MEDSYN: Benchmarking Multi-EviDence SYNthesis in Complex Clinical Cases for Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have shown great potential in medical applications, yet existing benchmarks inadequately capture real-world clinical complexity. We introduce MEDSYN, a multilingual, multimodal benchmark of highly complex clinical cases with up to 7 distinct visual clinical evidence (CE) types per case. Mirroring clinical workflow, we evaluate 18 MLLMs on differential diagnosis (DDx) generation and final diagnosis (FDx) selection. While top models often match or even outperform human experts on DDx generation, all MLLMs exhibit a much larger DDx–FDx performance gap compared to expert clinicians, indicating a failure mode in synthesis of heterogeneous CE types. Ablations attribute this failure to (i) overreliance on less discriminative textual CE (*e.g.*, medical history) and (ii) a cross-modal CE utilization gap. We introduce Evidence Sensitivity to quantify the latter and show that a smaller gap correlates with higher diagnostic accuracy. Finally, we demonstrate how it can be used to guide interventions to improve model performance. We will open-source our benchmark and code.

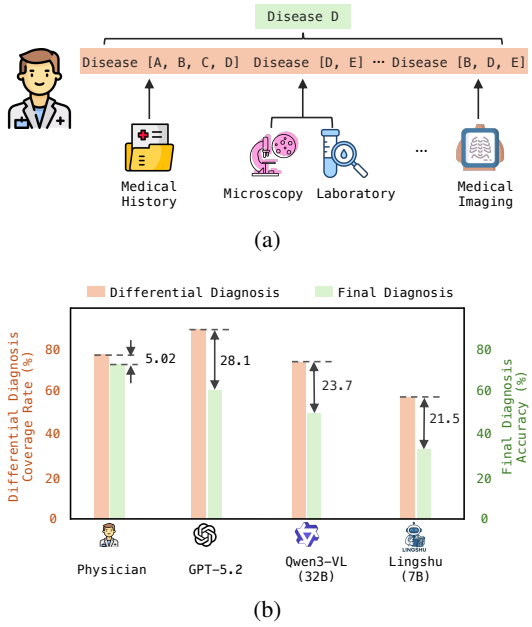


Figure 1: **(a)** Clinicians curate a broad differential diagnosis (DDx) list before determining a final diagnosis (FDx) via evidence synthesis. **(b)** Models exhibit a substantial gap between DDx coverage rate and FDx accuracy, far exceeding that observed in human experts.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated great potential in advancing clinical applications (Qiu et al., 2024), yet the benchmarks used to evaluate them remain limited and fragmented. Early benchmarks (Lau et al., 2018; Abacha et al., 2024; Liu et al., 2021; Zhang et al., 2023; Hu et al., 2024; Ye et al., 2024) primarily target single-image visual question answering (VQA) such as basic object recognition. More recent efforts (Zuo et al., 2025; Yu et al., 2025) move toward more realistic settings by requiring integrative reasoning over multiple images, but several limitations persist. First, despite including multiple images, individual questions draw images from the same clinical-evidence (CE) type, *e.g.*, cross-sectional CT scans. In clinical practice, clinicians typically leverage heterogeneous CE types, spanning laboratory tests, imaging from multiple modalities, microscopy images, and even omics data. This is particularly relevant for accurate diagnosis in complex clinical cases such as multimorbidity. While MedXpertQA MM (Zuo et al., 2025) contains a multi-CE subset, it constitutes only a small fraction of the benchmark, with an average of 2.74 CE types per case. Second, most benchmarks emphasize selecting a final diagnosis (FDx). However, real diagnostic workflows typically begin with generating a differential diagnosis (DDx), *i.e.*, a set of plausible conditions consistent with findings from one or more CE types, and then determine the FDx by synthesizing evidence across all available CE types. Finally, most ex-

isting benchmarks are English-only, limiting the assessment of MLLMs' multilingual capabilities in clinical contexts.

To address these limitations, we introduce MEDSYN, a multilingual, multimodal benchmark of complex clinical cases, where each question contains, on average, **3.97 CE types** and **8.42 images** (Table 1) drawn from up to **7 distinct CE types** (Figure 3b). Mirroring real-world diagnostic workflows, our benchmark evaluates models on two tasks: (i) **DDx generation** and (ii) **FDx selection**. We benchmark 18 state-of-the-art MLLMs, including both general-purpose models (proprietary and open-source) and domain-specific medical models, with open-source scales ranging from 2B to 78B parameters. We further conduct two ablation studies using two medical MLLMs and their base model: (i) perturb textual CE by removing it or replacing it with length-matched random token strings; and (ii) run a leave-one-out analysis where each CE type is withheld and measure the resulting update in the model's answer posterior, comparing the same CE provided as raw images versus expert-derived diagnostic findings from them. Our key findings are summarized as follows:

① We show that leading models outperform expert clinicians on DDx generation but underperform on FDx selection, suggesting a capability gap between identifying plausible conditions from heterogeneous CE types and synthesizing them into a singular, accurate diagnosis. This gap varies by language, highlighting concerning cross-lingual disparities;

② We find that MLLMs overrely on textual inputs, skewing evidence weighting toward less discriminative CE (*e.g.*, medical history). Removing such evidence increases attention to image tokens and, counterintuitively, improves diagnostic accuracy despite less CE being available;

③ We demonstrate that visual understanding remains a major bottleneck: cross-modal misalignment distorts how MLLMs calibrate different CE types, yielding a *cross-modal CE utilization gap*. We introduce a novel metric, termed *Evidence Sensitivity*, to quantify this gap and show that a smaller gap correlates with higher diagnostic accuracy. We further show that this metric provides actionable guidance for targeted interventions to improve model performance.

Benchmark	# Images	Multi-lingual	Avg. # Images per Case	Avg. # Evidence Types per Case
VQA-Rad	204	✗	1	1
VQA-Med	500	✗	1	1
Slake	96	✓	1	1
PMC-VQA	29k	✗	1	1
OmniMedVQA	118k	✗	1	1
GMAI-MMBench	21k	✗	1	1
MedXpertQA MM	2.8k	✗	2.1	2.74
MEDSYN	3.6k	✓	8.42	3.97

Table 1: Comparison of MEDSYN with existing multimodal medical benchmarks.

2 Related Work

Multimodal Large Language Models. General-purpose MLLMs (OpenAI, 2025; Anthropic, 2025; Comanici et al., 2025) have demonstrated remarkable zero-shot capabilities on medical tasks, including diagnosing complex clinical cases (Lam et al., 2025), owing to the extensive clinical knowledge encoded in their LLM backbones (Singhal et al., 2023). To further improve performance, recent work has explored domain-specific adaptation (Chen et al., 2024a; Xu et al., 2025) by fine-tuning MLLMs on specialized medical data spanning diverse CE types. Despite these advances, current MLLMs remain prone to hallucinations and biases that impede safe deployment in real-world clinical environments (Cross et al., 2024). These challenges highlight the need for comprehensive benchmarks that evaluate models' ability in complex and realistic diagnostic settings.

Multimodal Medical Benchmarks. Existing medical benchmarks fall short of reflecting real-world clinical demands. Early benchmarks primarily target single-image VQA in narrow domains such as radiology (Lau et al., 2018; Abacha et al., 2024; Liu et al., 2021) and pathology (He et al., 2020). Recently, more generalized benchmarks such as PMC-VQA (Zhang et al., 2023), OmniMedVQA (Hu et al., 2024), and GMAI-MMBench (Ye et al., 2024) have been proposed to assess MLLMs capabilities across diverse CE types. For instance, OmniMedVQA (Hu et al., 2024) covers 12 CE types, including multiple imaging modalities (*e.g.*, CT, MRI, X-ray, ultrasound), microscopy, and specialized imaging such as colonoscopy. However, individual questions in these benchmarks remain isolated single-image snapshots tied to a single CE type. Although MedXpertQA MM (Zuo et al., 2025) introduces multi-image, multi-CE VQA, this

constitutes a small fraction of the benchmark and includes a limited number of CE types per question. Finally, most benchmarks are English-only, limiting evaluation of multilingual capabilities of MLLMs in clinical contexts. A detailed comparison is provided in Table 1.

3 Benchmark

3.1 Data Collection and Preprocessing

Data Collection. We collect consecutive English cases from the *New England Journal of Medicine Case Challenge* series and Chinese case reports published in the *National Medical Journal of China* (see Appendix A.1 for examples) from November 2015 to October 2025. For each case, we extract background and initial discussion up to DDx, along with all referenced tables and figures. In particular, the background contains textual CE including medical history and physical examination findings. The referenced tables and figures comprise visual CE from diagnostic investigations (*i.e.*, objective, technical tests ordered by clinicians), such as laboratory studies, diagnostic imaging (*e.g.*, CT, MRI and X-ray), microscopy and electrophysiological measurements (*e.g.*, EEG). Finally, the initial discussion presents the expert-derived diagnostic findings from visual CE. The ground-truth (GT) FDx is taken from the FDx section or, if not explicitly stated, determined by a clinician based on the full report. All collected cases are subject to expert validation, and cases are excluded if (i) any individual visual CE is not diagnostically interpretable (*e.g.*, low image quality, incomplete anatomical coverage, or artifacts), or (ii) the complete set of CE is diagnostically insufficient (*i.e.*, the clinician is not able to confirm the FDx from all provided evidence). After validation, we obtain 452 cases in total (398 in English and 54 in Chinese). Figure 3 summarizes visual CE type distribution (Figure 3a) and counts per case (Figure 3b).

Data Preprocessing. For each case, we manually crop and label images by their CE type. For the same evidence from different time points, we add an additional reference when it is obtained. Image captions are added to the discussion, from which we employ GPT-5 (OpenAI, 2025) to extract and organize expert-derived diagnostic findings based on CE types through in-context learning (Dong et al., 2024). This establishes a one-to-one mapping from visual CE to textual interpretation. Since a single CE can be reviewed by multiple clinicians, we utilize GPT-5 to summarize individual interpretations into a singular, cohesive summary to mitigate redundancy. The template prompts and examples of processed data are in Appendix A.2. Finally, we conduct a manual quality check by cross-referencing the processed data with the original reports to ensure completeness and mitigate hallucinations, and a randomly selected 20% subset is further validated by clinicians.

3.2 Benchmark Design

To mirror real-world clinical workflow, we design two distinct tasks: (i) DDx generation, which involves information gathering and hypothesis generation, and (ii) FDx selection, which requires evidence synthesis and diagnostic verification.

DDx generation. The DDx generation is an open-ended generation task where MLLMs are asked to present a differential list ranked from the most probable to least probable diagnosis.

FDx selection. For FDx selection, we construct closed-ended multiple-choice questions (MCQs) where the single correct answer is the GT FDx and distractors are drawn from the model-generated DDx. We employ GPT-5 to generate a differential list and then select distractors according to two criteria: (i) distractors must exclude the FDx and any synonym or wording variant thereof; (ii) distractors should be clinically proximate to the FDx (*e.g.*, a closely related histologic subtype; see Appendix A.2 for details). We apply an adversarial refinement process where we iteratively identify and eliminate potential shortcuts in how models distinguish correct from incorrect diagnoses (Burgess et al., 2025). Finally, the MCQs are reviewed by expert clinicians based on content validity and clinical relevance to ensure that discriminative CE explicitly disqualifies all distractors, thereby eliminating diagnostic ambiguity among the options.

Figure 2 illustrates two cases from FDx selection tasks with color-coded CE types. Each visual CE is paired with an expert-derived summary of diagnostic findings from it. Duruing inference, a given CE type is provided as either visual or textual input, not both. Examples of DDx generation tasks are shown in Figure 8 in Appendix A.3.

Metrics. For DDx generation, we follow the evaluation framework of Kanjee et al. (Kanjee et al., 2023) and employ GPT-5 as an automated judge to score generated DDx on a 0–5 scale (details in Ap-

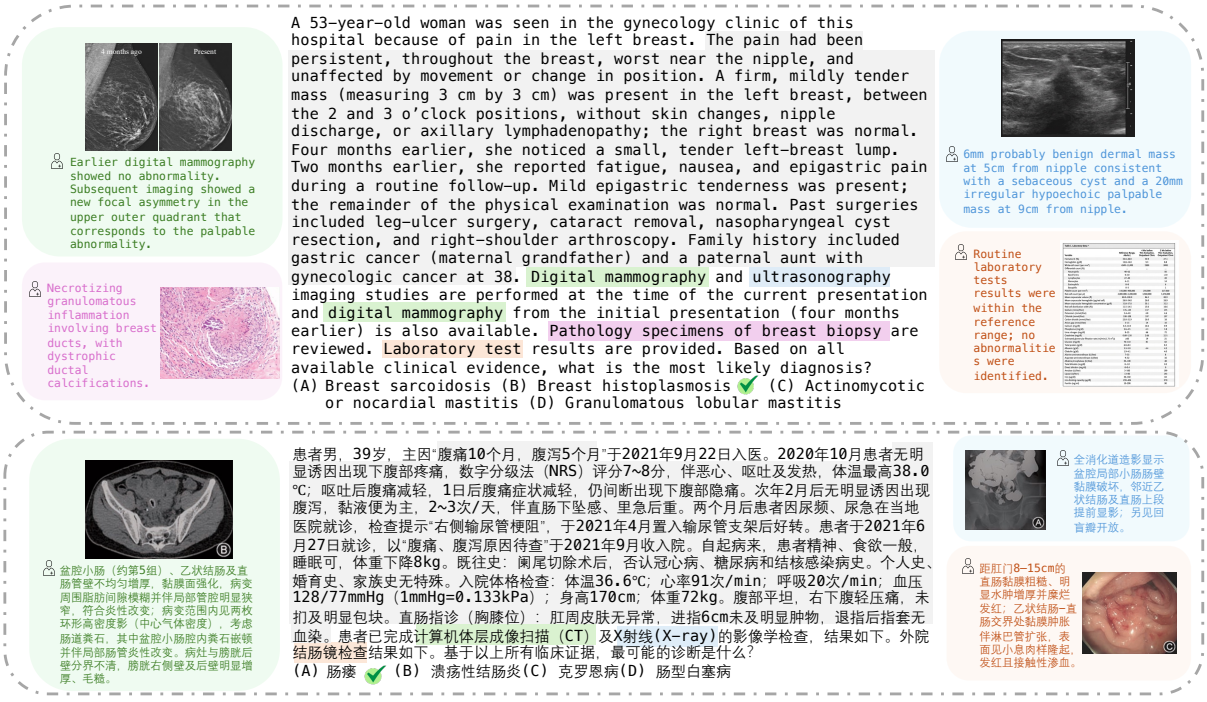


Figure 2: **Example final diagnosis selection tasks in English (top) and Chinese (bottom)**. Colors mark different visual clinical evidence (CE) types referenced in the question, with corresponding expert-derived diagnostic findings; gray denotes textual CE. In our experiment, each CE type is input as either raw images or text findings, not both.

pendix B.1). We report the coverage rate, defined as the percentage of cases whose DDx include the FDx, considering scores ≥ 4 (*i.e.*, DDx comprises the exact FDx or a highly synonymous condition) as a positive coverage based on clinician's suggestion. For FDx selection, we report overall accuracy.

4 Experiments

4.1 Evaluation

We evaluate a diverse set of MLLMs. Proprietary models include representative GPT (OpenAI, 2025), Gemini (Comanici et al., 2025), and Claude 4.5 Opus (Anthropic, 2025). Open-source models span 2B–72B parameters, covering widely used families such as Qwen (Yang et al., 2025a), InternVL (Chen et al., 2024b), and DeepSeek-VL (Wu et al., 2024). We also include three medical MLLMs: HuatuoGPT (Chen et al., 2024a), Lingshu (Xu et al., 2025), and Med-Mantis (Yang et al., 2025b). The evaluation is conducted using the VLMEvalKit (Duan et al., 2024) framework on 8 NVIDIA A6000 GPUs. We evaluate all models using a zero-shot setting. We also recruit two senior physicians to evaluate the English subset of the benchmark (see Appendix C for details).

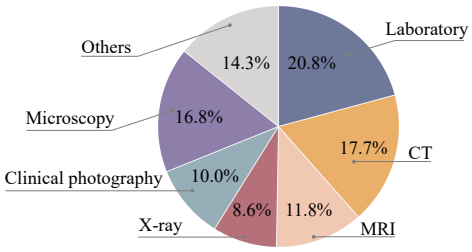
4.2 Main Results

Table 2 compare the performance of 18 MLLMs and senior physicians on DDx generation and FDx selection. Figure 4 further breaks down FDx performance across eight clinical specialties on the English subset. Overall, proprietary models consistently outperform open-source models, with an average gain of over 10 percentage points (pp) on both tasks. Claude Opus 4.5 achieves the highest FDx accuracy, while GPT-5.2 performs best on DDx generation. Among open-source models, Qwen3-VL (32B) demonstrates superior performance on both tasks, even surpassing models with significantly larger number of parameters. Among comparable settings, the domain-specific medical MLLMs: HuatuoGPT-Vision and Lingshu both improve FDx accuracy over their base model, Qwen2.5-VL (7B). Notably, senior physicians achieves a mean accuracy of 72.11% on FDx selection (English subset), which is 7.54 pp higher than the best MLLM, but score about 10 pp lower than the best model on DDx generation. Our key findings are as follows.

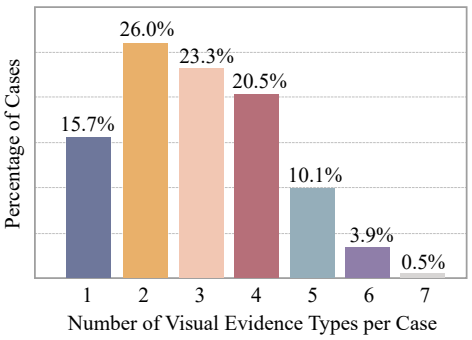
MLLMs Exhibit a Substantial Gap Between DDx Coverage and FDx Accuracy. Across models, we observe a large gap between DDx coverage

Model	English			Chinese			Overall	
	DDx Coverage Rate (%)	FDx Selection Acc. (%)	Rank (FDx)	DDx Coverage Rate (%)	FDx Selection Acc. (%)	Rank (FDx)	DDx Coverage Rate (%)	FDx Selection Acc. (%)
Senior Physician	77.13	72.11	-	-	-	-	-	-
Claude Opus 4.5	84.92	64.57	1	61.11	55.56	1	82.08	63.49
GPT-O3	86.15	64.07	2	64.81	48.15	4	83.60	62.17
Gemini 2.5 pro	80.60	62.81	3	64.81	53.70	3	78.72	61.72
GPT-5.2	88.55	60.45	4	66.04	55.56	1	85.86	59.87
Gemini 2.5 Flash	73.55	57.54	5	54.72	37.04	6	70.50	55.09
Qwen3-VL (32B)	73.55	49.87	6	48.00	33.33	8	70.50	47.89
Qwen2.5-VL (72B)	70.03	48.24	7	51.85	40.74	5	63.94	47.34
InternVL3 (78B)	66.25	44.72	8	38.89	31.48	10	61.63	43.14
InternVL3.5 (38B)	64.23	43.47	9	37.74	24.07	17	53.53	41.15
Qwen3-VL (8B)	62.47	40.20	10	39.22	31.48	10	66.08	39.16
HuatuoGPT-Vision (7B)	60.71	38.69	11	24.07	31.48	10	49.01	37.83
InternVL3.5 (8B)	59.45	38.19	12	35.19	25.93	14	51.91	36.73
Lingshu (7B)	58.44	36.93	13	22.22	33.33	8	49.90	36.50
Qwen2.5-VL (7B)	55.92	33.92	14	22.22	29.63	13	52.24	33.41
DeepSeek-VL2 (27B)	55.67	32.91	15	25.93	37.04	6	38.81	33.40
Qwen3-VL (2B)	52.14	30.90	16	20.37	25.93	14	34.82	30.31
DeepSeek-VL2 (3B)	51.89	30.90	16	20.37	24.07	17	21.46	30.08
Med-Mantis (8B)	47.61	27.89	18	9.26	25.93	14	27.79	27.66

Table 2: Comparison of differential diagnosis (DDx) coverage rate (%) and final diagnosis (FDx) selection accuracy (%) across models on the English and Chinese subsets, and overall. Ranks are computed within each subset based on FDx selection accuracy using competition ranking, excluding the physician baseline. Clinical evidence from diagnostic investigations is provided as raw images.



(a)



(b)

Figure 3: **(a)** Distribution of visual clinical evidence (CE) types. **(b)** Number of visual CE types per case.

rate and FDx accuracy (approximately 20 pp on average on the English subset), far exceeding that of senior physicians (5.02 pp). The gap becomes even more pronounced for smaller models. For example, For example, HuatuoGPT-Vision achieves 60.71% DDx coverage rate but only 38.69% FDx accuracy. This suggests that while MLLMs can effectively enumerate potential diagnoses, they cannot yet function at the level of human experts in calibrating and synthesizing heterogeneous CE to correctly arrive at the FDx. On the Chinese subset, a similar trend can be observed for larger models ($\geq$70B), whereas small and medium-sized models show a divergent trend, sometimes exhibiting a negative gap. This shift, however, likely reflects a floor effect: since lower-tier models perform poorly on Chinese tasks, their results are often close to random guessing, so the gaps are largely by chance.

Visual Understanding Remains a Diagnostic Bottleneck. Table 3 reports the accuracy of FDx when raw visual evidence from diagnostic investigations is replaced by expert interpretations. As shown in Table 3, both proprietary and open-source models demonstrate a significant accuracy gain of over 10 pp, suggesting that the vision encoder's inability to accurately capture granular details in complex clinical imagery remains a primary bottleneck for MLLMs in diagnostic tasks.

MLLMs Exhibits Cross-Lingual Performance Disparities. While proprietary models generally lead on both the English and Chinese subsets, individual model rankings vary significantly across languages. For instance, DeepSeek-VL2 (27B) moves from 15th in English to 6th in Chinese. The performance gap between DDx and FDx also varies considerably by language. For top-tier models, it narrows from approximately 20 pp to about 10 pp, and in some cases even reverses, with models performing better on FDx (*e.g.*, Med-Mantis).

Model	Raw Images	Expert-Derived Diagnostic Findings
GPT-O3	64.07	75.38 (+11.31)
GPT-5.2	60.45	75.13 (+14.68)
QWEN2.5-VL (72B)	48.24	65.83 (+17.59)
QWEN3-VL (32B)	49.87	65.58 (+15.71)
INTERNVL3 (78B)	44.72	63.82 (+19.10)
QWEN3-VL (8B)	40.20	59.05 (+18.85)
LINGSHU (7B)	36.93	54.27 (+17.34)
HUATUOGPT-VISION (7B)	38.69	52.26 (+13.57)
QWEN2.5-VL (7B)	33.92	50.50 (+16.58)

Table 3: Final diagnosis accuracy (%) across models on English subset. We compare performance when clinical evidence from diagnostic investigations is provided as raw images versus expert-derived text findings.

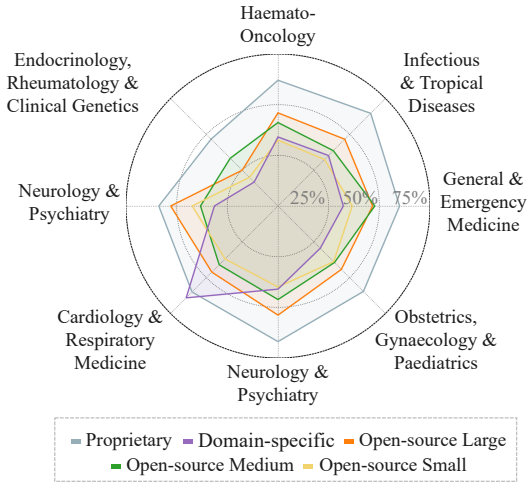


Figure 4: Mean final diagnosis accuracy (%) for proprietary models, domain-specific models, and open-source general-purpose models across clinical specialties.

Domain-Specific Training Can Outperform Parameter Scale in Specialized Clinical Tasks. We analyze FDx performance across 8 clinical specialties on the English subset. We report mean accuracy for proprietary models, domain-specific models, and open-source general-purpose models grouped by size: large (≈70B), medium (≈30B), and small (≈7B). As shown in Figure 4, proprietary models exhibit the most balanced performance across the clinical spectrum, notwithstanding a systematic performance deficit in obstetrics, gynaecology, and paediatrics. Notably, medical domain-specific models achieved a mean accuracy of 64.10% in cardiology and respiratory medicine, outperforming both proprietary models (60.00%) and general-purpose models with tenfold larger parameter counts (46.15%). We hypothesize that this localized dominance stems from specialized physiological mappings inherent to these disciplines such as correlation between ECG morphology and cardiac conduction events, which domain-specific finetuning more effectively captures than simple parameter scaling.

4.3 Analytical Results

Finding 1: Text bias miscalibrates evidence weighting in MLLMs.

Recent works show that MLLMs frequently exhibit a strong bias toward linguistic signals, placing "blind faith" in text even when visual evidence is presented (Deng et al., 2025; Lee et al., 2025). In clinical practice, textual CE such as medical history and physical findings is essential for DDx generation but often contains nonspecific features that are common to multiple diagnoses. Conversely, visual CE from diagnostic investigations (*e.g.*, diagnostic imaging and microscopy) is typically more discriminative, increasing physician confidence in FDx by over 30% (Peterson et al., 1992).

To investigate how this bias affects MLLM evidence calibration, we conducted two ablation studies on the English subset of the FDx selection task: (i) REMOVE-TEXT, where textual CE is omitted, and (ii) RANDOM-TEXT, where textual CE is replaced by length-matched, randomly sampled tokens to isolate the effect of token budget from semantic content. As shown in Table 4, both ablations yield consistent performance gains across all three models, with larger improvements for domain-specific models. This counterintuitive *less-is-more* effect that reducing available CE yet improves accuracy suggests textual CE acts as a distractor instead of helpful context for FDx selection due to inherent bias in MLLMs. For the RANDOM-TEXT ablation, we further analyze the Relative Attention per Token (RAPT) (Liu et al., 2025) (see Appendix B.2 for formal definition) for Lingshu on text and image tokens across layers (see Appendix B.4 for additional models). Our analysis (Figure 5) reveals that while the model generally allocates disproportionately high attention to text than image tokens, this gap narrows under the RANDOM-TEXT condition. In particular, we observe a more pronounced RAPT surge on image tokens within deeper layers which have been identified as visual grounders in recent work (Liu et al., 2025).

Overall, our results suggest MLLMs suffer from an evidence miscalibration failure mode where the models underweight visual CE despite its superior

Setting	HuatuoGPT-Vision	Lingshu	Qwen2.5-VL (7B)
BASELINE	38.69	36.93	33.92
REMOVE-TEXT	42.21 (+3.52 pp)	43.97 (+7.04 pp)	34.42 (+0.50 pp)
RANDOM-TEXT	39.70 (+1.01 pp)	41.71 (+4.78 pp)	34.67 (+0.75 pp)

Table 4: Final diagnosis accuracy (%) before and after REMOVE-TEXT and RANDOM-TEXT interventions.

diagnostic value due to inherent bias toward text. Such bias is semantics-sensitive: coherent clinical text diverts attention away from visual grounding, whereas removing or randomizing it attenuates this effect and restores the model's focus to more informative visual features.

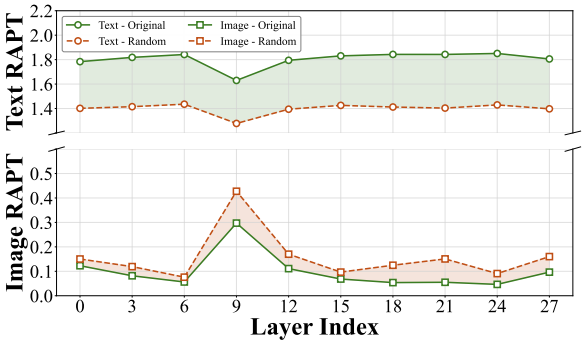


Figure 5: Layer-wise Relative Attention per Token (RAPT) for Lingshu on text (excluding question stem) versus image tokens, before and after the RANDOM-TEXT intervention.

Finding 2: Cross-modal misalignment distorts how MLLMs calibrate evidence.

Recent studies reveal a misalignment between text and image modalities in MLLMs, where visual encoding introduce noise or information loss when mapping raw images to image tokens (Jain et al., 2025; Shu et al., 2025; Li et al., 2025). We hypothesize this gap can lead to inconsistent weighting of identical CE presented in different modalities.

To investigate this hypothesis, we introduce *Evidence Sensitivity* to measure the impact of specific CE on the model's decision-making. Formally, let $\mathcal{E} = \{e_1, \ldots, e_M\}$ denote the set of CE for a given case, and let $p_\theta(y \mid \mathcal{E})$ represent the model's answer posterior over the answer space y. For each CE type $e_m \in \mathcal{E}$, we define its sensitivity as the shift in the model's answer posterior upon its removal, computed as the Jensen–Shannon divergence (JSD) between the full and ablated distributions:

$$S^{(m)} = \mathrm{JSD}(p_\theta(y \mid \mathcal{E}) \parallel p_\theta(y \mid \mathcal{E} \setminus \{e_m\})). \quad (1)$$

This metric captures the update in model's belief when evidence e_m is omitted. We compute the sensitivity for the same evidence presented in two distinct modalities: (i) raw image input ($S^{(m)}_{\text{image}}$) and (ii) expert-derived textual summaries of diagnostic findings from the image. Theoretically, identical evidence should yield comparable belief updates across modalities, *i.e.*, $S^{(m)}_{\text{image}} \approx S^{(m)}_{\text{text}}$, with data points $(S^{(m)}_{\text{text}}, S^{(m)}_{\text{image}})$ clustering along the identity line $y = x$. Deviations from this diagonal define a *cross-modal CE utilization gap*, reflecting the extent of evidence calibration distortion stemming from cross-modal misalignment. Figure 6 provides a qualitative comparison between two CE types (*i.e.,* CT and Microscopy) for Qwen2.5-VL (7B) and Lingshu (see Figure 13 and 14 in Appendix B.4 for results on more CE types and for HuatuoGPT-Vision). From the plot we can observe data points rarely align with the identity line. Instead, they predominantly cluster along either the horizontal or vertical axes, forming an L-shaped distribution. This pattern indicate a cross-modal distortion in evidence calibration: models remain relatively insensitive to evidence in one modality that they otherwise deem critical in another. Notably, for microscopy, most data points cluster below the identity line, suggesting a systematic undersensitivity to visual CE.

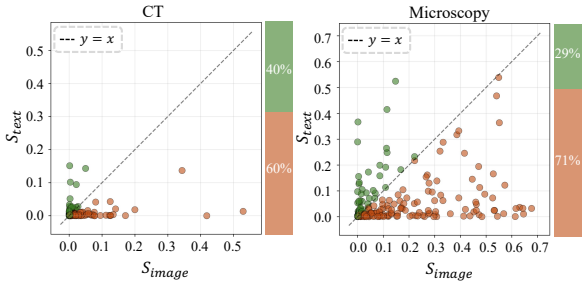
We further quantify this distortion by computing the Normalized Mean Squared Error (NMSE) relative to the identity line:

$$\mathrm{NMSE}_{y=x} = \frac{\sum_{i=1}^{N} \left(S^{(m)}_{\text{image},i} - S^{(m)}_{\text{text},i}\right)^2}{\sum_{i=1}^{N} \left(S^{(m)}_{\text{text},i}\right)^2} \quad (2)$$

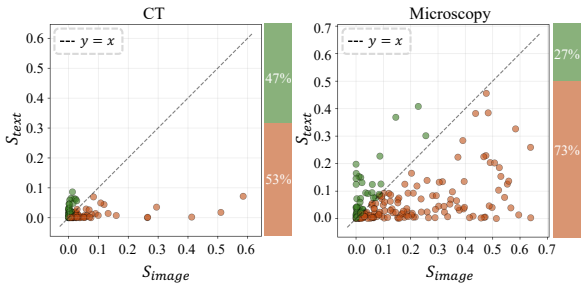
where i denotes individual cases and N is the total number of cases. Table 5 reports the NMSE for different evidence types across the three models (detailed results for each specific CE type are provided in Table 12 of Appendix B.4). Our analysis reveals that a lower mean $\mathrm{NMSE}_{y=x}$ correlates with higher overall accuracy.

Finally, we demonstrate that the cross-modal CE utilization gap is actionable through two strategies: (i) test-time prompt refinement and (ii) targeted supervised finetuning (SFT).

Test-time Prompt Refinement. To address the model's systematic undersensitivity to microscopy images, we redesign the system prompt to explicitly instruct the model to attend to it (see Appendix B.3



(a) Qwen2.5-VL (7B)



(b) Lingshu (7B)

Figure 6: Cross-modal sensitivity on CT and microscopy. Green and brown points indicate cases where $S^{(m)}$image $>$ $S^{(m)}$text and $S^{(m)}$image $<$ $S^{(m)}$text, respectively. The right-hand bar chart shows the proportion of cases in each category.

Model	Average NMSE ↓	Acc. ↑
QWEN2.5-VL (7B)	0.98	33.92
LINGSHU (7B)	0.95	36.93
HUATUOGPT-VISION (7B)	**0.93**	**38.69**

Table 5: Comparison of average normalized mean squared error (NMSE) and final diagnosis accuracy (%) across models. Best results are highlighted in **bold**.

for details). Following this intervention, we observe that, across models, the fraction of instances below the identity line decreases (see Figure 15 in Appendix B.4), indicating increased sensitivity to microscopy images. Consistently, the cross-modal gap narrows, reflected by lower microscopy NMSE (see Table 13 in Appendix B.4). Crucially, the reduced gap translates directly to downstream performance. As shown in Table 6, overall accuracy improves across three models after prompt refinement. These findings underscore that bridging the utilization gap between visual and textual evidence is essential for achieving reliable, high-performance clinical decision-making.

Targeted SFT. Another strategy to mitigate the utilization gap is to finetune the model on a curated dataset enriched with images from undersensitive CE types (*i.e.*, targeted SFT). Specifically, we curate a *microscopy-heavy* dataset comprising 50% microscopy images and 50% other medical images. As shown in Table 7, Qwen2.5-VL (7B) model fine-tuned on this curated dataset achieves an accuracy improvement of 1.51 pp, doubling the 0.75 pp gain from general SFT on non-curated data. These results demonstrate that cross-modal evidence utilization gap provides a strategic signal for guiding SFT data curation: by up-sampling CE types that are systematically underutilized by the base model, we can achieve better overall performance after SFT. We provide experiment details in Appendix D.

Model	Baseline Acc.	Prompt Refinement Acc.
QWEN2.5-VL (7B)	33.92	34.67 (+0.75 pp)
LINGSHU	36.93	38.44 (+1.51 pp)
HUATUOGPT-VISION	38.69	38.94 (+0.25 pp)

Table 6: Comparison of final diagnosis accuracy (%) before and after prompt refinement across models.

Setting	Acc.
BASELINE	33.92
GENERAL SFT	34.67 (+0.75)
TARGETED SFT	**35.43 (+1.51)**

Table 7: Final diagnosis accuracy (%) for Qwen2.5VL before and after supervised fine-tuning (SFT), comparing general with targeted SFT. Best results are in **bold**.

5 Conclusion

We introduce MEDSYN, a multilingual, multimodal benchmark of highly complex clinical cases that necessitate synthesis of multiple CE types for accurate diagnosis. We evaluate 18 MLLMs on both DDx generation and FDx selection tasks. Across models, we observe a critical gap between DDx coverage rate and FDx accuracy, substantially larger than that observed for human experts, indicating that MLLMs struggle to synthesize heterogeneous CE to arrive at the correct FDx. We further attribute it to two primary factors: overreliance on textual CE and cross-modal CE utilization gap. Together, these factors lead to an evidence-calibration failure mode, where models weight clinical evidence not only by its diagnostic value but also by their internal bias. We introduce Evidence Sensitivity to quantify this cross-modal gap, and show that a narrower gap correlates with better overall performance. Finally, we demonstrate that this gap provides actionable guidance for targeted interventions to improve model performance.

Limitations

Linguistic Imbalance. The Chinese subset constitutes only a smaller portion (∼12%) of the total benchmark. This disparity may limit the generalizability of our findings regarding model performance in non-English clinical contexts and diverse medical linguistic environments.

Absence of Chinese Physician Baseline. While we provide human expert benchmarks for the English subset, we currently lack a comparable physician baseline for the Chinese subset. Without this comparison, it is difficult to fully contextualize the gap between DDx coverage and FDx accuracy for models specifically on Chinese clinical cases.

Automated Evaluator Bias. Our evaluation framework utilizes GPT-5 as a judge to score generated DDx. While this automated approach enables scalable evaluation, it may introduce evaluator bias, potentially favoring models with similar linguistic patterns (*e.g.*, GPT-5.2 and GPT-o3) or over-indexing on specific medical terminologies.

References

Ben Abacha, Hasan Asma, A Sadid, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Muller. 2024. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *(No Title)*.

Anthropic. 2025. Introducing claude 4. Anthropic News.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, and 1 others. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19552–19564.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

James L Cross, Michael A Choma, and John A Onofrey. 2024. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 3(11):e0000651.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

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*.

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

Anubhooti Jain, Mayank Vatsa, and Richa Singh. 2025. Words over pixels? rethinking vision in multimodal large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25) Survey Track*, pages 10481–10489.

Zahir Kanjee, Byron Crowe, and Adam Rodman. 2023. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80.

Kyle Lam, Julia Calvo Latorre, Andrew Yiu, Grace Navin, Alexander Tan, and Jianing Qiu. 2025. Physician input improves generative artificial intelligence models' diagnostic performance in solving complex clinical cases. *The Lancet Digital Health*.

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.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulic, and Anders Søgaard. 2025. Lost in embeddings: Information loss in vision-language models. *arXiv preprint arXiv:2509.11986*, 2.

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.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, and 1 others. 2025. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

OpenAI. 2025. Introducing gpt-5. OpenAI.

MICHAEL C Peterson, JOHN H Holbrook, DE Von Hales, NL Smith, and LV Staker. 1992. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *Western Journal of Medicine*, 156(2):163.

Jianing Qiu, Wu Yuan, and Kyle Lam. 2024. The application of multimodal large language models in medicine. *The Lancet Regional Health–Western Pacific*, 45.

Mehmet Saygin Seyfioglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *Preprint*, arXiv:2312.04746.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

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, Jiaze 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, Ziyan 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*.

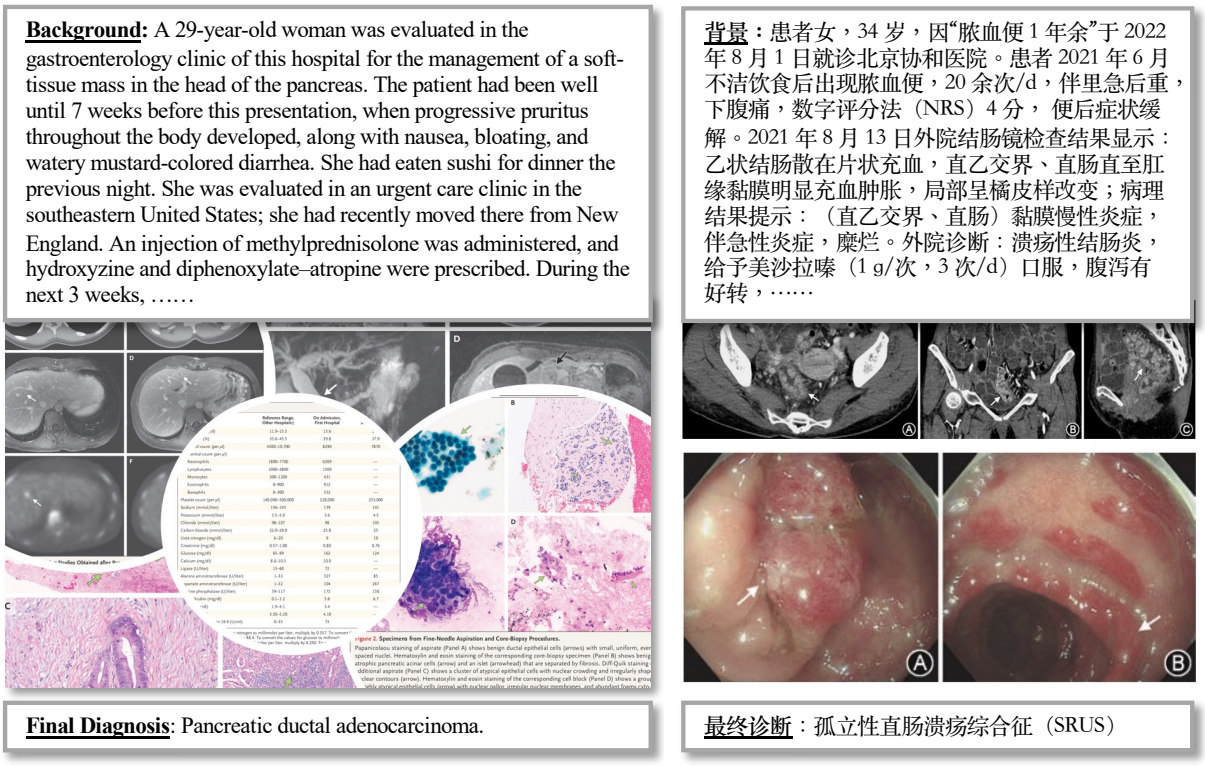
Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

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 Details of Data Collection and Processing

A.1 Examples of Case Reports

Figure 7 showcases two examples of case reports from the *New England Journal of Medicine Case Challenge* series (Figure 7a) and the *National Medical Journal of China* (Figure 7b).



(a) (b)

Figure 7: Examples of raw case reports. **(a)** Case records from Massachusetts General Hospital (Boston, MA); **(b)** Case reports published by National Medical Journal of China .

A.2 Prompt for Data Processing

The prompt for extracting textual evidence and clinician's interpretation of visual evidence based on evidence types:

```
You are an clinical information extraction
    assistant.

Task
Given a single medical case consisting of:
(1) Case ID
(2) Background text
(3) Figure caption text (captions extracted
    from figures)
return exactly ONE flat JSON object with
    verbatim (copy-pasted) content.
Do not paraphrase.

Inputs
- Case ID: {case_id}
- Background: {background}
- Caption: {caption}

Output format
Return only valid JSON (no prose, no markdown
    code blocks).

Required keys
- "Case ID": string
- "Modality": string (comma-separated)

Optional keys (include only if present)
- "Patient Background"
- "Laboratory Findings"
- "Medications"
- "Impression / Differential" (ONLY if
    explicitly labeled)
- "CT Findings"
- "MRI Findings"
- "PET Findings"
- "Ultrasound Findings"
- "X-ray Findings"
- "Pathology Findings"
- "Clinical Photography Findings"
- "Other Imaging Findings" (include the
    modality name in the value)

Value rules (for ALL optional keys)
- Each value is a single string made of one or
    more verbatim chunks.
- Join chunks with a DOUBLE line break: \n\n
- Preserve internal line breaks within a chunk.
- Keep temporal/procedural qualifiers inside
    the chunk (e.g., "after
  contrast", "hospital day 2", "1 year
    earlier", planes/reconstructions).
- Deduplicate materially identical statements
    (keep the clearest).
- If duplicates conflict, keep both (each with
    its own source tag).

Source tags (append when available; do NOT
    invent)
Allowed tag formats:
- (Dr. <Name>)
- (Figure <N> caption) OR (Figure caption) if
    unnumbered
```

```
If no source is available, omit the tag.

Modality extraction (concise rules)
- Detect modalities from Caption first. If
    Caption is entirely absent,
  infer modalities from Background.
- Allowed modalities (for "Modality" and
    imaging keys):
  CT, MRI, PET, Ultrasound, X-ray, Pathology,
    Clinical Photography, Other Imaging.
- Hybrids (e.g., PET-CT): split into PET and
    CT when feasible; duplicate
  only if a sentence truly spans both.
- Assign each imaging sentence to its matching
    modality key. If ambiguous,
  use "Other Imaging Findings" and name the
    modality exactly as written.

Rule for the "Modality" key
- List ONLY modalities present in FIGURES
    (from Caption).
- If Caption is entirely absent, list
    modalities discussed in Background.
- Do NOT include non-imaging categories (e.g.,
    labs, meds).

Caption handling
- If multiple figures are described, split by
    figure.
- Attribute each extracted chunk with the
    precise figure source tag.

Background handling
- Extract verbatim content for patient
    background, labs, medications, and any
  explicitly labeled impression/differential.
- Add imaging from Background ONLY if Caption
    is absent or clearly
  incomplete.
- Do not add interpretation unless "Impression
    / Differential" is
  explicitly labeled.

Hard constraints (do NOT do these)
- Do NOT paraphrase or normalize
    numbers/dates/units.
- Do NOT output nested JSON; output a single
    flat object.
- Do NOT include internal/system markers
    (e.g., :contentReference[...],
  oaicite).
- Do NOT use file names as sources.
- Do NOT output anything except the single
    JSON object.

Procedure
1. Parse Caption -> split per figure, identify
    modality cues, extract
   verbatim chunks with figure source tags.
2. Parse Background -> extract non-imaging
    (background, labs, meds) and any
   labeled impression/differential; add
    imaging only if needed.
3. Build "Modality" using the rule above.
4. Aggregate chunks per key with \n\n and
    return JSON only.

Here are some examples:
...
```

The prompt for summaize individual expert interpretations:

```
You are an expert medical summarization
    assistant.

Task
Given ONE case, produce concise summaries
    while preserving the case ID,
modality list, medications, and diagnosis
    information if present.
Use ONLY the provided text. Do not invent or
    infer missing details.

Inputs
- Case ID: {case_id}
- Modality: {modality_string}
    (comma-separated; provided by dataset)
- Patient Background: {patient_background}
- Findings sections (zero or more):
  - {findings_key_1}: {findings_text_1}
  - {findings_key_2}: {findings_text_2}
  - ...

Outputs
Return only valid JSON of summaries with the
    exact formats below.

A) Patient background summary (if Patient
    Background is provided)
Instructions
- Use professional medical language.
- Remove any explicit diagnosis names already
    stated in the background.
- Remove any imaging or laboratory results
    described in the background
  (including impressions/findings/conclusions).
- Focus on: presenting problem and context,
    key past medical/psychiatric
  history, major exam findings,
    complications/risk factors, current status
  and plan (including disposition and
    follow-up).
- Do not include author names or citations in
    parentheses.

Output format (exact labels)
Patient Background Summary: <single concise
    paragraph>

B) Imaging (or other) findings summary (for
    EACH provided findings section)
For each input section
    "{modality_or_section_name} Findings":
Instructions
- Include clinically meaningful abnormalities
    and anatomic locations
  (use specific structures when stated).
- Include interpretation ONLY if explicitly
    stated in the source text.
- Include comparisons to prior studies when
    stated.
- Keep modality/technique details only if they
    affect meaning.
- Exclude demographics, physician names, and
    extraneous technical details.
- Output ONLY the summary paragraph for that
    section.

Output format (per section)
```

```
<{modality_or_section_name} Findings Summary>:
    <single concise paragraph>

Hard constraints
- Do NOT add new diagnoses or conclusions.
- Do NOT normalize or change numbers, dates,
    or units.
- Do NOT output anything beyond the specified
    formats.

Here are some examples:
...
```

The prompt for generating distractors for multiple-choice questions (MCQs):

```
You are an expert clinician and adversarial
    exam-writer.

Goal
Given one case (history + imaging + labs) plus
    a final diagnosis and
a comma-separated differential list, produce
    three confusing but incorrect distractor
    diagnoses.

Inputs
- Patient history: {history_text}
- Imaging findings: {imaging_text}
- Laboratory findings: {labs_text}
- Ground-truth final diagnosis:
    {ground_truth_text}
- Differential diagnoses (comma-separated):
    {differential_diagnoses_text}

Core rules
1) Distractor constraints (must all be true)
- Output EXACTLY three distractors.
- ALL distractors MUST be drawn from the
    differential list.
  (You may normalize wording but must not
    change the underlying diagnosis.)
- Each distractor must be a real, recognized
    diagnosis.
- Each distractor must be genuinely plausible
    for THIS patient given the
  provided history, imaging, and labs.
- Do NOT output the ground truth or any
    synonym/trivial variant of it as a
  distractor.
- Avoid non-diagnoses (pure symptoms, risk
    factors, procedures).

2) Make distractors confusing
- Prefer distractors in the same organ system
    or a closely related one.
- Prefer distractors that share major features
    with the ground truth
  (symptoms, labs, imaging, histology).
- Each distractor should differ from the
    ground truth by subtle
  details (e.g., anatomic site, vascular
    territory, mechanism, histologic
  variant), such that there is still exactly
    ONE best answer.

3) Remove lexical giveaways
- Keep ground truth and distractors similar in
    "shape":
  same head noun/pattern when possible (e.g.,
    all "... adenocarcinoma",
  all "... ischemic stroke", all "...
    cardiomyopathy").
- Keep length and syntactic structure similar
    across the four options.
- Do not let only one option contain a
    uniquely specific keyword unless
  the others contain parallel, equally
    specific keywords.

Output format (JSON only)
Return ONLY valid JSON. No explanations. No
    comments. No code fences.
Use this exact schema:
{
  "ground_truth": "<string: normalized primary
    diagnosis>",
  "distractors": ["<string>", "<string>",
    "<string>"]
}

Here are some examples:
...
```

A.3 Examples of Differential Diagnosis Generation Cases

Figure 8 provides two example cases of differential diagnosis (DDx) generation.

A.4 Detailed Distribution of Clinical Specialties

We categorize cases into 17 clinical specialties based on UK's General Medical Council Specialties: General (internal) medicine, Infectious diseases, Tropical medicine, Emergency medicine, Haematology, Medical oncology, Endocrinology and diabetes mellitus, Rheumatology, Clinical genetics, Neurology, General psychiatry, Cardiology, Respiratory medicine, Renal medicine, Gastroenterology, Obstetrics and gynaecology, and Paediatrics. Figure 9a shows the distributions of 17 clinical specialties.

For analysis purposes, we further aggregate these specialties based on expert input to reflect patient management in clinical practice (*e.g.*, similar first-line investigations, inpatient vs. emergency workflows, and typical referral patterns) into 8 groups:

- **General & Emergency Medicine:** General (internal) medicine; Emergency medicine.
- **Infectious & Tropical Diseases:** Infectious diseases; Tropical medicine.
- **Haemato-Oncology:** Haematology; Medical oncology.
- **Endocrinology, Rheumatology & Clinical**

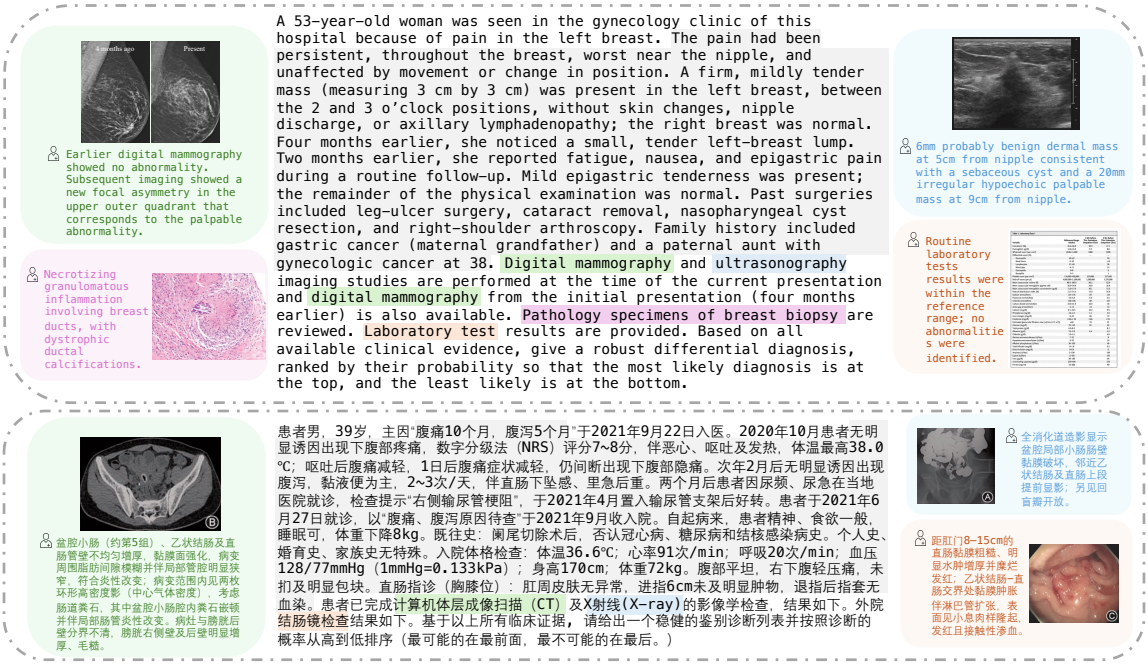


Figure 8: **Example differential diagnosis generation cases in English (top) and Chinese (bottom)**. Colors mark different visual clinical evidence (CE) types referenced in the question, with corresponding expert-derived diagnostic findings; gray denotes textual CE. In our experiment, each CE type is input as either raw images or text findings, not both.

Genetics: Endocrinology and diabetes mellitus; Rheumatology; Clinical genetics.

- **Neurology & Psychiatry:** Neurology; General psychiatry.
- **Cardiology & Respiratory Medicine:** Cardiology; Respiratory medicine.
- **Renal & Gastrointestinal Medicine:** Renal medicine; Gastro-enterology.
- **Obstetrics, Gynaecology & Paediatrics:** Obstetrics and gynaecology; Paediatrics.

Figure 9b demonstrates the distributions of these specialty groups in detail.

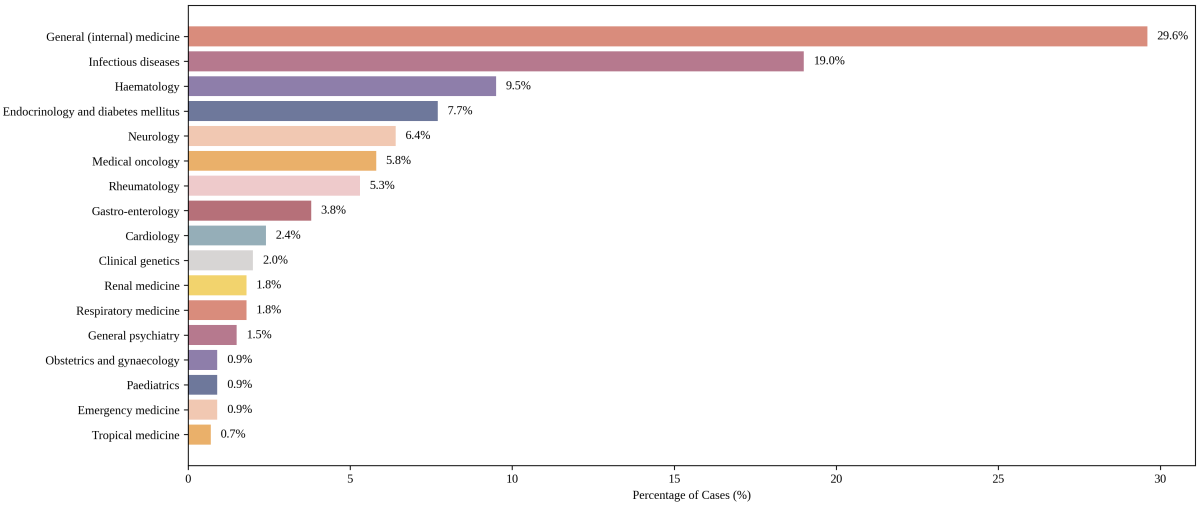
B Details of Experiments and Additional Results

B.1 Evaluation Metric

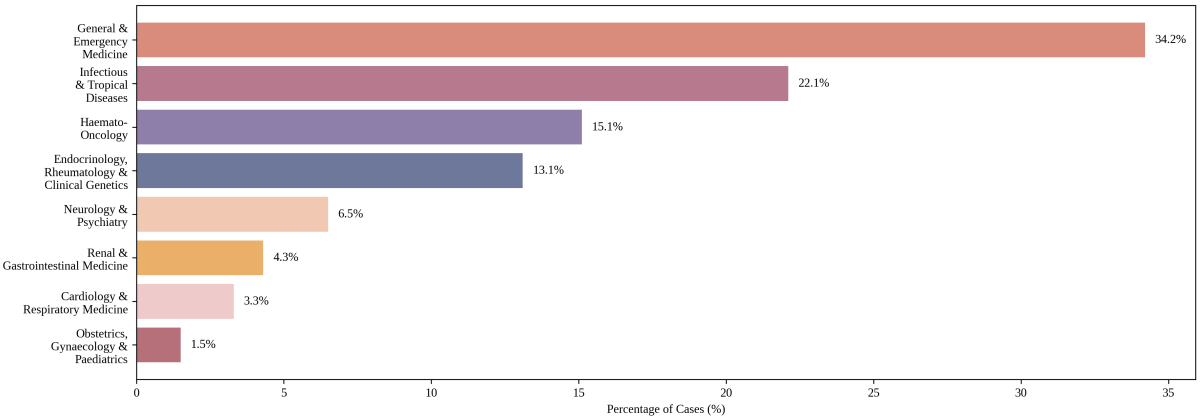
For open-ended generation, we utilize GPT-5 (OpenAI, 2025) for scoring based on the evaluation framework introduced by Kanjee et al. (Kanjee et al., 2023). The evaluation prompts for DDx and final diagnosis (FDx) are as follows:

- Prompt for scoring DDx in open-ended generation:

```
You are an expert clinician and expert
    exam-grader. You are grading the
    quality of the differential
    diagnosis list for a clinical case.

Inputs
1) Final confirmed diagnosis:
    {ground_truth}
2) Differential diagnoses: {differential
    list}

Task
Using medical knowledge, compare the
    differential diagnoses with the final
diagnosis and assign a quality score on a
    0--5 scale.

Scoring rules
5: The actual diagnosis appears
    explicitly in the differential.
4: A very close diagnosis is suggested
    (near-miss, closely adjacent).
3: A closely related diagnosis is
    suggested that could meaningfully
    help
   workup/management, but is not close in
    name.
2: A somewhat related diagnosis is
    suggested but is unlikely to be
   helpful for reaching the target.
0: No suggestions are close or clinically
    related to the target.

Output format
Score: X
Rationale: <one sentence>
```



(a) Clinical specialty distribution.



(b) Clinical specialty group distribution.

Figure 9: Distributions of clinical specialties and specialty groups.

```
Here are some examples:
...
```

- Prompt for scoring FDx in open-ended generation:

```
You are an expert clinician and expert
    exam-grader. You are grading the
    quality of the final diagnosis for a
    clinical case.

Inputs
1) Final confirmed diagnosis:
    {ground_truth}
2) Generated diagnosis: {diagnosis}

Task
Using medical knowledge, compare the
    generated diagnoses with the final
    confirmed diagnosis and assign a
    quality score on a 0--5 scale.

Scoring rules:
5: Exact match to the final diagnosis.
4: Very close diagnosis but not exact.
3: Closely related and plausibly helpful
    for reaching the final diagnosis
   (same disease family/syndrome;
    overlapping pathophysiology).
2: Related but unlikely to be clinically
    helpful (broad category or
   distant relative).
1: Unrelated to the final confirmed
    diagnosis.

Output format
Score: X
Rationale: <one sentence>

Here are some examples:
...
```

B.2 Relative Attention per Token (RAPT)

Relative Attention per Token (RAPT) is defined as the ratio of section-average attention per token to the input-average value in each layer (Liu et al., 2025). Formally, RAPT is defined as:

$$\begin{aligned}\mathrm{RAPT}_S^{(l)} &= \frac{\text{Average Attention per Token in } S}{\text{Uniform Attention per Token}} \\ &= \left(\sum_{i\in S} a_i^{(l)}\right) \times \frac{N}{N_S},\end{aligned}$$

where N is the total number of tokens in the input sequence, N_S is the number of tokens in section S, and $a^{(l)}$ is the aggregated attention vector for layer l averaged across all attention heads. For example, a RAPT of 0.1 means each token in that section receives 10% of the input-average attention.

B.3 Prompt Refinement

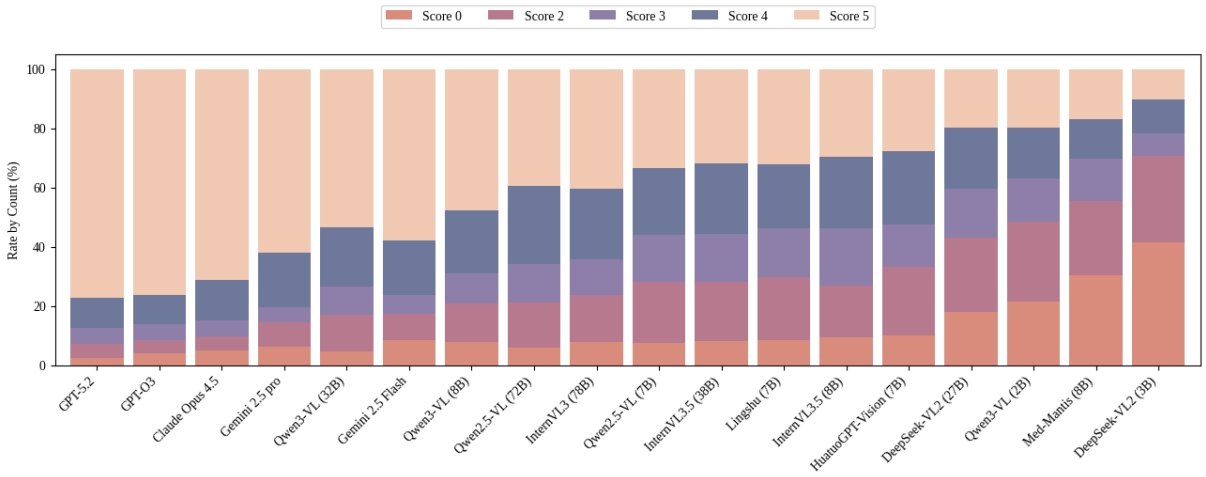
Prompt after the prompt refinement:

```
[SYSTEM]
You are an expert clinician tasked with
    synthesizing complex case files.
Use all provided information to select the
    most likely diagnosis from the given
    options.

CRITICAL INSTRUCTION: When pathology images
    are provided, you MUST prioritize and
    give PRIMARY WEIGHT to the pathology
    images in your diagnostic reasoning.
Pathology images are the gold standard for
    diagnosis and should take precedence over
    other imaging modalities.
Examine the pathology images carefully and
    base your diagnosis primarily on the
    pathological evidence shown in these
    images.
...
[USER]
...
Remember: Carefully examine and prioritize the
    pathology images in your diagnostic
    reasoning.
...
```

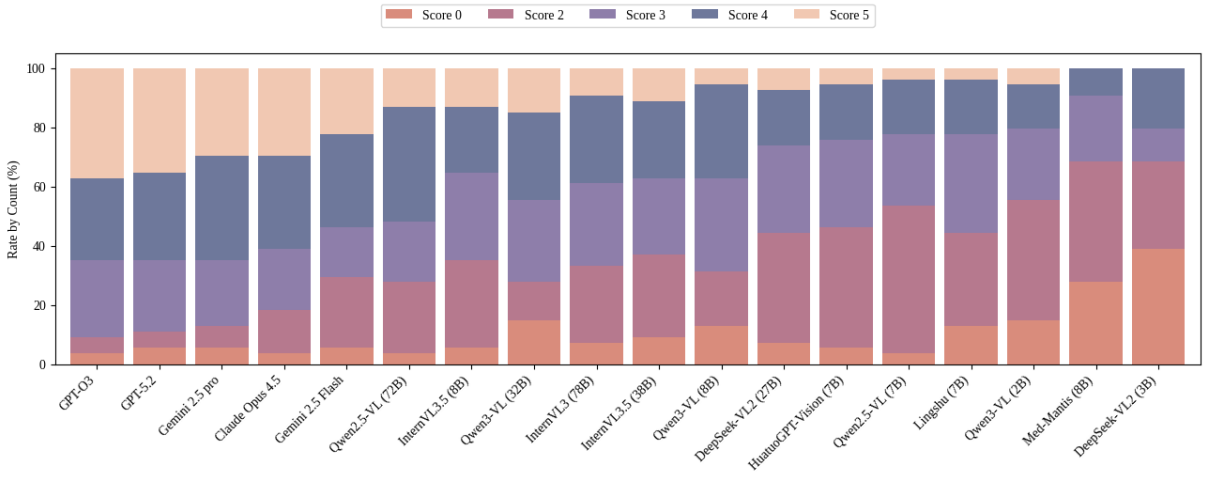
B.4 Additional Results

DDx Generation. Table 8 compares average scores of open-ended DDx generation across models. Figure 10 visualizes the distribution of scores. We can observe that top-performing models are characterized by a high density of scores of 5, whereas tiny models such as DeepSeek-VL2 (3B) show a higher frequency of scores of 0. We also notice nearly all models exhibit a performance gap between Chinese subset and the English subset, indicated by a significant drop of cases which achieves scores of 5.

Open-ended FDx Generation. In the main paper, we evaluate FDx selection based on MCQs, where models are required to select a FDx from a pre-defined differential list that contains both the correct FDx and distractors. Here, we further evaluate on FDx selection on the English subset of the benchmark by requiring models to output FDx based on their own generated differential lists, termed *open-ended* FDx generation. Outputs are evaluated by GPT-5 against ground-truth FDx (see Appendix B.1 for metric details). Table 9 reports the average performance across models. We can observe a similar gap between DDx and FDx performance, where the average score drops ~ 1 on average for leading proprietary models. Figure 11 details the score distributions. Compared with that



(a) English subset



(b) Chinese subset

Figure 10: Distributions of GPT-5 scores on differential diagnosis across models on the English and Chinese subsets. Clinical evidence from diagnostic investigations is provided as raw images.

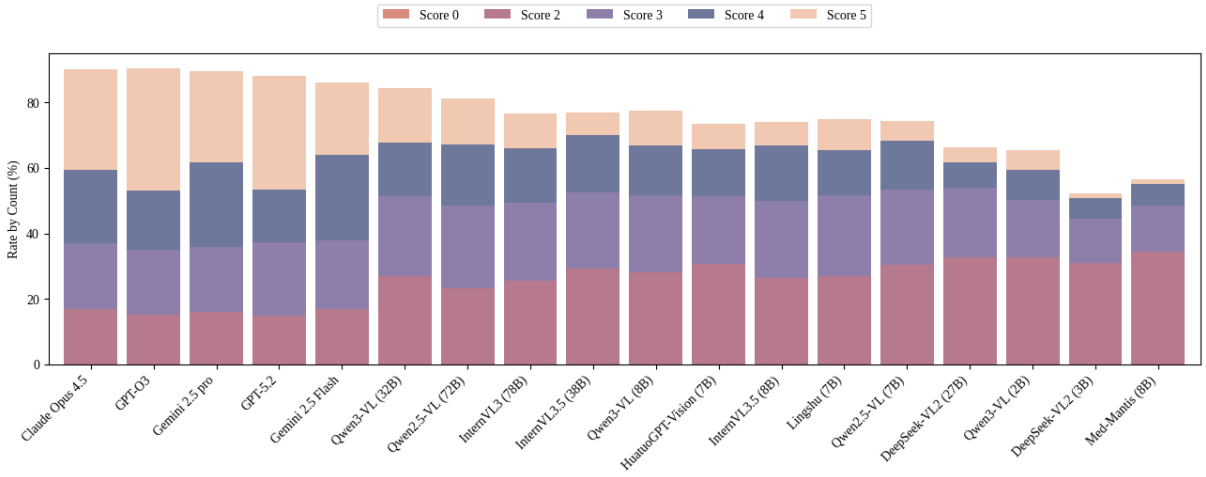


Figure 11: Distributions of GPT-5 scores on open-ended final diagnosis generation across models on the English subset. Clinical evidence from diagnostic investigations is provided as raw images.

Model	Average Score ↑		
	English	Chinese	Overall
Claude Opus 4.5	4.36	3.65	4.28
GPT-O3	4.46	3.85	4.38
Gemini 2.5 pro	4.15	3.70	4.10
GPT-5.2	4.52	3.78	4.43
Gemini 2.5 Flash	3.99	3.35	3.91
Qwen3-VL (32B)	4.00	3.02	3.88
Qwen2.5-VL (72B)	3.78	3.13	3.70
InternVL3 (78B)	3.70	2.96	3.61
InternVL3.5 (38B)	3.42	2.85	3.35
Qwen3-VL (8B)	3.35	2.85	3.29
HuatuoGPT-Vision (7B)	3.27	2.75	3.20
InternVL3.5 (8B)	3.22	2.68	3.15
Lingshu (7B)	3.18	2.66	3.12
Qwen2.5-VL (7B)	3.05	2.49	2.98
DeepSeek-VL2 (27B)	2.81	2.74	2.80
Qwen3-VL (2B)	2.65	2.41	2.62
DeepSeek-VL2 (3B)	1.78	1.74	1.78
Med-Mantis (8B)	2.31	1.85	2.25

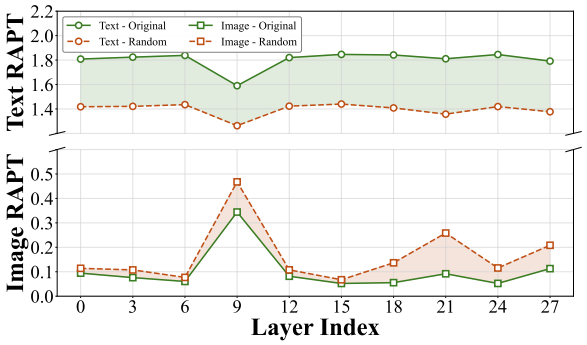
Table 8: Comparison of average GPT-5 scores on differential diagnosis generation across models on the English subset. Clinical evidence from diagnostic investigations is provided as raw images.

of DDx (Figure 10a), score distribution of FDx (Figure 11) exhibits a significant reduction in the proportion of samples with scores ≥ 4. This is consistent with our previous finding: while models excels at identifying possible medical conditions from heterogeneous clinical evidence (CE) types, they struggle to synthesize them to select the correct FDx.

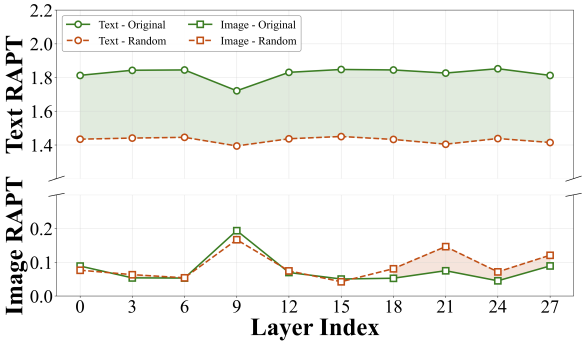
Model	Average Score ↑
Claude Opus 4.5	3.48
GPT-O3	3.59
Gemini 2.5 pro	3.45
GPT-5.2	3.47
Gemini 2.5 Flash	3.26
Qwen3-VL (32B)	2.92
Qwen2.5-VL (72B)	2.86
InternVL3 (78B)	2.66
InternVL3.5 (38B)	2.56
Qwen3-VL (8B)	2.63
HuatuoGPT-Vision (7B)	2.46
InternVL3.5 (8B)	2.53
Lingshu (7B)	2.55
Qwen2.5-VL (7B)	2.46
DeepSeek-VL2 (27B)	2.17
Qwen3-VL (2B)	2.19
DeepSeek-VL2 (3B)	1.83
Med-Mantis (8B)	1.88

Table 9: Comparison of average GPT-5 scores on open-ended final diagnosis generation across models on the English subset. Clinical evidence from diagnostic investigations is provided as raw images.

Detailed Results on FDx Selection across Clinical Specialties. Table 10 and 11 detail the accuracy on FDx selection (English subset) across 17 clinical specialties and 8 specialty groups, respectively.



(a) Qwen2.5-VL (7B)



(b) HuatuoGPT-Vision

Figure 12: Layer-wise Relative Attention per Token (RAPT) for Qwen2.5-VL (7B) and HuatuoGPT-Vision on text (excluding question stem) versus image tokens, before and after the RANDOM-TEXT intervention.

Relative Attention per Token (RAPT). Figure 12 visualizes the Relative Attention per Token (RAPT) on text (excluding the question stem) and image tokens of Qwen2.5-VL (7B) (Figure 12a) and HuatuoGPT-Vision (Figure 12b) before and after the RANDOM-TEXT intervention. We can observe similar trends as Lingshu (Figure 6b), discussed in Section 4.3 (Finding 1) in the main text.

Cross-modal Sensitivity Comparison. Table 12 shows Normalized Mean Squared Error (NMSE) (Equation 2) across 6 different clinical evidence (CE) types (*i.e.*, Laboratory, CT, Microscopy, MRI, Clinical Photography and X-ray). Figures 13 details cross-modal sensitivity comparison for Qwen2.5-VL (7B) (Figure 13a), Lingshu (Figure 13b), and HuatuoGPT-Vision (Figure 13c). We can observe that, beyond CT and Microscopy, the L-shaped distribution persists in Laboratory, MRI, Clinical Photography, and X-ray across all models. These additional results corroborate that the cross-

Model	Cardio.	Clin. Gen.	Emerg. Med.	Endo. & Diab.	Gastro.	Gen. Med.	Gen. Psych.	Haem.	Infect. Dis.	Med. Oncol.	Neuro.	Obs. & Gyn.	Paed.	Renal Med.	Resp. Med.	Rheum.	Trop. Med.
CLAUDE OPUS 4.5	71.43	44.44	100.00	68.18	69.23	65.15	71.43	64.86	64.71	52.17	84.21	0.00	33.33	50.00	83.33	57.14	66.67
GPT-O3	42.86	55.56	50.00	68.18	69.23	65.15	71.43	70.27	65.88	56.52	63.16	33.33	33.33	25.00	66.67	61.90	100.00
GEMINI 2.5 PRO	42.86	55.56	75.00	77.27	69.23	57.55	42.86	67.57	67.06	65.22	63.16	66.67	66.67	50.00	83.33	52.38	100.00
GPT-5.2	42.86	66.67	75.00	63.64	61.54	64.39	57.14	59.46	58.33	47.83	73.68	100.00	66.67	0.00	66.67	42.86	100.00
GEMINI 2.5 FLASH	42.86	22.22	75.00	77.27	61.54	56.06	14.29	45.95	62.35	65.22	78.95	33.33	33.33	50.00	66.67	47.62	100.00
QWEN3-VL (32B)	42.86	55.56	75.00	50.00	61.54	49.24	42.86	51.35	44.05	65.22	68.42	33.33	0.00	0.00	50.00	47.62	66.67
QWEN2.5-VL (72B)	28.57	33.33	50.00	54.55	69.23	46.21	28.57	43.24	51.76	56.52	63.16	33.33	0.00	25.00	66.67	38.10	66.67
INTERNVL3 (78B)	42.86	44.44	50.00	45.45	53.85	45.45	14.29	40.54	40.00	52.17	68.42	33.33	33.33	25.00	50.00	42.86	66.67
INTERNVL3.5 (38B)	42.86	0.00	50.00	59.09	38.46	43.18	14.29	51.35	38.82	52.17	57.89	100.00	0.00	25.00	33.33	42.86	66.67
QWEN3-VL (8B)	28.57	33.33	25.00	59.09	53.85	37.12	28.57	45.95	40.00	47.83	52.63	33.33	0.00	0.00	16.67	38.10	33.33
HUATUOGPT-V (7B)	71.43	22.22	50.00	45.45	38.46	38.64	57.14	29.73	41.18	34.78	42.11	66.67	0.00	25.00	83.33	19.05	33.33
INTERNVL3.5 (8B)	42.86	33.33	50.00	59.09	61.54	35.61	0.00	40.54	29.41	39.13	63.16	0.00	0.00	25.00	83.33	38.10	33.33
LINGSHU (7B)	57.14	44.44	50.00	31.82	30.77	36.36	42.86	35.14	35.29	34.78	42.11	33.33	0.00	25.00	100.00	33.33	33.33
QWEN2.5-VL (7B)	42.86	33.33	25.00	36.36	46.15	30.30	57.14	24.32	36.47	43.48	47.37	33.33	0.00	25.00	33.33	28.57	33.33
DEEPSEEK-VL2 (27B)	42.86	16.67	37.50	31.82	34.62	29.92	35.71	35.14	31.76	34.78	28.95	50.00	16.67	25.00	33.33	23.81	16.67
QWEN3-VL (2B)	28.57	22.22	25.00	36.36	46.15	25.00	28.57	32.43	32.94	43.48	26.32	66.67	33.33	25.00	33.33	38.10	0.00
DEEPSEEK-VL2 (3B)	14.29	33.33	75.00	36.36	46.15	33.33	14.29	29.73	25.88	26.09	36.84	0.00	33.33	0.00	50.00	33.33	0.00
MED-MANTIS (8B)	57.14	22.22	50.00	31.82	30.77	25.76	28.57	27.03	28.24	34.78	36.84	0.00	0.00	25.00	16.67	14.29	66.67

Table 10: Comparison of final diagnosis selection accuracy (%) across models on the English subset stratified by clinical specialties. Clinical evidence from diagnostic investigations is provided as raw images.

Model	Gen. & Emergency	Infectious & Trop.	Haemato-Oncol.	Endo, Rheum & Genetics	Neuro. & Psych.	Cardio. & Resp.	Renal & Gastro.	Obst., Gyn. & Paed.
CLAUDE OPUS 4.5	66.18	64.78	60.00	59.61	80.77	76.92	64.71	16.66
GPT-O3	64.70	67.04	65.00	63.46	65.39	53.85	58.82	33.33
GEMINI 2.5 PRO	58.06	68.18	66.67	63.46	57.69	61.54	64.71	66.67
GPT-5.2	64.70	59.75	55.00	55.77	69.23	53.85	47.06	83.33
GEMINI 2.5 FLASH	56.62	63.63	53.34	55.77	61.54	53.85	58.82	33.33
QWEN3-VL (32B)	50.00	44.82	56.67	50.00	61.54	46.16	47.06	16.66
QWEN2.5-VL (72B)	46.32	52.27	48.33	44.23	53.85	46.15	58.82	16.66
INTERNVL3 (78B)	45.58	40.91	45.00	44.23	53.85	46.16	47.06	33.33
INTERNVL3.5 (38B)	43.38	39.77	51.66	42.31	46.15	38.46	35.29	50.00
QWEN3-VL (8B)	36.76	39.77	46.67	46.15	46.15	23.08	41.18	16.66
HUATUOGPT-V (7B)	38.97	40.91	31.67	30.77	46.16	76.92	35.29	33.34
INTERNVL3.5 (8B)	36.03	29.54	40.00	46.15	46.16	61.54	52.94	0.00
LINGSHU (7B)	36.76	35.22	35.00	34.61	42.31	76.92	29.41	16.66
QWEN2.5-VL (7B)	30.14	36.36	31.66	32.69	50.00	38.46	41.17	16.66
DEEPSEEK-VL2 (27B)	30.14	31.25	35.00	25.96	30.77	38.46	32.36	33.34
QWEN3-VL (2B)	25.00	31.82	36.67	34.62	26.93	30.77	41.17	50.00
DEEPSEEK-VL2 (3B)	34.56	25.00	28.33	34.61	30.77	30.77	35.29	16.66
MED-MANTIS (8B)	26.47	29.55	30.00	23.08	34.61	38.46	29.41	0.00

Table 11: Comparison of final diagnosis selection accuracy (%) across models on the English subset stratified by clinical specialty groups. Clinical evidence from diagnostic investigations is provided as raw images.

modal utilization gap is a pervasive phenomenon across different CE types. Figures 14 visualizes the distributions of cases where $S_{\text{image}}^{(m)} < S_{\text{text}}^{(m)}$ and $S_{\text{image}}^{(m)} > S_{\text{text}}^{(m)}$. Notably, while Laboratory and X-ray distributions are more balanced, Microscopy remains a distinct outlier: $S_{\text{image}}^{(m)} < S_{\text{text}}^{(m)}$ in over 70% of cases, peaking at 77.6% in HuatuoGPT-Vision.

Modality	Qwen2.5-VL (7B)	Lingshu	HuatuoGPT-Vision
Laboratory	1.15	1.08	1.12
CT	0.90	0.88	0.91
Microscopy	0.73	0.63	0.74
MRI	1.00	0.99	0.99
Clinical Photography	0.85	0.91	0.90
X-ray	1.23	1.25	0.94
Average	0.98	0.95	0.93

Table 12: Comparison of normalized mean squared error for different clinical evidence types across models.

Prompt Refinement. Table 13 compares the microscopy NMSE before and after prompt refinement. Figure 15 visualizes the distributions of cases where $S_{\text{image}}^{(m)} < S_{\text{text}}^{(m)}$ and $S_{\text{image}}^{(m)} > S_{\text{text}}^{(m)}$ for microscopy before and after prompt refinement.

Modality	Qwen2.5-VL (7B)	Lingshu	HuatuoGPT-Vision
Baseline	0.73	0.63	0.74
Prompt Refinement	0.67 (-0.06)	0.61 (-0.02)	0.70 (-0.04)

Table 13: Comparison of normalized mean squared error (lower is better) for microscopy before and after prompt refinement across models.

C Human Evaluation Details

C.1 Physician Demographics

We recruited two volunteer senior physicians (one male, one female). Both are board-certified in their respective countries of practice and have 35 and 41 years of clinical experience. Participation was voluntary and uncompensated (*i.e.*, no monetary reward). Both participants use English as the language of instruction.

C.2 Study Procedure

Our study obtained ethical approval, and written consent was collected from each participant. The evaluation was conducted in a local, offline environment without Internet access. Both participants received a standardized briefing describing the study objective, task format, and evaluation interface, followed by a brief training phase (two example cases) to familiarize themselves with the interface and questions. The example cases were not included in the study. For each test case, the interface displays the question as well as multiple pieces of CE (*e.g.*, patient symptoms, laboratory findings, medical imaging scans and microscopy images), and records an answer determined via consensus between both participants to ensure reliability. Upon completing all cases, participants' responses were automatically logged and exported to a CSV file.

C.3 Evaluation Interface

Figure 16 illustrates the interface we used to assess expert clinicians' performance.

D Details on Supervised Fine-tuning

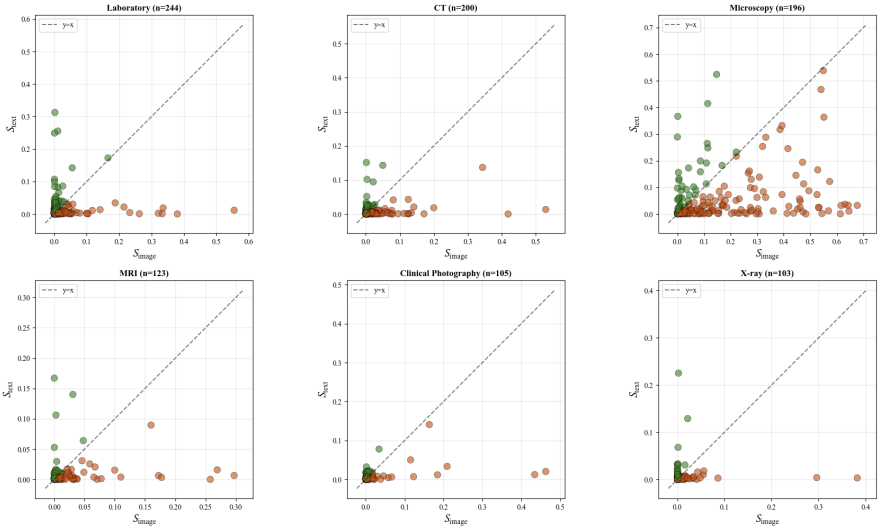
D.1 Data Construction

To ensure a fair comparison, we maintained a fixed dataset size of 3,000 samples for both the General and Targeted Supervised Fine-tuning (SFT) experiments.

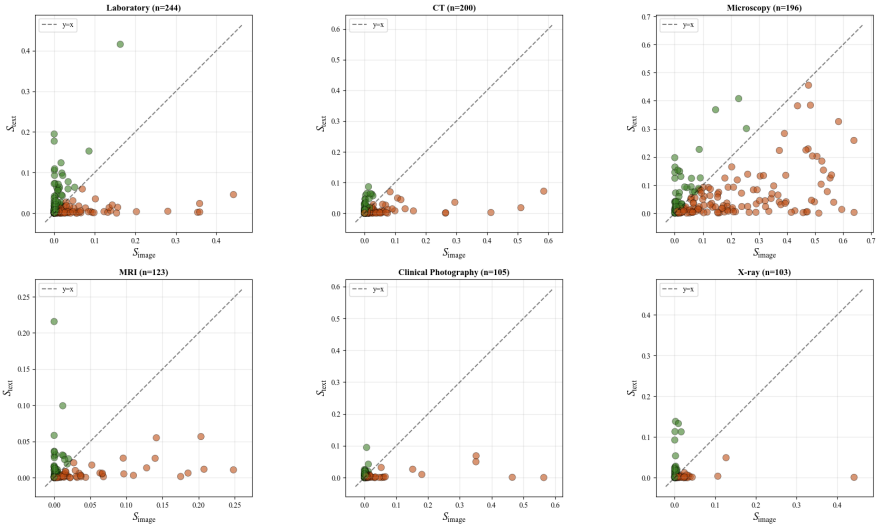
- **General SFT:** We constructed a baseline dataset by randomly sampling 3,000 instances from the PubMedVision dataset (Chen et al., 2024a), representing a general distribution of medical visual instruction tuning tasks following the distribution of the PubMedVision evidence types.
- **Targeted SFT:** We curated a "microscopy-heavy" dataset designed to enrich underutilized modalities. This dataset comprises a balanced mixture of 1,500 microscopy samples from the QUILT-LLaVA Visual Instruct 107K dataset (Seyfioglu et al., 2023) and 1,500 general medical samples from PubMedVision.

D.2 Training Configuration

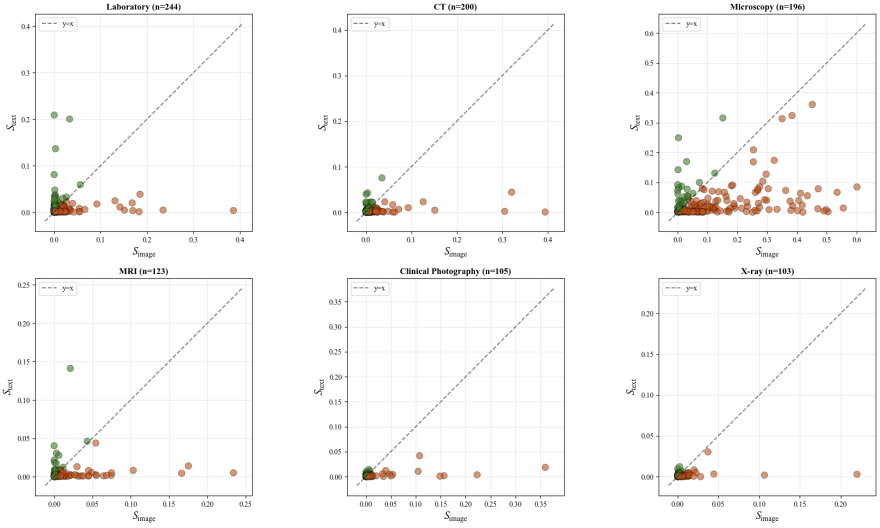
Both the General and Targeted SFT models were trained using an identical configuration to isolate the impact of data composition. We performed full-parameter fine-tuning on the Qwen2.5-VL (7B) (Yang et al., 2025a) model using the AdamW optimizer (Loshchilov and Hutter, 2017) paired with a cosine learning rate scheduler. The training was conducted on 8× NVIDIA A6000 GPUs with a global batch size of 128 and a maximum sequence length of 4,096 tokens. We utilized a peak learning rate of 5×10^{-6} and trained for 1 epoch.



(a) Qwen2.5-VL (7B)

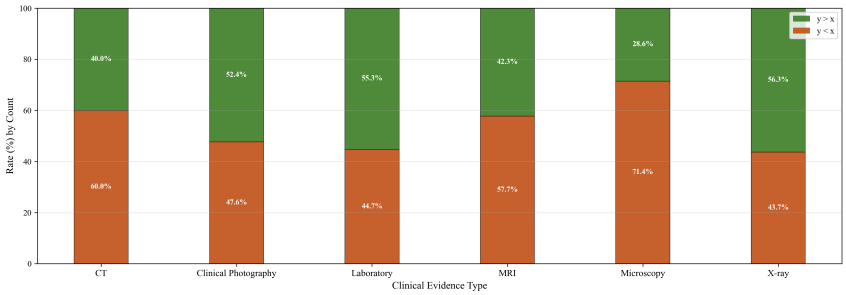


(b) Lingshu (7B)

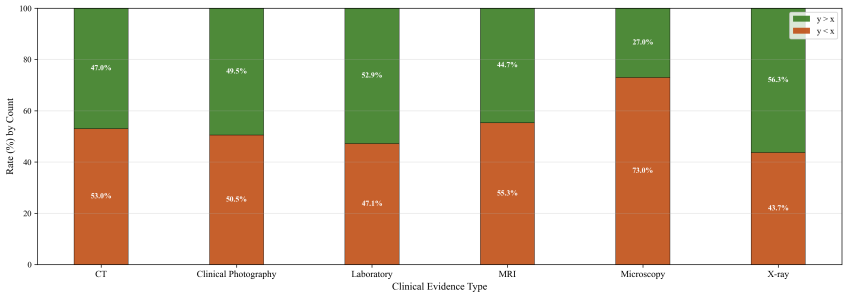


(c) HuatuoGPT-Vision (7B)

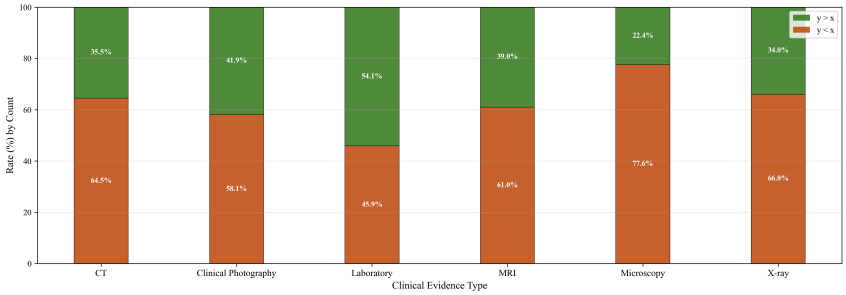
Figure 13: Cross-modal sensitivity across different clinical evidence types. Green and brown points indicate cases where $S^{(m)}\text{image} > S^{(m)}\text{text}$ and $S^{(m)}\text{image} < S^{(m)}\text{text}$, respectively.



(a) Qwen2.5-VL (7B)



(b) Lingshu (7B)



(c) HuatuoGPT-Vision (7B)

Figure 14: Distributions of cases where $S_{\text{image}}^{(m)} > S_{\text{text}}^{(m)}$ (green) and $S_{\text{image}}^{(m)} < S_{\text{text}}^{(m)}$ (brown).

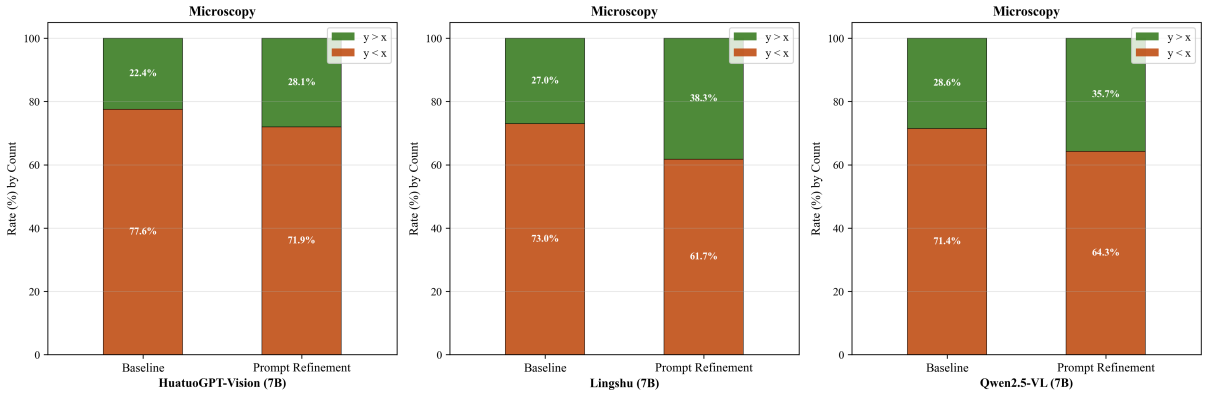


Figure 15: Distributions of cases where $S_{\text{image}}^{(m)} > S_{\text{text}}^{(m)}$ (green) and $S_{\text{image}}^{(m)} < S_{\text{text}}^{(m)}$ (brown) for microscopy before and after prompt refinement.

Case ID: NEJMcpc1616020

Clinical History

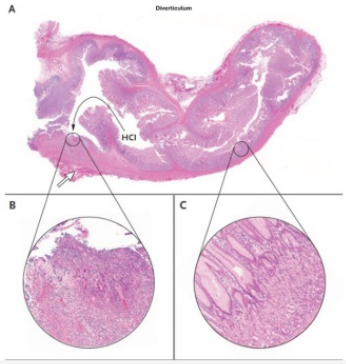
A 6-month-old boy presented after 2 days of intermittent, colicky abdominal pain progressing to large-volume gastrointestinal bleeding, including a dark red "currant jelly" stool, with associated pallor and diaphoresis; vital signs across two EDs showed persistent tachycardia (160–178 bpm) with blood pressures 98/47 to 94/36 mm Hg, initial diffuse abdominal tenderness greater on the right with no masses or fissures, and later a soft, nondistended abdomen with melena in the diaper. Past history includes infantile colic and gastroesophageal reflux previously treated with ranitidine, low-lactose cow's milk formula, recent introduction of pureed foods with constipation/straining, and immunizations current through 4 months including oral pentavalent rotavirus vaccine 6 weeks earlier; he received 5 mL/kg IV normal saline at an outside hospital, was transferred for further evaluation and treatment with ultrasonography performed, and remains under emergency department care with ongoing monitoring and hemodynamic assessment.

Diagnostic Investigations

Table 1. Laboratory Data.*

Variable	Reference Range, Age-Adjusted†	On Presentation, This Hospital
Hematocrit (%)	33.0–39.0	17.5
Hemoglobin (g/dl)	10.5–13.5	5.7
Reticulocyte count (%)	0.5–2.5	7.6
White-cell count (per mm^3)	6000–17,500	22,200
Differential count (%)		
Neutrophils	17–49	38
Lymphocytes	67–77	59
Monocytes	4–11	3
Red-cell count (per mm^3)	3,700,000–5,300,000	2,070,000
Prothrombin time (sec)	11.0–14.0	12.4
Prothrombin-time international normalized ratio	0.9–1.1	1.0
Activated partial thromboplastin time (sec)	22.1–37.0	19.4
Total protein (g/dl)	6.0–8.3	5.4
Albumin (g/dl)	3.3–5.0	4.1
Globulin (g/dl)	1.9–4.1	1.3
Iron (μg/dl)	45–160	19
Iron-binding capacity (μg/dl)	230–404	351

* To convert the values for iron and iron-binding capacity to micromoles per liter, multiply by 0.1791.
† Reference values are affected by many variables, including the patient population and the laboratory methods used. The ranges used at Massachusetts General Hospital are age-adjusted, for patients who are not pregnant and do not have medical conditions that could affect the results. They may therefore not be appropriate for all patients.



Based on this evidence, what is the most likely diagnosis?

A. Ileocolic intussusception

B. Small intestinal duplication cyst with ectopic gastric mucosa

C. Bleeding Meckel diverticulum

D. Acute infectious colitis

← Previous

Next →

Figure 16: Example of human evaluation interface.

D.3 Additional Results

We show addition results of targeted SFT on medical domain-specific model (*i.e.*, Lingshu). Table 14 demonstrates the average accuracy on FDx selection (English subset). We can observe that targeted SFT further improves performance of domain-specific model that has already undergone instruction tuning on medical data.

Setting	Acc.
BASELINE	36.93
TARGETED SFT	40.70 (+3.77)

Table 14: Final diagnosis accuracy (%) for Lingshu before and after targeted supervised fine-tuning (SFT).

E Ethical Consideration and Applications

E.1 Potential Risks

Our benchmark is constructed using real clinical cases from the *New England Journal of Medicine Case Challenge* series and the *National Medical Journal of China* journal. Releasing a dataset derived from these real clinical cases can introduce re-identification risk, particularly for patients with rare conditions or distinctive combinations of findings. Even after privacy safeguards are applied to minimize this risk, residual risk may remain.

E.2 Data Anonymization Procedures

Our data anonymization procedures include (i) manual screening for personally identifiable information (PII) and protected health information (PHI); (ii) redacting or generalizing high-risk quasi-identifiers (*e.g.*, exact ages, dates, locations, uncommon procedures); and (iii) removing identifiers embedded in images (*e.g.*, accession numbers, timestamps, or burned-in overlays). For clinical photographs, we additionally mask patient faces with black bars. The released data will exclude physician names and other extraneous identifying details. Where quasi-identifiers remain (*e.g.*, rare combinations of findings), we apply further redaction or generalization prior to release to reduce re-identification risk.

E.3 Instructions Given To Participants

E.3.1 Disclaimer for Annotators

Thank you for participating in our evaluation. Please read the following before you begin:

- **Voluntary participation:** Your participation is voluntary. You may stop at any time without penalty.
- **Confidentiality:** You will see anonymized materials that exclude personally identifiable information (PII) and protected health information (PHI). Your ratings and your responses will also be kept confidential.
- **Potential discomfort:** Although the task is low risk, some cases may include clinical content (*e.g.*, medical images or descriptions) that could be uncomfortable. You may skip any item or stop at any time.
- **Questions:** If you have questions or concerns during the task, please contact the study organizers.

E.3.2 Instructions for Annotation

Thank you for participating in our study. Please read the instructions below carefully before you begin.

Task A1: Case Verification and Evidence Completeness. For each case, review the clinical history and all associated clinical evidence (text, tables, and figures). Confirm that the evidence is complete and clinically interpretable, and flag any missing, low-quality, or uninterpretable items.

Task A2: Diagnosis Confirmation. Verify the final diagnosis reported in the source. If the final diagnosis is not explicitly stated, provide the most supported diagnosis based on the full case report and document the supporting evidence.

Task A3: Evidence–Interpretation Alignment. Check that each visual evidence item is correctly paired with its corresponding expert-derived diagnostic interpretation. Flag any misalignment, unsupported interpretation, or statement not grounded in the source.

E.3.3 Instructions for Experiments

Thank you for participating in our study. Please read the instructions below carefully.

Task 1: Interface Familiarization You will first complete a short training phase with two example cases to familiarize yourself with the interface, the case format, and how clinical evidence is presented. These examples are for practice only and are not included in our evaluation.

Task 2: Differential Diagnosis Generation and Final Diagnosis Selection. This task consists of two sub-tasks. Given the case history and all available evidence, you will first provide a ranked differential diagnosis list (from most likely to least likely). Enter all plausible diagnoses in the text box below, separated by commas. Next, click anywhere on the interface to reveal a multiple-choice question for selecting a single final diagnosis. Choose the option most strongly supported by the full set of clinical evidence considered collectively by clicking on your selected option. Your responses are saved automatically.

Task 3: Ambiguity Flags. If a case is particularly ambiguous or the evidence is deemed insufficient for a confident decision by you, flag it and provide a brief explanation (*e.g.*, missing key tests, low-quality images or multiple plausible diagnoses).

E.3.4 Data Consent

The information you provide in this study will be used only for academic research. Your responses will be stored securely and handled confidentially; we will remove direct identifiers and, where possible, report results only in aggregate to protect your privacy. Participation is voluntary. You may withdraw from the study at any time without penalty, and you may request that your data not be used in our analyses to the extent permitted after withdrawal. If you have any questions about data handling or use, please feel free to contact us.

E.4 Dataset License

- PubMedVision (Chen et al., 2024a): Apache license 2.0
- QUILT-LLaVA Visual Instruct 107K (Seyfioglu et al., 2023): Creative Commons Attribution Non Commercial No Derivatives 3.0

E.5 Use of AI Assistants in Research

In our study, generative AI assistants are used sparingly and in accordance with the guidelines on ACL's Policy on AI Writing Assistance. We utilize ChatGPT for basic paraphrasing and grammar checks. These tools are applied minimally to ensure the authenticity of our work and to adhere strictly to the regulatory standards set by ACL. Our use of these AI tools is focused, responsible, and aimed at supplementing rather than replacing human input and expertise in our research.