diagnostic reasoning such as differential correspondence or spatial reconstruction. Therefore, the lack of a systematic evaluation framework leaves the abilities of medical MLLMs in complex clinical reasoning unquantifiable.

To bridge this gap, we introduce **MedMultiBench**, a comprehensive benchmark specifically engineered to evaluate the multi-image understanding capabilities of MLLMs. Moving beyond simple image-text pairing, MedMultiBench requires the holistic synthesis of temporal, spatial, and comparative evidence across more than 11,000 expert-curated samples. We formalize medical multi-image reasoning into a novel taxonomy comprising four essential dimensions: (1) *Joint Reasoning* for cross-view feature fusion; (2) *Comparative Analysis* for differential diagnosis; (3) *Comprehensive Perception* for higher-order logic such as spatial ordering and temporal analysis; and (4) *In-Context Learning* for few-shot adaptability using interleaved clinical demonstrations. Our contributions are as follows:

(1) We establish a formal taxonomy for multi-image medical reasoning, defining inter-image relationships through four key dimensions: spatiotemporal continuity, cross-view heterogeneity, differential correspondence, and semantic reference.

(2) We introduce MedMultiBench, the first large-scale benchmark for multi-image clinical understanding, comprising 11,392 samples across 13 subtasks to rigorously evaluate holistic evidence synthesis in realistic scenarios.

(3) We conduct an extensive evaluation of 13 state-of-the-art MLLMs. Our experiments reveal that current open-source models suffer significantly from information overload in multi-image scenarios compared to closed-source counterparts, highlighting critical directions for future optimization.

## 2 Related Work

### 2.1 Multimodal Large Language Models

Recent multimodal large language models have achieved significant progress in vision tasks (Hurst

et al., 2024; Team et al., 2023; Yang et al., 2025a). However, the performance of these models in medical domain is constrained due to the lack of specialized medical knowledge. To bridge this gap, medical MLLMs such as LLaVA-Med (Li et al., 2023), HuatuoGPT-Vision (Chen et al., 2024), Med-Flamingo (Moor et al., 2023), Lingshu (Xu et al., 2025) and Fleming-VL (Shu et al., 2025) have been developed and achieve better scores on existing medical VQA benchmarks.

### 2.2 Medical Vision-Language Benchmarks

The landscape of medical VQA has evolved significantly. Early datasets like VQA-RAD (Lau et al., 2018) and PathVQA (He et al., 2020) established the baseline for closed-ended and open-ended QA on Radiology and Pathology images. Additionally, SLAKE (Liu et al., 2021) introduced bilingual annotations and richer semantic segmentation masks, while PMC-VQA (Zhang et al., 2023) scaled up the data magnitude by mining image-text pairs from open-access literature. More recently, works like OmniMedVQA (Hu et al., 2024) and MLEBench (Xu et al., 2024) expanded the scope by providing large-scale, comprehensive evaluation across more medical modalities and tasks. To move beyond simple perception, MedXpertQA (Zuo et al., 2025) introduced a benchmark specifically designed to test expert-level clinical reasoning, focusing on challenging cases that require deep medical knowledge. Despite their contributions, these benchmarks share a common limitation: inputs are restricted to single image. This restriction prevents the assessment of cross-referencing skills, which are vital for reducing diagnostic ambiguity.

### 2.3 Multi-Image Understanding Benchmark

In the general domain, evaluating the capability of MLLMs to reason across multiple visual inputs has become a pivotal research direction (Meng et al., 2024; Jiang et al., 2024; Liu et al., 2024). In the medical domain, MedMIM (Yang et al., 2025b) proposed a multi-image instruction-following benchmark to evaluate the multi-image understanding ability of MLLMs. MedFrameQA (Yu et al., 2025) extracts temporally coherent frames from videos to evaluate the reasoning capabilities of MLLMs in real-world clinical diagnosis. However, existing benchmarks focus on limited modalities or simple tasks and cannot effectively simulate real-world clinical scenarios. Our work aims to fill this gap by providing an expert-

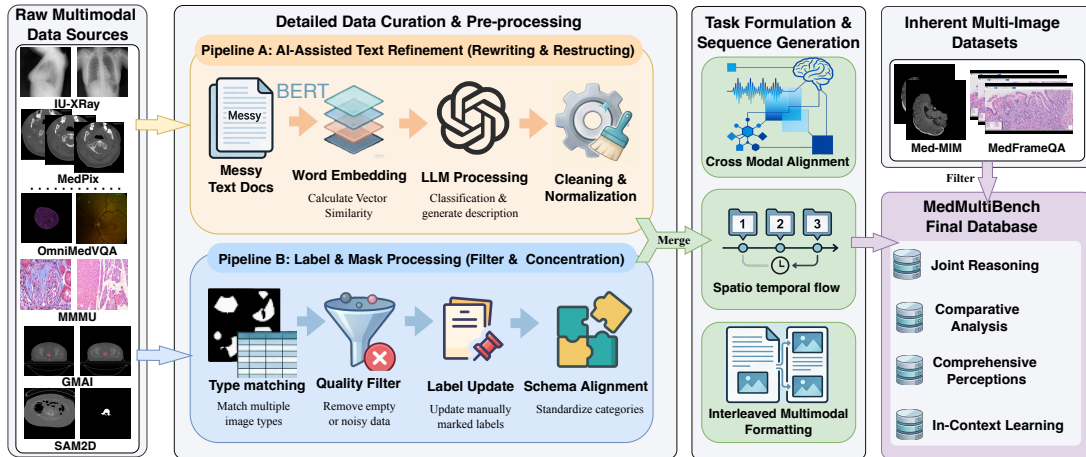


Figure 2: Construction pipeline of MedMultiBench.

curated, large-scale multi-image medical benchmark.

### 3 Preliminary

While standard MLLMs optimize the conditional probability  $P(y | I, T)$  for single image contexts (Huang et al., 2022), this paradigm often reduces medical diagnostics to static instance recognition. Such a simplified approach fails to account for the multi-source evidence synthesis essential to clinical decision-making. We therefore define *multi-image medical reasoning* as generating a response  $y$  given a textual instruction  $T$  and a set of  $N$  images  $\mathcal{I} = \{I_1, \dots, I_N\}$  where  $N \geq 2$ :

$$y = \operatorname{argmax}_y P(y | \mathcal{I}, T). \quad (1)$$

To succeed in this task, a model must learn an interaction function that captures the underlying relational structure of  $\mathcal{I}$ . We characterize these relationships through four primary dimensions.

**Spatiotemporal Continuity.** This category encompasses images constrained by physical or temporal proximity, simulating the foundational process of navigating volumetric or longitudinal medical data. In volumetric scenarios, the set  $\mathcal{I}$  consists of sequential 2D slices that represent a continuous 3D anatomical structure. The model is required to understand the spatial arrangement of these slices to maintain structural coherence. Similarly, for temporal data, the images capture the same anatomical region at different time points. Reasoning in this context shifts from static recognition to analyzing transitions, such as tracking tumor progression or evaluating treatment response over time.

**Cross-view Heterogeneity.** Images in this category represent the same subject through diverse

imaging physics or viewpoints. This includes multi-modal correlation between MRI and CT scans or multi-angle projections such as frontal and lateral X-rays. Unlike continuity-based tasks, this dimension requires synergistic integration; the model must perform cross-modal alignment to resolve ambiguities, effectively triangulating the diagnosis through diverse sensory inputs.

**Differential Correspondence.** This category focuses on image pairs curated for comparative analysis to identify subtle pathological or physiological shifts. Typical clinical scenarios include evaluating contrast enhancement between pre- and post-contrast phases. The model must distinguish persistent anatomical structures from interval changes.

**Semantic Reference.** Unlike other categories, these images are linked by analogical relevance rather than patient identity. This dimension encompasses visual retrieval based on shared attributes and in-context learning, where the image set  $\mathcal{I}$  contains support demonstrations consisting of image-text pairs  $\{I_k, T_k\}$ . In these cases, the model must perform analogical reasoning to map patterns from reference examples to the target query, necessitating a meta-level understanding of diagnostic logic.

## 4 MedMultiBench Construction

To systematically evaluate MLLMs in complex clinical scenarios, we introduce **MedMultiBench**, a large-scale benchmark comprising 11,392 samples. As illustrated in the construction pipeline shown in Figure 2, MedMultiBench integrates diverse high-quality sources such as IU-Xray (Demner-Fushman et al., 2015), GMAI (Ye et al., 2024), SAM-Med2D (Cheng et al., 2023), MedPix (Siragusa et al., 2025), OmniMedVQA (Hu et al., 2024), MMMU (Yue et al.,

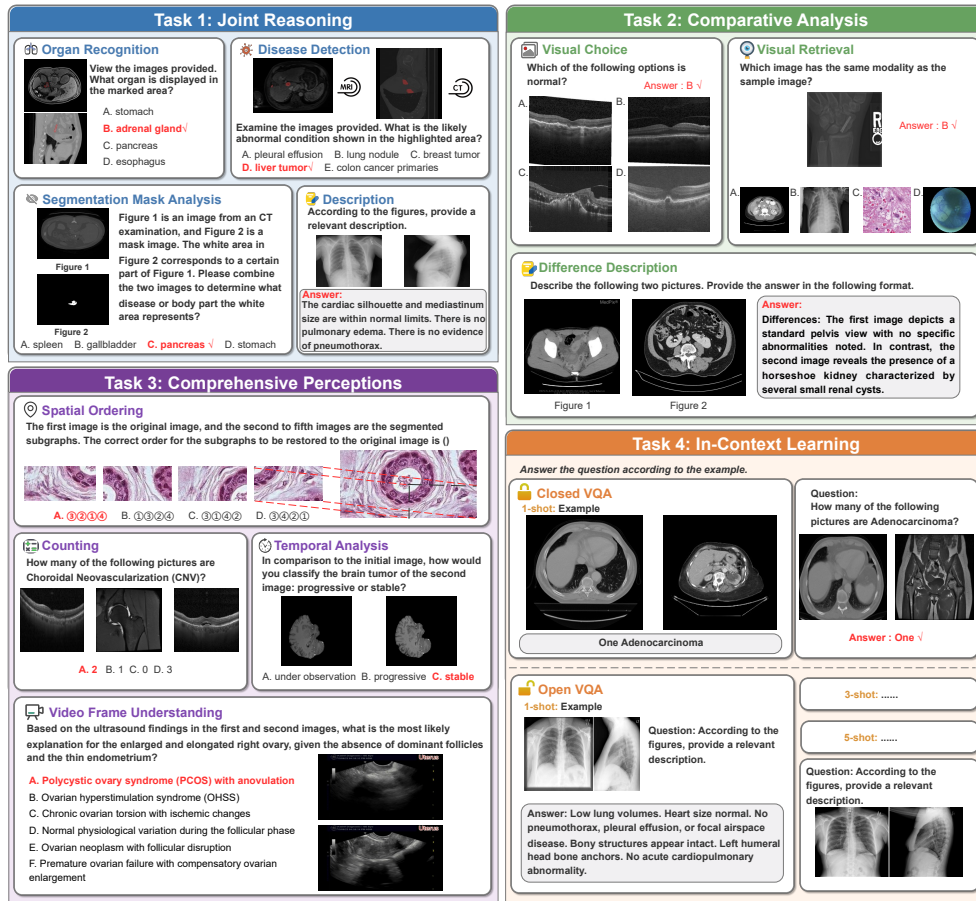


Figure 3: An illustration of MedMultiBench, including 4 tasks and 13 subtasks.

2024), Med-MIM (Yang et al., 2025b) and Med-FrameQA (Yu et al., 2025). Our curation process ensures clinical relevance through rigorous data selection, expert-verified quality control, and task-specific formatting. These core capability dimensions are structured to align directly with the inter-image relationships defined in the previous section.

#### 4.1 Task 1: Joint Reasoning

This task evaluates the ability to synthesize semantic information from multiple visual inputs, primarily targeting the cross-view heterogeneity.

**Organ Recognition and Disease Detection.** These subtasks assess fine-grained identification across different planes, such as axial and coronal views, or pathological diagnoses requiring feature fusion from multiple complementary modalities. We constructed 1019 organ recognition examples and 703 disease detection examples from GMAI.

**Segmentation Mask Analysis.** This subtask aims to explore the ability to visually align the original image  $I$  with a binary mask  $M \in \{0, 1\}^{H \times W}$ . Unlike standard VQA, the model must map the anatomical structures represented by  $M$  onto  $I$  to identify highlighted regions of interest without

pixel-level preprocessing. We screened 978 samples from SAM-Med2D, containing 25 classes of anatomical structures.

**Multi-image Description.** Real-world clinical scenarios require models to generate clinical reports based on multi-view inputs, such as frontal and lateral X-rays, or CT modalities and X-ray modalities. This necessitates the model’s ability to synthesize multi-view evidence into a structured narrative. We filtered 992 samples from IU-Xray and MedPix to obtain a standard, high-quality report generation test set with multi-image inputs.

#### 4.2 Task 2: Comparative Analysis

Clinical diagnosis is inherently differential. This category maps to the differential correspondence by evaluating the discriminative power between related visual instances.

**Visual Choice and Visual Retrieval.** Visual choice employs a reverse VQA paradigm, requiring the model to select the correct image ( $I_k$ ) from a candidate image set  $O = \{I_1, I_2, \dots, I_M\}$  that best matches the clinical description  $q$ . The core objective is to evaluate how the model can differentiate between multiple visual candidates and make

decisions based on professional medical cues. Visual Retrieval focuses on semantic attribute consistency retrieval across images. Specifically, given a query image  $I_{query}$  and a text description  $q$  for a specific semantic attribute, the model must identify the target image from the candidate image set  $O$  that is consistent with  $I_{query}$  in the dimension specified by  $q$ . We construct 989 visual choice examples and 835 visual retrieval examples from OmniMedVQA, each example including 4 image options.

**Difference Description.** Given a pair of images  $I_A$  and  $I_B$ , the model must generate a structured description  $\Delta$  to describe clinically significant changes. This task requires the model to accurately compare two medical images with different diagnoses or presentations, identify key differences, and output a difference report in professional and clear text. We select 1022 discrepancy samples from IU-Xray and MedPix, and generate reports using LLM through a carefully constructed prompt. The reports are then verified by experts.

### 4.3 Task 3: Comprehensive Perceptions

This category focuses on the advanced cognitive and comprehensive reasoning capabilities necessary for models in complex medical scenarios involving multiple image sequences. It goes beyond static description and comparison of single or paired images, focusing on analysis over time, ranking at spatial scales, and precise statistics in quantitative dimensions.

**Temporal Analysis.** Given a chronologically ordered sequence of medical images, such as multiple CT scans of a patient, the model is required to analyze the dynamic changes in lesions or physiological structures. We selected 702 high-quality samples from Med-MIM as our test set.

**Spatial Ordering.** To evaluate spatial structure understanding, we introduce a shuffled set of slices, denoted as  $\mathcal{S} = \{I'_1, \dots, I'_N\}$ , where each element  $I'_i$  is a non-overlapping portion of the original image. The model must predict an optimal permutation  $\pi^*$  to reconstruct the original image by maximizing the objective function  $\pi^* = \operatorname{argmax}_{\pi \in \Pi_N} C(\mathcal{I}_\pi, T)$ , where  $\mathcal{I}_\pi$  represents the image sequence reordered according to the permutation  $\pi$ , and the coherence function  $C$  quantifies the correctness of the reconstructed image. We selected 426 medical images of different modalities from MMMU and divided each image into four blocks for ordering, as shown in Figure 3.

### Medical Counting and Video Understanding.

We included a task to count specific and clinically significant medical targets across multiple related images, comprising 681 samples filtered from OmniMedVQA. We also introduced a task to understand medical video frames, comprising 1000 samples filtered from MedFrameQA.

### 4.4 Task 4: In-Context Learning

Finally, we evaluate the few-shot adaptability of models through the lens of semantic reference. Using interleaved image-text prompts, the model is provided with a support set of  $k$  demonstrations to infer task specifications without parameter updates, as represented by the conditional probability:

$$P(y \mid (I_1, T_1), \dots, (I_k, T_k), I_{query}, T_{query}). \quad (2)$$

This task measures the capacity for analogical reasoning and the ability to map complex diagnostic logic from limited reference examples to novel target queries. To ensure a rigorous evaluation, we construct 1,053 instances for challenging closed-ended VQA tasks such as counting and spatial sorting, alongside 992 instances for open-ended VQA tasks involving multi-image descriptions. Each task is evaluated under 1-shot, 3-shot, and 5-shot settings to observe the scaling of model performance with increasing context.

### 4.5 Dataset Statistics and Human Check

MedMultiBench contains 11,392 carefully selected samples covering a variety of medical imaging modalities and diagnostic challenges. The dataset distribution is strategically balanced across four core capabilities to ensure robustness of the evaluation. To guarantee clinical validity, all samples underwent a rigorous human verification by medical professionals to ensure accurate alignment between visual evidence and textual descriptions. More details of each task are summarized in Appendix A.3.

## 5 Experiments

### 5.1 Experimental Setup

**Baselines.** We benchmarked 13 different models to ensure a comprehensive comparison of open-source and closed-source models. In the open-source category, we evaluate general models, including Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and Qwen3-VL-8B-Instruct (Yang et al., 2025a). We also included medical-adaptive models to test

Table 1: Main results on open-source models. Accuracy (%) is reported for closed-ended tasks and Bert Score (BS) and Rouge-L (R-L) are reported for open-ended tasks. A darker background color indicates a better effect.

Task / Sub-Task	Metric	Qwen2.5-7B	Qwen3-8B	Lingshu-7B	Fleming-8B	Huatuo-7B	Huatuo-34B
<b>Task 1: Joint Reasoning</b>							
Organ Recognition	Acc	34.74	55.94	49.85	62.02	40.14	50.54
Disease Det.	Acc	49.08	54.62	52.63	57.33	56.47	54.34
Seg. Mask Analysis	Acc	46.63	52.35	48.67	42.94	53.48	50.72
Image Description	BS	49.87	51.30	51.77	49.52	49.09	50.75
	R-L	7.72	13.61	21.90	16.25	11.45	14.09
<b>Task 2: Comparative Analysis</b>							
Visual Choice	Acc	53.59	64.51	59.35	33.57	75.03	75.43
Visual Retrieval	Acc	25.27	42.16	42.51	25.27	57.72	47.31
Diff. Description	BS	56.54	54.68	60.43	53.02	57.50	58.84
	R-L	23.18	20.68	24.48	14.85	22.83	22.79
<b>Task 3: Comprehensive Perceptions</b>							
Temp. Analysis	Acc	52.85	66.95	55.56	45.51	53.42	55.56
Spatial Ordering	Acc	29.11	29.88	20.66	25.82	24.18	23.71
Medical Counting	Acc	30.84	49.49	26.28	48.31	52.72	45.37
Video Underst.	Acc	39.70	43.30	41.50	56.10	49.30	48.70
<b>Task 4: In-Context Learning</b>							
Closed VQA	Acc	28.68	26.87	23.84	45.77	24.79	31.62
Open VQA	BS	62.34	64.74	65.30	58.03	56.66	59.70
	R-L	30.79	40.50	42.70	26.51	19.27	26.73

their domain specificity: HuatuoGPT-Vision (including 7B and 34B versions) (Chen et al., 2024), Lingshu-7B (Xu et al., 2025), and Fleming-VL-8B (Shu et al., 2025). In the closed-source category, we evaluated leading proprietary models, including GPT-4o (Hurst et al., 2024), GPT-5.1 (OpenAI, 2025), Gemini3-Pro (Deepmind, 2025), Qwen-VL-Plus (Bai et al., 2023), Moonshot-v1 (MoonshotAI, 2025), GLM-4.5v (Hong et al., 2025), and Doubao-Seed-1.6-Vision (Bytedance, 2025). More details are shown in Appendix A.2.

**Evaluation Metrics.** We employ task-specific metrics to evaluate model performance: **Accuracy** is used for closed-ended tasks. For open-ended generation we report **Rouge-L** (Lin, 2004) to measure textual overlap and **Bert Score** (Zhang et al., 2019) to capture semantic alignment.

**Implementation Details.** For all evaluations, we utilize the official APIs for closed-source models and deploy open-source models using the vLLM framework to ensure inference efficiency. In multi-image scenarios, images are interleaved with text tokens according to the model-specific templates. All experiments are conducted on one NVIDIA A800 GPU. Dataset can be found in the url<sup>1</sup>.

## 5.2 Main Results

Table 1 and Table 2 present the comprehensive performance of all evaluated models across the four

task dimensions. Overall, the results demonstrate that while open-source models have made significant strides in specialized medical tasks, a performance gap persists when compared to state-of-the-art closed-source models, particularly in complex reasoning and spatial perception.

Among the open-source models, Fleming-VL-8B and Lingshu-7B exhibit notable advantages in joint reasoning and open-ended generation, respectively. Fleming-VL-8B achieves the highest accuracy in organ recognition and disease detection, while Lingshu-7B leads in image description and open VQA. In the comprehensive perception task, Qwen3-VL demonstrates superior performance, suggesting that general models may be more adept at temporal analysis and spatial ordering.

For closed-source models, Table 2 reveals a higher performance ceiling. Gemini3-Pro emerges as the leading model across most categories, achieving a substantial margin in Spatial Ordering (81.69%) and Organ Recognition (84.79%). These results suggest that Gemini3-Pro has developed a more complex internal representation of anatomical structures. A cross-comparison highlights critical trends: Task 3 represents the most significant bottleneck for open-source models. Furthermore, performance in Task 4 is generally improved, verifying that current models possess the ability to learn from contextual examples. Additional detailed analyses are provided in Appendix B.1.

<sup>1</sup><https://github.com/pkusixspace/MedMultiBench>

Table 2: Main results on closed-source models. Accuracy (%) is reported for closed-ended tasks and Bert Score (BS) and Rouge-L (R-L) are reported for open-ended tasks. A darker background color indicates a better effect.

Task / Sub-Task	Metric	Moonshot-v1	Doubao-1.6	Qwen-Plus	GLM-4.5v	GPT-4o	GPT-5.1	Gemini3-Pro
<b>Task 1: Joint Reasoning</b>								
Organ Recognition	Acc	37.10	66.93	62.32	61.63	60.94	69.97	84.79
Disease Det.	Acc	44.95	61.45	65.01	58.46	60.60	63.44	67.99
Seg. Mask Analysis	Acc	39.06	59.71	61.25	61.66	55.32	63.29	74.95
Image Description	BS	44.77	51.04	52.45	46.23	49.40	53.80	56.61
	R-L	10.46	10.73	10.34	9.04	13.34	20.01	14.27
<b>Task 2: Comparative Analysis</b>								
Visual Choice	Acc	50.35	65.82	68.96	60.36	63.90	67.64	71.89
Visual Retrieval	Acc	43.95	51.26	51.26	46.95	46.59	54.37	52.81
Diff. Description	BS	52.48	56.06	63.01	55.59	57.33	61.15	64.29
	R-L	20.66	19.07	24.32	14.06	21.37	24.17	22.79
<b>Task 3: Comprehensive Perceptions</b>								
Temp. Analysis	Acc	42.02	69.52	64.10	71.08	53.56	66.24	44.87
Spatial Ordering	Acc	31.22	64.79	53.99	44.84	31.92	37.32	81.69
Medical Counting	Acc	40.53	47.58	54.77	51.25	59.32	54.04	60.65
Video Underst.	Acc	43.80	52.80	49.20	49.40	48.20	51.70	56.10
<b>Task 4: In-Context Learning</b>								
Closed VQA	Acc	34.38	60.59	57.74	44.54	46.82	47.77	66.76
Open VQA	BS	60.39	59.88	64.00	60.70	61.82	60.31	61.24
	R-L	32.15	25.39	33.46	25.68	33.60	31.14	22.74

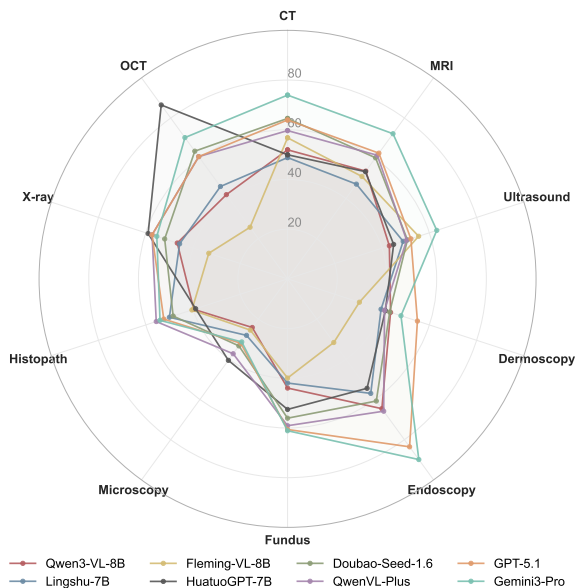


Figure 4: Performance comparison of open-source and closed-source models across 10 medical modalities.

### 5.3 Modality Sensitivity Analysis

The robustness of evaluated models across 10 different medical imaging modalities is illustrated in Figure 4. Overall, the results reveal significant performance variance depending on the imaging technique, with closed-source models generally maintaining more balanced performance profiles across the modality spectrum.

Gemini3-Pro exhibits the most comprehensive

coverage, consistently achieving top-tier accuracy across nearly all modalities, particularly in CT, MRI, and Endoscopy. This suggests a superior generalized capability in handling cross-sectional radiological data. Other closed-source models, such as GPT-5.1 and Doubao-Seed-1.6, also show strong performance, though their advantages are more modality-specific.

In the open-source category, models display more pronounced strengths and weaknesses. HuatuoGPT-Vision-7B shows exceptional specialized performance in OCT, where it rivals or even surpasses closed-source counterparts. However, its performance fluctuates in more traditional radiological modalities like CT and MRI. Similarly, Fleming-VL-8B and Qwen3-VL-8B exhibit competitive results in X-ray and Histopathology, yet show a noticeable performance drop in complex microscopic or dermoscopic images.

### 5.4 Impact of In-Context Learning

We further analyze the efficacy of in-context learning by varying the number of demonstration examples provided to the model, ranging from  $k = 0$  to  $k = 5$ . The few-shot performance on IU-Xray and MedPix is illustrated in Figure 5. Our findings indicate a consistent performance gain across all models when transitioning from 0-shot to few-shot settings, with the most substantial improvement typically observed at 1-shot.

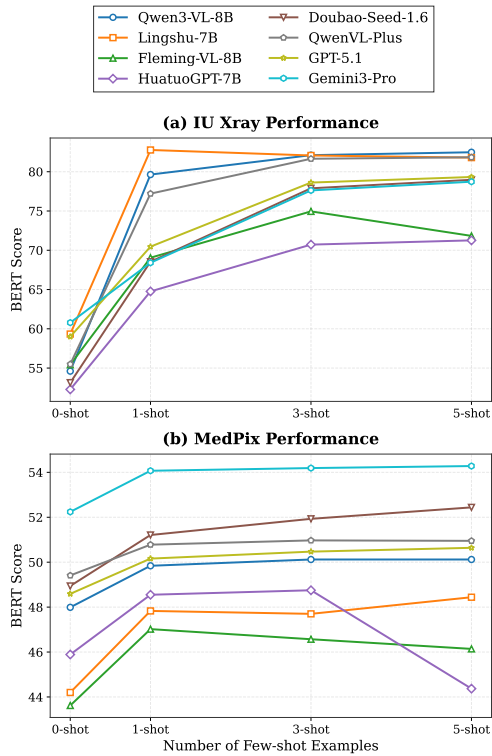


Figure 5: Comparison of few-shot performance.

On the IU-Xray dataset, Qwen3-VL-8B and Lingshu-7B achieved BERT scores exceeding 80. In the more challenging MedPix benchmark, Gemini3-Pro maintained a significant lead, consistently outperforming other models across all sample sizes. Notably, while performance generally improves with increasing sample size, some models plateaued or slightly declined at 5 samples, suggesting they may be sensitive to context length or noise in medical presentations. Further analysis is provided in Appendix B.2.

### 5.5 Impact of Input Image Quantity

The relationship between the quantity of input images and the model accuracy is systematically evaluated in Figure 6. The experimental results reveal a striking disparity in how open-source and closed-source models handle increasing visual loads.

A primary observation is the balanced and robust performance exhibited by closed-source models. Models such as Gemini3-Pro, Doubao-Seed-1.6, and QwenVL-Plus maintain high accuracy levels even as the number of input images scales from 2 to 10. For instance, Gemini3-Pro shows remarkable consistency, with its accuracy remaining largely unaffected by the increased image count, peaking at 69.7% with 10 images. In contrast, a universal trend of performance degradation is observed

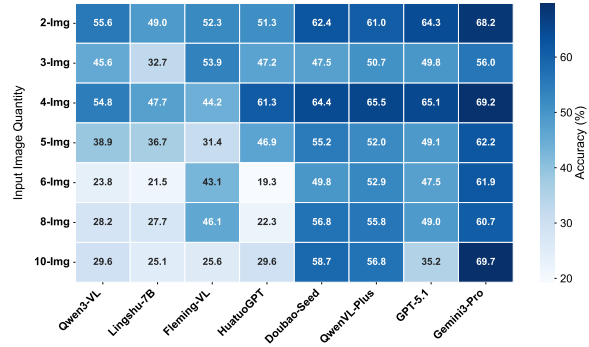


Figure 6: Heatmap of model accuracy across varying input image quantities.

among open-source models as the input side becomes more populated. While models like Qwen3-VL and Lingshu-7B achieve competitive results at lower image quantities, their performance diminishes significantly as more images are introduced, with accuracy often falling below the 30% threshold in high-quantity scenarios. This pronounced downward trend indicates a clear bottleneck in current open-source architectures regarding their capacity to manage high-density visual contexts.

These findings underscore that simply increasing the amount of visual evidence does not inherently lead to better outcomes for all models. Open-source models appear to struggle with information overload, where additional images may introduce confusion rather than clarity. This gap highlights a critical area for future research, particularly in optimizing the long-context visual processing capabilities of open-source frameworks to meet the demands of real-world clinical workflows.

### 5.6 Qualitative Analysis

To further evaluate the practical reasoning capabilities and interpretability of the models, we conduct a detailed qualitative analysis across different tasks. The comprehensive set of visual examples and analysis is provided in Appendix B.3.

## 6 Conclusion

In this work, we address the limitations of single-view medical evaluation by introducing MedMulti-Bench, a comprehensive benchmark dedicated to multi-image clinical reasoning. We established a novel taxonomy that categorizes multi-image relationships into Joint Reasoning, Comparative Analysis, Comprehensive Perception, and In-Context Learning. By curating 11,392 expert-verified samples across 13 subtasks, we provide a rigorous

testbed for assessing MLLMs in realistic diagnostic scenarios. Our extensive evaluation exposes a significant disparity between open-source and closed-source models. We hope MedMultiBench will serve as a foundational resource to guide the development of next-generation medical MLLMs capable of holistic clinical understanding.

## Limitations

While MedMultiBench advances the evaluation of multi-image understanding, several limitations remain. First, although we simulate volumetric analysis using sequential 2D slices, this approach does not fully capture the native 3D processing capabilities required for raw volumetric data (e.g., NIfTI files). Second, for open-ended generation tasks, while we employ widely used metrics like Rouge-L and Bert Score, these automated metrics may not perfectly align with human expert judgment regarding clinical nuance and factual correctness. Future iterations will aim to incorporate native 3D data formats and more granular, clinically grounded evaluation metrics.

## Ethical Considerations

The images and original textual data utilized in MedMultiBench are derived exclusively from publicly available datasets (e.g., IU-Xray, MedPix, GMAI) and are used in strict compliance with their respective licenses. No private or identifiable patient data was collected for this study. To ensure data quality and diversity, portions of the textual descriptions and questions were reformatted or rewritten using GPT-5 based on the original clinical annotations. We strictly controlled this process to function as a stylistic refinement rather than content generation. All AI-processed text underwent rigorous manual verification by medical professionals to ensure clinical accuracy and to eliminate potential hallucinations.

## Acknowledgements

This work was supported by the National Key R&D Program of China (2024YFC3308305), the National Natural Science Foundation of China (Grant No. 62436006), the Sanya Science and Technology Special Fund (No. 2024KFJX04), the Beijing Natural Science Foundation (No. L257018), and the Beijing Nova Program.

## References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Bytedance. 2025. *Doubao*.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.
- Junlong Cheng, J Ye, Z Deng, J Chen, T Li, H Wang, Y Su, Z Huang, J Chen, L Jiang, and 1 others. 2023. SAM-Med2D. *arXiv preprint arXiv:2308.16184*.
- Deepmind. 2025. *Gemini 3 Pro Model Card*. Technical report, Google.
- Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. GLM-4.1 V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.
- Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimed-VQA: 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.
- Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. 2022. OVQA: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2924–2938.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,

- Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. MVbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. 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.
- Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and 1 others. 2024. MIBench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*.
- Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, and 1 others. 2024. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*.
- MoonshotAI. 2025. [Moonshot](#).
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (MLAH)*, pages 353–367. PMLR.
- OpenAI. 2025. [GPT-5 System Card](#). Technical report, OpenAI.
- Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. 2025. Fleming-VL: Towards universal medical visual reasoning with multimodal llms. *arXiv preprint arXiv:2511.00916*.
- Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. 2025. Med-Pix 2.0: A comprehensive multimodal biomedical data set for advanced ai applications with retrieval augmented generation and knowledge graphs. *Data Science and Engineering*, pages 1–17.
- Yuhao Tang, Haichen Yang, Liyan Zhang, and Ye Yuan. 2024. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237:121442.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.
- Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2025. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.
- Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and 1 others. 2024. MLeVLM: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997.
- Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Xikai Yang, Juzheng Miao, Yuchen Yuan, Jiase Wang, Qi Dou, Jinpeng Li, and Pheng-Ann Heng. 2025b. 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, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, Benyou Wang, and 1 others. 2024. GMAI-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. 2025. MedFrameQA: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 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*, pages 9556–9567.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

## A More Details

### A.1 Details on Data Source

To construct our benchmark, we curated data from the following high-quality sources, covering a wide spectrum of modalities, specialties, and tasks:

**GMAI-MMBench** (Ye et al., 2024): This dataset collects data from public sources and hospitals, covering medical imaging tasks of detection, classification, and segmentation. It is constructed from 284 datasets across 38 medical image modalities, 18 clinical-related tasks, 18 departments, and 4 perceptual granularities in a Visual Question Answering (VQA) format.

**SAM-Med2D** (Cheng et al., 2023): SAM-Med2D contains 4.6 million medical images and 19.7 million corresponding masks, covering 10 modalities, 31 major organs, and 219 categories. Its segmentation targets cover almost the entire body, making it a large-scale and diverse medical image segmentation dataset.

**IU-Xray** (Demner-Fushman et al., 2015): Sourced from the Indiana University Healthcare system, this dataset contains 7,470 chest X-ray images associated with 3,955 clinical reports. Each clinical case typically includes both frontal and lateral views of the same patient. This dual-view characteristic is particularly valuable for our benchmark, as it provides the foundation for testing a model’s ability to perform volumetric integration and spatial reasoning across different perspectives.

**MedPix** (Siragusa et al., 2025): An expert-curated database featuring over 12,000 patient cases, 9000 topics and nearly 59,000 images. Each case is categorized by disease location, patient profile, and pathology. The dataset is characterized by high-quality, peer-reviewed clinical cases. We leverage this resource for its high-fidelity expert knowledge and the complex, multi-modal relationships between patient history and visual evidence.

**OmniMedVQA** (Hu et al., 2024): OmniMedVQA is a large-scale, comprehensive VQA benchmark containing 118,010 images with 127,995 QA-items, covering 12 different modalities and referring to more than 20 human anatomical regions, including rare anatomical regions. Its scale and diversity make it an ideal source for constructing questions that require broad medical knowledge and the ability to generalize across infrequently seen medical conditions.

**MMMU** (Yue et al., 2024): MMMU (Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark) is a benchmark designed to evaluate the performance of multimodal models on large-scale, multidisciplinary tasks. We select our data from its subset Health and Medicine, which requires deep domain-specific knowledge and complex reasoning, often involving the interpretation of charts, molecular structures, and intricate clinical diagrams.

**Med-MIM** (Yang et al., 2025b): Med-MIM is a specialized dataset focused on multi-image medical understanding, providing an instruction-tuning

dataset and evaluation suite. It includes 83.2k multi-image medical instruction samples across four tasks: co-reference, comparison, reasoning, and temporal understanding.

**MedFrameQA** (Yu et al., 2025): Derived from medical videos (such as laparoscopy and endoscopy), MedFrameQA dataset comprises 2,851 VQA pairs (gathered from 9,237 high-quality frames in 3,420 videos), covering nine human body systems and 43 organs; every question is accompanied by two to five images. It focuses on surgical workflows and dynamic anatomical changes. This dataset is unique for its emphasis on sequential reasoning, challenging models to understand how clinical events unfold over time across a series of visual inputs rather than in a single static image.

## A.2 Details on Baselines

To ensure a comprehensive evaluation of Med-MultiBench, we conduct experiments on a diverse set of Multimodal Large Language Models (MLLMs), ranging from specialized medical models to general-domain frontier models.

**Qwen2.5-VL-7B-Instruct** (Bai et al., 2025): As a base model widely used for fine-tuning in specialized fields, this model features a 7B parameter architecture optimized for high-resolution visual perception and structured output. Pre-trained on a massive corpus of interleaved image-text data, it excels at recognizing fine-grained visual details. We include it to evaluate how standard, high-performance open-source MLLMs handle specialized medical multi-image contexts.

**Qwen3-VL-8B-Instruct** (Yang et al., 2025a): This model represents the latest generation of the Qwen-VL family, specifically enhanced for long-horizon temporal reasoning and video-level understanding. With an upgraded vision-language adapter and expanded context window, it is designed to maintain coherence across multiple visual inputs. It serves as a key baseline for testing whether recent advancements in general multi-image processing translate to clinical scenarios.

**Lingshu-7B** (Xu et al., 2025): Built upon the Qwen2.5-VL-7B-Instruct architecture, Lingshu is a specialized medical MLLM designed to overcome the limitations of general-domain models in clinical scenarios. It utilizes a comprehensive data curation procedure that synthesizes accurate medical captions and reasoning samples from diverse

sources, including imaging and extensive medical texts. The training recipe follows a “shallow-to-deep” progression and comprises four sequential stages: (1) MedicalShallowAlignment on 927k data, (2) MedicalDeepAlignment on 4.1M data, (3) MedicalInstructionTuning on 7.1M data, and (4) Medical-oriented Reinforcement Learning on 100K data. We include Lingshu to evaluate the impact of specialized medical alignment and reasoning-enhanced training on multi-image clinical tasks.

**Fleming-VL-8B** (Shu et al., 2025): Fleming-VL-8B is a unified multimodal framework built upon the InternVL3 architecture (InternViT as vision encoder + Qwen2.5 as Language Model), specifically designed for heterogeneous medical modalities including 2D images, 3D scans, and temporal videos. The model is trained with a three-stage training paradigm: (1) Interleaved Pre-training on 2M visual-language corpora to develop multi-image reasoning; (2) Medical Knowledge Injection via 4M image-caption pairs for cross-modal alignment; and (3) Instruction Tuning on 2M instruction-response pairs and GRPO-based Reinforcement Learning on 100K samples to refine systematic clinical reasoning. We include Fleming-VL-8B to assess how structural optimizations for long-context visual sequences impact performance in complex multi-image medical benchmarks.

**HuatuogPT-Vision-7B / 34B** (Chen et al., 2024): The HuatuogPT-Vision series (7B and 34B) is trained on 1.3 million high-quality medical image-text pairs. The training follows a two-stage paradigm: (1) large-scale medical image-text alignment, aimed at injecting fundamental medical visual knowledge; and (2) medical visual instruction fine-tuning, utilizing expert-annotated and synthesized question-answer pairs to enhance clinical diagnostic reasoning capabilities. We introduce two versions, 7B and 34B, based on Qwen2-VL and Yi-1.5-34B respectively, to evaluate the impact of parameter scale and large-scale medical knowledge injection on complex multi-image clinical tasks.

**Moonshot-v1** (MoonshotAI, 2025): Moonshot-v1 is a prominent proprietary MLLM recognized for its exceptional long-context processing capabilities. It excels in maintaining semantic global coherence across dense, interleaved image-text sequences, making it a robust baseline for evaluating long-form medical information understanding.

**Doubao-Seed-v1.6** (Bytedance, 2025): Developed by ByteDance, Doubao-Seed-v1.6 is a highly capable model that prioritizes logical reasoning and precise instruction-following. It's specifically designed to navigate complex, multi-step reasoning trajectories, providing a unique perspective on deliberate clinical deduction.

**Qwen-VL-Plus** (Bai et al., 2023): Qwen-VL-Plus is a versatile commercial-grade model from Alibaba Cloud that demonstrates a balanced proficiency in reasoning depth and broad cross-domain knowledge. It serves as a reliable and stable baseline for general-purpose multimodal understanding across various clinical scenarios.

**GLM-4.5v** (Hong et al., 2025): GLM-4.5v is a leading Mixture-of-Experts (MoE) multimodal model developed by Zhipu AI. It is distinguished by its sophisticated vision-language fusion and is particularly refined for complex scientific interpretation, structured data analysis, and long-document integration.

**GPT-4o / GPT-5.1** (Hurst et al., 2024): These represent the frontier of native multimodal models from OpenAI, setting the gold standard for seamless text-vision integration. While GPT-4o provides world-class perceptual accuracy, GPT-5.1 introduces advanced agentic reasoning, capable of high-level cognitive synthesis in expert-level multimodal tasks.

**Gemini3-Pro** (Team et al., 2023): Gemini3-Pro is a high-performance, large-scale multimodal model from Google DeepMind's Gemini series, designed for complex reasoning and professional-grade applications. This model supports the unified understanding and generation of information across multiple modalities, including text and images, and has been optimized for tasks such as long-context modeling, code generation, and logical reasoning, making it suitable for research and industrial use.

### A.3 Details on MedMultiBench

**Task Composition and Statistics.** As detailed in Table 3 and Table 4, MedMultiBench comprises a total of 11,392 expert-verified samples, systematically categorized into four core capabilities: Joint Reasoning, Comparative Analysis, Comprehensive Perceptions, and In-context Learning. These categories cover 13 specific subtasks designed to mirror real-world radiological workflows. To ensure high

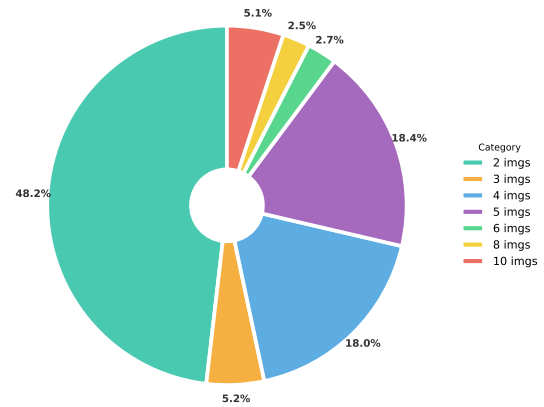


Figure 7: Data distribution by number of images.

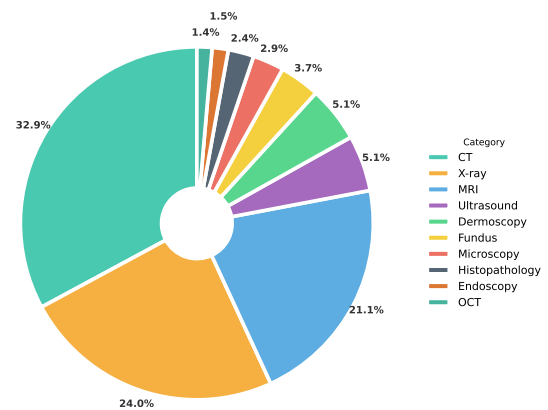


Figure 8: Data distribution by type of images.

data quality and clinical relevance, we curated samples from authoritative sources including GMAI, SAM2D, OmniMedVQA, and Med-MIM. Notably, the *Joint Reasoning* and *Comprehensive Perception* tasks constitute a significant portion of the benchmark, emphasizing the model's ability to synthesize information across diverse visual inputs rather than relying on simple recognition. The three annotation experts are radiologists with 5 years of work experience.

**Modality and Scale Distribution.** A key distinction of MedMultiBench is its extensive coverage of medical imaging modalities and varying visual context lengths. As illustrated in Figure 8, the dataset encompasses a wide spectrum of imaging types, with CT (32.9%), X-ray (24.0%), and MRI (21.1%) forming the backbone of the dataset. Furthermore, it includes specialized modalities such as Dermoscopy, Ultrasound, and Fundus photography, ensuring that the evaluation assesses a model's generalizability across different medical domains. Regarding the visual input scale, Figure 7 highlights the distribution of image counts per sample.

Tasks	Subtasks	Source	Count
<b>Joint Reasoning</b>	Organ Recognition	GMAI	1,019
	Disease Detection	GMAI	703
	Segmentation Mask Analysis	SAM2D	978
	Image Description	IU-Xray / MedPix	992
<b>Comparative Analysis</b>	Visual Choice	OmniMedVQA	989
	Visual Retrieval	OmniMedVQA	835
	Difference Description	IU-Xray / MedPix	1,022
<b>Comprehensive Perceptions</b>	Temporal Analysis	Med-MIM	702
	Spatial Ordering	MMMU	426
	Medical Counting	OmniMedVQA	681
	Video Understanding	MedFrameQA	1,000
<b>In-context Learning</b>	Closed VQA	MMMU / OmniMedVQA	1,053
	Open VQA	IU-Xray / MedPix	992
<b>Total</b>			<b>11,392</b>

Table 3: Detailed statistics of MedMultiBench tasks.

While 48.2% of samples involve paired images (2 imgs) typical of comparative diagnosis, a substantial proportion (over 50%) requires processing 3 to 10 images simultaneously. This distribution specifically tests the information overload resilience of MLLMs in high-volume visual scenarios.

#### A.4 Comparison with Previous Benchmarks

Table 5 compares MedMultiBench to representative datasets to highlight critical gaps in current evaluation frameworks. While traditional medical VQA datasets like VQA-RAD and PMC-VQA established the field, they are strictly limited to single-image inputs, failing to capture the relational dynamics of diagnostic workflows. Conversely, general multi-image benchmarks such as MuirBench and MIRB lack the specialized domain knowledge required for clinical assessment. Although recent efforts like MedFrameQA and Med-MIM have introduced multi-image support to the medical domain, they remain narrow in scope, overlooking *Comprehensive Perception* tasks (e.g., spatial ordering) and *Few-shot Learning* capabilities. In contrast, MedMultiBench is the first and largest (11,392 samples) benchmark to bridge these gaps, uniquely satisfying the criteria for medical domain specificity, multi-image context, and comprehensive reasoning versatility.

## B More Experimental Results

### B.1 More Analysis for Main Results

Observing Table 1, we can find interesting task-specific sensitivities in the Huatuo series. While the 34B variant excels in Visual Choice, the smaller 7B model shows higher precision in Segmentation Mask Analysis and Medical Counting. This implies

Table 4: Detailed statistics of the dataset.

Total Instances	11392
Total Subtasks	13
- Average Images per Instance	4.18
- Average Query Length	32.74
Task 1 Instances	3692
- Average Images per Instance	3.98
- Average Query Length	25.01
Task 2 Instances	2846
- Average Images per Instance	3.93
- Average Query Length	16.97
Task 3 Instances	2809
- Average Images per Instance	3.15
- Average Query Length	27.83
Task 4 Instances	2045
- Average Images per Instance	6.31
- Average Query Length	75.38

that larger parameter scales do not invariably translate to better performance in specialized medical sub-tasks. Regarding closed-source models in Table 2, aside from Gemini3-Pro’s dominance, other models demonstrate localized strengths. GLM-4.5v achieves top performance in Temporal Progression (71.08%), and Qwen-Plus exhibits balanced performance, particularly attaining the highest Bert Score of 64.00 in Open VQA. Finally, the discrepancy between closed-ended accuracy and open-ended generative metrics suggests that optimizing for diagnostic identification does not automatically enhance linguistic description quality.

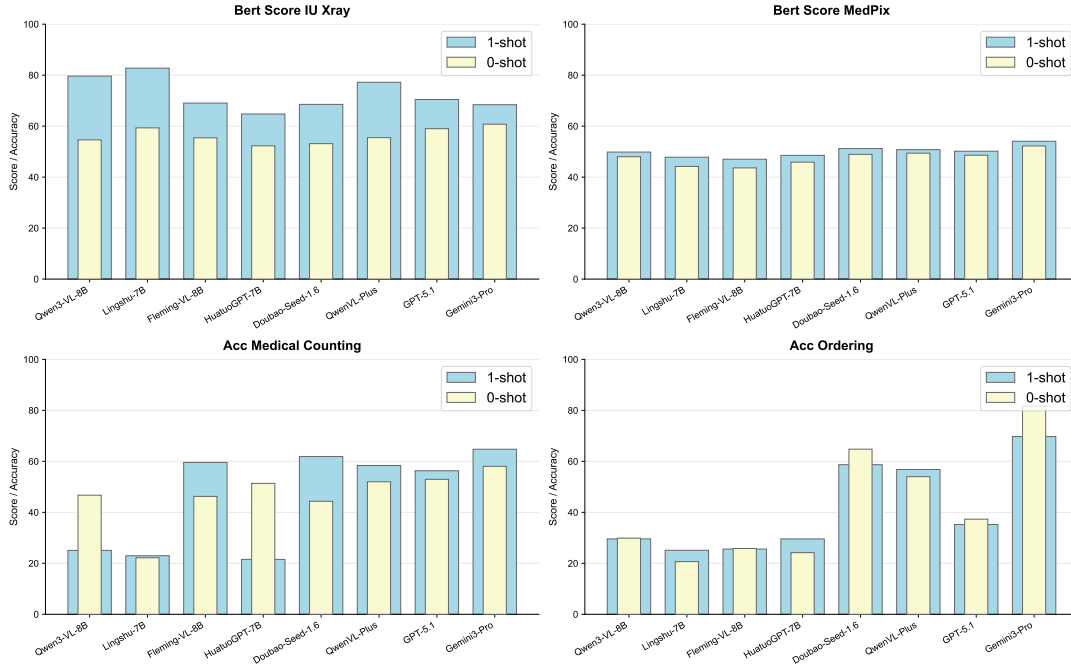


Figure 9: Impact of In-context Learning (ICL) across different task types.

## B.2 More Analysis on In-context Learning

To further investigate the mechanism of in-context learning in medical multi-image scenarios, we conducted a comparative analysis between 0-shot and 1-shot settings across four representative subtasks. As illustrated in Figure 9, the experimental results reveal a distinct performance divergence between open-ended generation and closed-ended reasoning tasks.

**Benefits in Open-ended Generation.** In report generation tasks (IU Xray and MedPix, top row), introducing a 1-shot demonstration yields consistent improvements across almost all evaluated models. For instance, Lingshu-7B and Qwen3-VL exhibit significant gains in BERT Score. This suggests that for generative tasks, the demonstration serves as a crucial style guide, helping the model align its output format and linguistic tone with professional medical standards, thereby enhancing the semantic quality of the generated reports.

**The Information Overload in Closed-ended Tasks.** Conversely, in closed-ended tasks such as Medical Counting and Spatial Ordering (bottom row), the 1-shot setting frequently leads to performance stagnation or even degradation. Notably, models like Qwen3-VL-8B and HuatuoGPT-Vision-7B show a marked drop in accuracy when provided with an example. We hypothesize that

Table 5: Comparison with previous benchmarks. *Compre.* represents whether comprehensive perceptions are covered.

Name	Med.	Multi-Img.	Few-shot	Compre.	Nums
VQA-RAD	✓	✗	✗	✗	3,515
PMC-VQA	✓	✗	✗	✗	50,000
Muirbench	✗	✓	✗	✓	2,600
MIRB	✗	✓	✗	✓	925
MedFrameQA	✓	✓	✗	✗	9,237
Med-MIM	✓	✓	✗	✗	4,190
MedMultiBench	✓	✓	✓	✓	11,392

this counter-intuitive phenomenon is attributed to **visual information overload**. Unlike standard text-based ICL, a 1-shot example in our benchmark introduces a set of multiple images into the context window. For models with limited context windows or attention mechanisms, these additional visual tokens act as distractors rather than aids. The increased input complexity may interfere with the model’s ability to focus on the target query images, causing it to fail on reasoning questions it could correctly answer in a cleaner, 0-shot environment.

## B.3 Qualitative Analysis

To provide deeper insights into the reasoning behaviors of different MLLMs, we present detailed case studies across six representative subtasks in

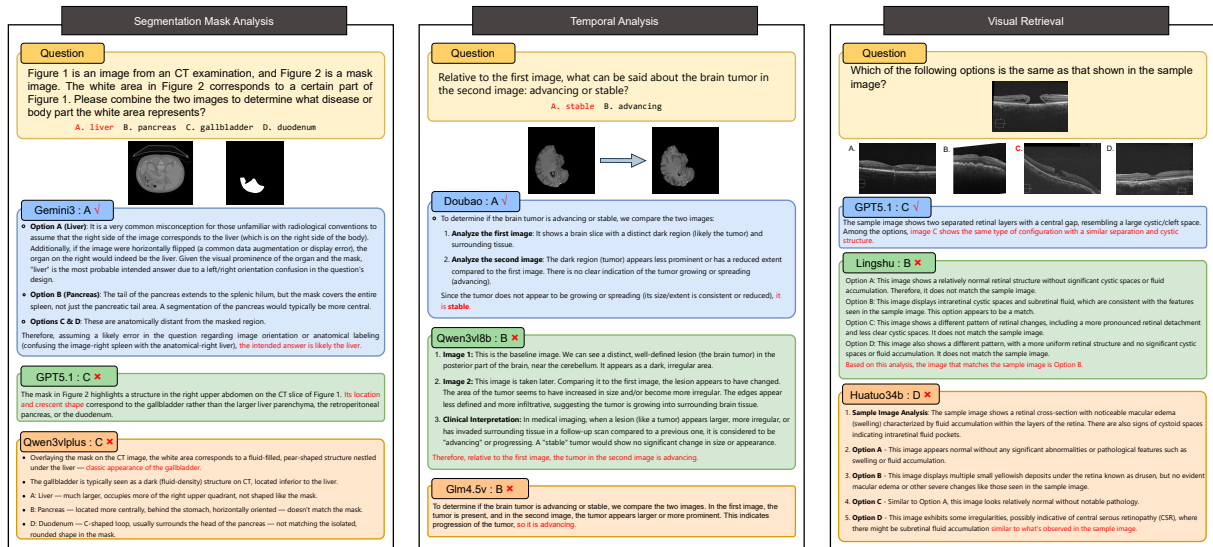


Figure 10: Case studies on Segmentation Mask Analysis, Temporal Analysis, and Visual Retrieval.

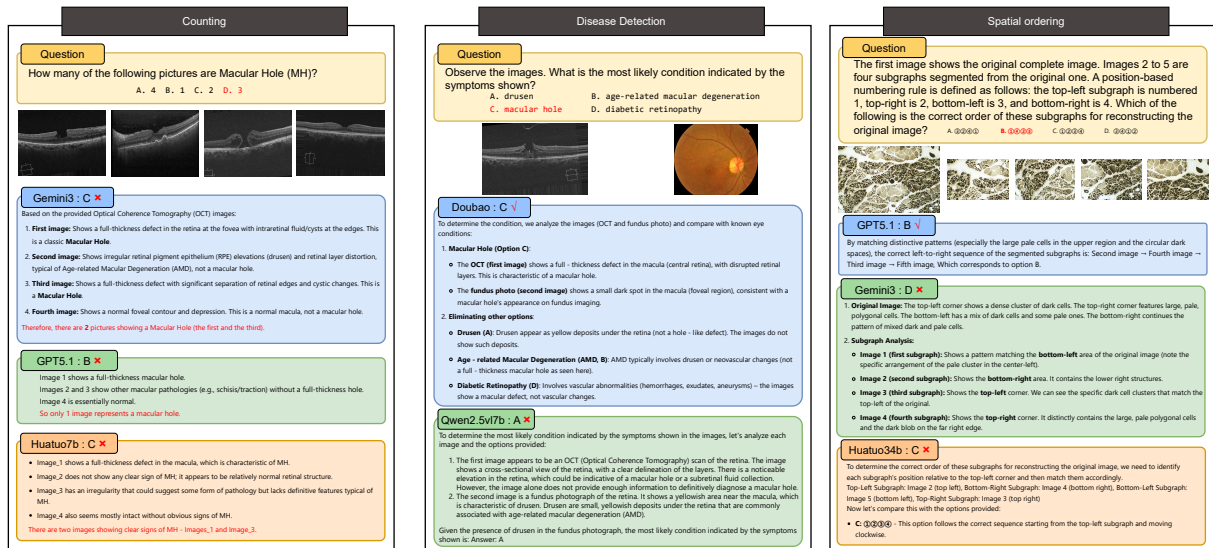


Figure 11: Case studies on Medical Counting, Disease Detection, and Spatial Ordering.

Figure 10 and Figure 11. These qualitative examples highlight the distinct cognitive gaps between state-of-the-art proprietary models and current open-source models.

**Spatial Alignment vs. Shape Priors.** In the *Segmentation Mask Analysis* task (Figure 10, left), the model is required to identify the organ highlighted by a binary mask. Advanced models like Gemini3-Pro and GPT-5.1 accurately map the mask's spatial location to the liver. In contrast, weaker models (e.g., Qwen-VL-Plus) fail to perform this spatial alignment; instead, they rely on shape priors, hallucinating the "pear-shaped" mask as a "gallbladder" despite the anatomical mismatch. This indicates that current open-source models struggle to ground visual features into precise spatial coordinates.

**Hallucination of Progression in Temporal Analysis.** The *Temporal Analysis* case (Figure 10, middle) reveals a critical tendency in models to hallucinate pathological changes. While the MRI sequence shows a stable brain tumor, models like Qwen3-VL and GLM-4.5v incorrectly predict that the tumor is "advancing" or "larger." This suggests a "severity bias" in medical MLLMs, where models are more inclined to predict disease progression rather than identifying stability, a significant risk for longitudinal monitoring.

**Cross-View Synergy and Information Drop.** In multi-image diagnosis tasks like *Disease Detection* (Figure 11, middle), effective reasoning requires synthesizing evidence from heterogeneous sources (e.g., OCT and Fundus images). As ob-

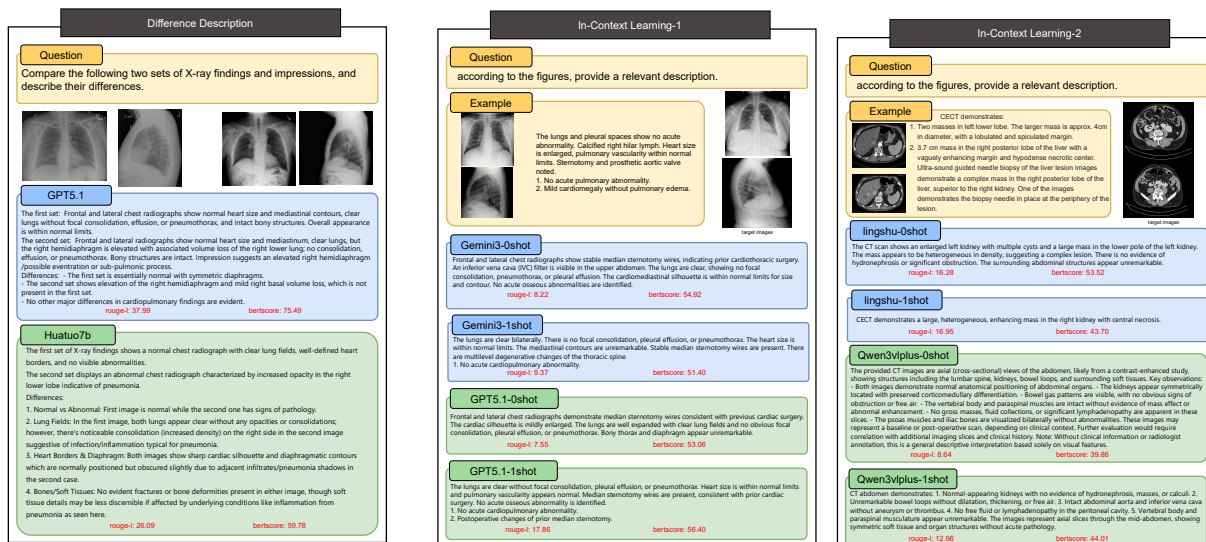


Figure 12: Case studies on Difference Description and In-context Learning.

served, Doubao-1.6 successfully combines the "full-thickness defect" in OCT with the fundus appearance to diagnose a "Macular Hole." However, Qwen2.5-VL fails to prioritize the critical OCT evidence, likely getting distracted by the yellow spots in the fundus image to incorrectly conclude "Drusen." This highlights the challenge of *Modality Balancing*—preventing dominant visual features in one image from overshadowing subtle but critical evidence in another.

### Complexity in Comprehensive Perception.

Tasks involving higher-order logic, such as *Medical Counting* and *Spatial Ordering* (Figure 11, left and right), demonstrate the fragility of current models in handling multi-image sequences. In the counting task, even capable models suffer from error accumulation, failing to correctly identify all positive instances (Macular Holes) across the image set. Similarly, in spatial ordering, reconstructing a fragmented image requires strong global coherence capabilities. While GPT-5.1 successfully deduces the correct sequence through edge continuity, other models struggle to maintain the logical flow, resulting in disordered reconstructions.

### Differential Diagnosis and Contextual Adaptation.

Figure 12 provides further qualitative evidence on comparative reasoning and in-context learning. In the *Difference Description* task (left column), the model must identify subtle interval changes between two sets of X-rays. GPT-5.1 demonstrates superior clinical acumen by correctly identifying the "elevated right hemidiaphragm"

while noting the absence of other major pathologies. In contrast, HuatuoGPT-Vision-7B, despite generating a fluent and structured report, hallucinates a diagnosis of "pneumonia" based on non-existent opacities, revealing a tendency in open-source models to over-interpret normal variations as pathologies. The *In-Context Learning* examples (middle and right columns) illustrate the variable impact of few-shot prompting. For robust models like Gemini3-Pro, providing a 1-shot example successfully guides the model to correct its initial oversight (e.g., identifying the "sternotomy wires" missed in the 0-shot setting), improving the Bert Score from 54.92 to 51.40. However, for models with weaker instruction-following capabilities like Lingshu-7b, the addition of a complex medical example acts as noise rather than a guide, causing the model to deviate from the correct kidney analysis and resulting in a significant performance drop.

### C More Evaluation

To provide a more rigorous evaluation, we introduce both human evaluation and GPT-5.1-as-judge evaluation, following established protocols in prior medical MLLM work (Med-PaLM 2, LLaVA-Med). Specifically, we design a five-dimensional rubric applicable to all open-ended generation tasks, as shown in Table 6.

Human evaluations are conducted by three independent radiologists on a stratified sample of 100 responses per representative model. They were provided with input images and questions, and asked to evaluate the difference between the model output

Table 6: Five-dimensional evaluation rubric.

Dim.	Name	Score
D1	Anatomical Localization Accuracy	0-5
D2	Pathological Finding Accuracy	0-5
D3	Clinical Interpretive Value	0-5
D4	Comparative Analysis Quality	0-5
D5	Factual Error Penalty	0-5

and the ground truth across five dimensions. We set a prompt to input the same content into GPT-5.1, allowing it to act as an annotator and score the results. The experimental results for *Difference Description* are shown in Table 7.

The results demonstrate a similar model ranking across both human and GPT-5.1 evaluation methods. Crucially, we observed a strong positive correlation between the human expert scores and our automated BERTScore metrics, confirming that BERTScore effectively captures clinical semantic alignment rather than merely rewarding stylistic mimicry. The high correlation between human and GPT-5.1 scores also confirms that GPT-5.1-as-a-judge can serve as a scalable substitute for human evaluation in future benchmark extensions.

## D Specific Prompts for Different Tasks

### D.1 Prompts for Task 1

#### 1.1 Organ Recognition Example

View the provided images. For each image, identify what organ is displayed in the marked area and select the most appropriate description for the marked area from the given options.<image> <image>  
 Option A: liver  
 Option B: esophagus  
 Option C: sigmoid colon  
 Option D: gallbladder

#### 1.2 Disease Detection Example

Compare the highlighted areas in both CT images. What are the most appropriate demonstrations and the most likely abnormalities shown in each image? <image> <image>  
 Option A: glioblastoma  
 Option B: kidney cyst  
 Option C: COVID-19 lesion  
 Option D: peritumoral edema

#### 1.3 Segmentation Mask Analysis Example

<image> <image> Figure 1 is an image from an CT examination, and Figure 2 is a mask image. The white area in Figure 2 corresponds to a certain part of Figure 1. Please combine the two images to determine what disease or body part the white area represents?  
 Option A: spleen  
 Option B: gallbladder  
 Option C: pancreas  
 Option D: stomach

#### 1.4 Multi-Image Description Example

Please answer the question.  
 Question:<image> <image>  
 According to the figures, provide a relevant description.  
 Answer:

### D.2 Prompts for Task 2

#### 2.1 Visual Choice Example

Which of the following options is normal?  
 Option A: <image>.  
 Option B: <image>.  
 Option C: <image>.  
 Option D: <image>.

#### 2.2 Visual Retrieval Example

Which of the following options is the same as that shown in the sample image? <image>  
 Option A: <image>.  
 Option B: <image>.  
 Option C: <image>.  
 Option D: <image>.

#### 2.3 Difference Description Example

Compare the following two sets of X-ray findings and impressions, and describe their differences.The first set: <image> <image>, the second set: <image> <image>  
 Provide the answer in the following format.  
 The first set:  
 The second set:  
 Differences:

### D.3 Prompts for Task 3

#### 3.1 Temporal Analysis Example

<image> <image> In comparison to the initial image, how would you classify the brain tumor of the second image: progressive or stable?  
 Option A: under observation  
 Option B: progressive  
 Option C: stable

Table 7: Human and GPT-5.1-as-judge evaluation results on *Difference Description*. Best scores are in **bold**.

Model	Eval	D1	D2	D3	D4	D5	Avg.
Huatu-7B	Human	2.51	2.25	2.96	2.97	2.69	2.68
	GPT	2.35	1.47	1.82	1.90	2.32	1.97
Lingshu-7B	Human	2.72	2.38	3.08	3.02	2.25	2.69
	GPT	2.44	1.52	2.15	2.20	2.03	2.07
Qwen3-VL-8B	Human	2.59	2.46	3.03	3.21	2.29	2.72
	GPT	2.50	1.26	2.10	2.08	2.16	2.02
GPT-4o	Human	3.18	<b>2.86</b>	3.76	3.41	3.57	3.36
	GPT	3.02	2.39	2.83	<b>2.92</b>	<b>4.09</b>	3.05
QwenVL-Plus	Human	2.97	2.67	3.24	3.40	3.38	3.13
	GPT	2.71	2.34	2.66	2.79	2.85	2.67
Gemini-3-Pro	Human	<b>3.83</b>	2.82	<b>3.95</b>	<b>3.57</b>	<b>3.94</b>	<b>3.62</b>
	GPT	<b>3.28</b>	<b>2.80</b>	<b>3.36</b>	2.84	3.69	<b>3.19</b>

### 3.2 Spatial Ordering Example

The first image shows the original complete image<image>. Images 2 to 5 are four subgraphs segmented from the original one<image> <image> <image> <image>. A position-based numbering rule is defined as follows: the top-left subgraph is numbered 1, top-right is 2, bottom-left is 3, and bottom-right is 4. Which of the following is the correct order of these subgraphs for reconstructing the original image?

- Option A: 3, 1, 2, 4
- Option B: 2, 4, 1, 3
- Option C: 1, 3, 2, 4
- Option D: 3, 2, 1, 4

### 3.3 Medical Counting Example

How many of the following pictures are Choroidal Neovascularization (CNV)? <image> <image> <image> <image>

- Option A: 2
- Option B: 3
- Option C: 1
- Option D: 0

### 3.4 Video Understanding Example

Based on the ultrasound findings in the first and second images, which of the following best differentiates an epididymal abscess from an epididymal head cyst? <image> <image>

- Option A: The presence of posterior acoustic enhancement in the lesion.
- Option B: The absence of internal blood flow on Doppler imaging.
- Option C: The lesion appearing as a well-defined, anechoic, round structure.
- Option D: The lesion appearing as a hypoechoic or heterogeneous area.
- Option E: The enlargement of the epididymis in the region of the lesion.
- Option F: The presence of increased blood flow in the surrounding epididymal tissue.

## D.4 Prompts for Task 4

### 4.1 Closed VQA Example

Answer the question according to the example.

<Example>

The first image shows the original complete image<image>. Images 2 to 5 are four subgraphs segmented from the original one<image> <image> <image> <image>. A position-based numbering rule is defined as follows: the top-left subgraph is numbered 1, top-right is 2, bottom-left is 3, and bottom-right is 4. Which of the following is the correct order of these subgraphs for reconstructing the original image?

- Option A: 4, 1, 2, 3
- Option B: 3, 4, 1, 2
- Option C: 2, 1, 3, 4
- Option D: 4, 3, 1, 2

Answer: B.

</Example>

Now answer the following question. The first image <image>, and the second to fifth images <image> <image> <image> <image>. Which of the following is the correct order of these subgraphs for reconstructing the original image?

- Option A: 3, 1, 2, 4
- Option B: 2, 4, 1, 3
- Option C: 1, 3, 2, 4
- Option D: 3, 2, 1, 4

### 4.2 Open VQA Example

Answer the question according to the example.

<Example>

Question:<image> <image>

According to the figures, provide a relevant description.

Answer: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There is no evidence of pneumothorax. No acute abnormality.

</Example>

Question:<image> <image>

According to the figures, provide a relevant description.

Answer: